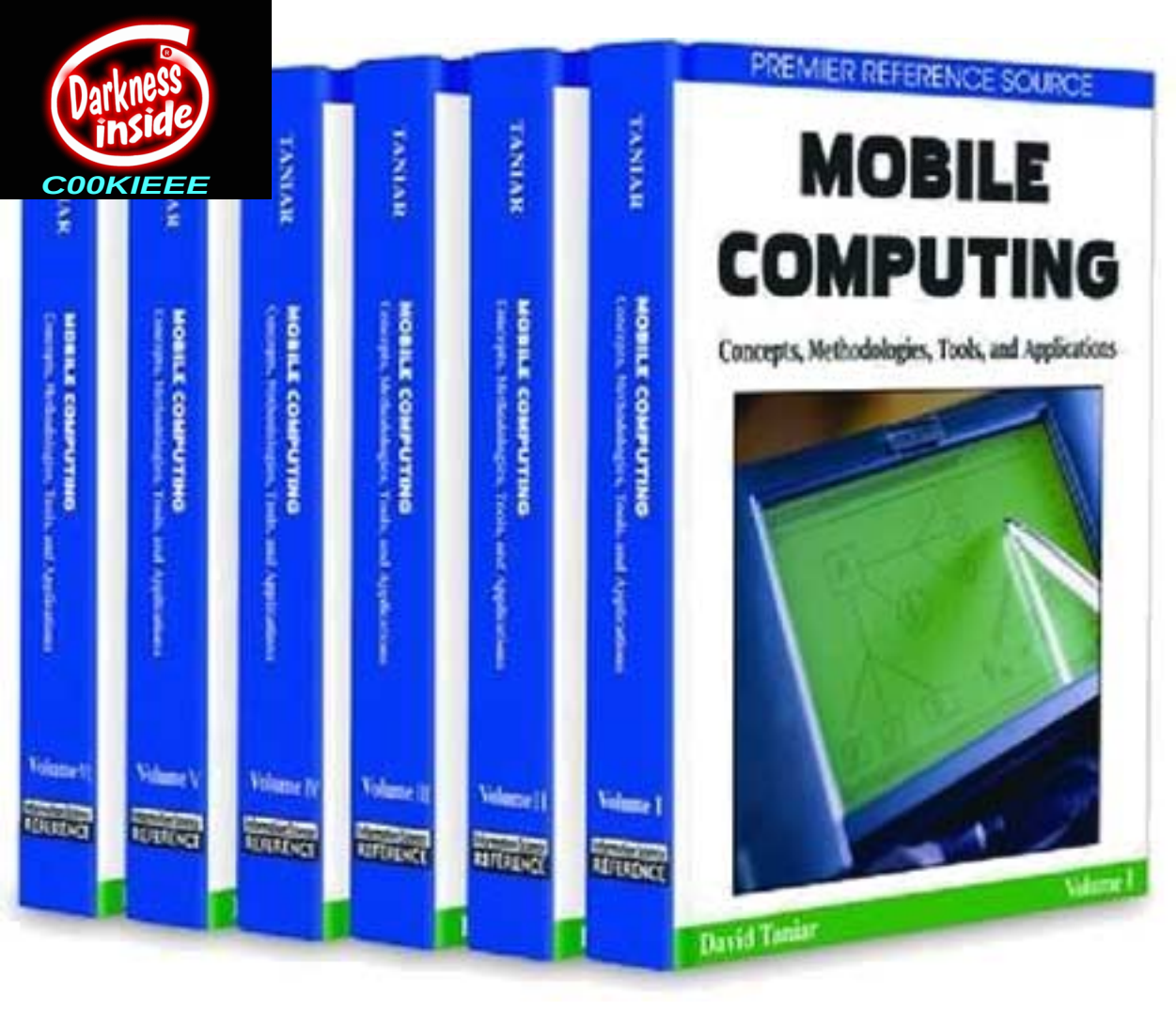


Mobile Computing:
Concepts, Methodologies, Tools, and Applications

David Taniar
Monash University, Australia

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Jeff Ash, Michael Brehm, Carole Coulson, Elizabeth Duke, Jennifer Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Mobile computing : concepts, methodologies, tools, and applications / David Taniar, editor.
v. cm.
Includes bibliographical references and index.
Summary: "This multiple-volume publication advances the emergent field of mobile computing offering research on approaches, observations and models pertaining to mobile devices and wireless communications from over 400 leading researchers"--Provided by publisher.
ISBN 978-1-60566-054-7 (hardcover) -- ISBN 978-1-60566-055-4 (ebook)
1. Mobile computing. 2. Wireless communication systems. I. Taniar, David.
QA76.59.M636 2009
004.165--dc22
2008037391

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Editor-in-Chief

Associate Editors

Editorial Advisory Board

Additional Research Collections found in the "Contemporary Research in Information Science and Technology" Book Series

Data Mining and Warehousing: Concepts, Methodologies, Tools, and Applications
John Wang, Montclair University, USA • 6-volume set • ISBN 978-1-60566-056-1

Electronic Business: Concepts, Methodologies, Tools, and Applications
In Lee, Western Illinois University • 4-volume set • ISBN 978-1-59904-943-4

Electronic Commerce: Concepts, Methodologies, Tools, and Applications
S. Ann Becker, Florida Institute of Technology, USA • 4-volume set • ISBN 978-1-59904-943-4

Electronic Government: Concepts, Methodologies, Tools, and Applications
Ari-Veikko Anttiroiko, University of Tampere, Finland • 6-volume set • ISBN 978-1-59904-947-2

Knowledge Management: Concepts, Methodologies, Tools, and Applications
Murray E. Jennex, San Diego State University, USA • 6-volume set • ISBN 978-1-59904-933-5

Information Communication Technologies: Concepts, Methodologies, Tools, and Applications
Craig Van Slyke, University of Central Florida, USA • 6-volume set • ISBN 978-1-59904-949-6

Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications
Vijayan Sugumaran, Oakland University, USA • 4-volume set • ISBN 978-1-59904-941-0

Information Security and Ethics: Concepts, Methodologies, Tools, and Applications
Hamid Nemati, The University of North Carolina at Greensboro, USA • 6-volume set • ISBN 978-1-59904-937-3

Medical Informatics: Concepts, Methodologies, Tools, and Applications
Joseph Tan, Wayne State University, USA • 4-volume set • ISBN 978-1-60566-050-9

Mobile Computing: Concepts, Methodologies, Tools, and Applications
David Taniar, Monash University, Australia • 6-volume set • ISBN 978-1-60566-054-7

Multimedia Technologies: Concepts, Methodologies, Tools, and Applications
Syed Mahbubur Rahman, Minnesota State University, Mankato, USA • 3-volume set • ISBN 978-1-60566-054-7

Virtual Technologies: Concepts, Methodologies, Tools, and Applications
Jerzy Kisielnicki, Warsaw University, Poland • 3-volume set • ISBN 978-1-59904-955-7

List of Contributors

Abdel Samad, Yara / *Ministry of Information & Communication Technologies, Jordan* 1543
Abdul Razak, Aishah / *Multimedia University, Malaysia* 3511
Abramowicz, Witold / *The Poznan University of Economics, Poland* 565, 1562
Abu-Samaha, Ala M. / *Amman University, Jordan* 1543
Ahmed Farrag, Tamer / *Mansoura University, Egypt* 3151
Ahn, Kyungmo / *Kyunghee University, Korea* 152
Ahrens, Martin / *Inductis India Pvt. Ltd., India* 2862
Al Haj Ali, Eman / *Higher Colleges of Technology, UAE* 1466
Alexander, Thomas / *FGAN—Research Institute for Communication, Information Processing, and Ergonomics, Germany* 206
Ali, Hesham A. / *Mansoura University, Egypt* 3151
Ally, Mohamed / *Athabasca University, Canada* 776
Almeida, Hyggo / *Federal University of Campina Grande, Brazil* 1763, 3212
AlMidfa, K.O. / *Etisalat University College, UAE* 558
Al-Marri, A. / *Etisalat University College, UAE* 558
Al-Nuaimi, M. / *Etisalat University College, UAE* 558
Andersson, Christer / *Combitech, Sweden* 2696
Angelides, Marios C. / *Brunel University, UK* 1584
Antunes, Pedro / *University of Lisboa, Portugal* 518
Arunatileka, Dinesh / *University of Western Sydney, Australia* 2188, 2289
Atiquzzaman, Mohammed / *University of Oklahoma, USA* 3130
Avouris, Nikolaos / *University of Patras, Greece* 3251, 3282
Baber, Chris / *The University of Birmingham, UK* 225
Balakrishnan, Vimala / *Multimedia University, Malaysia* 1984
Ballon, Pieter / *Vrije Universiteit Brussel, Belgium* 1143
Banaśkiewicz, Krzysztof / *The Pozañ University of Economics, Poland* 565
Bandyopadhyay, Subir K. / *Indiana University Northwest, USA* 38
Baousis, Vasileios / *University of Athens, Greece* 2936
Bardají, Antonio Valdovinos / *University of Zaragoza, Spain* 419
Barkhuus, Louise / *University of Glasgow, UK* 2130
Barnes, Stuart J. / *University of East Anglia, UK* 257, 1810
Bassara, Andrzej / *The Poznan University of Economics, Poland* 1562
Beckerman, Barbara G. / *Oak Ridge National Laboratory, USA* 1442
Beekhuyzen, Jenine / *Griffith University, Australia* 1351
Beer, David / *University of York, UK* 1168

Beer, Martin / *Sheffield Hallam University, UK* 1960
Bellotti, Francesco / *University of Genoa, Italy* 3387
Berger, Stefan / *Universität Passau, Germany* 2496
Berger, Stefan / *Detecon International GmbH, Germany* 188, 1359
Berry, Marsha / *RMIT University, Australia* 817
Berta, Riccardo / *University of Genoa, Italy* 3387
Billinghurst, Mark / *Human Interface Technology Laboratory, New Zealand* 984
Bina, Maria / *Athens University of Economics and Business, Greece* 1296
Black, Jason T. / *Florida A&M University, USA* 3540
Blandford, Ann / *University College London, UK* 2027
Bose, Indranil / *University of Hong Kong, Hong Kong* 870, 2179
Bourgoin, David L. / *University of Hawaii at Manoa, USA* 1665
Bozanis, Panayiotis / *University of Thessaly, Greece* 313
Bradley, John F. / *University College Dublin, Ireland* 850
Braet, Olivier / *Vrije Universiteit Brussel, Belgium* 1143
Brenner, Walter / *University of St. Gallen, Switzerland* 2257
Brodt, Torsten / *University of St. Gallen, Switzerland* 1867
Brown-Martin, Graham / *Handheld Learning, London, UK* 144
Burmester, Mike / *Florida State University, USA* 2827
Burstein, F. / *Monash University, Australia* 3552
Cardell, Nicholas Scott / *Salford Systems, USA* 2871
Carrasco, Rolando A. / *University of Newcastle-upon-Tyne, UK* 1408
Carroll, Amy / *Victoria University of Wellington, New Zealand* 1810
Caudill, Jason / *Independent Consultant, USA* 835
Chan, Shirley / *City University of Hong Kong, Hong Kong* 2124
Chan, Susy S. / *DePaul University, USA* 526, 2212
Chand, Narottam / *Indian Institute of Technology Roorkee, India* 3012
Chang, Elizabeth / *Curtin University of Technology, Australia* 546
Chang, Jun-Yang / *National Kaohsiung University of Applied Sciences, Taiwan* 3361
Chao, Han-Chieh / *National Dong Hwa University, Taiwan* 117, 3349
Charaf, Wissam / *American University of Sharjah, UAE* 1771
Chatzinotas, Symeon / *University of Surrey, UK* 2766
Chen, Charlie / *Appalachian State University, USA* 1615
Chen, Thomas M. / *Southern Methodist University, USA* 3588
Chen, Zhengxin / *University of Nebraska at Omaha, USA* 3021
Chen, Xi / *University of Hong Kong, Hong Kong* 2179
Chochliouros, Ioannis P. / *Hellenic Telecommunications Organization S.A. (OTE), Greece* 47
Chokvasin, Theptawee / *Suranaree University of Technology, Thailand* 2066
Christopoulou, Eleni / *University of Patras & Ionian University, Greece* 65
Chuang, Li-Yeh / *I-Shou University, Taiwan* 3361
Chun, Heasun / *The State University of New York at Buffalo, USA* 2509
Ciganek, Andrew P. / *University of Wisconsin-Milwaukee, USA* 2092
Cing, Tay Joc / *Nanyang Technological University, Singapore* 2896
Coaker, Ben / *Whiting-Turner Contracting Company, USA* 1530
Constantiou, Ioanna D. / *Copenhagen Business School, Denmark* 1296
Costa, Evandro / *Federal University of Alagoas, Brazil* 1763

Cowie, J. / *University of Stirling, UK* 3552
Crease, Murray / *National Research Council of Canada, Canada* 2042
Crowther, Paul / *Sheffield Hallam University, UK* 1960
Cunningham, Sally Jo / *University of Waikato, New Zealand* 3529
Dahlberg, Tomi / *Helsinki School of Economics, Finland* 1626
Dananjayan, P. / *Pondicherry Engineering College, India* 961
de Amescua-Seco, Antonio / *Universidad Carlos III de Madrid, Spain* 729
de Fátima Queiroz Vieira Turnell, Maria / *Universidade Federal de Campina Grande (UFCG), Brazil* 3168
De Gloria, Alessandro / *University of Genoa, Italy* 3387
de Haro, Guillermo / *Instituto De Empresa, Spain* 1738
de Queiroz, José Eustáquio Rangel / *Universidade Federal de Campina Grande (UFCG), Brazil* 3168
de Sousa Ferreira, Danilo / *Universidade Federal de Campina Grande (UFCG), Brazil* 3168
de Vries, Imar / *Utrecht University, The Netherlands* 1946
Deans, Candace / *University of Richmond, USA* 1530
Deek, Fadi P. / *New Jersey Institute of Technology, USA* 589
Derballa, Volker / *University of Augsburg, Germany* 197, 2169
Dey, Anind K. / *Carnegie Mellon University, USA* 3222
Dhar, Subhankar / *San Jose State University, USA* 952
Dholakia, Nikhilesh / *University of Rhode Island, USA* 27, 1331
Di Noia, Tommaso / *Politecnico di Bari, Italy* 2957
Di Sciascio, Eugenio / *Politecnico di Bari, Italy* 2957
Dietze, Claus / *The European Telecommunications Standards Institute (ETSI), France* 1004
Dillon, Tharam / *University of Technology, Australia* 546
Donini, Francesco Maria / *Università della Tuscia, Italy* 2957
Doolin, Bill / *Auckland University of Technology, New Zealand* 1466
El Morr, Christo / *York University, Canada* 1771
El-Said, Mostafa / *Grand Valley State University, USA* 3204
Enders, Albrecht / *Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany* 1653
Erlandson, Benjamin E. / *Arizona State University, USA* 3333
Fafali, P. / *National Technical University of Athens, Greece* 1
Fang, Chua Fang / *Multimedia University, Malaysia* 2600
Fang, Xiaowen / *DePaul University, USA* 526
Fernández Navajas, Julián / *University of Zaragoza, Spain* 419
Ferreira, Glauber / *Federal University of Campina Grande, Brazil* 1763
Fields, Bob / *Middlesex University, UK* 2027
Filipowska, Agata / *The Poznan University of Economics, Poland* 1562
Fiotakis, Georgios / *University of Patras, Greece* 3251
Fischer-Hübner, Simone / *Karlstad University, Sweden* 2696
Fong, Michelle W. L. / *Victoria University, Australia* 1312
Forsberg, Kim / *Intrum Justitia Finland, Finland* 1911
Fortino, Giancarlo / *DEIS, University of Calabria, Italy* 1226
Fouliras, Panayotis / *University of Macedonia, Greece* 3068
Fraser, Gordon / *Institute for Software Technology, Graz University of Technology, Austria* 597
Fraunholz, Bardo / *Deakin University, Australia* 2323

Fu, Lixin / *The University of North Carolina at Greensboro, USA* 534
Fu, Yongjian / *Cleveland State University, USA* 2969
Gallo, Jason / *Northwestern University, USA* 1096
Galloway, Chris / *Monash University, Australia* 240
Gan, Jason / *University of Technology, Australia* 383
García Moros, José / *University of Zaragoza, Spain* 419
García-Guzmán, Javier / *Universidad Carlos III de Madrid, Spain* 729
Garret, Bernie / *University of British Columbia, Canada* 998
Garrett, Bernard Mark / *University of British Columbia, Canada* 1256
Garro, Alfredo / *DEIS, University of Calabria, Italy* 1226
Gartmann, Rüdiger / *University of Münster, Germany* 3404
Gayeski, Diane M. / *Ithaca College, USA* 811
Ghanbary, Abbass / *University of Western Sydney, Australia* 785, 2289, 2368
Ghosh, Sutirtha / *Inductis India Pvt. Ltd., India* 2862
Giroux, Sylvain / *Université de Sherbrooke, Canada* 1069
Golovnya, Mikhaylo / *Salford Systems, USA* 2871
Gomathy, C. / *Deemed University, India* 2996
Grahn, Kaj / *Arcada University of Applied Sciences, Finland & Arcada Polytechnic, Finland* 2660, 2766
Grillo, Antonio / *Universita di Roma "Tor Vergata", Italy* 1237
Gruber, Franz / *RISC Software GmbH, Austria* 459
Gruijters, Dominic / *University of Cape Town, South Africa* 1396
Guah, Matthew W. / *School of Business Economics, Erasmus University Rotterdam, The Netherlands* 403
Guan, Sheng-Uei / *National University of Singapore, Singapore; Brunel University, UK* 305, 881, 1640, 2715
Hadjiefthymiades, Stathes / *University of Athens, Greece* 2936
Haghirian, Parissa / *Sophia University, Japan* 1893
Häkkilä, Jonna / *University of Oulu, Finland & Nokia Research Center, Finland* 1351, 3222
Halid Kuscu, M. / *Mobile Government Consortium International, UK* 248
Hallin, Anette / *Royal Institute of Technology (KTH), Sweden* 3455
Hameurlain, A. / *IRIT—Paul Sabatier University, France* 334
Hamidi, H. / *Iran University of Science & Technology, Iran-Tehran* 2739
Hamilton, Margaret / *RMIT University, Australia* 817
Han, Song / *Curtin University of Technology, Australia* 546
Harno, Jarmo / *Nokia Research Center, Finland* 2475
Harris, Catherine / *Gloucestershire Hospitals NHS Foundation Trust, UK* 3529
Harrison Jr., William T. / *University of West Florida, USA & U.S. Navy, USA* 1381
Hartmann, Werner / *FAW Software Engineering GmbH, Austria* 459
Heinonen, Kristina / *HANKEN—Swedish School of Economics and Business Administration, Finland* 2233
Heinonen, Sirkka / *VTT Building and Transport, Finland* 2061
Henrysson, Anders / *Norrköping Visualisation and Interaction Studio, Sweden* 984
Hernández Ramos, Carolina / *University of Zaragoza, Spain* 419
Hertweck, Dieter / *University for Applied Sciences Heilbronn, Germany* 2391
Herzog, Naomi / *RMIT University, Australia* 817

Hoffman, Holger / *Technische Universität München, Germany* 1499
Holtkamp, Bernhard / *Fraunhofer Institute for Software and Systems Engineering, Germany* 3404
Horvat, Marko / *Croatian Railways Ltd., Croatia* 1516
Hsieh, Pei-Hung / *STPRIC, National Science Council, Taiwan* 2530
Hsu, HY Sonya / *Southern Illinois University, USA* 1886
Hu, Weihong / *Auburn University, USA* 1204
Hu, Wen-Chen / *University of North Dakota, USA* 534, 909, 1183, 1204, 2614
Hua, Zhigang / *Chinese Academy of Sciences, China* 497
Hurson, A.R. / *The Pennsylvania State University, USA* 1442, 3079
Hwang, Chong-Sun / *Korea University, Republic of Korea* 2982
Iglesias, Álvaro Alesanco / *University of Zaragoza, Spain* 419
Imre, Sándor / *Budapest University of Technology and Economics, Hungary* 682
Isomäki, Hannakaisa / *University of Jyväskylä, Finland* 1967
Istepanian, Robert S. H. / *Kingston University, UK* 419
Jain, Ankur / *Inductis India Pvt. Ltd., India* 2862
Jane, F. Mary Magdalene / *P. S. G. R. Krishnammal, India* 2568
Jefferies, Laura / *University of Gloucestershire, UK* 3529
Jelassi, Tawfik / *Ecole Nationale des Ponts et Chaussées, France* 1653
Jentzsch, Ric / *Compucat Research Pty Limited, Australia* 3368
Jeong, Eui Jun / *Michigan State University, USA* 289
Jiao, Y. / *The Pennsylvania State University, USA* 3079
Jiao, Yu / *Oak Ridge National Laboratory, USA* 1442
Jih, Wen-Jang (Kenny) / *Middle Tennessee State University, USA* 1823, 1840
Jones, Matt / *University of Waikato, New Zealand* 1396
Jones, Matthew R. / *University of Cambridge, UK* 1429
Joshi, R. C. / *Indian Institute of Technology Roorkee, India* 3012
Kaasinen, Eija / *VTT Technical Research Centre of Finland, Finland* 1996
Kallio, Jukka / *Helsinki School of Economics, Finland* 506
Kálmán, György / *University Graduate Center – UniK, Norway* 2725, 2792
Kamthan, Pankaj / *Concordia University, Canada* 372, 796, 1937
Kangasharju, Jaakko / *Helsinki Institute for Information Technology, Finland* 2633
Kao, I-Lung / *IBM, USA* 909
Kao, Michelle T.C. / *National Dong Hwa University, Taiwan R.O.C.* 117
Karlsson, Jonny / *Arcada University of Applied Sciences, Finland & Arcada Polytechnic, Finland* 2660, 2766
Karnouskos, Stamatis / *SAP Research, Germany; Fraunhofer Institute FOKUS, Germany* 642, 2280
Karoui, Kamel / *Institut National des Sciences Appliquées de Tunis, Tunisia* 296
Katsianis, Dimitris / *National and Kapodistrian University of Athens, Greece* 2475
Kawash, Jalal / *American University of Sharjah, UAE* 1771
Kela, Juha / *Finwe Ltd., Finland* 1029
Khashchanskiy, Victor I. / *First Hop Ltd., Finland* 1135
Khattab, Ishraga / *Brunel University, UK* 2110
Kim, Dan J. / *University of Houston Clear Lake, USA* 289, 2807
Kim, Jin Ki / *Korea Advanced University, Korea* 2509
Kini, Ranjan B. / *Indiana University Northwest, USA* 38

Kitisin, Sukumal / *Kasetsart University, Thailand* 269
Klemmer, Scott / *Stanford University, USA*..... 920
Kolbe, Lutz M. / *University of St. Gallen, Switzerland*..... 2257
Komis, Vassilis / *University of Patras, Greece* 3251, 3282
Komiya, Ryoichi / *Multimedia University, Malaysia* 3511
Kong, Ki-Sik / *Korea University, Republic of Korea*..... 2982
Korpipää, Panu / *Finwe Ltd., Finland* 1029
Kotulski, Zbigniew / *Polish Academy of Sciences, Warsaw & Warsaw University of Technology, Poland*..... 2583
Koubaa, Hend / *Norwegian University of Science and Technology (NTNU), Norway*..... 1103
Koukia, Spiridoula / *University of Patras, Greece & Research Academic Computer Technology Institute, Greece* 1064
Kourbelis, N. / *National Technical University of Athens, Greece* 1
Krcmar, Helmut / *Technische Universität München, Germany*..... 1499
Kshetri, Nir / *University of North Carolina at Greensboro, USA*..... 1665
Kulviwat, Songpol / *Hofstra University, USA*..... 1886
Kundu, Suddha Sattwa / *Inductis India Pvt. Ltd., India* 2862
Kuppuswami, Anand / *University of Western Sydney, Australia*..... 618
Kushchu, Ibrahim / *Mobile Government Consortium International, UK* 248
Kustov, Andrei L. / *First Hop Ltd., Finland* 1135
Kwok, Sai Ho / *California State University, Long Beach, USA* 1117
Kwon, Youngsun / *Information and Communications University, Republic of Korea*..... 1699
Lahti, Janne / *VTT Technical Research Centre of Finland, Finland*..... 1080
Lalopoulos, George K. / *Hellenic Telecommunications Organization S.A. (OTE), Greece* 47
Lam, Jean / *IBM, USA*..... 2212
Landay, James A. / *University of Washington & Intel Research Seattle, USA* 920
Lee, Cheon-Pyo / *Mississippi State University, USA; Carson-Newman College, USA* 1246, 2163
Lee, Chung-wei / *Auburn University, USA*..... 1183, 2614
Lee, Dennis / *The University of Queensland, Australia & The Australian CRC for Interaction Design, Australia* 279
Lee, Jason Chong / *Virginia Polytechnic Institute and State University (Virginia Tech), USA* 3320
Lee, Kun Chang / *Sungkyunkwan University, Korea* 3421
Lee, Namho / *Sungkyunkwan University, Korea* 3421
Lee, Sheng-Chien / *University of Florida, USA*..... 534
Lee, Su-Fang / *Overseas Chinese Institute of Technology, Taiwan* 1823
Lehmann, Hans / *Victoria University of Wellington, New Zealand* 188, 1359
Leimeister, Jan Marco / *Technische Universität München, Germany*..... 1499
Lentini, Alessandro / *Universita di Roma "Tor Vergata", Italy* 1237
Leow, Chye-Huang / *Singapore Polytechnic, Republic of Singapore*..... 2343
Leow, Winnie C. H. / *Singapore Polytechnic, Singapore*..... 1713
Leyk, Dieter / *German Sport University Cologne, Germany & Central Institute of the Federal Armed Forces Medical Services, Koblenz, Germany*..... 206
Li, Xining / *University of Guelph, Guelph, Canada* 858
Li, Yang / *University of Washington, USA* 920
Li, Yuan-chao / *China University of Petroleum, P.R. China* 473
Liljander, Veronica / *Swedish School of Economics and Business Administration, Finland* 1911

Lim, Say Ying / *Monash University, Australia* 350, 3185
Lindholm, Tancred / *Helsinki Institute for Information Technology, Finland* 2633
Linjama, Jukka / *Nokia, Finland* 1029
Longworth, Robert / *University of New Brunswick, Canada* 2042
Loureiro, Emerson / *Federal University of Campina Grande, Brazil* 3212
Love, Steve / *Brunel University, UK* 2110
Lu, Hanqing / *Chinese Academy of Sciences, China* 497
Lundevall, Kristina / *The City of Stockholm, Sweden* 3455
Luo, Xin / *Virginia State University, USA* 2203
Ma, Louis C. K. / *City University of Hong Kong, Hong Kong* 2124
Ma, Wei-Ying / *Microsoft Research Asia, China* 497
Maamar, Zakaria / *Zayed University, UAE* 388, 451, 891
Madlberger, Maria / *Vienna University of Economics and Business Administration, Austria* 1893
Mahmoud, Qusay H. / *University of Guelph, Canada* 388, 451
Maitland, Carleen / *Pennsylvania State University, USA* 1721, 2440
Mäkinen, Sari / *University of Tampere, Finland* 968
Mallat, Niina / *Helsinki School of Economics, Finland* 1626
Mammeri, Z. / *IRIT—Paul Sabatier University, France* 334
Maniraj Singh, Anesh / *University of KwaZulu-Natal, South Africa* 1690
Margarone, Massimiliano / *University of Genoa, Italy* 3387
María García, José / *Instituto De Empresa, Spain* 1738
Marques, Paulo / *University of Coimbra, Portugal* 3300
Marsden, Gary / *University of Cape Town, South Africa* 1396
Marsit, N. / *IRIT—Paul Sabatier University, France* 334
Martins, Henrique M. G. / *University of Cambridge, UK* 1429
Martucci, Leonardo A. / *Karlstad University, Sweden* 2696
McCrickard, D. Scott / *Virginia Polytechnic Institute and State University (Virginia Tech), USA* 3320
McManus, Patricia / *Edith Cowan University, Australia* 1788
Me, Gianluigi / *Universita di Roma "Tor Vergata", Italy* 1237
Melliar-Smith, P. M. / *University of California, Santa Barbara, USA* 3494
Merakos, Lazaros / *University of Athens, Greece* 2936
Merten, Patrick S. / *University of Fribourg, Switzerland* 10
Minogiannis, N. / *National Technical University of Athens, Greece* 1
Misra, Manoj / *Indian Institute of Technology Roorkee, India* 3012
Mittal, Nitin / *Nokia Pte Ltd, Singapore* 1194
Mohamedally, Dean / *City University London, UK* 2019
Mohammadi, K. / *Iran University of Science & Technology, Iran-Tehran* 2739
Mohammadian, Masoud / *University of Canberra, Australia* 3368
Moore Olmstead, Paul / *Atos Research and Innovation, Spain* 1562
Moreau, Jean-François / *Université de Sherbrooke, Canada* 1069
Morvan, F. / *IRIT—Paul Sabatier University, France* 334
Moser, Louise E. / *University of California, Santa Barbara, USA* 3494
Muhlberger, Ralf / *The University of Queensland, Australia & The Australian CRC for Interaction Design, Australia* 279
Muldoon, Conor / *University College Dublin, Ireland* 850

Nam, Changi / *Information and Communications University, Republic of Korea* 1699
Nand, Sashi / *Rushmore University, Grand Cayman, BWI* 2784
Navarro, Mariano / *TRAGSA Group Information, Spain* 729
Ni, Jingbo / *University of Guelph, Guelph, Canada* 858
Noll, Josef / *University Graduate Center – UniK, Norway* 2725, 2792
Northrup, Pamela T. / *University of West Florida, USA* 1381
Nösekabel, Holger / *University of Passau, Germany* 122, 1125
Ntantogian, Christoforos / *University of Athens, Greece* 2674
O'Grady, Michael J. / *University College Dublin, Ireland* 850, 1047, 3442
O'Hare, Gregory M. P. / *University College Dublin, Ireland* 850, 1047, 3442
Oddershede, Astrid M. / *University of Santiago of Chile, Chile* 1408
Okazaki, Shintaro / *Autonomous University of Madrid, Spain* 1975
Oliveira, Loreno / *Federal University of Campina Grande, Brazil* 3212
Olla, Phillip / *Madonna University, USA* 432
Ollila, Mark / *Norrköping Visualisation and Interaction Studio, Sweden* 984
Ong, Chee Chye / *Nanyang Technological University, Singapore* 1713
Paay, Jeni / *Aalborg University, Denmark* 3333
Padgham, Lin / *RMIT University, Australia* 817
Palola, Marko / *VTT Technical Research Centre of Finland, Finland* 1080
Papadimitriou, Ioanna / *University of Patras, Greece* 3251, 3282
Parker, Shin / *University of Nebraska at Omaha, USA* 3021
Parsons, David / *Massey University, New Zealand* 805
Patrikakis, Ch. Z. / *National Technical University of Athens, Greece* 1
Paul, Hironmoy / *Cleveland State University, USA* 2969
Päykkönen, Kirsi / *University of Lapland, Finland* 1967
Peikari, Cyrus / *Airscanner Mobile Security Corporation, USA* 3588
Peinel, Gertraud / *Fraunhofer FIT, Germany* 1562
Peltola, Johannes / *VTT Technical Research Centre of Finland, Finland* 1080
Perkusich, Angelo / *Federal University of Campina Grande, Brazil* 1763, 3212
Petrie, Helen / *City University London, UK* 2019
Petrova, Krassie / *Auckland University of Technology, New Zealand* 1593
Piekarski, Wayne / *University of South Australia, Australia* 937
Pierre, Samuel / *École Polytechnique de Montréal, Canada* 18, 650, 2653
Pigot, Hélène / *Université de Sherbrooke, Canada* 1069
Ping, Wang / *University of Hong Kong, Hong Kong* 870
Piscitelli, Giacomo / *Politecnico di Bari, Italy* 2957
Polsa, Pia / *Swedish School of Economics and Business Administration, Finland* 1911
Polyzos, George C. / *Athens University of Economics and Business, Greece* 1754
Poole, Marshall Scott / *Texas A&M University, USA* 56
Potdar, Vidyasagar / *Curtin University of Technology, Australia* 546
Potok, Thomas E. / *Oak Ridge National Laboratory, USA* 1442
Pousttchi, Key / *University of Augsburg, Germany* 197, 2169
Prasad, Rohit / *Management Development Institute, India* 2306
Pulkkis, Göran / *Arcada Polytechnic, Finland & Arcada University of Applied Sciences, Finland* 2660, 2766
Pura, Minna / *HANKEN—Swedish School of Economics and Business Administration, Finland* 2233

Quah, Jon T. S. / *Nanyang Technological University, Singapore* 1713
Quah, Tong-Seng / *Nanyang Technological University, Republic of Singapore* 2343
Radhamani, G. / *Multimedia University, Malaysia* 2600
Räisänen, Hanna / *University of Lapland, Finland* 1967
Rajala, Risto / *Helsinki School of Economics, Finland* 2463
Rajeev, S. / *PSG College of Technology, India* 3236
Ramamurthy, K. / *University of Wisconsin-Milwaukee, USA* 2092
Ranft, Anne-Marie / *University of Technology, Australia* 1857
Rantakokko, Tapani / *Finwe Ltd., Finland* 1029
Rao, N. Raghavendra / *SSN School of Management & Computer Applications, India* 1602
Rao, Ranjan / *Inductis India Pvt. Ltd., India* 2862
Rao Hill, Sally / *University of Adelaide, Australia* 84
Raptis, Dimitrios / *University of Patras, Greece* 3251, 3282
Rashid, Asarnusch / *Research Center for Information Technology Karlsruhe, Germany* 2391
Rask, Morten / *Aarhus School of Business, Denmark* 27
Reeves, Nina / *University of Gloucestershire, UK* 3529
Remus, Ulrich / *University of Erlangen-Nuremberg, Germany* 188, 1359, 2496
Rigou, Maria / *University of Patras, Greece & Research Academic Computer Technology Institute, Greece* 1064
Röckelein, Wolfgang / *EMPRISE Consulting Düseldorf, Germany* 1125
Roggenkamp, Klas / *Dipl. Designer Electronic Business, Germany* 756
Roh, Sung-Ju / *Technology R&D Center, LG Telecom Co., Republic of Korea* 2982
Rohling, Hermann / *Hamburg University of Technology, Germany* 3561
Rokkas, Theodoros / *National and Kapodistrian University of Athens, Greece* 2475
Rossi, Matti / *Helsinki School of Economics, Finland* 2463
Roy Dholakia, Ruby / *University of Rhode Island, USA* 27
Ruhi, Umar / *Wilfrid Laurier University, Canada* 1483
Ruiz Mas, José / *University of Zaragoza, Spain* 419
Russo, Wilma / *DEIS, University of Calabria, Italy* 1226
Ruta, Michele / *Politecnico di Bari, Italy* 2957
Saha, Debashis / *Indian Institute of Management (IIM) Calcutta, India* 488
Salam, A. F. / *University of North Carolina at Greensboro, USA* 1053
Salo, Jari / *University of Oulu, Finland* 1878
Sampat, Miten / *Feeva Techology, Inc., USA* 3320
Samuelsson, Mats / *Mobio Networks, USA* 1331
Samundeeswari, E. S. / *Vellalar College for Women, India* 2568
Sánchez-Segura, María-Isabel / *Universidad Carlos III de Madrid, Spain* 729
Saravanan, I. / *Pondicherry Engineering College, India* 961
Sardana, Sanjeev / *Mobio Networks, USA* 1331
Sardar, Bhaskar / *Jadavpur University, India* 488
Savary, Jean-Pierre / *Division R&D CRD, France* 1069
Sawyer, Steve / *The Pennsylvania State University, USA* 2079
Schierholz, Ragnar / *University of St. Gallen, Switzerland* 2257
Schilhavy, Richard / *University of North Carolina at Greensboro, USA* 1053
Schizas, Christos / *University of Cyprus, Cyprus* 1584
Schlick, Christopher / *RWTH Aachen University, Germany* 206

Schnelle, Dirk / *Technische Universität Darmstadt, Germany* 3468
Scornavacca, Eusebio / *Victoria University of Wellington, New Zealand* 257, 1810
Seah, Winston K. G. / *Institute for Infocomm Research, Singapore* 2833
Serenko, Alexander / *Lakehead University, Canada* 171, 181, 1929
Shan, Mok Wai / *University of Hong Kong, Hong Kong* 870
Shanmugavel, S. / *Anna University, India* 2996
Sharda, Nalin / *Victoria University, Australia* 2843
Shetty, Namita / *Cleveland State University, USA* 2969
Shim, J. P. / *Mississippi State University, USA* 152
Shim, Julie M. / *Soldier Design LLC, USA* 152
Shing, Wong Ka / *University of Hong Kong, Hong Kong* 870
Shing, Yip Yee / *University of Hong Kong, Hong Kong* 870
Shubair, R.M. / *Etisalat University College, UAE* 558
Siek, Katie A. / *University of Colorado at Boulder, USA* 3270
Sievert, Alexander / *German Sport University Cologne, Germany* 206
Silva, Luís / *University of Coimbra, Portugal* 3300
Simon, Vilmos / *Budapest University of Technology and Economics, Hungary* 682
Sirmakessis, Spiros / *Technological Institution of Messolongi, Greece & Research Academic Computer Technology Institute, Greece* 1064
Sivagurunathan, Surendra Kumar / *University of Oklahoma, USA* 3130
Sivanandam, S. N. / *PSG College of Technology, India* 3236
Sivaradje, G. / *Pondicherry Engineering College, India* 961
So, Simon / *Hong Kong Institute of Education, Hong Kong* 1344
Sofokleous, Anastasis / *Brunel University, UK* 1584
Song, Lei / *University of Guelph, Guelph, Canada* 858
Sphicopoulos, Thomas / *National and Kapodistrian University of Athens, Greece* 2475
Spiliopoulos, Vassilis / *University of the Aegean and National Centre of Scientific Research "Demokritos", Greece* 2936
Spiliopoulou-Chochliourou, Anastasia S. / *Hellenic Telecommunications Organization S.A. (OTE), Greece* 47
Sreenaath, K. V. / *PSG College of Technology, India* 3236
Sridhar, Varadharajan / *Management Development Institute, India* 2306
Srikhutkhao, Nopparat / *Kasetsart University, Thailand* 269
Srinivasan, Bala / *Monash University, Australia* 350, 3185
Standing, Craig / *Edith Cowan University, Australia* 1788
Statica, Robert / *New Jersey Institute of Technology, USA* 589
Steinbauer, Gerald / *Institute for Software Technology, Graz University of Technology, Austria* 597
Steinberg, Dan / *Salford Systems, USA* 2871
Steinert, Martin / *University of Fribourg, Switzerland* 10
Stelmaszewska, Hanna / *Middlesex University, UK* 2027
Stoica, Adrian / *University of Patras, Greece* 3251
Sun, Jun / *Texas A&M University, USA & University of Texas–Pan American, USA* 56, 1780
Swan, Karen / *Kent State University, USA* 144
Taha, Hamza / *American University of Sharjah, UAE* 1771
Tähtinen, Jaana / *University of Oulu, Finland* 1878
Tan, Hwee-Xian / *National University of Singapore, Singapore* 2833

Tan, Joseph / *Wayne State University, USA* 432
Taniar, David / *Monash University, Australia* 350, 3185
Tapia, Andrea / *Pennsylvania State University, USA* 2079
Tarkoma, Sasu / *Helsinki Institute for Information Technology, Finland* 2633
Tarnacha, Ankur / *The Pennsylvania State University, USA* 1721
Teoh, Jenny / *K & J Business Solutions, Australia* 2145
Terziyan, Vagan / *University of Jyvaskyla, Finland* 630
Teufel, Stephanie / *University of Fribourg, Switzerland* 10
Tilsner, Dirk / *EDISOFT, Portugal* 1562
Tin, Chan Lit / *University of Hong Kong, Hong Kong* 870
Tinnilä, Markku / *Helsinki School of Economics, Finland* 506
Tomak, Kerem / *University of Texas at Austin, USA* 1796
Tong, Carrison K. S. / *Pamela Youde Nethersole Eastern Hospital, Hong Kong* 1261
Tran, Dai / *Arcada Polytechnic, Finland* 2660
Tran, Thomas / *University of Ottawa, Canada* 712
Trifonova, Anna / *University of Trento, Italy* 1367
Troshani, Indrit / *University of Adelaide, Australia* 84
Tsai, Yuan-Cheng / *Da-Yeh University, Taiwan* 1823
Tselios, Nikolaos / *University of Patras, Greece* 3282
Tseng, Anne / *Helsinki School of Economics, Finland* 506
Turel, Ofir / *California State University, Fullerton, USA & McMaster University, Canada* 171, 181, 1483, 1929
Turowski, Klaus / *Universität Augsburg, Germany* 2169
Tuunainen, Virpi Kristiina / *Helsinki School of Economics, Finland* 2463
Unhelkar, Bhuvan / *University of Western Sydney, Australia* 2289, 2368
Unnithan, Chandana / *Deakin University, Australia* 2323
Vaghjiani, Khimji / *K & J Business Solutions, Australia* 2145
van 't Hooft, Mark / *Kent State University, USA* 144
van de Kar, Els / *Delft University of Technology, The Netherlands* 2440
Van Schyndel, Ron / *RMIT University, Australia* 817
Varoutas, Dimitris / *National and Kapodistrian University of Athens, Greece* 2475
Veijalainen, Jari / *University of Jyvaskyla, Finland* 2908
Verkasalo, Hannu / *Helsinki University of Technology, Finland* 1273
Ververidis, Christopher / *Athens University of Economics and Business, Greece* 1754
Vesa, Jarkko / *Helsinki School of Economics, Finland* 696
Vihinen, Janne / *Helsinki School of Economics, Finland* 2463
Vildjiounaite, Elena / *VTT Technical Research Centre of Finland, Finland* 1080
Vilmos, András / *SafePay Systems, Ltd., Hungary* 2280
Viruete Navarro, Eduardo Antonio / *University of Zaragoza, Spain* 419
Vogel, Douglas / *City University of Hong Kong, Hong Kong* 2124
Vyas, Amrish / *University of Maryland, Baltimore County, USA* 573
Wagner, Roland / *Johannes Kepler University Linz, Austria* 459
Wai, Shiu Ka / *University of Hong Kong, Hong Kong* 870
Wang, Fu Lee / *City University of Hong Kong, Hong Kong* 2418
Wang, Hsiao-Fan / *National Tsing Hua University, Taiwan ROC* 2924
Wang, Miao-Ling / *Minghsin University of Science & Technology, Taiwan* 2924

Wang, Zhou / *Fraunhofer Integrated Publication and Information Systems Institute (IPSI), Germany* 1103
Wangikar, Lalit / *Inductis India Pvt. Ltd., India* 2862
Warkentin, Merrill / *Mississippi State University, USA* 1246, 2203
Watkins, Andrew / *Mississippi State University, USA* 2896
Weber, Jörg / *Institute for Software Technology, Graz University of Technology, Austria* 597
Wehn Montalvo, Uta / *TNO Strategy, Technology and Policy, The Netherlands* 2440
Weißenberg, Norbert / *Fraunhofer Institute for Software and Systems Engineering, Germany* ... 3404
Welling, Ilary / *Nokia Research Center, Finland* 2475
Weng, Zhiyong / *University of Ottawa, Canada* 712
Westermann, Utz / *VTT Technical Research Centre of Finland, Finland* 1080
Wieloch, Karol / *The Pozañ University of Economics, Poland* 565
Williamson, Nicholas / *University of North Carolina at Greensboro, USA* 1665
Willis, Robert / *Lakehead University, Canada* 1929
Wiśniewski, Marek / *The Poznan University of Economics, Poland* 1562
Wojciechowski, Manfred / *Fraunhofer Institute for Software and Systems Engineering, Germany* 3404
Wong, Eric T. T. / *The Hong Kong Polytechnic University, Hong Kong* 1261
Wotawa, Franz / *Institute for Software Technology, Graz University of Technology, Austria* 597
Wright, David / *University of Ottawa, Canada* 976, 1175
Wright Hawkes, Lois / *Florida State University, USA* 3540
Wu, Ming-Chien / *University of Western Sydney, Australia* 2368
Wu, Tin-Yu / *I-Shou University, Taiwan; National Dong Hwa University, Taiwan* 117, 3349
Wyse, James E. / *Memorial University of Newfoundland, Canada* 3040
Xenakis, Christos / *University of Piraeus, Greece* 2674, 2752
Xie, Xing / *Microsoft Research Asia, China* 497
Xu, Jianliang / *Hong Kong Baptist University, Hong Kong* 3031
Yang, Cheng-Hong / *National Kaohsiung University of Applied Sciences, Taiwan* 3361
Yang, Cheng-Huei / *National Kaohsiung Marine University, Taiwan* 3361
Yang, Christopher C. / *Chinese University of Hong Kong, Hong Kong* 2418
Yang, Chyuan-Huei Thomas / *Hsuan-Chuang University, Taiwan* 1204
Yang, Hung-Jen / *National Kaohsiung Normal University, Taiwan* 534, 1183
Yang, Samuel C. / *California State University, Fullerton, USA* 1615
Yeh, Jyh-haw / *Boise State University, USA* 909, 1183, 1204, 2614
Yeow, P. H. P. / *Multimedia University, Malaysia* 1984
Yiannoutsou, Nikoletta / *University of Patras, Greece* 3282
Yoon, Victoria / *University of Maryland, Baltimore County, USA* 573
Yow, Kin Choong / *Nanyang Technological University, Singapore* 1194
Yu, Betty / *The Chinese University of Hong Kong, Hong Kong* 248
Yuan, Soe-Tsyr / *National Chengchi University, Taiwan* 2530
Yuen, Patrivan K. / *William Carey University, USA* 108
Yuen, Steve Chi-Yin / *The Univeristy of Southern Mississippi, USA* 108
Yulius Limanto, Hanny / *Nanyang Technological University, Singapore* 2896
Žagar, Mario / *University of Zagreb, Croatia* 1516
Zainal Abidin, Mohamad Izani / *Multimedia University, Malaysia* 3511

Zaphiris, Panayiotis / *City University London, UK* 2019
Zavitsanos, Elias / *University of the Aegean and National Centre of Scientific Research "Demokritos", Greece* 2936
Żebrowski, Pawel / *The Poznan University of Economics, Poland* 565, 1562
Zeng, Guangping / *University of Science and Technology of Beijing, China* 473
Zhang, Degan / *University of Science and Technology of Beijing, China* 473
Zhang, Huaiyu / *Northwest University, China* 473
Zhang, Xinshang / *Jidong Oilfield, China* 473
Zhong, Yapin / *Shandong Institute of Physical Education and Sport, China* 909
Zhu, Fangming / *National University of Singapore, Singapore* 881
Zwick, Detlev / *Schulich School of Business, York University, Canada* 1675
Zwierko, Aneta / *Warsaw University of Technology, Poland* 2583

Contents

Volume I

Section I. Fundamental Concepts and Theories

This section serves as the foundation for this exhaustive reference tool by addressing crucial theories essential to the understanding of mobile computing. Chapters found within these pages provide an excellent framework in which to position mobile computing within the field of information science and technology. Individual contributions provide overviews of mobile learning, mobile portals, and mobile government, while also exploring critical stumbling blocks of this field. Within this introductory section, the reader can learn and choose from a compendium of expert research on the elemental theories underscoring the research and application of mobile computing.

Chapter 1.1. Ubiquitous Access to Information Through Portable, Mobile and Handheld Devices /
Ch. Z. Patrikakis, National Technical University of Athens, Greece;
P. Fafali, National Technical University of Athens, Greece;
N. Minogiannis, National Technical University of Athens, Greece;
N. Kourbelis, National Technical University of Athens, Greece.....1

Chapter 1.2. Mobile Computing and Commerce Framework /
Stephanie Teufel, University of Fribourg, Switzerland;
Patrick S. Merten, University of Fribourg, Switzerland;
Martin Steinert, University of Fribourg, Switzerland.....10

Chapter 1.3. Mobile Electronic Commerce /
Samuel Pierre, École Polytechnique de Montréal, Canada 18

Chapter 1.4. Mobile Communications and Mobile Commerce: Conceptual Frames
to Grasp the Global Tectonic Shifts / *Nikhilesh Dholakia, University of Rhode Island, USA;*
Morten Rask, Aarhus School of Business, Denmark;
Ruby Roy Dholakia, University of Rhode Island, USA.....27

Chapter 1.5. Adoption and Diffusion of M-Commerce /
Ranjan B. Kini, Indiana University Northwest, USA;
Subir K. Bandyopadhyay, Indiana University Northwest, USA.....38

Chapter 1.6. Evolution of Mobile Commerce Applications /
George K. Lalopoulos, Hellenic Telecommunications Organization S.A. (OTE), Greece;
Ioannis P. Chochliouros, Hellenic Telecommunications Organization S.A. (OTE), Greece;
Anastasia S. Spiliopoulou-Chochliourou, Hellenic Telecommunications Organization S.A. (OTE), Greece...47

Chapter 1.7. Context-Awareness in Mobile Commerce / *Jun Sun, Texas A&M University, USA;*
Marshall Scott Poole, Texas A&M University, USA.....56

Chapter 1.8. Context as a Necessity in Mobile Applications /
Eleni Christopoulou, University of Patras & Ionian University, Greece 65

Chapter 1.9. A Proposed Framework for Mobile Services Adoption:
A Review of Existing Theories, Extensions, and Future Research Directions /
Indrit Troshani, University of Adelaide, Australia;
Sally Rao Hill, University of Adelaide, Australia 84

Chapter 1.10. Mobile Learning: Learning on the Go / *Steve Chi-Yin Yuen, The Univeristy of*
Southern Mississippi,USA; Patrivan K. Yuen, William Carey University, USA 108

Chapter 1.11. Environments for Mobile Learning /
Han-Chieh Chao, National Dong Hwa University, Taiwan, R.O.C.;
Tin-Yu Wu, National Dong Hwa University, Taiwan, R.O.C.;
Michelle T.C. Kao, National Dong Hwa University, Taiwan, R.O.C......117

Chapter 1.12. Mobile Education: Lessons Learned /
Holger Nösekabel, University of Passau, Germany 122

Chapter 1.13. Anywhere, Anytime Learning Using Highly Mobile Devices /
Mark van 't Hooft, Kent State University, USA;
Graham Brown-Martin, Handheld Learning, London, UK;
Karen Swan, Kent State University, USA.....144

Chapter 1.14. Current Status of Mobile Wireless Technology and Digital Multimedia Broadcasting /
J. P. Shim, Mississippi State University, USA; Kyungmo Ahn, Kyunghee University, Korea;
Julie M. Shim, Soldier Design LLC, USA.....152

Chapter 1.15. Mobile Portals / *Ofir Turel, California State University, USA;*
Alexander Serenko, Lakehead University, Canada.....171

Chapter 1.16. Mobile Portals as Innovations / *Alexander Serenko, Lakehead University, Canada;*
Ofir Turel, California State University, Fullerton, USA 181

Chapter 1.17. Mobile Portals for Knowledge Management /
Hans Lehmann, Victoria University of Wellington, New Zealand;
Ulrich Remus, University of Erlangen-Nuremberg, Germany;
Stefan Berger, Detecon International GmbH, Germany 188

Chapter 1.18. Mobile Knowledge Management / *Volker Derballa, University of Augsburg, Germany; Key Pousttchi, University of Augsburg, Germany* 197

Chapter 1.19. Assessing Human Mobile Computing Performance by Fitts' Law /
Thomas Alexander, FGAN - Research Institute for Communication, Information Processing, and Ergonomics, Germany; Christopher Schlick, RWTH Aachen University, Germany;
Alexander Sievert, German Sport University Cologne, Germany;
Dieter Leyk, German Sport University Cologne, Germany & Central Institute of the Federal Armed Forces Medical Services, Koblenz, Germany 206

Chapter 1.20. Evaluating Mobile Human-Computer Interaction /
Chris Baber, The University of Birmingham, UK 225

Chapter 1.21. Mobile Public Relations Strategies / *Chris Galloway, Monash University, Australia* 240

Chapter 1.22. Introducing Mobile Government /
M. Halid Kuscu, Mobile Government Consortium International, UK;
Ibrahim Kushchu, Mobile Government Consortium International, UK;
Betty Yu, The Chinese University of Hong Kong, Hong Kong 248

Chapter 1.23. Key Issues in Mobile Marketing: Permission and Acceptance /
Stuart J. Barnes, University of East Anglia, UK;
Eusebio Scornavacca, Victoria University of Wellington, New Zealand 257

Chapter 1.24. Dynamic Pricing Based on Net Cost for Mobile Content Services /
Nopparat Srikhutkhao, Kasetsart University, Thailand;
Sukumal Kitisin, Kasetsart University, Thailand 269

Chapter 1.25. A Technology Intervention Perspective of Mobile Marketing /
Dennis Lee, The University of Queensland, Australia & The Australian CRC for Interaction Design, Australia; Ralf Muhlberger, The University of Queensland, Australia & The Australian CRC for Interaction Design, Australia 279

Chapter 1.26. Definitions, Key Characteristics, and Generations of Mobile Games /
Eui Jun Jeong, Michigan State University, USA;
Dan J. Kim, University of Houston Clear Lake, USA 289

Chapter 1.27. Mobile Agents /
Kamel Karoui, Institut National des Sciences Appliquées de Tunis, Tunisia 296

Chapter 1.28. Protection of Mobile Agent Data / *Sheng-Uei Guan, Brunel University, UK* 305

Chapter 1.29. Indexing Mobile Objects: An Overview of Contemporary Solutions /
Panayiotis Bozanis, University of Thessaly, Greece 313

Chapter 1.30. Database Queries in Mobile Environments /
N. Marsit, IRIT—Paul Sabatier University, France;
A. Hameurlain, IRIT—Paul Sabatier University, France;
Z. Mammeri, IRIT—Paul Sabatier University, France;
F. Morvan, IRIT—Paul Sabatier University, France 334

Chapter 1.31. A Taxonomy of Database Operations on Mobile Devices /
Say Ying Lim, Monash University, Australia;
David Taniar, Monash University, Australia;
Bala Srinivasan, Monash University, Australia 350

Chapter 1.32. Addressing the Credibility of Mobile Applications /
Pankaj Kamthan, Concordia University, Canada 372

Section II. Development and Design Methodologies

This section provides in-depth coverage of conceptual architectures, frameworks and methodologies related to the design and implementation of mobile systems and technologies. Throughout these contributions, research fundamentals in the discipline are presented and discussed. From broad examinations to specific discussions on particular frameworks and infrastructures, the research found within this section spans the discipline while also offering detailed, specific discussions. Basic designs, as well as abstract developments, are explained within these chapters, and frameworks for designing successful mobile applications, interfaces, and agents are discussed.

Chapter 2.1. Developing Smart Client Mobile Applications /
Jason Gan, University of Technology, Australia 383

Chapter 2.2. Engineering Wireless Mobile Applications /
Qusay H. Mahmoud, University of Guelph, Canada;
Zakaria Maamar, Zayed University, UAE 388

Chapter 2.3. Conceptual Framework for Mobile-Based Application in Healthcare /
Matthew W. Guah, School of Business Economics, Erasmus University Rotterdam, The Netherlands 403

Chapter 2.4. Design of an Enhanced 3G-Based Mobile Healthcare System /
Julián Fernández Navajas, University of Zaragoza, Spain;
Antonio Valdovinos Bardají, University of Zaragoza, Spain;
Robert S. H. Istepanian, Kingston University, UK;
José García Moros, University of Zaragoza, Spain
José Ruiz Mas, University of Zaragoza, Spain;
Eduardo Antonio Viruete Navarro, University of Zaragoza, Spain;
Carolina Hernández Ramos, University of Zaragoza, Spain;
Álvaro Alesanco Iglesias, University of Zaragoza, Spain; 419

Chapter 2.5. The M-Health Reference Model: An Organizing Framework
for Conceptualizing Mobile Health Systems / *Phillip Olla, Madonna University, USA;*
Joseph Tan, Wayne State University, USA 432

Chapter 2.6. Design Methodology for Mobile Information Systems /
Zakaria Maamar, Zayed University, UAE;
Qusay H. Mahmoud, University of Guelph, Canada 451

Chapter 2.7. Distribution Patterns for Mobile Internet Applications /
Roland Wagner, Johannes Kepler University Linz, Austria;
Franz Gruber, RISC Software GmbH, Austria;
Werner Hartmann, FAW Software Engineering GmbH, Austria.....459

Chapter 2.8. Web-Based Seamless Migration for Task-Oriented Mobile Distance Learning /
Degan Zhang, University of Science and Technology of Beijing, China;
Yuan-chao Li, China University of Petroleum, P.R. China;
Huaiyu Zhang, Northwest University, China;
Xinshang Zhang, Jidong Oilfield, P.R. China;
Guangping Zeng, University of Science and Technology of Beijing, China.....473

Chapter 2.9. TCP Enhancements for Mobile Internet / *Bhaskar Sardar, Jadavpur University, India;*
Debashis Saha, Indian Institute of Management (IIM) Calcutta, India 488

Chapter 2.10. A Cooperative Framework for Information Browsing in Mobile Environment /
Zhigang Hua, Chinese Academy of Sciences, China; Xing Xie, Microsoft Research Asia, China;
Hanqing Lu, Chinese Academy of Sciences, China; Wei-Ying Ma, Microsoft Research Asia, China 497

Chapter 2.11. Describing the Critical Factors for Creating Successful Mobile Data Services /
Anne Tseng, Helsinki School of Economics, Finland;
Jukka Kallio, Helsinki School of Economics, Finland;
Markku Tinnilä, Helsinki School of Economics, Finland 506

Chapter 2.12. A Design Framework for Mobile Collaboration /
Pedro Antunes, University of Lisboa, Portugal 518

Chapter 2.13. Interface Design Issues for Mobile Commerce /
Susy S. Chan, DePaul University, USA; Xiaowen Fang, DePaul University, USA.....526

Chapter 2.14. Handheld Computing and Palm OS Programming for Mobile Commerce /
Wen-Chen Hu, University of North Dakota, USA;
Lixin Fu, The University of North Carolina at Greensboro, USA;
Hung-Jen Yang, National Kaohsiung Normal University, Taiwan;
Sheng-Chien Lee, University of Florida, USA.....534

Chapter 2.15. Privacy-Preserving Transactions Protocol Using Mobile Agents with Mutual Authentication / *Song Han, Curtin University of Technology, Australia; Vidyasagar Potdar, Curtin University of Technology, Australia; Elizabeth Chang, Curtin University of Technology, Australia; Tharam Dillon, University of Technology, Australia* 546

Chapter 2.16. Robust Algorithms for DOA Estimation and Adaptive Beamforming in Wireless Mobile Communications / *R.M. Shubair, Etisalat University College, UAE; K.O. AlMidfa, Etisalat University College, UAE; A. Al-Marri, Etisalat University College, UAE; M. Al-Nuaimi, Etisalat University College, UAE* 558

Chapter 2.17. Mobile Information Filtering / *Witold Abramowicz, The Pozañ University of Economics, Poland; Krzysztof Banaśkiewicz, The Pozañ University of Economics, Poland; Karol Wieloch, The Pozañ University of Economics, Poland; Paweł Żebrowski, The Pozañ University of Economics, Poland* 565

Volume II

Chapter 2.18. Information Management in Mobile Environments Using a Location-Aware Intelligent Agent System / *Amrish Vyas, University of Maryland, Baltimore County, USA; Victoria Yoon, University of Maryland, Baltimore County, USA* 573

Chapter 2.19. Topology for Intelligent Mobile Computing / *Robert Statica, New Jersey Institute of Technology, USA; Fadi P. Deek, New Jersey Institute of Technology, USA* 589

Chapter 2.20. Robust Intelligent Control of Mobile Robots / *Gordon Fraser, Institute for Software Technology, Graz University of Technology, Austria; Gerald Steinbauer, Institute for Software Technology, Graz University of Technology, Austria; Jörg Weber, Institute for Software Technology, Graz University of Technology, Austria; Franz Wotawa, Institute for Software Technology, Graz University of Technology, Austria* 597

Chapter 2.21. A Neural Network-Based Mobile Architecture for Mobile Agents / *Anand Kuppuswami, University of Western Sydney, Australia* 618

Chapter 2.22. Semantic Web Services for Smart Devices Based on Mobile Agents / *Vagan Terziyan, University of Jyvaskyla, Finland* 630

Chapter 2.23. Towards Autonomic Infrastructures via Mobile Agents and Active Networks / *Stamatis Karnouskos, SAP Research, Germany* 642

Chapter 2.24. Mobility Management in Mobile Computing and Networking Environments / *Samuel Pierre, Ecole Polytechnique de Montreal, Canada* 650

Chapter 2.25. Location Area Design Algorithms for Minimizing Signalling Costs in Mobile Networks / *Vilmos Simon, Budapest University of Technology and Economics, Hungary; Sándor Imre, Budapest University of Technology and Economics, Hungary* 682

Chapter 2.26. Market Configuration and the Success of Mobile Services: Lessons From Japan and Finland / *Jarkko Vesa, Helsinki School of Economics, Finland*.....696

Chapter 2.27. A Mobile Intelligent Agent-Based Architecture for E-Business / *Zhiyong Weng, University of Ottawa, Canada; Thomas Tran, University of Ottawa, Canada*.....712

Chapter 2.28. A Framework for Information Systems Integration in Mobile Working Environments / *Javier García-Guzmán, Universidad Carlos III de Madrid, Spain; María-Isabel Sánchez-Segura, Universidad Carlos III de Madrid, Spain; Antonio de Amescua-Seco, Universidad Carlos III de Madrid, Spain; Mariano Navarro, TRAGSA Group Information, Spain* 729

Chapter 2.29. "It's the Mobility, Stupid": Designing Mobile Government / *Klas Roggenkamp, Dipl. Designer Electronic Business, Germany* 756

Chapter 2.30. Design of Government Information for Access by Wireless Mobile Technology / *Mohamed Ally, Athabasca University, Canada* 776

Section III. Tools and Technologies

This section presents extensive coverage of the technology that both derives from and informs mobile computing. These chapters provide an in-depth analysis of the use and development of innumerable devices and tools, while also providing insight into new and upcoming technologies, theories, and instruments that will soon be commonplace. Within these rigorously researched chapters, readers are presented with examples of the tools that facilitate and support mobile computing. In addition, the successful implementation and resulting impact of these various tools and technologies are discussed within this collection of chapters.

Chapter 3.1. Evaluation of Mobile Technologies in the Context of Their Applications, Limitations, and Transformation / *Abbass Ghanbary, University of Western Sydney, Australia*.....785

Chapter 3.2. Knowledge Representation in Semantic Mobile Applications / *Pankaj Kamthan, Concordia University, Canada* 796

Chapter 3.3. Mobile Portal Technologies and Business Models / *David Parsons, Massey University, New Zealand*.....805

Chapter 3.4. Mobile Learning Technologies / *Diane M. Gayeski, Ithaca College, USA*.....811

Chapter 3.5. Enhancing Learning Through Mobile Computing /
Marsha Berry, RMIT University, Australia;
Margaret Hamilton, RMIT University, Australia;
Naomi Herzog, RMIT University, Australia;
Lin Padgham, RMIT University, Australia;
Ron Van Schyndel, RMIT University, Australia 817

Chapter 3.6. Mobile Technology and its Applications in Instructional Conversation /
Jason Caudill, Independent Consultant, USA..... 835

Chapter 3.7. Embedded Agents for Mobile Services / *John F. Bradley, University College Dublin, Ireland; Conor Muldoon, University College Dublin, Ireland;*
Gregory M. P. O'Hare, University College Dublin, Ireland;
Michael J. O'Grady, University College Dublin, Ireland..... 850

Chapter 3.8. A Database Service Discovery Model for Mobile Agents /
Lei Song, University of Guelph, Guelph, Canada;
Xining Li, University of Guelph, Guelph, Canada;
Jingbo Ni, University of Guelph, Guelph, Canada..... 858

Chapter 3.9. Databases for Mobile Applications / *Indranil Bose, University of Hong Kong, Hong Kong;*
Wang Ping, University of Hong Kong, Hong Kong;
Mok Wai Shan, University of Hong Kong, Hong Kong;
Wong Ka Shing, University of Hong Kong, Hong Kong;
Yip Yee Shing, University of Hong Kong, Hong Kong;
Chan Lit Tin, University of Hong Kong, Hong Kong;
Shiu Ka Wai, University of Hong Kong, Hong Kong..... 870

Chapter 3.10. A Virtual Community for Mobile Agents / *Sheng-Uei Guan, Brunel University, UK;*
Fangming Zhu, National University of Singapore, Singapore..... 881

Chapter 3.11. Concepts and Operations of Two Research Projects
on Web Services and Mobile Web Services /
Zakaria Maamar, Zayed University, United Arab Emirates 891

Chapter 3.12. Handheld Computing and J2ME Programming for Mobile Handheld Devices /
Wen-Chen Hu, University of North Dakota, USA;
Jyh-haw Yeh, Boise State University, USA;
I-Lung Kao, IBM, USA;
Yapin Zhong, Shandong Institute of Physical Education and Sport, China 909

Chapter 3.13. Tools for Rapidly Prototyping Mobile Interactions /
Yang Li, University of Washington, USA; Scott Klemmer, Stanford University, USA;
James A. Landay, University of Washington & Intel Research Seattle, USA 920

Chapter 3.14. Real-Time 3D Design Modelling of Outdoor Structures
Using Mobile Augmented Reality Systems /
Wayne Piekarski, University of South Australia, Australia 937

Chapter 3.15. Mobile Ad Hoc Network / *Subhankar Dhar, San Jose State University, USA*.....952

Chapter 3.16. Convergence Technology for Enabling Technologies /
G. Sivaradje, Pondicherry Engineering College, India;
I. Saravanan, Pondicherry Engineering College, India;
P. Dananjayan, Pondicherry Engineering College, India 961

Chapter 3.17. Document Management, Organizational Memory, and Mobile Environment /
Sari Mäkinen, University of Tampere, Finland.....968

Chapter 3.18. Business and Technology Issues in Wireless Networking /
David Wright, University of Ottawa, Canada.....976

Chapter 3.19. Mobile Phone Based Augmented Reality /
Anders Henrysson, Norrköping Visualisation and Interaction Studio, Sweden;
Mark Ollila, Norrköping Visualisation and Interaction Studio, Sweden;
Mark Billinghurst, Human Interface Technology Laboratory, New Zealand 984

Chapter 3.20. Pen-Based Mobile Computing /
Bernie Garret, University of British Columbia, Canada 998

Chapter 3.21. The Smart Card in Mobile Communications:
Enabler of Next-Generation (NG) Services /
Claus Dietze, The European Telecommunications Standards Institute (ETSI), France 1004

Chapter 3.22. Unobtrusive Movement Interaction for Mobile Devices /
Panu Korpipää, Finwe Ltd., Finland; Jukka Linjama, Nokia, Finland;
Juha Kela, Finwe Ltd., Finland; Tapani Rantakokko, Finwe Ltd., Finland 1029

Chapter 3.23. Positioning Technologies for Mobile Computing /
Michael J. O'Grady, University College Dublin, Ireland;
Gregory M. P. O'Hare, University College Dublin, Ireland.....1047

Chapter 3.24. Emerging Mobile Technology and Supply Chain Integration:
Using RFID to Streamline the Integrated Supply Chain /
Richard Schilhavy, University of North Carolina at Greensboro, USA;
A. F. Salam, University of North Carolina at Greensboro, USA.....1053

Chapter 3.25. Content Personalization for Mobile Interfaces /
Spiridoula Koukia, University of Patras, Greece;
Maria Rigou, University of Patras, Greece & Research Academic Computer Technology Institute, Greece; Spiros Sirmakessis, Technological Institution of Messolongi, Greece & Research Academic Computer Technology Institute, Greece 1064

Chapter 3.26. Distributed Mobile Services and Interfaces
for People Suffering from Cognitive Deficits / *Sylvain Giroux, Université de Sherbrooke, Canada;*
Hélène Pigot, Université de Sherbrooke, Canada;
Jean-François Moreau, Université de Sherbrooke, Canada;
Jean-Pierre Savary, Division R&D CRD, France 1069

Chapter 3.27. Context-Aware Mobile Capture and Sharing of Video Clips /
Janne Lahti, VTT Technical Research Centre of Finland, Finland;
Utz Westermann, VTT Technical Research Centre of Finland, Finland;
Marko Palola, VTT Technical Research Centre of Finland, Finland;
Johannes Peltola, VTT Technical Research Centre of Finland, Finland;
Elena Vildjiounaite, VTT Technical Research Centre of Finland, Finland 1080

Chapter 3.28. From CCTV to Biometrics through Mobile Surveillance /
Jason Gallo, Northwestern University, USA..... 1096

Chapter 3.29. Discovering Multimedia Services and Contents in Mobile Environments /
Zhou Wang, Fraunhofer Integrated Publication and Information Systems Institute (IPSI),
Germany; Hend Koubaa, Norwegian University of Science and Technology (NTNU), Norway 1103

Chapter 3.30. DRM Technology for Mobile Multimedia /
Sai Ho Kwok, California State University, Long Beach, USA..... 1117

Chapter 3.31. V-Card: Mobile Multimedia for Mobile Marketing /
Holger Nösekabel, University of Passau, Germany;
Wolfgang Röckelein, EMPRISE Consulting Düseldorf, Germany..... 1125

Chapter 3.32. Acoustic Data Communication with Mobile Devices /
Victor I. Khashchanskiy, First Hop Ltd., Finland;
Andrei L. Kustov, First Hop Ltd., Finland 1135

Chapter 3.33. The Design of Mobile Television in Europe /
Pieter Ballon, Vrije Universiteit Brussel, Belgium;
Olivier Braet, Vrije Universiteit Brussel, Belgium 1143

Chapter 3.34. The MP3 Player as a Mobile Digital Music Collection Portal /
David Beer, University of York, UK 1168

Volume III

Chapter 3.35. Wireless Technologies for Mobile Computing and Commerce /
David Wright, University of Ottawa, Canada..... 1175

Chapter 3.36. Mobile Handheld Devices for Mobile Commerce /
Wen-Chen Hu, University of North Dakota, USA; Jyh-haw Yeh, Boise State University, USA;
Hung-Jen Yang, National Kaohsiung Normal University, Taiwan;
Chung-wei Lee, Auburn University, USA..... 1183

Chapter 3.37. Mobile Commerce Multimedia Messaging Peer /
Kin Choong Yow, Nanyang Technological University, Singapore;
Nitin Mittal, Nokia Pte Ltd, Singapore 1194

Chapter 3.38. Mobile and Electronic Commerce Systems and Technologies /
Wen-Chen Hu, University of North Dakota, USA;
Chyuan-Huei Thomas Yang, Hsuan-Chuang University, Taiwan;
Jyh-haw Yeh, Boise State University, USA; Weihong Hu, Auburn University, USA 1204

Chapter 3.39. E-Commerce Services Based on Mobile Agents /
Giancarlo Fortino, DEIS, University of Calabria, Italy;
Alfredo Garro, DEIS, University of Calabria, Italy;
Wilma Russo, DEIS, University of Calabria, Italy 1226

Chapter 3.40. B-POS Secure Mobile Payment System /
Antonio Grillo, Universita di Roma "Tor Vergata", Italy;
Alessandro Lentini, Universita di Roma "Tor Vergata", Italy;
Gianluigi Me, Universita di Roma "Tor Vergata", Italy 1237

Chapter 3.41. Mobile Banking Systems and Technologies /
Cheon-Pyo Lee, Mississippi State University, USA;
Merrill Warkentin, Mississippi State University, USA 1246

Chapter 3.42. Mobile Clinical Learning Tools Using Networked Personal Digital Assistants (PDAs) / *Bernard Mark Garrett, University of British Columbia, Canada* 1256

Chapter 3.43. 3G Mobile Medical Image Viewing /
Eric T. T. Wong, The Hong Kong Polytechnic University, Hong Kong;
Carrison K. S. Tong, Pamela Youde Nethersole Eastern Hospital, Hong Kong 1261

Section IV. Utilization and Application

This section introduces and discusses the ways in which information technology has been used to shape the realm of mobile computing and proposes new ways in which IT-related innovations can be implemented within organizations and in society as a whole. These particular selections highlight, among other topics, the implementation of mobile technology in healthcare settings, and the evolution of mobile commerce. Contributions included in this section provide excellent coverage of today's mobile environment and insight into how mobile computing impacts the fabric of our present-day global village.

Chapter 4.1. Dynamics of Mobile Service Adoption /
Hannu Verkasalo, Helsinki University of Technology, Finland 1273

Chapter 4.2. Exploring the Use of Mobile Data Services in Europe:
The Cases of Denmark and Greece / *Ioanna D. Constantiou, Copenhagen Business School, Denmark; Maria Bina, Athens University of Economics and Business, Greece* 1296

Chapter 4.3. The Mobile Phone Telecommunications Service Sector in China /
Michelle W. L. Fong, Victoria University, Australia 1312

Chapter 4.4. United States of America: Renewed Race for Mobile Services /
Mats Samuelsson, Mobio Networks, USA; Nikhilesh Dholakia, University of Rhode Island, USA;
Sanjeev Sardana, Mobio Networks, USA 1331

Chapter 4.5. M-Learning with Mobile Phones /
Simon So, Hong Kong Institute of Education, Hong Kong 1344

Chapter 4.6. Using Mobile Communication Technology in Student Mentoring /
Jonna Häkkilä, University of Oulu, Finland; Jenine Beekhuyzen, Griffith University, Australia 1351

Chapter 4.7. A Mobile Portal for Academe /
Hans Lehmann, Victoria University of Wellington, New Zealand;
Stefan Berger, Detecon International GmbH, Germany;
Ulrich Remus, University of Erlangen-Nuremberg, Germany 1359

Chapter 4.8. Accessing Learning Content in a Mobile System: Does Mobile Mean Always Connected? / *Anna Trifonova, University of Trento, Italy* 1367

Chapter 4.9. Using Learning Objects for Rapid Deployment to Mobile Learning Devices for the U.S. Coast Guard /
Pamela T. Northrup, University of West Florida, USA;
William T. Harrison Jr., University of West Florida, USA & U.S. Navy, USA 1381

Chapter 4.10. Using Mobile Phones and PDAs in Ad Hoc Audience Response Systems /
Matt Jones, University of Waikato, New Zealand;
Gary Marsden, University of Cape Town, South Africa;
Dominic Gruijters, University of Cape Town, South Africa 1396

Chapter 4.11. Perception of Mobile Technology Provision in Health Service /
Astrid M. Oddershede, University of Santiago of Chile, Chile;
Rolando A. Carrasco, University of Newcastle-upon-Tyne, UK 1408

Chapter 4.12. Relevance of Mobile Computing in the Field of Medicine /
Henrique M. G. Martins, University of Cambridge, UK;
Matthew R. Jones, University of Cambridge, UK 1429

Chapter 4.13. Integrating Mobile-Based Systems with Healthcare Databases /
Yu Jiao, Oak Ridge National Laboratory, USA;
Ali R. Hurson, Pennsylvania State University, USA;
Thomas E. Potok, Oak Ridge National Laboratory, USA;
Barbara G. Beckerman, Oak Ridge National Laboratory, USA 1442

Chapter 4.14. Adoption of Mobile Technology in the Supply Chain:
An Exploratory Cross-Case Analysis / *Bill Doolin, Auckland University of Technology, New Zealand; Eman Al Haj Ali, Higher Colleges of Technology, UAE* 1466

Chapter 4.15. Enabling the Glass Pipeline: The Infusion of Mobile Technology Applications in Supply Chain Management / *Umar Ruhi, Wilfrid Laurier University, Canada; Ofir Turel, McMaster University, Canada* 1483

Chapter 4.16. Mobile Automotive Cooperative Services (MACS):
Systematic Development of Personalizable Interactive Mobile Automotive Services / *Holger Hoffman, Technische Universität München, Germany; Jan Marco Leimeister, Technische Universität München, Germany; Helmut Krcmar, Technische Universität München, Germany* 1499

Chapter 4.17. Using the Railway Mobile Terminalsin the Process of Validation and Vending Tickets / *Marko Horvat, Croatian Railways Ltd., Croatia; Mario Žagar, University of Zagreb, Croatia* 1516

Chapter 4.18. An Evaluation of U.S. City Government Wireless Networks for Mobile Internet Access / *Ben Coaker, Whiting-Turner Contracting Company, USA; Candace Deans, University of Richmond, USA* 1530

Chapter 4.19. The Prospects of Mobile Government in Jordan:
An Evaluation of Different Delivery Platforms / *Ala M. Abu-Samaha, Amman University, Jordan; Yara Abdel Samad, Ministry of Information & Communication Technologies, Jordan* 1543

Chapter 4.20. Usability Driven Open Platform for Mobile Government (USE-ME.GOV) / *Paul Moore Olmstead, Atos Research and Innovation, Spain; Gertraud Peinel, Fraunhofer FIT, Germany; Dirk Tilsner, EDISOFT, Portugal; Witold Abramowicz, The Poznan University of Economics, Poland; Andrzej Bassara, The Poznan University of Economics, Poland; Agata Filipowska, The Poznan University of Economics, Poland; Marek Wiśniewski, The Poznan University of Economics, Poland; Pawel Żebrowski, The Poznan University of Economics, Poland* 1562

Chapter 4.21. Mobile Computing for M-Commerce / *Anastasis Sofokleous, Brunel University, UK; Marios C. Angelides, Brunel University, UK; Christos Schizas, University of Cyprus, Cyprus* 1584

Chapter 4.22. Mobile Commerce Applications and Adoption / *Krassie Petrova, Auckland University of Technology, New Zealand* 1593

Chapter 4.23. Mobile Computing: An Enabler in International Financial Services / *N. Raghavendra Rao, SSN School of Management & Computer Applications, India* 1602

Chapter 4.24. E-Commerce and Mobile Commerce Applications Adoptions / *Charlie Chen, Appalachian State University, USA; Samuel C. Yang, California State University, Fullerton, USA* 1615

Chapter 4.25. Consumer and Merchant Adoption of Mobile Payment Solutions /
Niina Mallat, Helsinki School of Economics, Finland;
Tomi Dahlberg, Helsinki School of Economics, Finland 1626

Chapter 4.26. An Electronic Auction Service Framework Based on Mobile Software Agents /
Sheng-Uei Guan, National University of Singapore, Singapore 1640

Chapter 4.27. Mobile Advertising: A European Perspective /
Tawfik Jelassi, Ecole Nationale des Ponts et Chaussées, France;
Albrecht Enders, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany 1653

Chapter 4.28. China: M-Commerce in World's Largest Mobile Market /
Nir Kshetri, University of North Carolina at Greensboro, USA;
Nicholas Williamson, University of North Carolina at Greensboro, USA;
David L. Bourgoin, University of Hawaii at Manoa, USA 1665

Chapter 4.29. Canada: Mobile Commerce Under Construction / *Detlev Zwick, Schulich School of Business, York University, Canada* 1675

Chapter 4.30. Mobile Commerce in South Africa /
Anesh Maniraj Singh, University of KwaZulu-Natal, South Africa 1690

Chapter 4.31. Mobile Payment Issues and Policy Implications: The Case of Korea /
Youngsun Kwon, Information and Communications University, Republic of Korea;
Changi Nam, Information and Communications University, Republic of Korea 1699

Chapter 4.32. Payment Mechanism of Mobile Agent-Based Restaurant Ordering System /
Jon T. S. Quah, Nanyang Technological University, Singapore;
Winnie C. H. Leow, Singapore Polytechnic, Singapore;
Chee Chye Ong, Nanyang Technological University, Singapore 1713

Chapter 4.33. Structural Effects of Platform Certification on a Complementary Product Market: The Case of Mobile Applications / *Ankur Tarnacha, Pennsylvania State University, USA;*
Carleen Maitland, Pennsylvania State University, USA 1721

Chapter 4.34. Buongiorno! My Alert: Creating a Market to Develop a Mobile Business /
Guillermo de Haro, Instituto De Empresa, Spain;
José María García, Instituto De Empresa, Spain 1738

Chapter 4.35. Location-Based Services in the Mobile Communications Industry /
Christopher Ververidis, Athens University of Economics and Business, Greece;
George C. Polyzos, Athens University of Economics and Business, Greece 1754

Section V. Organizational and Social Implications

This section includes a wide range of research pertaining to the social and organizational impact of mobile computing around the world. Chapters introducing this section analyze mobile virtual communities and consumer attitudes toward mobile marketing, while later contributions offer an extensive analysis of the accessibility of mobile applications and technologies. The inquiries and methods presented in this section offer insight into the implications of mobile computing at both a personal and organizational level, while also emphasizing potential areas of study within the discipline.

Chapter 5.1. Mobile Virtual Communities / *Glauber Ferreira, Federal University of Campina Grande, Brazil; Hyggo Almeida, Federal University of Campina Grande, Brazil; Angelo Perkusich, Federal University of Campina Grande, Brazil; Evandro Costa, Federal University of Alagoas, Brazil* 1763

Chapter 5.2. Mobile Virtual Communities of Commuters / *Jalal Kawash, American University of Sharjah, UAE; Christo El Morr, York University, Canada; Hamza Taha, American University of Sharjah, UAE; Wissam Charaf, American University of Sharjah, UAE* 1771

Chapter 5.3. Wireless Local Communities in Mobile Commerce / *Jun Sun, University of Texas–Pan American, USA* 1780

Volume IV

Chapter 5.4. From Communities to Mobile Communities of Values / *Patricia McManus, Edith Cowan University, Australia; Craig Standing, Edith Cowan University, Australia* 1788

Chapter 5.5. Economics of Immediate Gratification in Mobile Commerce / *Kerem Tomak, University of Texas at Austin, USA* 1796

Chapter 5.6. Consumer Perceptions and Attitudes Towards Mobile Marketing / *Amy Carroll, Victoria University of Wellington, New Zealand; Stuart J. Barnes, University of East Anglia, UK; Eusebio Scornavacca, Victoria University of Wellington, New Zealand* 1810

Chapter 5.7. An Empirical Examination of Customer Perceptions of Mobile Advertising / *Su-Fang Lee, Overseas Chinese Institute of Technology, Taiwan; Yuan-Cheng Tsai, Da-Yeh University, Taiwan; Wen-Jang (Kenny) Jih, Middle Tennessee State University, USA* 1823

Chapter 5.8. Effects of Consumer-Perceived Convenience on Shopping Intention in Mobile Commerce: An Empirical Study / *Wen-Jang (Kenny) Jih, Middle Tennessee State University, USA* 1840

Chapter 5.9. Factors Influencing Segmentation and Demographics of Mobile-Customers /
Anne-Marie Ranft, University of Technology, Australia 1857

Chapter 5.10. Identified Customer Requirements in Mobile Video Markets – A Pan-European Case /
Torsten Brodt, University of St. Gallen, Switzerland 1867

Chapter 5.11. Special Features of Mobile Advertising and Their Utilization /
Jari Salo, University of Oulu, Finland; Jaana Tähtinen, University of Oulu, Finland 1878

Chapter 5.12. Personalization and Customer Satisfaction in Mobile Commerce /
HY Sonya Hsu, Southern Illinois University, USA; Songpol Kulviwat, Hofstra University, USA 1886

Chapter 5.13. Cross-Cultural Consumer Perceptions of Advertising via Mobile Devices:
Some Evidence from Europe and Japan / *Parissa Haghirian, Sophia University, Japan;*
Maria Madlberger, Vienna University of Economics and Business Administration, Austria 1893

Chapter 5.14. Do Mobile CRM Services Appeal to Loyalty Program Customers? /
Veronica Liljander, Swedish School of Economics and Business Administration, Finland;
Pia Polsa, Swedish School of Economics and Business Administration, Finland;
Kim Forsberg, Intrum Justitia Finland, Finland 1911

Chapter 5.15. Contractual Obligations Between Mobile Service Providers and Users /
Robert Willis, Lakehead University, Canada; Alexander Serenko, Lakehead University, Canada;
Ofir Turel, McMaster University, Canada 1929

Chapter 5.16. Accessibility of Mobile Applications /
Pankaj Kamthan, Concordia University, Canada 1937

Chapter 5.17 Propagating the Ideal: The Mobile Communication Paradox /
Imar de Vries, Utrecht University, The Netherlands 1946

Chapter 5.18. Portals Supporting a Mobile Learning Environment /
Paul Crowther, Sheffield Hallam University, UK;
Martin Beer, Sheffield Hallam University, UK 1960

Chapter 5.19. Secure Collaborative Learning Practices and Mobile Technology /
Hannakaisa Isomäki, University of Jyväskylä, Finland;
Kirsi Päykkönen, University of Lapland, Finland;
Hanna Räisänen, University of Lapland, Finland 1967

Chapter 5.20. Gender Difference in the Motivations of Mobile Internet Usage /
Shintaro Okazaki, Autonomous University of Madrid, Spain 1975

Chapter 5.21. Hand Measurements and Gender Effect on Mobile Phone Messaging Satisfaction:
A Study Based on Keypad Design Factors / *Vimala Balakrishnan, Multimedia University,*
Malaysia; P. H. P. Yeow, Multimedia University, Malaysia 1984

Chapter 5.22. User Acceptance of Mobile Services /
Eija Kaasinen, VTT Technical Research Centre of Finland, Finland.....1996

Chapter 5.23. User-Centered Mobile Computing / *Dean Mohamedally, City University London, UK; Panayiotis Zaphiris, City University London, UK; Helen Petrie, City University London, UK*.....2019

Chapter 5.24. User Experience of Camera Phones in Social Contexts /
Hanna Stelmaszewska, Middlesex University, UK; Bob Fields, Middlesex University, UK; Ann Blandford, University College London, UK.....2027

Chapter 5.25. Mobile Evaluations in a Lab Environment /
Murray Crease, National Research Council of Canada, Canada; Robert Longworth, University of New Brunswick, Canada.....2042

Chapter 5.26. Mobile E-Work to Support Regional and Rural Communities /
Sirkka Heinonen, VTT Building and Transport, Finland.....2061

Chapter 5.27. Mobile Phone and Autonomy / *Theptawee Chokvasin, Suranaree University of Technology, Thailand*.....2066

Chapter 5.28. The Sociotechnical Nature of Mobile Computing Work: Evidence from a Study of Policing in the United States / *Steve Sawyer, The Pennsylvania State University, USA; Andrea Tapia, The Pennsylvania State University, USA*.....2079

Chapter 5.29. Social Context for Mobile Computing Device Adoption and Diffusion: A Proposed Research Model and Key Research Issues /
Andrew P. Ciganek, University of Wisconsin-Milwaukee, USA; K. Ramamurthy, University of Wisconsin-Milwaukee, USA.....2092

Chapter 5.30. Mobile Phone Use Across Cultures: A Comparison Between the United Kingdom and Sudan / *Ishraga Khattab, Brunel University, UK; Steve Love, Brunel University, UK*.....2110

Chapter 5.31. Mobile Phone Communication Innovation in Multiple Time and Space Zones: The Case of Hong Kong Culture / *Shirley Chan, City University of Hong Kong, Hong Kong; Douglas Vogel, City University of Hong Kong, Hong Kong; Louis C. K. Ma, City University of Hong Kong, Hong Kong*.....2124

Chapter 5.32. Mobile Networked Text Communication: The Case of SMS and Its Infuence on Social Interaction / *Louise Barkhuus, University of Glasgow, UK*.....2130

Section VI. Managerial Impact

This section presents contemporary coverage of the managerial implications of mobile computing. Particular contributions address business strategies for mobile marketing, mobile customer services, and mobile service business opportunities. The managerial research provided in this section allows executives, practitioners, and researchers to gain a better sense of how mobile computing can inform their practices and behavior.

Chapter 6.1. Comprehensive Impact of Mobile Technology on Business /
Khimji Vaghjiani, K & J Business Solutions, Australia;
Jenny Teoh, K & J Business Solutions, Australia 2145

Chapter 6.2. Mobile Business Applications / *Cheon-Pyo Lee, Carson-Newman College, USA* 2163

Chapter 6.3. Business Model Typology for Mobile Commerce /
Volker Derballa, Universität Augsburg, Germany;
Key Pousttchi, Universität Augsburg, Germany;
Klaus Turowski, Universität Augsburg, Germany 2169

Chapter 6.4. Business Strategies for Mobile Marketing /
Indranil Bose, University of Hong Kong, Hong Kong;
Chen Xi, University of Hong Kong, Hong Kong 2179

Chapter 6.5. Applying Mobile Technologies to Banking Business Processes /
Dinesh Arunatileka, University of Western Sydney, Australia 2188

Chapter 6.6. Consumers' Preferences and Attitudes Toward Mobile Office Use:
A Technology Trade-Off Research Agenda / *Xin Luo, Virginia State University, USA;*
Merrill Warkentin, Mississippi State University, USA 2203

Chapter 6.7. Customer Relationship Management on Internet and Mobile Channels:
An Analytical Framework and Research Directions / *Susy S. Chan, DePaul University, USA;*
Jean Lam, IBM, USA 2212

Chapter 6.8. Exploring Mobile Service Business Opportunities from a Customer-Centric
Perspective / *Minna Pura, HANKEN—Swedish School of Economics and Business*
Administration, Finland; Kristina Heinonen, HANKEN—Swedish School of Economics and
Business Administration, Finland 2233

Chapter 6.9. Strategy Aligned Process Selection for Mobile Customer Services /
Ragnar Schierholz, University of St. Gallen, Switzerland;
Lutz M. Kolbe, University of St. Gallen, Switzerland;
Walter Brenner, University of St. Gallen, Switzerland 2257

Chapter 6.10. Universal Approach to Mobile Payments /
Stamatis Karnouskos, Fraunhofer Institute FOKUS, Germany;
András Vilmos, SafePay Systems, Ltd., Hungary 2280

Chapter 6.11. Influence of Mobile Technologies on Global Business Processes in Global Organizations / *Dinesh Arunatileka, University of Western Sydney, Australia; Abbass Ghanbary, University of Western Sydney, Australia; Bhuvan Unhelkar, University of Western Sydney, Australia* 2289

Chapter 6.12. Optimal Number of Mobile Service Providers in India: Trade-Off between Efficiency and Competition / *Rohit Prasad, Management Development Institute, India; Varadharajan Sridhar, Management Development Institute, India* 2306

Chapter 6.13. Evolution of Telecommunications and Mobile Communications in India: A Synthesis in the Transition from Electronic to Mobile Business / *Chandana Unnithan, Deakin University, Australia; Bardo Fraunholz, Deakin University, Australia* 2323

Chapter 6.14. Linking Businesses for Competitive Advantage: A Mobile Agent-Based Approach / *Tong-Seng Quah, Nanyang Technological University, Republic of Singapore; Chye-Huang Leow, Singapore Polytechnic, Singapore* 2343

Chapter 6.15. Integrating Mobile Technologies in Enterprise Architecture with a Focus on Global Supply Chain Management Systems / *Bhuvan Unhelkar, University of Western Sydney, Australia; Ming-Chien Wu, University of Western Sydney, Australia; Abbass Ghanbary, University of Western Sydney, Australia*..... 2368

Chapter 6.16. Mobile Business Process Reengineering: How to Measure the Input of Mobile Applications to Business Processes in European Hospitals / *Dieter Hertweck, University for Applied Sciences Heilbronn, Germany; Asarnusch Rashid, Research Center for Information Technology Karlsruhe, Germany*..... 2391

Volume V

Chapter 6.17. Information Delivery for Mobile Business: Architecture for Accessing Large Documents through Mobile Devices / *Christopher C. Yang, Chinese University of Hong Kong, Hong Kong; Fu Lee Wang, City University of Hong Kong, Hong Kong*..... 2418

Chapter 6.18. Resource-Based Interdependencies in Value Networks for Mobile E-Services / *Uta Wehn Montalvo, TNO Strategy, Technology and Policy, The Netherlands; Els van de Kar, Delft University of Technology, The Netherlands; Carleen Maitland, Pennsylvania State University, USA*..... 2440

Chapter 6.19. Channel Choices and Revenue Logistics of Software Companies Developing Mobile Games / *Risto Rajala, Helsinki School of Economics, Finland; Matti Rossi, Helsinki School of Economics, Finland; Virpi Kristiina Tuunainen, Helsinki School of Economics, Finland; Janne Vihinen, Helsinki School of Economics, Finland*..... 2463

Chapter 6.20. 3G Mobile Virtual Network Operators (MVNOs):
Business Strategies, Regulation, and Policy Issues /
Dimitris Katsianis, National and Kapodistrian University of Athens, Greece;
Theodoros Rokkas, National and Kapodistrian University of Athens, Greece;
Dimitris Varoutas, National and Kapodistrian University of Athens, Greece;
Thomas Sphicopoulos, National and Kapodistrian University of Athens, Greece;
Jarmo Harno, Nokia Research Center, Finland;
Ilary Welling, Nokia Research Center, Finland 2475

Chapter 6.21. A Mobile Portal Solution for Knowledge Management /
Stefan Berger, Universität Passau, Germany;
Ulrich Remus, University of Erlangen-Nuremberg, Germany 2496

Chapter 6.22. Strategies of Mobile Value-Added Services in Korea /
Jin Ki Kim, Korea Aerospace University, Korea; Heasun Chun, The State University of New York at Buffalo, USA 2509

Chapter 6.23. Semantic Location Modeling for Mobile Enterprises /
Soe-Tsyr Yuan, National Chengchi University, Taiwan;
Pei-Hung Hsieh, STPRIC, National Science Council, Taiwan 2530

Section VII. Critical Issues

This section addresses conceptual and theoretical issues related to the field of mobile computing, which include security issues in numerous facets of the discipline including mobile agents, mobile commerce, and mobile networks. Within these chapters, the reader is presented with analysis of the most current and relevant conceptual inquires within this growing field of study. Particular chapters also address quality of service issues in mobile networks, mobile ontologies and mobile web mining for marketing. Overall, contributions within this section ask unique, often theoretical questions related to the study of mobile computing and, more often than not, conclude that solutions are both numerous and contradictory.

Chapter 7.1. Mobile Code and Security Issues /
E. S. Samundeeswari, Vellalar College for Women, India;
F. Mary Magdalene Jane, P. S. G. R. Krishnammal, India 2568

Chapter 7.2. Security of Mobile Code /
Zbigniew Kotulski, Polish Academy of Sciences, Warsaw & Warsaw University of Technology, Poland;
Aneta Zwierko, Warsaw University of Technology, Poland 2583

Chapter 7.3. Security in Mobile Agent Systems /
Chua Fang Fang, Multimedia University, Malaysia;
G. Radhamani, Multimedia University, Malaysia 2600

Chapter 7.4. Security Issues and Possible Countermeasures
for a Mobile Agent Based M-Commerce Application / *Jyh-haw Yeh, Boise State University, USA;*
Wen-Chen Hu, University of North Dakota, USA; Chung-wei Lee, Auburn University, USA 2614

Chapter 7.5. XML Security with Binary XML for Mobile Web Services /
Jaakko Kangasharju, Helsinki Institute for Information Technology, Finland;
Tancred Lindholm, Helsinki Institute for Information Technology, Finland;
Sasu Tarkoma, Helsinki Institute for Information Technology, Finland 2633

Chapter 7.6. Security Issues Concerning Mobile Commerce /
Samuel Pierre, École Polytechnique de Montréal, Canada 2653

Chapter 7.7. Security Architectures of Mobile Computing / *Kaj Grahn, Arcada Polytechnic, Finland; Göran Pulkkis, Arcada Polytechnic, Finland; Jonny Karlsson, Arcada Polytechnic, Finland; Dai Tran, Arcada Polytechnic, Finland* 2660

Chapter 7.8. Security Architectures for B3G Mobile Networks /
Christoforos Ntantogian, University of Athens, Greece;
Christos Xenakis, University of Piraeus, Greece 2674

Chapter 7.9. Privacy and Anonymity in Mobile Ad Hoc Networks /
Christer Andersson, Combitech, Sweden; Leonardo A. Martucci, Karlstad University, Sweden;
Simone Fischer-Hübner, Karlstad University, Sweden 2696

Chapter 7.10. Integrity Protection of Mobile Agent Data / *Sheng-Uei Guan, Brunel University, UK* 2715

Chapter 7.11. Key Distribution and Management for Mobile Applications /
György Kálmán, University Graduate Center–UniK, Norway;
Josef Noll, University Graduate Center–UniK, Norway 2725

Chapter 7.12. Modeling Fault Tolerant and Secure Mobile Agent Execution in Distributed Systems / *H. Hamidi, Iran University of Science & Technology, Iran;*
K. Mohammadi, Iran University of Science & Technology, Iran 2739

Chapter 7.13. Security in 2.5G Mobile Systems / *Christos Xenakis, University of Piraeus, Greece* 2752

Chapter 7.14. Evaluation of Security Architectures for Mobile Broadband Access /
Symeon Chatzinotas, University of Surrey, UK;
Jonny Karlsson, Arcada University of Applied Sciences, Finland;
Göran Pulkkis, Arcada University of Applied Sciences, Finland;
Kaj Grahn, Arcada University of Applied Sciences, Finland 2766

Chapter 7.15. Developing a Theory of Portable Public Key Infrastructure (PORTABLEPKI) for Mobile Business Security / *Sashi Nand, Rushmore University, Grand Cayman, BWI* 2784

Chapter 7.16. Authentication, Authorisation, and Access Control in Mobile Systems /
Josef Noll, University Graduate Center–UniK, Norway;
György Kálmán, University Graduate Center–UniK, Norway 2792

Chapter 7.17. Antecedents of Consumer Trust in B2C Electronic Commerce and Mobile Commerce / *Dan J. Kim, University of Houston Clear Lake, USA* 2807

Chapter 7.18. Trust Models for Ubiquitous Mobile Systems /
Mike Burmester, Florida State University, USA 2827

Chapter 7.19. Quality of Service in Mobile Ad Hoc Networks /
Winston K. G. Seah, Institute for Infocomm Research, Singapore;
Hwee-Xian Tan, National University of Singapore, Singapore 2833

Chapter 7.20. Quality of Service Issues in Mobile Multimedia Transmission /
Nalin Sharda, Victoria University, Australia 2843

Chapter 7.21. Classification of 3G Mobile Phone Customers /
Ankur Jain, Inductis India Pvt. Ltd., India; Lalit Wangikar, Inductis India Pvt. Ltd., India;
Martin Ahrens, Inductis India Pvt. Ltd., India; Ranjan Rao, Inductis India Pvt. Ltd., India;
Suddha Sattwa Kundu, Inductis India Pvt. Ltd., India; Sutirtha Ghosh, Inductis India Pvt. Ltd., India .. 2862

Chapter 7.22. Mobile Phone Customer Type Discrimination via Stochastic Gradient Boosting /
Dan Steinberg, Salford Systems, USA; Mikhaylo Golovnya, Salford Systems, USA;
Nicholas Scott Cardell, Salford Systems, USA 2871

Chapter 7.23. An Immune Systems Approach for Classifying Mobile Phone Usage /
Hanny Yulius Limanto, Nanyang Technological University, Singapore;
Tay Joc Cing, Nanyang Technological University, Singapore;
Andrew Watkins, Mississippi State University, USA 2896

Chapter 7.24. Mobile Ontologies: Concept, Development, Usage, and Business Potential /
Jari Veijalainen, University of Jyvaskyla, Finland 2908

Chapter 7.25. Web Mining System for Mobile-Phone Marketing /
Miao-Ling Wang, Minghsin University of Science & Technology, Taiwan, ROC;
Hsiao-Fan Wang, National Tsing Hua University, Taiwan, ROC 2924

Chapter 7.26. Semantic Web Services and Mobile Agents Integration for Efficient Mobile Services /
Vasileios Baousis, University of Athens, Greece;
Vassilis Spiliopoulos, University of the Aegean and National Centre of Scientific Research "Demokritos", Greece; Elias Zavitsanos, University of the Aegean and National Centre of Scientific Research "Demokritos", Greece; Stathes Hadjiefthymiades, University of Athens, Greece;
Lazaros Merakos, University of Athens, Greece 2936

Chapter 7.27. Advanced Resource Discovery Protocol for Semantic-Enabled M-Commerce /
Michele Ruta, Politecnico di Bari, Italy; Tommaso Di Noia, Politecnico di Bari, Italy;
Eugenio Di Sciascio, Politecnico di Bari, Italy;
Francesco Maria Donini, Università della Tuscia, Italy;
Giacomo Piscitelli, Politecnico di Bari, Italy 2957

Chapter 7.28. Improving Mobile Web Navigation Using N-Grams Prediction Models /
Yongjian Fu, Cleveland State University, USA; Hironmoy Paul, Cleveland State University, USA;
Namita Shetty, Cleveland State University, USA 2969

Chapter 7.29. A Study on the Performance of IPv6-Based Mobility Protocols:
Mobile IPv6 vs. Hierarchical Mobile IPv6 / *Ki-Sik Kong, Korea University, Republic of Korea;*
Sung-Ju Roh, Technology R&D Center, LG Telecom Co., Republic of Korea; Chong-Sun Hwang,
Korea University, Republic of Korea 2982

Chapter 7.30. A Novel Fuzzy Scheduler for Mobile Ad Hoc Networks /
S. Shanmugavel, Anna University, India; C. Gomathy, Deemed University, India 2996

Chapter 7.31. Energy-Efficient Cache Invalidation in Wireless Mobile Environment /
R. C. Joshi, Indian Institute of Technology Roorkee, India;
Manoj Misra, Indian Institute of Technology Roorkee, India;
Narottam Chand, Indian Institute of Technology Roorkee, India 3012

Chapter 7.32. Ensuring Serializability for Mobile-Client Data Caching /
Shin Parker, University of Nebraska at Omaha, USA;
Zhengxin Chen, University of Nebraska at Omaha, USA 3021

Volume VI

Chapter 7.33. Mobile Caching for Location-Based Services /
Jianliang Xu, Hong Kong Baptist University, Hong Kong 3031

Chapter 7.34. Location-Aware Query Resolution for Location-Based Mobile Commerce:
Performance Evaluation and Optimization /
James E. Wyse, Memorial University of Newfoundland, Canada 3040

Chapter 7.35. Data Dissemination in Mobile Environments /
Panayotis Fouliras, University of Macedonia, Greece 3068

Chapter 7.36. Data Broadcasting in a Mobile Environment /
A.R. Hurson, The Pennsylvania State University, USA;
Y. Jiao, The Pennsylvania State University, USA 3079

Chapter 7.37. Multimedia over Wireless Mobile Data Networks /
Surendra Kumar Sivagurunathan, University of Oklahoma, USA;
Mohammed Atiquzzaman, University of Oklahoma, USA 3130

Chapter 7.38. High Performance Scheduling Mechanism for Mobile Computing Based
on Self-Ranking Algorithm / *Hesham A. Ali, Mansoura University, Egypt;*
Tamer Ahmed Farrag, Mansoura University, Egypt 3151

Chapter 7.39. Multilayered Approach to Evaluate Mobile User Interfaces /
Maria de Fátima Queiroz Vieira Turnell, Universidade Federal de Campina Grande (UFCG), Brazil;
José Eustáquio Rangel de Queiroz, Universidade Federal de Campina Grande (UFCG), Brazil;
Danilo de Sousa Ferreira, Universidade Federal de Campina Grande (UFCG), Brazil 3168

Chapter 7.40. Mobile Information Processing Involving Multiple Non-Collaborative Sources / *Say Ying Lim, Monash University, Australia; David Taniar, Monash University, Australia; Bala Srinivasan, Monash University, Australia* 3185

Chapter 7.41. A Bio-Inspired Approach for the Next Generation of Cellular Systems / *Mostafa El-Said, Grand Valley State University, USA* 3204

Section VIII. Emerging Trends

This section highlights research potential within the field of mobile computing while exploring uncharted areas of study for the advancement of the discipline. Chapters within this section highlight evolutions in mobile services, frameworks, and interfaces. These contributions, which conclude this exhaustive, multi-volume set, provide emerging trends and suggestions for future research within this rapidly expanding discipline.

Chapter 8.1. Bridging Together Mobile and Service-Oriented Computing / *Loreno Oliveira, Federal University of Campina Grande, Brazil; Emerson Loureiro, Federal University of Campina Grande, Brazil; Hyggo Almeida, Federal University of Campina Grande, Brazil; Angelo Perkusich, Federal University of Campina Grande, Brazil* 3212

Chapter 8.2. Context-Awareness and Mobile Devices / *Anind K. Dey, Carnegie Mellon University, USA; Jonna Häkkilä, Nokia Research Center, Finland* 3222

Chapter 8.3. Policy-Based Mobile Computing / *S. Rajeev, PSG College of Technology, India; S. N. Sivanandam, PSG College of Technology, India; K. V. Sreenaath, PSG College of Technology, India* 3236

Chapter 8.4. Field Evaluation of Collaborative Mobile Applications / *Adrian Stoica, University of Patras, Greece; Georgios Fiotakis, University of Patras, Greece; Dimitrios Raptis, University of Patras, Greece; Ioanna Papadimitriou, University of Patras, Greece; Vassilis Komis, University of Patras, Greece; Nikolaos Avouris, University of Patras, Greece* 3251

Chapter 8.5. Mobile Design for Older Adults / *Katie A. Siek, University of Colorado at Boulder, USA* 3270

Chapter 8.6. Design for Mobile Learning in Museums / *Nikolaos Tselios, University of Patras, Greece; Ioanna Papadimitriou, University of Patras, Greece; Dimitrios Raptis, University of Patras, Greece; Nikoletta Yiannoutsou, University of Patras, Greece; Vassilis Komis, University of Patras, Greece; Nikolaos Avouris, University of Patras, Greece* 3282

Chapter 8.7. Component Agent Systems: Building a Mobile Agent Architecture That You Can Reuse / *Paulo Marques, University of Coimbra, Portugal; Luís Silva, University of Coimbra, Portugal* 3300

Chapter 8.8. Building Applications to Establish Location Awareness:
New Approaches to Design, Implementation, and Evaluation of Mobile and Ubiquitous Interfaces /
D. Scott McCrickard, Virginia Polytechnic Institute and State University (Virginia Tech), USA;
Miten Sampat, Feeva Technoology, Inc., USA;
Jason Chong Lee, Virginia Polytechnic Institute and State University (Virginia Tech), USA 3320

Chapter 8.9. From Ethnography to Interface Design / *Jeni Paay, Aalborg University, Denmark;*
Benjamin E. Erlandson, Arizona State University, USA.....3333

Chapter 8.10. Mobile e-Learning for Next Generation Communication Environment /
Tin-Yu Wu, I-Shou University, Taiwan; Han-Chieh Chao, National Dong Hwa
University, Taiwan.....3349

Chapter 8.11. An Interactive Wireless Morse Code Learning System /
Cheng-Huei Yang, National Kaohsiung Marine University, Taiwan;
Li-Yeh Chuang, I-Shou University, Taiwan;
Cheng-Hong Yang, National Kaohsiung University of Applied Sciences, Taiwan;
Jun-Yang Chang, National Kaohsiung University of Applied Sciences, Taiwan 3361

Chapter 8.12. A Mobile Computing Framework for Passive RFID Detection System in
Health Care / *Masoud Mohammadian, University of Canberra, Australia;*
Ric Jentzsch, Compucat Research Pty Limited, Australia 3368

Chapter 8.13. Widely Usable User Interfaces on Mobile Devices with RFID /
Francesco Bellotti, University of Genoa, Italy; Riccardo Berta, University of Genoa, Italy;
Alessandro De Gloria, University of Genoa, Italy; Massimiliano Margarone,
University of Genoa, Italy 3387

Chapter 8.14. Matching Dynamic Demands of Mobile Users with Dynamic Service Offers /
Bernhard Holtkamp, Fraunhofer Institute for Software and Systems Engineering, Germany;
Norbert Weißenberg, Fraunhofer Institute for Software and Systems Engineering, Germany;
Manfred Wojciechowski, Fraunhofer Institute for Software and Systems Engineering, Germany;
Rüdiger Gartmann, University of Münster, Germany.....3404

Chapter 8.15. A Multi-Agent System Approach to Mobile Negotiation Support Mechanism
by Integrating Case-Based Reasoning and Fuzzy Cognitive Map /
Kun Chang Lee, Sungkyunkwan University, Korea;
Namho Lee, Sungkyunkwan University, Korea 3421

Chapter 8.16. Intelligent User Interfaces for Mobile Computing /
Michael J. O'Grady, University College Dublin, Ireland;
Gregory M. P. O'Hare, University College Dublin, Ireland.....3442

Chapter 8.17. mCity: User Focused Development of Mobile Services
Within the City of Stockholm / *Anette Hallin, Royal Institute of Technology (KTH), Sweden;*
Kristina Lundevall, The City of Stockholm, Sweden.....3455

Chapter 8.18. Mobile Speech Recognition /
Dirk Schnelle, Technische Universität Darmstadt, Germany 3468

Chapter 8.19. Voice-Enabled User Interfaces for Mobile Devices /
Louise E. Moser, University of California, Santa Barbara, USA;
P. M. Melliar-Smith, University of California, Santa Barbara, USA 3494

Chapter 8.20. Voice Driven Emotion Recognizer Mobile Phone: Proposal and Evaluations /
Aishah Abdul Razak, Multimedia University, Malaysia;
Mohamad Izani Zainal Abidin, Multimedia University, Malaysia;
Ryoichi Komiya, Multimedia University, Malaysia 3511

Chapter 8.21. Mobile Multimedia for Speech and Language Therapy /
Nina Reeves, University of Gloucestershire, UK;
Sally Jo Cunningham, University of Waikato, New Zealand;
Laura Jefferies, University of Gloucestershire, UK;
Catherine Harris, Gloucestershire Hospitals NHS Foundation Trust, UK 3529

Chapter 8.22. A Proposed Tool for Mobile Collaborative Reading /
Jason T. Black, Florida A&M University, USA;
Lois Wright Hawkes, Florida State University, USA 3540

Chapter 8.23. Mobile Decision Support for Time-Critical Decision Making /
F. Burstein, Monash University, Australia; J. Cowie, University of Stirling, UK 3552

Chapter 8.24. OFDM Transmission Technique: A Strong Candidate for the Next Generation Mobile Communications / *Hermann Rohling, Hamburg University of Technology, Germany* 3561

Chapter 8.25. Malicious Software in Mobile Devices /
Thomas M. Chen, Southern Methodist University, USA;
Cyrus Peikari, Airscanner Mobile Security Corporation, USA 3588

Preface

In many ways, motion and computing are the two advances that define our modern age. The ability to move previously unimaginable distances and electronically perform complex tasks, both in breathtakingly short amounts of time, has revolutionized and opened up the entire globe. Mobile computing sits in the vibrant junction of these two defining advances. As this modern world demands more mobility and a greater range of computing options, a keen understanding of the issues, theories, strategies and emerging trends associated with this rapidly developing field is becoming more and more important to researchers, professionals and all users alike.

In recent years, the applications and technologies generated through the study of mobile computing have grown in both number and popularity. As a result, researchers, practitioners, and educators have devised a variety of techniques and methodologies to develop, deliver, and, at the same time, evaluate the effectiveness of their use. The explosion of methodologies in the field has created an abundance of new, state-of-the-art literature related to all aspects of this expanding discipline. This body of work allows researchers to learn about the fundamental theories, latest discoveries, and forthcoming trends in the field of medical informatics.

Constant technological and theoretical innovation challenges researchers to remain informed of and continue to develop and deliver methodologies and techniques utilizing the discipline's latest advancements. In order to provide the most comprehensive, in-depth, and current coverage of all related topics and their applications, as well as to offer a single reference source on all conceptual, methodological, technical, and managerial issues in medical informatics, Information Science Reference is pleased to offer a six-volume reference collection on this rapidly growing discipline. This collection aims to empower researchers, practitioners, and students by facilitating their comprehensive understanding of the most critical areas within this field of study.

This collection, entitled **Mobile Computing: Concepts, Methodologies, Tools, and Applications**, is organized into eight distinct sections which are as follows: (1) Fundamental Concepts and Theories, (2) Development and Design Methodologies, (3) Tools and Technologies, (4) Utilization and Application, (5) Organizational and Social Implications, (6) Managerial Impact, (7) Critical Issues, and (8) Emerging Trends. The following paragraphs provide a summary of what is covered in each section of this multi-volume reference collection.

Section One, **Fundamental Concepts and Theories,** serves as a foundation for this exhaustive reference tool by addressing crucial theories essential to understanding mobile computing. Some basic topics impacted by this field are examined in this section through articles such as "A Mobile Computing and Commerce Framework" by Stephanie Teufel, Patrick S. Merten and Martin Steinert. This selection introduces a key topic further developed and discussed through later selections, the intersection of mobile computing and its commercial implications. The selection "Environments for Mobile Learning" by Han-Chieh Chao, Tin-Yu Wu, and Michelle T.C. Kao provides a sampling of how mobility impacts

educational trendsetting. Another important basic topic in this section is ushered in by M. Halid Kuscu, Ibrahim Kushchu, and Betty Yu and their contribution entitled "Introducing Mobile Government," which introduces the concept of mobile government and creates a context for discussing various applications, services, and the relevant technologies. "A Taxonomy of Database Operations on Mobile Devices" by Say Ying Lim, David Taniar and Bala Srinivasan informs the vital area of databases while grounding the reader in its possible operations, and "Mobile Portals" by Ofir Turel and Alexander Serenko offers a explanation and exploration of the ability of mobile portals to diffuse and penetrate even remote populations. These are only some of the elemental topics provided by the selections within this comprehensive, foundational section that allow readers to learn from expert research on the elemental theories underscoring mobile computing.

Section Two, **Development and Design Methodologies,** contains in-depth coverage of conceptual architectures and frameworks, providing the reader with a comprehensive understanding of emerging theoretical and conceptual developments within the development and utilization of mobile computing. In opening this section, "Developing Smart Client Mobile Applications" by Jason Gan exemplifies the issues addressed in this section by examining the usability and accessibility of mobile applications and services and suggesting development. The development of mobile applications is also discussed in "Location Area Design Algorithms for Minimizing Signalling Costs in Mobile Networks" by J. Gutierrez, Vilmos Simon and Sándor Imre. Also included in this section is the selection "'It's the Mobility, Stupid': Designing Mobile Government" by Klas Roggenkamp, which lays out the challenges and possibilities of designing for mobile government. Overall, these selections outline design and development concerns and procedures, advancing research in this vital field.

Section Three, **Tools and Technologies**, presents extensive coverage of various tools and technologies and their use in creating and expanding the reaches of mobile computing. The multitude of mobile business applications, their uses and their individual efficiency is explored in such articles as "Evaluation of Mobile Technologies in the Context of Their Applications, Limitations, and Transformation" by Abbass Ghanbary, "Knowledge Representation in Semantic Mobile Applications" by Pankaj Kamthan, and "Mobile Portal Technologies and Business Models" by David Parsons. With a look toward the near future, the selection "A Virtual Community for Mobile Agents" by Sheng-Uei Guan and Fangming Zhu features in-depth discussions of the probable uses and benefits of mobile agents in a variety of fields. The ever-developing culture of mobile multimedia is represented in this section as well, with "Discovering Multimedia Services and Contents in Mobile Environments" by Zhou Wang and Hend Koubaa, "V-Card: Mobile Multimedia for Mobile Marketing" by Holger Nösekabel and Wolfgang Röckelein, and "The Design of Mobile Television in Europe" by Pieter Ballon and Olivier Braet. The rigorously researched chapters contained in this section offer readers countless examples of modern tools and technologies that emerge from or can be applied to mobile computing.

Section Four, **Utilization and Application**, investigates the use and implementation of mobile technologies and informatics in a variety of contexts. One prominent context is mobile phone use, thoroughly analyzed in throughout the world in the articles "Exploring the Use of Mobile Data Services in Europe: The Cases of Denmark and Greece" by Ioanna D. Constantiou and Maria Bina, "The Mobile Phone Telecommunications Service Sector in China" by Michelle W. L. Fong, "United States of America: Renewed Race for Mobile Services" by Mats Samuelsson, Nikhilesh Dholakia and Sanjeev Sardana, and "M-Learning with Mobile Phones" by Simon So. This latter topic, m-learning, is continued by Hans Lehmann, Stefan Berger and Ulrich Remus in "A Mobile Portal for Academe," while Anna Trifonova reaches the root of two vital issues in "Accessing Learning Content in a Mobile System: Does Mobile Mean Always Connected?" Questioning applications and technology is rarely more important than in the health sector, the focus of a number of articles beginning with "Perception of Mobile Technology Provi-

sion in Health Service" by Astrid M. Oddershede and Rolando A. Carrasco and ending with "Integrating Mobile-Based Systems with Healthcare Databases" by Yu Jiao, Ali R. Hurson, Thomas E. Potok and Barbara G. Beckerman. This section ends with articles pertaining to commerce, business and government, providing a complete understanding of the successes and limitations of mobile computing.

Section Five, **Organizational and Social Implications**, includes a wide range of research pertaining to the organizational and cultural implications of mobile computing. The section begins with "Mobile Virtual Communities" by Glauber Ferreira, Hyggo Almeida, Angelo Perkusich, and Evandro Costa, a selection explores virtual communities, describing the main issues that have culminated in the creation of this research area, also the topic of "Mobile Virtual Communities of Commuters" by Jalal Kawash, Christo El Morr, Hamza Taha, and Wissam Charaf, "Wireless Local Communities in Mobile Commerce" by Jun Sun, and "From Communities to Mobile Communities of Values" by Patricia McManus and Craig Standing. Akin the idea of community is the topic of trust, the subject of "Consumer Perceptions and Attitudes Towards Mobile Marketing" by Amy Carroll, Stuart J. Barnes and Eusebio Scornavacca. Lastly, "Mobile Networked Text Communication: The Case of SMS and Its Influence on Social Interaction" by Louise Barkhuus provides insight into how certain functions of mobile technology affect social interaction – an important consideration to end this section detailing how mobile computing shapes and is shaped by human culture and logic.

Section Six, **Managerial Impact**, presents contemporary coverage of the managerial applications and implications of mobile computing. Core concepts covered include the impact of mobile computing on business practices, customer and business interaction, and business communication, policies and strategies. "Comprehensive Impact of Mobile Technology on Business" by Khimji Vaghjiani and Jenny Teoh begins the section with an insightful introduction. Also included are the articles "Consumers' Preferences and Attitudes Toward Mobile Office Use: A Technology Trade-Off Research Agenda" by Xin Luo and Merrill Warkentin, "Customer Relationship Management on Internet and Mobile Channels: An Analytical Framework and Research Directions" by Susy S. Chan and Jean Lam, and "Exploring Mobile Service Business Opportunities from a Customer-Centric Perspective" by Minna Pura and Kristina Heinonen, which expound on the concerns surrounding customer relationships with mobile technologies. This section concludes with a insights on topics including telecommunications, the media and gaming industries, knowledge management and mobile enterprising—a few of the subjects necessary to understand managing and mobile computing.

Section Seven, **Critical Issues**, presents readers with an in-depth analysis of the more theoretical and conceptual issues within this growing field of study by addressing topics such as the quality and security of mobile computing. "Mobile Code and Security Issues" by E. S. Samundeeswari and F. Mary Magdalene Jane, "Security of Mobile Code" by Zbigniew Kotulski, and "Security in Mobile Agent Systems" by Chua Fang Fang and G. Radhamani, address necessary security considerations. The article "Privacy and Anonymity in Mobile Ad Hoc Networks" by Christer Andersson, Leonardo A. Martucci and Simone Fischer-Hübner raises similar concerns. The quality of mobile computing services is pondered in articles such as "Quality of Service in Mobile Ad Hoc Networks" by Winston K. G. Seah and Hwee-Xian Tan, "Quality of Service Issues in Mobile Multimedia Transmission" by Nalin Sharda, and "A Study on the Performance of IPv6-Based Mobility Protocols: Mobile IPv6 vs. Hierarchical Mobile IPv6" by Ki-Sik Kong, Sung-Ju Roh and Chong-Sun Hwang. Further discussion of critical issues includes obstacles surrounding ad hoc networking, database querying and management, data dissemination, broadcasting and processing. In all, the theoretical and abstract issues presented and analyzed within this collection form the backbone of revolutionary research in and evaluation of mobile computing.

The concluding section of this authoritative reference tool, **Emerging Trends**, highlights research potential within the field of mobile computing while exploring uncharted areas of study for the advance-

ment of the discipline. The development and deployment of new forms of mobile computing is explored in selections entitled “Bridging Together Mobile and Service-Oriented Computing” by Loreno Oliveira, Emerson Loureiro, Hyggo Almeida and Angelo Perkusich, “Context-Awareness and Mobile Devices” by Anind K. Dey and Jonna Häkkilä, “Component Agent Systems: Building a Mobile Agent Architecture That You Can Reuse” by Paulo Marques and Luís Silva, and “Voice Driven Emotion Recognizer Mobile Phone: Proposal and Evaluations” by Aishah Abdul Razak, Mohamad Izani Zainal Abidin, and Ryoichi Komiya. Other new trends, such as developments concerning RFID, time-critical decisions, collaboration between mobile devices and new concepts supporting user collaboration are included in and stretch our concept of what mobile computing can be. This final section demonstrates that mobile computing, with its propensity for constant change and evolution, will continue to both shape and define the modern face of business, health, culture and human interaction.

Although the contents of this multi-volume book are organized within the preceding eight sections which offer a progression of coverage of important concepts, methodologies, technologies, applications, social issues, and emerging trends, the reader can also identify specific contents by utilizing the extensive indexing system listed at the end of each volume. Furthermore, to ensure that the scholar, researcher, and educator have access to the entire contents of this multi-volume set, as well as additional coverage that could not be included in the print version of this publication, the publisher will provide unlimited, multi-user electronic access to the online aggregated database of this collection for the life of the edition free of charge when a library purchases a print copy. In addition to providing content not included within the print version, this aggregated database is also continually updated to ensure that the most current research is available to those interested in mobile computing.

As mobile computing continues to expand, both in variety and usefulness, this exciting and revolutionary field will prove even more necessary to everyday life. Intrinsic to our ever-modernizing, ever-expanding, global economy is mobility and technology, the two aspects that define the contents of these articles. Continued progress and innovation, driven by a mobile, demanding consumer base, will only further establish how necessary and vital a sure understanding of mobile computing and the changes and challenges influencing today’s modern, dynamic world.

The diverse and comprehensive coverage of mobile computing in this six-volume, authoritative publication will contribute to a better understanding of all topics, research, and discoveries in this developing, significant field of study. Furthermore, the contributions included in this multi-volume collection series will be instrumental in the expansion of the body of knowledge in this enormous field, resulting in a greater understanding of the fundamentals while also fueling the research initiatives in emerging fields. We at Information Science Reference, along with the editor of this collection, hope that this multi-volume collection will become instrumental in the expansion of the discipline and will promote the continued growth of mobile computing.

An Introduction to Mobile Technology and Its Applications: A Data-Centric Perspective

David Taniar
Monash University, Australia

ABSTRACT

The emergence of mobile computing provides the ability to access information at anytime and place. However, as mobile computing environments have inherent factors like power, storage, asymmetric communication cost, and bandwidth limitations, efficient mobile information access has become a challenging application of mobile technology. This paper introduces wireless technology and environment, and discusses one of the main applications of mobile technology, especially in a data-centric domain. A framework consisting of three main elements of mobile data processing applications, namely (i) server strategy, (ii) on-air strategy, and (iii) client strategy, is presented. The main aspect of mobile data processing lies in the various types of mobile queries, including mobile location-based queries. As the queries, as well as the interest objects, may be moving, it is critical to understand the complexity of mobile location-based queries. Finally, an application of mobile data processing, specifically in the context of mobile e-health is presented. The application development tools used in the application case study are described.

INTRODUCTION

In recent years, the use of wireless technology devices has been growing at an exponential rate. Most people are now able to access information systems located in wired networks anywhere and anytime using portable size wireless computing devices powered by batteries (e.g. notebooks, tablet PCs, personal digital assistants (PDAs) and GPRS-enabled cellular phones). These portable computing devices communicate with a central stationary server via a wireless channel and become the integrated part of the existing distributed computing environment. Subsequently, mobile users can have access to informa-

tion located at a static network while they are travelling and this type of computing is known as *mobile computing* (Barbara, 1999; Myers & Beigl, 2003; Waluyo, Srinivasan, & Taniar, 2005e). Mobile computing provides data-intensive applications with useful aspects of wireless technology, and the mobile technology to support such applications is referred to as *mobile databases* (Barbara, 1999; Malladi, et al 2002; Waluyo, Srinivasan, & Taniar, 2004b).

Mobile service providers are establishing a number of information services including weather information or weather forecast services, news, stock indices information, foreign exchange, election results, tourist services, airline schedules, location-dependent query, and route guidance, to name a few (http://www.wapforum.org/). In order to realize the potential of wireless information services, a number of issues and challenges need to be addressed including mobile data management (Barbara, 1999), cache management (Barbara & Imielinski, 1994), wireless network infrastructure (Bria et al, 2001), location-dependent data management (Lee et al, 2002), power management issues (Stan & Skadron, 2003) and data broadcasting issues (Imielinski, Viswanathan & Badrinath, 1994). Location-dependent queries will soon become common and of great interest. Consequently, providing efficient and effective location-dependent mobile information services will be highly desirable.

Despite the complexity involved in processing mobile location-dependent information services, we need to understand the unique characteristics of a mobile computing environment, covering:

- **Resource constrained mobile devices:** To provide better portability and improve attractiveness, mobile devices are becoming smaller and lighter. However, such designs usually involve some trade-offs including low battery life, low computational power and smaller storage capacity. Especially with battery power, the life expectancy of a battery (e.g. nickel-cadmium, lithium ion) was estimated to increase the time of effective use by only another 15% in several years to come (Paulson, 2003). Furthermore, it should be noted that wireless data transmission requires a greater amount of power or up to 10 times as much power as the reception operation (Zaslavsky & Tari, 1998; Xu et al, 2002).
- **Low network bandwidth:** Mobile users can connect to the fixed network via various wireless communication networks including wireless radio, wireless Local Area Network (LAN), wireless cellular, satellite, etc. Each of the wireless networks provides a different bandwidth capacity. However, this wireless bandwidth is too small compared with a fixed network such as ATM (Asynchronous Transfer Mode) that can provide speeds of up to 155Mbps (Elmasri & Navathe, 2003). Designing a high network utilization data access method to provide an acceptable response time becomes an important issue in the mobile computing literature.
- **Asymmetric communication cost:** The different bandwidth capacity between the downstream communication and upstream communication has created a new environment called *Asymmetric Communication Environment*. In fact, there are two situations that can lead to communication asymmetry (Acharya, et al, 1995). One is due to the capability of physical devices. For example, servers have powerful broadcast transmitters, while mobile clients have little transmission capability. The other is due to the patterns of information flow in the applications. For instance, in a situation where the number of servers is far fewer than the number of clients, it is asymmetric because there is not enough capacity to handle simultaneous requests from multiple clients.
- **Heterogeneity of mobile devices:** Mobile telecommunication industries have developed a large variety of mobile devices such as Laptops, Tablet PC, Handheld PCs, Pocket PC, and Mobile Phones. However, these mobile devices have also various features and capabilities such as operating system, computational power, display and network capability. Consequently, this heterogeneity raises some challenges in content management and content delivery to the mobile service providers.

- **Mobility:** Wireless technology enables mobile users to move freely and independently from one place to another. A service handoff occurs when a user moves from one network service area into another. It is essential to ensure service handoffs seamlessly and transparently to the users.
- **Frequent Disconnections:** Mobile users are frequently disconnected from the network. This may be due to several reasons including signal failures, empty network coverage, and power saving. The later reason is advantageous since active mode requires thousand times more power than doze or power saving mode (Imielinski, Viswanathan & Badrinath, 1994). Wireless radio signals may also be weakened due to the client's further distance from the base station or the speed at which the client is moving.

In the light of the characteristics of the mobile environment, it is essential to have an effective mobile data processing mechanism. This includes understanding the full spectrum of mobile computing technologies and mobile data environment. Having an extensive framework for mobile data processing is therefore critical, including understanding the complexity of mobile queries. Finally, in order to put these into perspective, it is important to see how mobile data processing is put into an application, including the use of various mobile application development tools.

MOBILE COMPUTING TECHNOLOGIES

Mobile data processing is made available due to the advances in wireless technologies, and location-detection and positioning systems. This section gives an overview of these foundation technologies.

Wireless Technologies

This section summarizes the current wireless technologies, covering the wireless technologies used for indoor and outdoor networks. These include (*i*) In-room Networks, (*ii*) Wireless LAN (WLAN), (*iii*) Broadband Wireless Networks, (*iv*) Wide Area Wireless/Radio Networks, (*v*) Satellite-based Networks, and (*vi*) Cellular Networks.

In-Room Networks

In-room networks provide mobile devices to communicate with others using a short-range wireless. In general, there are two types of in-room networks (Helal, et al, 2002): *infrared* and *radio frequency*.

Using infrared, the wireless network coverage can be up to 50 metres with a supported bandwidth of about 1Mbps. The most common standard used for infrared network technology is the *Infrared Data Association* (IrDA), an industry-sponsored organization set up in 1993 to create international standards for the hardware and software used in infrared communication links (Williams, 2000; Vitsas & Boucouvalas, 2003). IrDA is a point-to-point, narrow angle, ad-hoc data transmission standard designed to operate over a distance of 0 to 1 meter and at speeds of 9600 bps to 16 Mbps. In the IrDA-1.1 standard, the maximum data size that may be transmitted is 2048 bytes and the maximum transmission rate is 4 Mbps (Vitsas & Boucouvalas, 2002). IrDA is the same technology used to control a TV set with a remote control. In general, it is used to provide wireless connectivity technologies for devices that would normally use cables for connectivity (Robertson, Hansen, Sorensen, & Knutson, 2001).

Another in-room network is based on radio frequency. The most common standard used for this technology is *Bluetooth*. The BlueTooth Special Interest Group established the in-room radio frequency

in 1998 (Bluetooth, 2008). Bluetooth (Chiasserini, Marsan, Baralis, & Garza, 2003) is a high-speed, low-power microwave wireless link technology, designed to connect phones, laptops, PDAs and other portable equipment together with little or no work by the user. Unlike infrared, Bluetooth does not require line-of-sight positioning of connected units. The technology uses modifications of existing wireless LAN techniques but is most notable for its small size and low cost. Whenever any Bluetooth-enabled devices come within range of each other, they instantly transfer address information and establish small networks between each other, without the user being involved. The wireless network coverage ranges from 1 meter up to 100 meters, and the data transfer rate is up to 3Mbps.

Wireless LAN (WLAN)

A wireless local area network is a network that provides wide wireless bandwidth to low mobility clients. This technology expands the range of the infrared and the Bluetooth technologies by improving the network diameter to about 200m (Helal et al., 2002). It provides low-mobility, high-data-rate data communications within a confined region (Zaslavsky & Tari, 1998). The aim of WLANs is to provide a wireless bridge to conventional wired networks rather than supporting true mobility (Pitoura & Samaras, 1998).

Amongst several available standards for WLAN, IEEE 802.11 standard for wireless LANs is the most successful standard today and it is superficially similar to Ethernet (Gast, 2005). The IEEE 802.11 standard has a number of protocols (The IEEE 802.11 Standards, 2008). However, there are only three types of IEEE 802.11 that have been widely used, namely IEEE 802.11a, IEEE 802.11b, IEEE 802.11g (Gast, 2005). The 802.11 specifications are part of an evolving set of wireless network standards known as the 802.11 family. The particular specification under which a Wi-Fi network operates is called the "flavour" of the network.

Wi-Fi (short for "wireless fidelity") is a term for certain types of wireless local area network (WLAN) that uses specifications in the 802.11 family (Vaughan-Nichols, 2003). The term Wi-Fi was created by an organization called the Wi-Fi Alliance, which oversees tests that certify product interoperability. A product that passes the alliance tests is given the label "Wi-Fi certified" (a registered trademark). Originally, Wi-Fi certification was applicable only to products using the 802.11b standard (Ferro & Potorti, 2005). Today, Wi-Fi can apply to products that use any 802.11 standard.

Wi-Fi has gained acceptance in many businesses, agencies, schools, and homes as an alternative to a wired LAN. Many airports, hotels, and fast-food facilities offer public access to Wi-Fi networks. These locations are known as hot spots. Many charge a daily or hourly rate for access, but some are free. An interconnected area of hot spots and network access points is known as a hot zone.

The current IEEE 802.11 (IEEE, 1999) is known to lack a viable security mechanism. Unless adequately protected (Hole, Dyrnes, & Thorsheim, 2005), a Wi-Fi network can be susceptible to access by unauthorized users who use the access as a free Internet connection. Any entity that has a wireless LAN should use security safeguards such as the wired equivalent privacy (WEP) encryption standard, the more recent Wi-Fi protected access (WPA), Internet protocol security (IPsec), or a virtual private network (VPN).

There is another wireless LAN standard, called *HiperLAN*, which is primarily used in the European countries. There are two specifications: HiperLAN/1 and HiperLAN/2. Both have been adopted by the European Telecommunications Standards Institute (ETSI). The HiperLAN standards provide features and capabilities similar to 802.11. HiperLAN/1 provides communications at up to 20 Mbps in the 5-GHz range of the radio frequency (RF) spectrum. HiperLAN/ 2 is defined as a flexible Radio LAN standard designed to provide high speed access up to 54 Mbps to a variety of networks including 3G mobile core

networks, ATM networks, and IP-based networks, and also for private use as a wireless LAN system. Basic applications include data, voice and video, with specific Quality of Service (QoS) parameters taken into account. HiperLAN/2 systems can be deployed in offices, classrooms, homes, factories, hot spot areas like exhibition halls, and more generally where radio transmission is an efficient alternative or a complement to wired technology. It is worth noting that HiperLAN/2 has been developed in conjunction with the Japanese standards body, the Association of Radio Industries and Broadcasting.

Broadband Wireless Networks

Broadband wireless is a wireless technology that allows simultaneous wireless delivery of voice, data, and video has appeared recently in metropolitan areas (Overview of Wireless Technologies, 2004). This wireless technology is mainly available in metropolitan areas with a requirement of clear sight between the transmitter and the mobile computing devices. Two types of this technology are: *Local Multi-point Distribution Service* (LMDS) and *Multi-channel Multi-point Distribution Service* (MMDS). LMDS uses a high bandwidth wireless frequency within a range of 20-31 GHz, whereas MMDS uses a lower bandwidth wireless frequency within 2 GHz and has coverage of up to 50 kilometres.

Broadband wireless network, which is built using the IEEE 802 standard, is WiMAX. *WiMAX*, formed in April 2001, is a wireless industry coalition whose members organized to advance IEEE 802.16 standards for broadband wireless access networks (Ghosh et al., 2005; Vaughan-Nichols, 2004; Hamalainen et al., 2002; Giuliano & Mazzenga, 2005). The WiMAX 802.16 technology is expected to enable multimedia applications with wireless connection and, with a range of up to 50 kilometres (Giuliano & Mazzenga, 2005; Ghavami, Michael, & Kohno, 2005). The main aim of WiMAX is to promote and certify compatibility and interoperability of devices based on the 802.16 specification, and to develop such devices for the marketplace.

Wireless Wide Area/Radio Networks

Wireless Wide Area Network is designed to provide data transmission and its infrastructure consists of base stations, network control centres and switches to transmit the data (Zaslavsky & Tari, 1998). The characteristics of Wireless wide area network are high mobility, wide ranging and low data rate digital communication (Pitoura & Samaras, 1998; Zaslavsky & Tari, 1998). This network type can be categorised into *public* and *private radio network* (Pitoura & Samaras, 1998). The first category is the wireless data communications supplied to the public by service providers and the average data rate is 4800 bps to 19.2 Kbps (Zaslavsky & Tari, 1998), whereas the private radio network is provided by private companies for their own purposes.

Satellite-Based Networks

The *satellite network* has been used to relay voice, video or data, since the 1960s (DeRose, 2002). The characteristics of the satellite-based network are that it has wide range coverage, expensive, two-way communication and low quality voice. It has wide area coverage, which spans the ocean as well as remote land areas (Lodge, 1991). It provides two-way communications, however, it has low quality voice or limited data (Zaslavsky & Tari, 1998; Pitoura & Samaras, 1998). It is also expensive to provide this type of network (El-Ghazaly & Golio, 1996).

There are three common terms used for these satellites based on their distance and spatial relationship with the earth, namely *GEOstationary Satellites* (GEOS), *Medium Earth Orbit Satellites* (MEOS)

and *Low Earth Orbit Satellites* (LEOS) (Pitoura & Samaras, 1998; El-Ghazaly & Golio, 1996; Toh & Li, 1998). GEOS, MEOS and LEOS are located at altitudes of 35,786 km, 10,000 km and 1,000 km respectively.

Cellular Networks

The cellular network has evolved from first generation (1G) up to fourth generation (4G).

The *first generation* (1G) of cellular systems appeared in the early 1980s and is based on analog technology. The main characteristics are low capacity, lack of security, and unsuitable for non-voice applications (Agrawal & Famolari, 1999b). Voice is transmitted using Frequency Modulation (FM), and the data transfer rate is 1.2-9.6 Kbps (Pitoura & Samaras, 1998).

After the first-generation analogue mobile systems, the *second-generation* (2G) mobile digital systems were introduced around 1991 offering higher capacity and lower costs for network operators, while for the users, they offered short messages and low-rate data services added to speech services. 2G marked the arrival of digital modulation techniques that increase capacity, have a better speech quality, enhance security features, and offer more efficient terminals (Agrawal & Famolari, 1999). It has a data transfer rate from 9 to 14 Kbps (Pitoura & Samaras, 1998). Presently, the 2G systems are Global System for Mobile Communications (GSM), Time Division Multiple Access (TDMA), Personal Digital Cellular (PDC), and Code Division Multiple Access (CDMA). GSM is used in most parts of the world except in Japan, were PDC is the second-generation system used (Dixit, Guo, & Antoniou, 2001).

The *second and a half generation* (2.5G) is an enhancement of the second generation. It has the ability to use packet-switched solution in GPRS (General Packet Radio System). GPRS offers the possibility to always be online and only pay for the data actually transferred. Data rates of up to 20 kbps per used time slot will be offered, and with multiple time-slots per user in the downlink, attractive services can be offered (Stallings, 2001).

The *third generation* (3G) was developed in 1992. The shift to 3G in the radio access networks is demanding a lot of efforts. The ITU efforts through IMT-2000 have led to a number of recommendations. These recommendations address areas such as user bandwidth, richness of service offerings (multimedia services), and flexibility (networks that can support small or large numbers of subscribers) (UMTS Forum, 2000a,b). The examples of third generation include the Universal Mobile Telecommunications System (UMTS), the Code Division Multiple Access (CDMA2000). This generation has three categories of data rates as follow (Agrawal & Famolari, 1999):

- 2.4 Mbps to stationary users (fixed location)
- 384 Kbps to pedestrian users (travel speed: 3 m/hr)
- 144 Kbps to vehicular users (travel speed: 60 m/hr)

The next generation of 3G wireless network is *3.5G* with 3Mbits/secs data rates (Dulaney, 2008).

The *fourth generation* (4G) has not officially been released yet, but it is expected that this generation will support applications up to 1 Gbps (Kim & Prasad, 2006).

Table 1 shows a comparison matrix among the four generations of cellular networks.

Location Positioning Systems

As most data processing in a wireless and mobile environment is location-dependent query processing, location-positioning systems form an integral part of mobile data processing. This section summarizes

Table 1. From 1G to 4G

1G	2G	3G	4G
Basic mobility	Advance mobility “roaming”	Seamless roaming	IP based mobility
Basic service	Various services “data exchange”	Service concept and model	Extremely high data rates
Incompatibility	Headed for global solution	Global solution	Perfect telecom and data communication convergence

the current location positioning systems, covering the location positioning technologies used for indoor and outdoor networks. These include (*i*) Satellite positioning systems, (*ii*) Cellular position systems, and (*iii*) Indoor positioning systems.

Satellite Positioning Systems

The *GPS* system is the most popular satellite positioning system. It provides two basic types of services: the *Standard Positioning Service* (SPS) and the *Precise Positioning Service* (PPS) (Kpper, 2005). SPS is a positioning and timing service focusing on the civilian user, whereas PPS is a positioning, velocity, and timing service for military applications, which is restricted to authorized users only. There is another satellite positioning system called Gallileo, which is due to start its operation in the year 2009.

Cellular Positioning Systems

This cellular positioning system integrates GPS so that the cellular network provides terminals with assistance and correction of the satellites (Kpper, 2005).

Examples of the cellular positioning for the second generation cellular network (GSM) are Cell-Id in combination with timing advance, Enhanced Observed Time Difference (E-OTD), Uplink Time Difference of Arrival (U-TDoA), and Assisted GPS (A-GPS). The introduction of Cell-Id and A-GPS into existing GSM networks is comparatively simple, while E-OTD and U-TDoA comprise essential modifications and extensions.

Examples of the cellular positioning for the third generation cellular network are Cell-based methods, Observed time difference of arrival with idle period downlink (OTDoA-IPDL), and Assisted GPS (A-GPS).

Assisted GPS (A-GPS) is a hybrid solution to use information from both the satellites and network (Agrawal & Zeng, 2006). It enables a mobile terminal including GPS receiver to be positioned faster and more accurately (Tsalgatidou, et al., 2003). The A-GPS is located at the base station and feeds information to mobile computing devices. It is also accurate, increases sensitivity, reduces position acquisition time, and uses less power at the GPS server (Agrawal & Zeng, 2006).

Table 2 shows a comparison among the abovementioned cellular positioning systems (Kpper, 2005). The table shows that A-GPS performance is the most accurate and consistent, although the service area is the smallest.

Indoor Positioning Systems

Indoor positioning system operates within an indoor or local environment, such as shopping centres or buildings. There are four indoor-based positioning systems: (*i*) WLan-based, (*ii*) Radio Frequency Identification (RFID)-based, (*iii*) infrared-based, and (*iv*) ultrasound-based.

Table 2. Cellular positioning systems

	Accuracy			Consistency	Yield
	Rural	**Suburban**	**Urban**		
Cell-Id	> 10km	2-10km	50-1000m	Poor	Good
E-OTD and OTDoA	50-150m	50-250m	50-300m	Average	Average
U-TDoA	50-120m	40-50m	40-50m	Average	Average
A-GPS	10-40m	20-100m	30-150m	Good	Good

WLan-based indoor positioning system is the most popular, and it uses IEEE 802.11 devices.

The *RFID-based* is an emerging technology that is growingly popular for used in applications like asset management, product identification, and factory automation (Kpper, 2005).

The *infrared-based* positioning systems use the infrared technology. An example of an infrared positioning system is the Xerox ParcTab (Want et al, 1996), and the WIPS project (Wireless Indoor Positioning Systems) (WIPS, 2007).

The *ultra-sound* positioning system uses ultrasounds, and combines ultra-sounds and radio networks. Active Bat (Ward et al, 1997) is an example of an ultra-sound positioning system using the ultra-sound technology, whereas the Cricket system (Priyantha et al., 2000) combines the ultra-sounds with the radio networks.

MOBILE DATA ENVIRONMENT

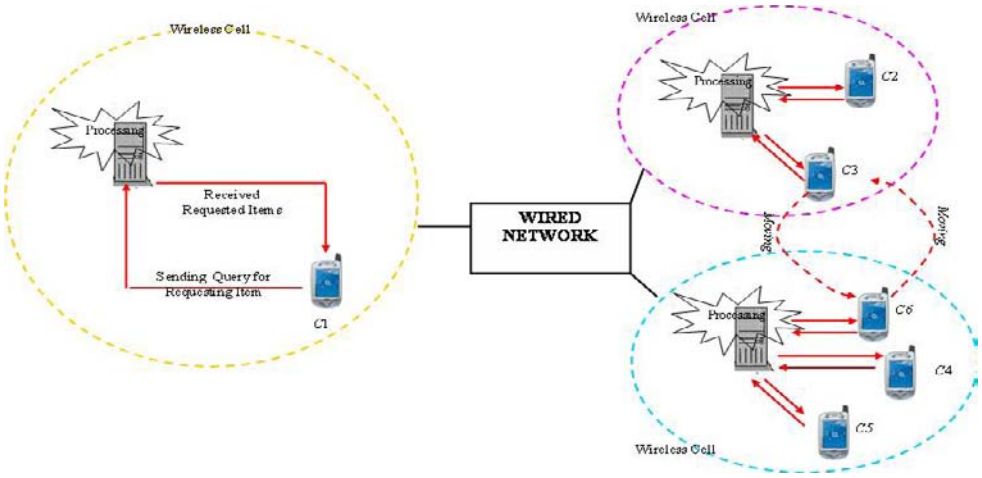
Generally, mobile devices are defined as electronic equipment which operate without cables for the purposes of communication, data processing and exchange, which can be carried by their users and can receive, send or transmit information anywhere, anytime due to their mobility and portability (Bose et al., 2005). In particular, mobile devices include mobile phones, Personal Digital Assistants (PDAs), laptops that can be connected to a network, and PDA-mobile phones that add mobile phone functionalities to a PDA (Waluyo, Goh, Taniar & Srinivasan, 2005; Waluyo, Goh, Srinivasan, & Taniar, 2005; Waluyo, Hsieh, Taniar, Rahayu, & Srinivasan, 2004).

Mobile users, with their mobile devices and servers that store data, are involved in a typical mobile data environment (Wolfson et al., 2006). As wireless architecture is fundamentally different from the wired environment, the type of queries, query processing mechanisms as well as communication technology, also differs accordingly. Wireless networking infrastructure provides ubiquitous wireless communication coverage. This coverage will assist mobile users to have access to network resources via a different type of communication media and independent from the location of the user or the information being accessed.

In general, each mobile user communicates with a *Mobile Base Station* (MBS) in order to carry out any activities such as transaction and information retrieval. MBS has a wireless interface to establish communication with mobile clients and it serves a large number of mobile users in a specific region called *cell*.

A *cell* is a service area for one MBS where each cell may have the same or different size. According to Lunde & Mjøvik (2000), and Feuerstein & Rappaport (1993), cells are classified into three types: Macro, Micro and Pico cells. A Macro cell is a cell which has a radius of 700-8000 metres, a data transfer rate of 144-384 Kbps with bandwidth frequency of 11.34 Mhz. A Micro cell has a radius of 75-700 metres with a data transfer rate of 384 Kbps and bandwidth frequency of 1.26 Mhz. A Pico cell is an area with a radius of 20-75 metres, a 384 Kbps-2 Mbps data transfer rate and 1.26 Mhz bandwidth frequency.

Figure 1. A mobile database environment



Each MBS is connected to a fixed network. Mobile clients can move between cells while being active and the inter-cell movement is known as a handoff process (Imielinski & Badrinath, 1994; Trivedi, Dharmaraja & Ma, 2002). Each client in a cell can connect to the fixed network via wireless radio, wireless Local Area Network (LAN), wireless cellular, or satellite. Each of the wireless networks provides a different bandwidth capacity. However, this wireless bandwidth is too small compared with the fixed network such as ATM (Asynchronous Transfer Mode) that can provide speed of up to 155Mbps (Elmasri & Navathe, 2003).

Communication between mobile users and servers is required in order to carry out any transactions and information retrieval. Basically, the servers are normally static, whereas mobile users can move from one place to another and are therefore dynamic. Nevertheless, mobile users have to be within a specific region so as to be able to receive signals in order to connect to the servers (Lim, Taniar, & Srinivasan, 2005,b).

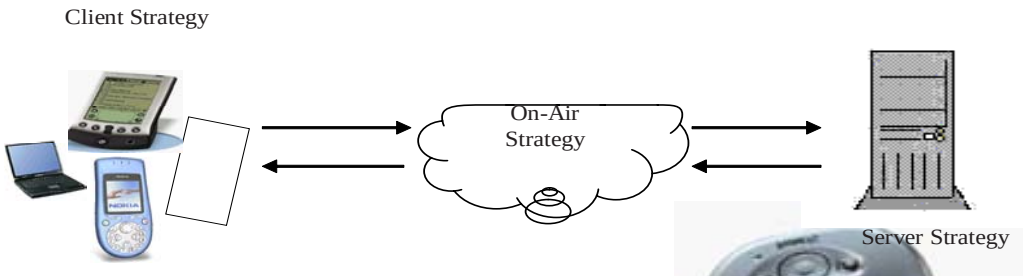
Figure 1 illustrates a scenario of a mobile database environment, which involves mobile users moving from one location to another location. Whenever mobile users are within a specific region or cell, they can access information provided by the servers within that region. And the mobile users will obtain the desired data and downloaded into their mobile device. When they move to a different location, the information that they access may have been changed due to the change of the region. So in a new region, mobile users can download other data. Assuming *C*3 is currently in Location *B* and is accessing *S*2. Once he receives the desired data, he moves to Location *C* and now accesses *S*3 which provides different data (Lim, Taniar, & Srinivasan, 2006, 2007a,b).

MOBILE PROCESSING STRATEGIES

There are three strategies to process mobile queries: mobile query processing via (*a*) *server strategy*, (*b*) *on-air strategy* and (*c*) *client strategy* (Acharya, Kumar & Yang, 2007; Bose et al., 2005; Chan, Si & Leong, 1998; Chang & Yang, 2002; Wolfson et al., 2006). Figure 2 depicts the architecture of the available strategies of query processing in a mobile environment.

In general, the *server strategy* refers to mobile users sending a query to the server for processing and then the results being returned to the user (Jayaputera & Taniar, 2005a,b; Wolfson et al., 2006). Issues,

Figure 2. A mobile query architecture



such as location dependency, should be taken into account since different locations will be accessing different servers, and subsequently it impacts on the processing by the server and the return of the results based on the new location of the mobile user (Pissinou, Makki & Campbell, 1999).

The *on-air strategy* which is also known as the *broadcasting strategy* is basically where the server broadcasts data to the air, and mobile users tune into a channel to download the necessary data (Waluyo, Srinivasan, & Taniar, 2003, 2004a,c, 2005a). This broadcasting technique broadcasts a set of database items to the air to a large number of mobile users over a single channel or multiple channels (Datta et al., 1999; Huang & Chen, 2004; Waluyo, Srinivasan, & Taniar, 2004d; Waluyo, Srinivasan, Taniar, & Rahayu, 2005). This strategy strongly addresses the problem of channel distortion and fault transmission. With the set of data on the air, mobile users can tune into one or more channels to get the data. This, subsequently, improves query performance.

The *client strategy* is where the mobile user downloads multiple lists of data from the server and processes them locally on their mobile device (Lo et al., 2004; Ozakar, Morvan, & Hameurlein, 2005; Lim, Taniar, & Srinivasan, 2007a,b). This strategy deals with processing locally in the mobile devices itself, such as when data are downloaded from remote databases and need to be processed to return a join result. Downloading both independent relations entirely may not be a good method due to the limitations of mobile devices which have limited memory space to hold large volume of data and small display screens which limit the visualization (Lo et al., 2004). Thus, efficient space management of output contents has to be taken into account. In addition, this strategy also relates to maintaining cached data in the local storage, since efficient cache management is critical in mobile query processing (Cao, 2003; Elmagarmid et al., 2003; Xu et al., 2003; Zheng, Xu & Lee, 2002).

Each of the above query processing strategies will be explored in more detail in the following sections.

Server Strategy

In general, the *server strategy* refers to mobile users sending a query to the server for processing and then the results being returned to the mobile user (Lee & Chen, 2001; Waluyo, Srinivasan, & Taniar, 2005c; Jayaputera & Taniar, 2004a,b). The problem with this strategy is promulgated by any disconnections, which may occur especially during the transmission of the query. The sudden disconnection may result in loss of information. There are several related issues that have been investigated in the server strategy related work. There are two main issues, including data placements and data scheduling. By determining how data will be placed on the server disk, this will affect the query processing in terms of having frequently accessed data being able to be retrieved more quickly thereby improving query response time.

Also, deciding which data should be given priority is important in relation to the server strategy. This aspect regarding query scheduling, which concerns issues of finding how the query is scheduled, is important, as is the issue of finding how to process the query in an efficient manner. Possible ways to process the query being issued by the mobile user can occur either on the server side or client side depending on the processing side, which will incur a lower transfer cost as well as less memory consumption.

Issues, such as location dependency, are taken into account since different locations will be accessing different servers, and subsequently this affects the processing by the server and the query results being returned based on the new location of the mobile user (Kottkamp & Zukunft, 1998). Various techniques have been developed to update the new location movement of the mobile user. The time function method provides an estimation of the location of mobile users at different times (Cao, Wang & Li, 2003). The limitation of this technique is that it avoids excessive location updates due to no explicit update being necessary. By indexing the locations, we can predict the future movement of the mobile users (Chen, Wu & Yu, 2003; Hung & Leu, 2003).

On-Air Strategy

The *on-air strategy* which is also known as the *broadcasting strategy* is basically where the server broadcasts data to the air and mobile users tune into a channel to download the necessary data (Peng & Chen, 2003; Waluyo, Srinivasan, & Taniar, 2005a). This broadcasting technique broadcasts a set of database items to the air to a large number of mobile users over a single channel or multiple channels and is related to activities that take place on-air (Chang & Chiu, 2002; Huang, Chen & Peng, 2003; Waluyo, Srinivasan, Taniar, Rahayu, Apduhan, 2006). This strategy strongly addresses the problem of channel distortion and fault transmission. With the set of data on the air, mobile users can tune into one or more channels to get the desired data items. However, unexpected situations may arise such as when a mobile user may not have enough memory to cache the desired data items, or s/he may experience a lengthy tune-in time for the desired data. Other important issues include data organization, data selection and data indexing. All these take into account optimizing the capacity, response time as well as bandwidth usage.

The question of the organization of data items is related to matching the order of the broadcast data with the order of data required by the query (Huang & Chen, 2004). The basic idea is to be able to allocate related data items in such a way that a mobile user does not have to wait for a substantial amount of time for the desired data items. This helps to reduce waiting and downloading time. Related work addresses this issue by examining the query access patterns and semantics of the queries that are being issued by the user (Lee, Leong & Si, 2002). However, a complex cost model has to be used in order to decide the best organization of data. And this raises another important issue that needs to be addressed, especially when a large number of database items are to be broadcast.

Also, by looking at recent access, the contents and organization of data broadcast can be determined (Waluyo, Srinivasan, & Taniar, 2005b,d, 2007). However, this technique appears not to be taking into account the handling of a request that has been around for quite some time. It only reduces the probability of occurrence, but if it occurs, then it is not taken into consideration. Also, whenever there is too many data in the broadcast cycle, a decision has to be made regarding prioritization.

One of the core issues is determining the priorities of the data items to be broadcast. This refers to which data items should be broadcast in the next period (Huang & Chen, 2004). Several scheduling algorithms have been proposed which include '*First Come First Served*' which sequences the data items according to their requested time and this demonstrates that any access request would receive a response after waiting a substantial amount of time (Chang & Chiu, 2002). Although there is no endless waiting,

this algorithm has the disadvantage of having low average performance because it takes into account only the requested time and ignores the difference in access frequency of various data items. Another proposed algorithm includes *'Most Request First'* which prioritizes those data items with most requests (Datta et al., 1999). However, the shortcoming of this algorithm is that those data items that have few requests would always be lined up behind the data items that have more frequent requests, and therefore those less frequent request items would have an endless waiting period. An improved algorithm from the same related work would be '*Long Wait First*' which gives priority to the data items that have the longest waiting time (Acharya, Kumar & Yang, 2007). This algorithm considers both the number of requests as well as the waiting time so as to reduce the occurrence of endless waiting.

A selection mechanism is designed to reduce the broadcast cycle length, which can reduce the query response time. During each broadcast cycle, the data items can be qualified as either hot or cold data items. Hot data items are data items that are accessed by most mobile users; conversely, cold data items are those that are less in demand. It is often important to replace the cold data items with the new hot data items, which are believed to be more in demand. Based on several existing works regarding the selection of data items, several replacement algorithms that deal with replacing cold data items with hot data items have been investigated. The proposed algorithms namely 'Mean', 'Window' and 'Exponentially Weighted Moving Average' maintain a score for each data item in order to estimate the access probability (Lee, Leong & Si, 2002). The scores are obtained by measuring the cumulative access frequencies.

Another issue that has been investigated in relation to the on-air strategy would be data indexing. Data indexing is believed to lower the tuning time by providing information for the mobile user to tune into the broadcast channel at an appropriate time when the desired data items arrive (Lee, Leong & Si, 2002). The use of indexing helps mobile users to search desired data items by determining when the data arrives, thereby reducing query processing time which benefits mobile users, as well as utilizing power more efficiently (Cao, 2002). Along with the broadcast data, some form of directory, which is known as the index, is attached. The information consists of the exact time of the data to be broadcast. And thus, while waiting for the desired data items to arrive, the mobile users can switch to "doze" mode and switch back to "active" mode when the desired data items arrived. Indicating the time that the indexed data record will be broadcast is one of the current approaches in the existing related work (Chen, Wu & Yu, 2003).

However, there are several trade-offs in the current approach. If the index is broadcast sparsely, the client might miss the index records in his first attempt and would have to keep tuning until the desired index record or the real data record are obtained. This increases the tuning time tremendously. On the other hand, broadcasting an index too frequently will increase the size of the broadcast data and thus leads to increase duration of the broadcast cycle, which eventually leads to higher database access time. However, if we eliminate the index completely, although it will yield minimal access time, the mobile user would have to listen to every single broadcast data item until the desired one is obtained. Therefore, when designing the index directory, several concerns including finding the optimal balance between tuning time and response time must be taken into account, as both will be greatly affected due to the occupancy of the index in the broadcast cycle.

One limitation of existing broadcasting mechanisms is their inability to efficiently recover from faults induced by unreliable wireless transmission which forces mobile users to wait for the next broadcast cycle if the required data item or index records is damaged during transmission. Other limitations of this existing on-air strategy involves their main concern in using a single broadcast channel and trying to limit the broadcast data by selecting the data items that are more in demand.

Client Strategy

The *client strategy* relates to data caching in mobile databases, which allows mobile users to obtain as high a computing speed as the server by involving a much smaller volume of data items. It also maintains cached data in the local storage since efficient cache management is critical in mobile query processing (Cao, 2003; Elmagarmid et al, 2003; Xu et al, 2003; Zheng, Xu & Lee, 2002). It deals with the question of how mobile users maintain and manipulate the data in its local cache in an efficient manner. The advantages are amplified since each mobile user is likely to initiate queries frequently within a short time span. On most occasions, with the inherent characteristics of a mobile environment that suffers from narrow bandwidth and frequent disconnection, caching the frequently accessed data in the local cache will aid in enhancing the performance, especially data availability. Most existing work on client strategies, discusses in particular, issues relating to caching replacement, granularity and coherency or invalidation (Cao, 2003; Chuang & Hsu, 2004; Elmagarmid et al., 2003; Hu & Lee, 1998).

Due to limited memory capacity, cache replacement needs investigation. The replacement policy discards old cache data items that are no longer relevant or are out of date and replaces them with the newly obtained data items. The issue that needs to be addressed with respect to this policy is to determine which data items are no longer needed and need to be replaced. This has to be addressed carefully because if a bad replacement policy is being used, then it may result in waste of energy as well as memory space since the mobile user may not be able to use the cached data but will still have to send a query to obtain the desired data items. Thus, the effectiveness of a caching replacement will affect the performance of the queries: if an effective cache replacement is used, the processing of a query will be much better, and this enhanced performance will also allow a greater number of cold queries to be served, especially during a disconnection situation. Most of the replacement policies that were investigated involve utilizing access probability as the primary factor when determining which data items are to be replaced.

For caching in respect to location dependency, the distance between the mobile user's current location and that of the cache data needs to be considered and often it is associated with semantic caching. Semantic caching stores semantic descriptions and associated answers for the previously issued queries. Due to the rapid movement of mobile users as well as the location parameter, by having semantic descriptions of the previous queries, the performance of future queries will, supposedly, be enhanced (Lee, Leong, & Si, 2002).

Cache granularity relates to determining the physical form of cached data items. It appears to be one of the key issues in cache management systems. There are three different level of caching granularities in object-oriented databases which include (*a*) attribute caching, (*b*) object caching and (*c*) hybrid caching (Chan, Si, & Leong, 1998). Attribute caching refers to frequently accessed attributes that are stored in the client's local storage. As for object caching, instead of the attribute itself being cached, the object is cached. However, attribute caching creates undesirable overheads due to the large number of independent cache attributes. Thus, hybrid caching appears to be a better approach, since it takes advantage of both granularities.

Cache coherence, also known as invalidation strategy, involves cache invalidation and update schemes to invalidate and update out-dated or non-valid cached items (Chan, Si & Leong, 1998; Cao, 2003). After a certain period, a cached data may no longer be valid and therefore mobile users should obtain a newer cache before retrieving the data (Tan, 2001). There are several techniques that have been proposed to overcome this issue. These include (*a*) stateful server, (*b*) stateless server, (Barbara & Imielinski, 1994) and (*c*) leases file aching mechanism (Lee, Leong & Si, 2002). A stateful server refers to the server having an obligation to its clients; that is, it is responsible for notifying the users about changes, if there are any. In contrast, a stateless server refers to the server not being aware of its clients. Therefore, the

server broadcasts a report, which contains the updated item either asynchronously or synchronously. The leases files mechanism, also known as the lazy invalidation approach, assigns each mobile user the responsibility for invalidating its cached items. The main concern of this strategy lies is in determining the refresh time for the cached data.

Currently, the client strategy focuses mainly on traditional queries and is not applicable to a wireless communication environment, especially when this involves more complex queries such as location dependent queries that can be either a series of continuous queries or on-demand queries. Thus, the main drawback of the client strategy relates to the new nomadic queries.

MOBILE QUERIES

Queries play an important part in mobile processing strategies. This section describes query types classification in a mobile environment. The general query types are divided into two classes: (*i*) Traditional and (*ii*) Mobile Queries. The traditional query type category contains common query types that exist in a wired network database, whereas the mobile query contains queries that exist only in a wireless environment.

Traditional queries, typical database queries, normally contain some form of spatial and/or temporal. Hence, they are normally based on the geographical presentation, which can be location-aware or non-location. In the mobile computing environment, the location of mobile users is dynamic and the query results often depend on this dynamic location. Therefore, this situation creates another additional class, which is called Location-Dependent queries. Hence, mobile queries are generally location-dependent and/or location-aware queries.

Traditional Queries

Traditional query is the most widely known query used in a database. The query types of traditional query can be classified as: *Spatial, Temporal, Spatio-Temporal* (Hybrid), and *Others*.

A *Spatial* query performs operations, which include spatial searches and map overlay, as well as distance-related operations (Gaede & Günther, 1998). A spatial query always requests for spatial data information. Spatial data means that the requested data have a complex structure, are often dynamic and no standard algebra are defined.

A *Temporal* query specifies a validity or deadline for the query results to be returned. Example: "A student retrieves a subject timetable for this year". The subject timetable will not be valid for the past or future year.

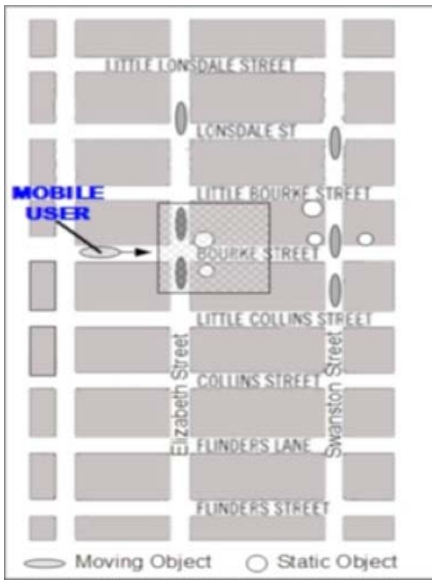
A *Spatial-Temporal* query requests for a spatial search and specifies the validity or deadline for the query results to be received. For example: "Retrieve the five ambulances that were nearest to the location of the accident between 4-5pm." (Porkaew et al., 2001).

The last category is *Others*. It implies that the other remaining queries do not belong to one of the classifications above. For examples: A tourist requests restaurant information, or students request their academic records or contact details.

Mobile Location-Dependent Queries

Imielinski & Badrinath (1992) were the first authors to introduce the idea of queries with location constraints. These types of queries have one parameter: location. It implies that the query result is related to or depends on, that parameter.

Figure 3. Location-dependent query



Location-Dependent Query (Zheng, Xu & Lee et al., 2002; Lee, Xu, Zheng, Lee, 2002; Ren & Dunham, 2000) is a type of query where the answers depend on the current location of the requesters. For example, "Select all restaurants within 500 metres from my location". The answer should give a list of restaurants within 500 metres from the current location of the requester as illustrated in Figure 3. If the requester moves to a new location, the list of restaurants will be changed. A location is an important field in this type of query and this field can be implicitly or explicitly mentioned in the query (Ren & Dunham, 2000).

These types of queries can be further categorised into two groups. The first group is based on *sources and objects*, and the second one is based on *query retrieval* (Waluyo, Srinivasan & Taniar, 2005e). The sources and objects are represented as users while sending the query and the searched objects. Their states can be either static or moving. The second state is based on the states of the query retrieval either one-time or continuous. A one-time query is a query that expects query result in one-time. On the other hand, a continuous query, as the name implies, is a query that receives a query result, which is based on the current location of the source at some moment in time. This query is sent only once and updated location information is sent to notify the server that the client has moved to a different location. Both groups mentioned above can be further elaborated as follows.

Data Sources and Objects States

This group focuses on states of location for either users or objects while a user query is being proceed. The states of location for both can be static or dynamic during the query processing.

As we can see from Figure 4, this kind of query is divided into four subgroups. The first subgroup is a static user probes for static object/s. This subgroup does not involve a mobility factor for either users or objects. Whenever the query is sent, the query result returned will always be the same. Therefore, the first subgroup cannot be included as a Location-Dependent Query. The rest of the three subgroups are: (*i*) moving user probing static object/s, (*ii*) moving user probing moving object/s, and (*iii*) static user probing moving object/s.

Figure 4. Location query category based on data sources and object states

	Static User	Moving User
Static Object	✗	✓
Moving Object	✓	✓

Moving user searching for static object/s is where the user or requester is moving while issuing a query and the requested query results are static. Examples of this type of query are: While a taxi driver is driving, the driver requests a list of restaurants within 500 metres from the current location. A tour guide in a moving car requests information about tourist attractions nearby. In the first example, the searching distance is explicitly mentioned, whereas in the second one, the searching distance is not mentioned. This situation is not only applied for this type; it can also be applied to the other two types. Seydim, Dunham & Kumar (2001) and Trajcevski et al (2004) give the common operators for constrained location-dependent queries, which can be applied to location-dependent queries in both groups.

Moving user searches for moving object/s is where both users and objects are moving. Some of the examples are: A walking person is searching for an available taxi close to his location. A police in a patrol vehicle are pursuing a running thief.

Static user searching for moving object/s is where the user remains in the same position while asking for moving object/s. Some examples of this query type are: A security officer in a control room is searching for a fleeing thief. An officer in a control room is asking for landing time when an aircraft is landing.

Query Retrieval States

This type of query relates to how often the query result is expected to be received; that is, whether it is periodic or one-time. Figure 5 incorporate the previous query types with the query retrieval states.

One-time Query expects a query result to be received once. It means that this query does not depend on the time interval. All the query types in the previous section are one-time queries if their results are received once.

Periodic Query is similar to one-time query, except query results are received at every time interval and the time interval is specified in periodic query. Periodic query is also called range-monitoring query (Cai & Hua, 2002). It is used for monitoring query continuously. The returned query results of periodic query may be the same as or different from the previous query results in a past interval time. For example: "A moving car is asking for traffic conditions within 500 metres for every 5 minutes".

Figure 5. Location query category based on query retrieval states

	Periodic	One-time
Static User – Moving Object	✓	✓
Moving User – Static Object	✓	✓
Moving User – Moving Object	✓	✓

MOBILE APPLICATION DEVELOPMENT TOOLS

Mobile application development tools include hardware, software, and network. In this section, these development tools are highlighted. These tools will be used in the case study section following this section. Hence, only the development tools applicable to the case study will be included.

Hardware

There are three important physical aspects need to be addressed, which are as follows: (1) a suitable server device; (2) a suitable client device; and (3) a suitable transmission medium. The server device acts as a data source and as an intermediary service provider to the client device, and subsequently the client device accesses data from the server device over a wireless network. The server device and client device generally follows a traditional client/server architecture (Stamper, 2001). This is especially true when the client device requests data or information from the server device. However, when the server device pushes data to the client, the client device can be viewed as a 'server' to the server device. Regardless of whether the data flow is pull-based or push-based, the server device acts as a data source and as a means of conveying the data to clients.

The server device is a desktop computer, whereas the client device is a Pocket PC-based Personal Digital Assistant (PDA). The PDA communicates with the server device over a wireless LAN. Since the transmission medium is wireless, the nominated wireless transmission standard is 802.11b as defined by the Institute of Electrical and Electronic Engineers (IEEE) (Blake, 2002).

Server Device

The server pushes data or responds to pull-based requests generated by the client. Since the nominated wireless transmission standard is 802.11b, the server consequently requires an 802.11b-compliant wireless network interface device. The wireless network interface device is either a built-in device or an external peripheral device.

Client Device

The Personal Digital Assistant acts as a portable client device to the server. It communicates with the laptop server program via an IEEE 802.11b-compliant wireless network interface device. The wireless network interface device can be a built-in device or an external peripheral device. Newer varieties of PDAs often have built-in IEEE 802.11b-compliant wireless network interface devices. Older varieties of PDAs do not include built-in IEEE 802.11b-compliant wireless network interface devices. To overcome this limitation, we can add the PDA with a sleeve-device that enables additional peripherals (e.g. wireless network interface cards) to be installed.

Wireless Network Interface Card

In the event that neither server nor the PDA includes a built-in wireless network interface, the installation of a peripheral wireless network interface card will be necessary. Another wireless technology, known as Bluetooth, also offers wireless connectivity between Bluetooth-enabled devices. Like 802.11b, Bluetooth technology can be used to set up a wireless network. The main difference between Bluetooth and 802.11

is in the data transmission rate and signal range (Blake, 2002). Hence, 802.11b is the favoured wireless medium, especially in our case study.

SOFTWARE

In addition to the appropriate hardware, suitable software also needs to be acquired and installed. This is especially true for the server device and client device. Four important software aspects need to be considered and evaluated. The four aspects are as follows: (1) the development software; (2) the operating system software; (3) the transmission protocol; and (4) the database software. Each software aspect is equally important and influences the type of hardware required to create the test environment.

Operating Systems Software

The operating system software serves two fundamental purposes: to controls and operate hardware in an efficient manner; and to empower the user with various facilities and services (Englander, 2000). In our case, the minimum required operating system software for the server device is Microsoft® Windows™ 2000. The reason for this requirement is due to programming language environment, Microsoft® Embedded Visual Basic® 3.0, requires a minimum of Microsoft® Windows™ 2000 as the operating system on the server device. Furthermore, Microsoft® Windows™ 2000 features a set of network-related services required by the proposed model.

With regards to the client device, the minimum required operating system is Microsoft® Pocket PC 2002. It is a Windows™ CE-based operating system designed for portable devices. The Hewlett Packard iPAQ™ H5450 Pocket PC Personal Digital Assistant used for the proposed model has a pre-installed copy of Microsoft® Pocket PC 2002. Furthermore, Microsoft® Pocket PC 2002 also features a set of network-related services required by the proposed model.

Development Software

In our case study, we use two software development products: Microsoft® Visual Basic® 6.0 and Microsoft® eMbedded Visual Basic® 3.0. Microsoft® Visual Basic® 6.0 is an object-oriented/event-driven high-level programming language and is a convenient a software development tool for creating desktop Windows applications (Zak, 1999).

Microsoft® eMbedded Visual Basic® 3.0 is an object-oriented/event-driven high-level programming language. It is similar to Microsoft® Visual Basic® 6.0; except, it is a software development tool for Microsoft® Windows™ CE-based devices. The Pocket PC-based Personal Digital Assistant is an example of a Windows™ CE-based device. Like Microsoft® Visual Basic® 6.0, portable Windows™ CE-based applications can be created using Microsoft® eMbedded Visual Basic® 3.0. The Microsoft® eMbedded Visual Basic® 3.0 software development tool also enables the creation of Windows-style applications for Pocket PC. The Microsoft® eMbedded Visual Basic® 3.0 software is part of the Microsoft® Embedded Visual Tools 3.0 software package. Microsoft® Embedded Visual Tools 3.0 is freely available for download from the Microsoft website.

Both Microsoft® Visual Basic® 6.0 and Microsoft® eMbedded Visual Basic® 3.0 are installed on the server device. The server device acts as server in the proposed model; however, it also acts as a software development platform for creating server and client applications.

Database Software and Data Access

Our case study uses a relational database as the data source. To create and maintain a relational database, a Relational Database Management Software (RDBMS) is required, which in this case Microsoft® Access® is adopted for convenient purposes.

In order for the server component to access the database, the server application uses version 2.7 of the Microsoft® ActiveX Data Objects (ADO) (Gunderloy, 2002). ADO 2.7 is a necessary software component that enables Visual Basic® 6.0 applications to connect to various data sources such as a Microsoft® Access® database. Pre-Windows™ 2000 computer systems require the installation of Microsoft® Data Access Components (MDAC) to obtain a copy of Microsoft® ActiveX Data Objects. The server application uses Structured Query Language (SQL) strings and ADO 2.7 to search and manipulate the database. To perform database searching and manipulation, the server application creates an ADODB Recordset object (which implicitly creates an ADODB connection) using the SQL string to retrieve the necessary records from the database. With the creation of the Recordset object, the server application can manipulate the database by: inserting new records; updating records; or deleting records.

Transmission Protocol, Winsock, and Network Settings

To create a wireless environment, two issues need to be addressed. One issue relates to the type of hardware required. Another important issue relates to the type of wireless network and associated transmission protocol. The nominated transmission protocol for the proposed model is actually two protocols; namely the Transmission Control Protocol and Internet Protocol (TCP/IP). Many Wide Area Networks (WANs) and the Internet are based on TCP/IP (Blake, 2002); however, TCP/IP is also used for wired IEEE 802.3 Ethernet Local Area Networks (LANs). The difference between a WAN and a LAN is largely geographic.

Our case study uses IEEE 802.11b wireless network interface cards to create a wireless local area network. The service device and the Personal Digital Assistant both use 802.11b to create a wireless ad-hoc network. The two protocols TCP/IP are used to manage the transmission of messages between the server device and client device over the 802.11b wireless network. The Transmission Control Protocol (TCP) is responsible for assembling messages into datagrams and ensures that the messages are properly delivered to the correct destination. The Internet Protocol handles the routing of datagrams through the network to the desired destination.

TCP/IP is a necessity on the server device and client device because the two software development programs use a control called *Winsock*. The Winsock control is added to both the server application and client application at design time (Zak, 1999). This enables the server application and client application to communicate messages over a TCP/IP network. Either device can initiate a connection with each other; however, only the client device is required to create the connection. It is noteworthy that the Winsock control offered in Microsoft® Visual Basic® 6.0 differs slightly, in terms of features, from the Winsock control (referred to as Winsock CE) offered in Microsoft® eMbedded Visual Basic® 3.0. However, the differences between the two controls are negligible in regards to the proposed model.

Finally, the TCP/IP network settings on the server device and client device need to be set to the correct address range and subnet mask (Stamper, 2001). The TCP/IP settings for the 802.11b-compliant wireless interface device on the server are set to: 192.168.0.1 for the IP address; and 255.255.255.0 for the subnet mask. The TCP/IP settings for the 802.11b-compliant wireless interface device on the client device are set to: 192.168.0.2 for the IP address; and 255.255.255.0 for the subnet mask. These settings enable the server application and client application to communicate with each other over the TCP/IP network.

MOBILE E-HEALTH APPLICATION

In this case application, we apply (*i*) *pull-based*, and (*ii*) *push-based* mechanisms in a wireless environment. We use a simplified e-health (hospital) context to develop a hospital server and client application (Waluyo, Taniar, and Srinivasan, 2007). The hospital context relates to doctors as the principal clients to a server application. The application will demonstrate the usability of wireless networks, and improve the mobility of doctors through wireless data dissemination.

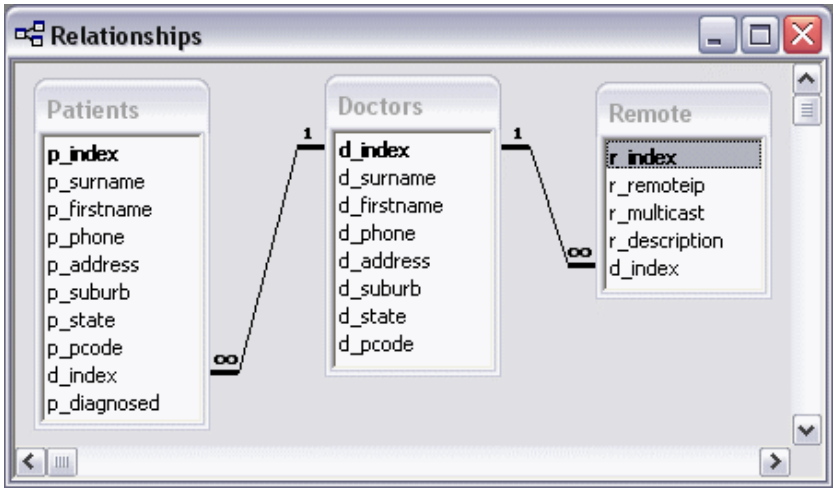
There are two ways of data delivery in wireless environment. One is called *pull* mechanism, and the other is *push* mechanism (Aksoy, et al, 1999, Waluyo et al, 2004). Pull mechanism is when the data are delivered on a demand basis. In the e-health context, we apply this mechanism for doctors to retrieve his/her patients. In a push mechanism, the server initiates the delivery of data without a specific request from the client. We apply this mechanism to send a direct message to a specific doctor, and to distribute information to all or selective doctors such as news bulletin.

The push mechanism can be categorized into 1-1 (unicast) and 1-*N* communication type. Unicast communication involves a server and a client, and the data is sent from the server to the client. 1-*N* communication can be either multicast or broadcast mode. In multicast mode, the recipients are known and the data are delivered only to those recipients. For example, the information is delivered to doctors and nurses that are registered in a specific domain. On the contrary, the broadcast mode simply sent the data without knowing the number of clients who might receive the data. This case study deals only with 1-*N* (multicast mode) communication type.

Database Setup

The database utilised by the case study comprises of three related tables. The three tables are entitled *Doctors*, *Patients*, and *Remote*. The Doctors table stores records of doctors employed by the hospital. The Patients table stores records of patient details; including the doctor assigned to care for the patient. Finally, the Remote table stores a list of IP addresses. Each IP address stored in the Remote table is assigned to a doctor. The following diagram shown in Figure 6 depicts the relationship between the three tables.

Figure 6. Table relationships for the hospital database



Server: Application Overview

The server component of the case study application is programmed in Microsoft® Visual Basic 6. Like the test application, the server component of the application utilises Microsoft® Data Access Components (MDAC) 2.7. The server application connects to the data source via an ADODB connection. The server application uses Structured Query Language (SQL) strings to search and manipulate the database. The server component also utilises the Winsock control to enable network connections over wireless TCP/IP a network.

Visually, the server application form comprises of: a command button entitled '*Send Message*'; two text boxes; one list box; three timers; a Winsock control; a Common Dialog control; and a menu item '*New → Client*'. Figure 7 depicts the server form during design time.

When the server application is executed, the image shown in Figure 8 is displayed. The server application responds to four human-generated events: (1) opening a dummy client; and (2) sending a message to a specific doctor; (3) multicasting a global message to all doctors connected to the server application; and (4) closing the server application. The command button '*Send Message*' is used to send a message to a specific doctor selected from the list box. The list box is populated when a doctor (or doctors) connects to server application. The '*Send Message*' command button is only enabled when at least one doctor is connected to the server application. Finally, the text box labelled '*News Bulletin*' is used to multicast a global message to all doctors connected to the server application.

Figure 7. Server form at design time

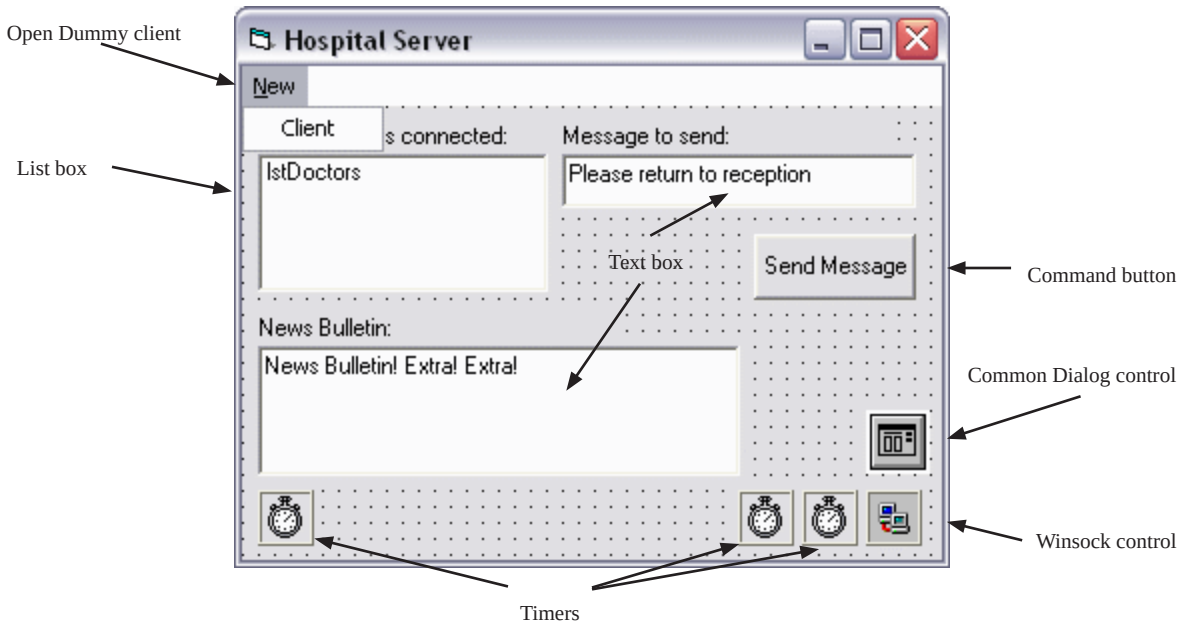
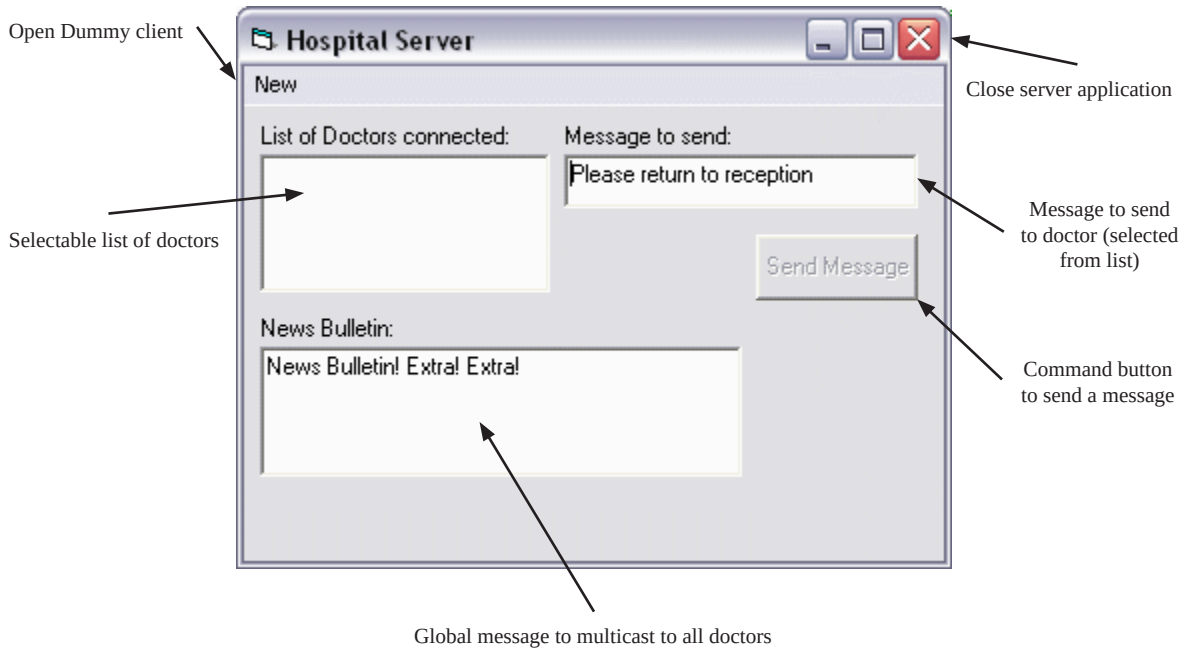


Figure 8. Server form at run time



Client: Application Overview

The client component of the test application is programmed in eMbedded Visual Basic 3. The eMbedded Visual Basic program utilises the Winsock CE control to enable network connections. The application form comprises of: four command buttons; two text boxes; one list box; one check box; one timer; and a Winsock CE control. Figure 9 depicts the client form at design time.

When the client application is executed on the Personal Digital Assistant, the image shown in Figure 10 is displayed to the client. The client application responds to events, as follows: (1) clicking any of four command buttons; (2) clicking the check box; and (3) closing the client application.

When the doctor clicks the command button '*Connect to Server*', the application attempts to connect to the server. Once connected, the doctor can choose to retrieve relevant information about patients from the database (via the server application). This is achieved by clicking the '*My Patients*'. Once a list of patients is returned to the list box labelled '*My Patients*', the doctor can choose to diagnose a patient. The doctor can diagnose a patient by: selecting the desired patient; and then clicking the '*Diagnosed*' button. The client can also enable or disable the multicasting of a global news bulletin (sent by the server application). The doctor simply enables or disables the multicasting feature by selecting or deselecting the check box labelled '*Receive News Bulletin?*'. The doctor can choose to disconnect from the server application by clicking the '*Disconnect from Server*' command button. Finally, clicking the '*OK*' button situated to the top right of the application window causes the client application to close.

The objective of the client application is: to display the results client-specific messages; a global news bulletin; retrieving patient records; and diagnosing patients. The client-specific messages and the global news bulletin result from push-based communications. Retrieving patient records and diagnosing patients result from pull-based communications with the server application. The specific details of the

Figure 9. Client form at design time

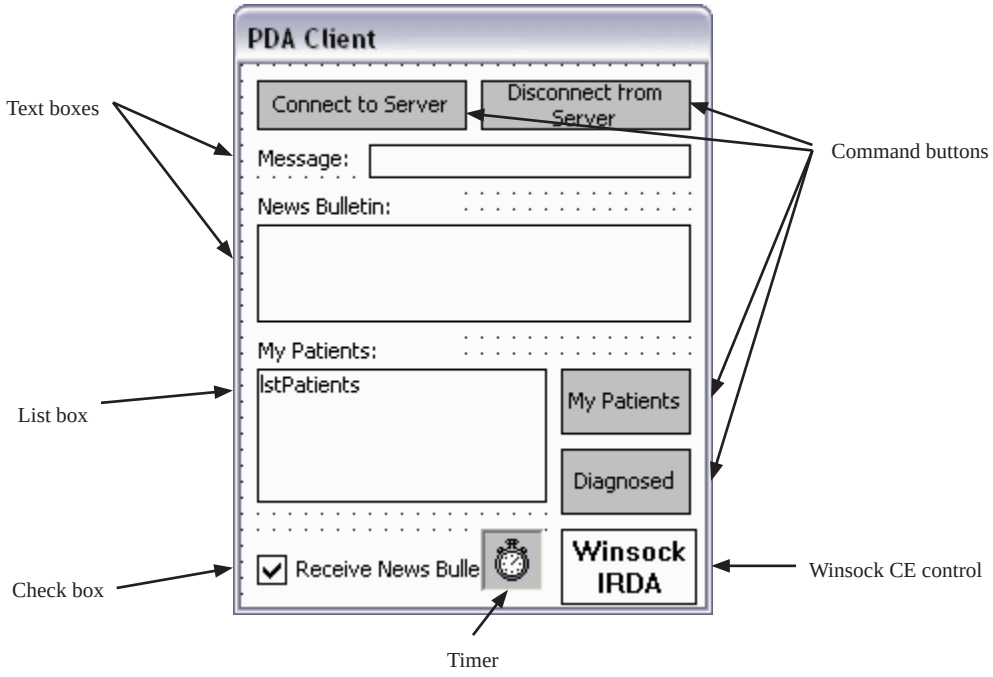
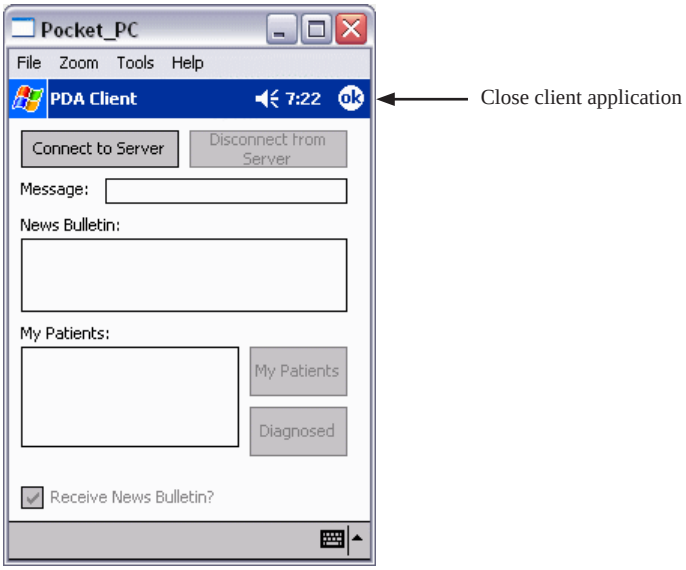


Figure 10. Client form at run time



techniques employed to achieve the pull-based and push-based communications are discussed in later subchapters.

Pull-Based #1: Client Retrieving Data from the Database

When a doctor connects to the server application, the doctor may retrieve a list of patients from the database. Firstly, the doctor needs to establish a connection from his/her PDA to the server. When the connection is established, a specific identity value is automatically sent by the server application to the client application. The specific identity value sent by the server application is the primary key value assigned to the doctor in the Doctors table. The server application uses the Doctors table in conjunction with the Remote table to determine the necessary identity value required by the client program. Figure 11 highlights the foreign key of the Remote table that relates to the Doctors table.

The relationship between the Doctors and Remote table is used, in conjunction with the IP addresses stored in the `r_remoteip` field of the Remote table, to determine the correct identity value to send to the client device.

Figure 11. The Remote Table showing the identity field for doctors (d_index)

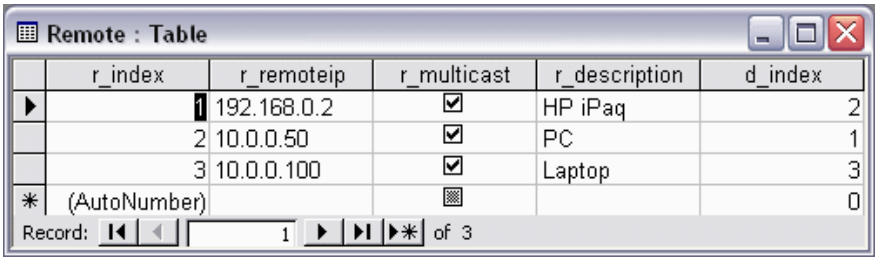


Figure 12. The 'My Patients' button

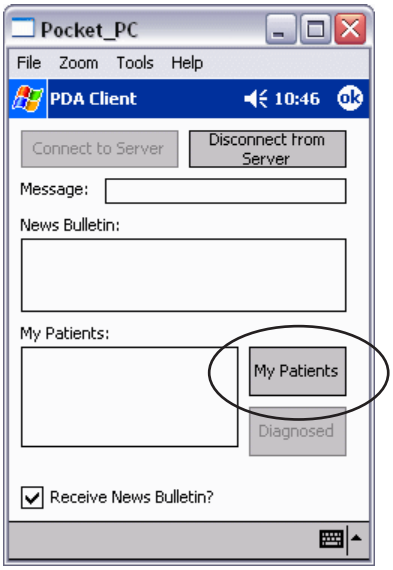
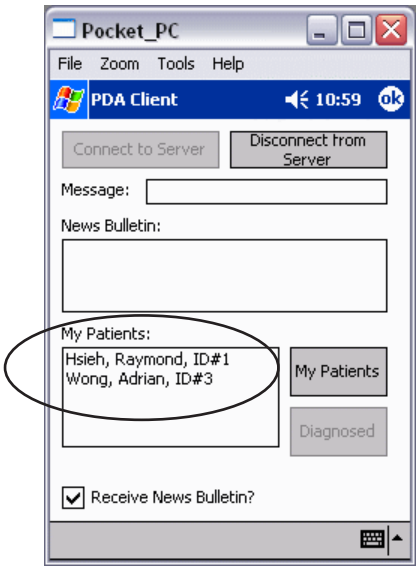


Figure 13. Client details returned for display



If a doctor wishes to retrieve a list of patients, the doctor simply clicks the '*My Patients*' button on the client application. Clicking the '*My Patients*' button triggers an event procedure on the client application. Figure 12 shows the 'My Patients' button on the client program.

This sends the identity value (supplied as a result of the initial connection request) to the server application. Since the client generates and sends the message, the transmission is a *pull-based*. When the server application receives the complete message, it sends an SQL query to the database to process the request. It basically creates a query string, which searches the Patients table for undiagnosed patients. Furthermore, the query string returns patient records relevant to the doctor.

The complete message containing the requested records is sent by the server application to the client application. The client application parses the complete message and processes the complete message for display. The client application updates the list box entitled '*My Patients*' by displaying the surname, first name and identity value of each patient (see Figure 13).

Pull-Based #2: Client Updating the Database

As an extension to retrieving patient information from the database, the client application also enables the doctor to diagnose a patient. Once a list of patients is retrieved from the server application, the doctor may click the '*Diagnosed*' button to remove a selected patient from the list of patients. Figure 14 illustrates a selected patient and the relevant '*Diagnosed*' button

The '*Diagnosed*' button sends the identity value of the patient to the server application for processing. To obtain the necessary identity value, of the selected patient in the list box is evaluated. When the '*Diagnosed*' button is clicked, the procedure extracts the identity value of the patient by parsing text of the selected patient (in the list box).

The complete message is sent by the client application to the server application. Once again, the transmission is a *pull-based*. This is due to the message originating from the client application. The server application receives the message and after determining the nature of the message from the keyword, the server application extracts the patient and doctor identity values from the message. Then it uses the patient identity value to update the relevant record in the Patients table.

Figure 14. Selected patient and 'Diagnosed' button

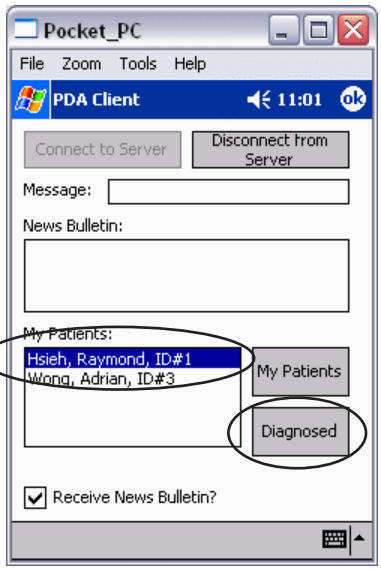
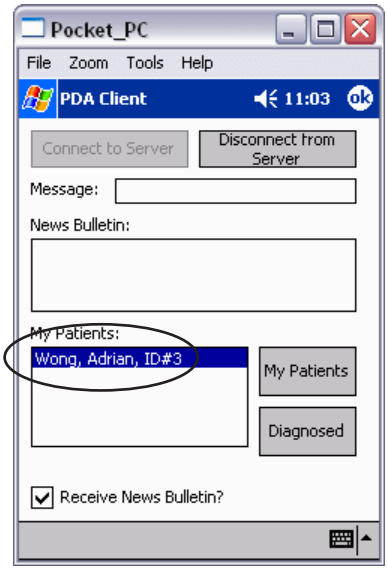


Figure 15. Patient removed from client display



Once a patient is diagnosed, the list of patients on the client application is updated to reflect the changes made to the Patients table. Figure 15 illustrates the updated client display.

Push-based #1: Sending a Message from Server to Client

Whenever a doctor connects to the server, a list of all doctors currently connected is displayed on the server application. This is shown in Figure 16. The list is created or updated whenever a doctor connects or disconnects from the server application. When a doctor attempts to connect the server, a connection request is sent. The event procedure creates a connection for each connection request; and adds the details of the doctor to the list.

Figure 16. One client connected to the server

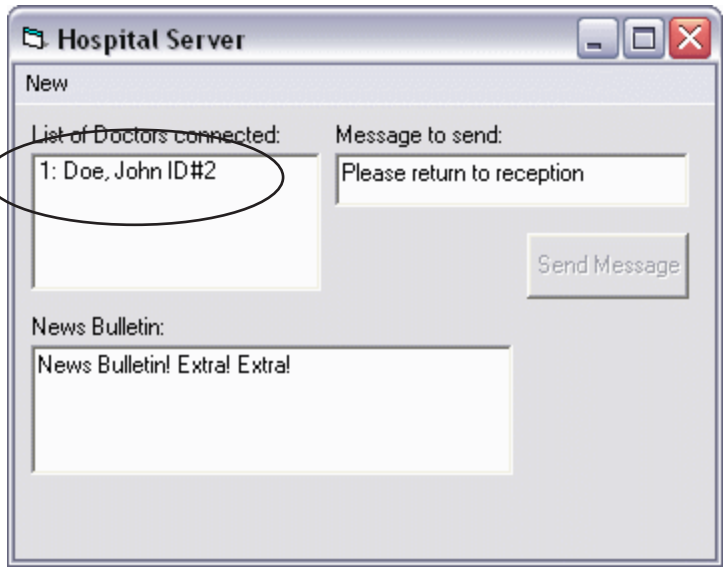
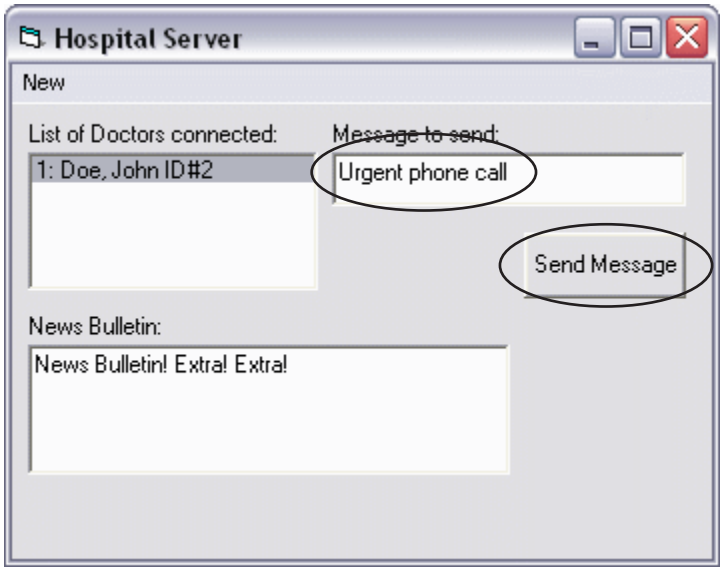


Figure 17. Sending a message from server to client

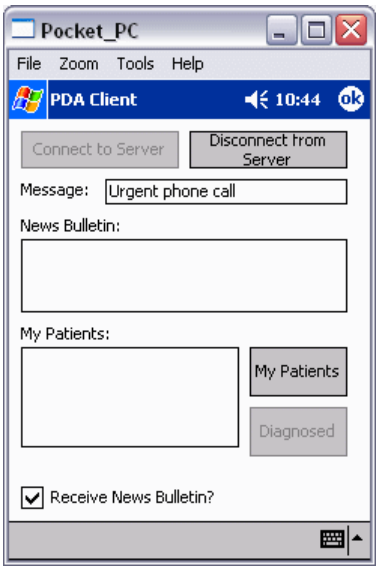


To send a message to the client, the server operator simply: selects the desired client from the list; types the desired message into the '*Message to send*' text box; and clicks '*Send Message*'. This is shown in Figure 17.

When the server operator clicks the '*Send Message*' button, the selected doctor in the list box is evaluated to determine the necessary connection identifier value required by the Winsock control. The Winsock control uses the connection identifier value and a method to send the message to the correct doctor.

When the client application receives a message, the entire message is processed to determine the nature of the message. It then extracts the message and displays it to the '*Message*' text box on the client

Figure 18. Result of sending message



application. Figure 18 demonstrates the result of the server application sending a message to the client application.

Push-Based #2: Multicasting a Message from Server to Client(s)

The server application may automatically send a news bulletin to each doctor connected to the server. The server application uses a timer to periodically send the news bulletin to each doctor. Whenever a client connects to the server application, the IP address of the client is stored in a form-level string array on the server application. This occurs during the connection on the server application. The server application determines which clients are the recipients of the news bulletin. Furthermore, the server application also determines if the doctor wishes to receive a copy of the news bulletin. This is determined by the `r_multicast` field in the Remote table. If the checkbox field in `r_multicast` is checked, then the doctor requires a copy of the news bulletin. Figure 19 depicts the table and fields used by the server application.

When the timer interval expires, the event procedure performs a check to determine which doctors require a copy of the news bulletin. If a match is determined, the event procedure retrieves the text message in the '*News Bulletin*' text box (see Figure 20). The news bulletin is then sent to each client that requires a copy of the news bulletin. Since the message originates from the server application, the transmission is push-based.

To modify the contents of the news bulletin, the server operator simply changes the text in the '*News Bulletin*' text box. When the twenty-five second interval expires, the new contents of the '*News Bulletin*' text box are sent to each doctor who requires a copy of the news bulletin.

When the client application receives a message, the entire message is processed to determine the nature of the message, which in this case it extracts the message and displays it to the '*News Bulletin*' box on the client application. Figure 21 demonstrates the result of the server application sending a message to the client application.

Figure 19. Remote table and the fields used for multicasting the news bulletin

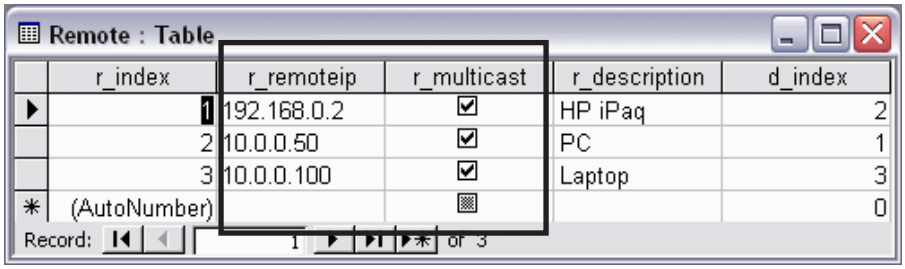


Figure 20. Sending a news bulletin to doctors connected to the server application

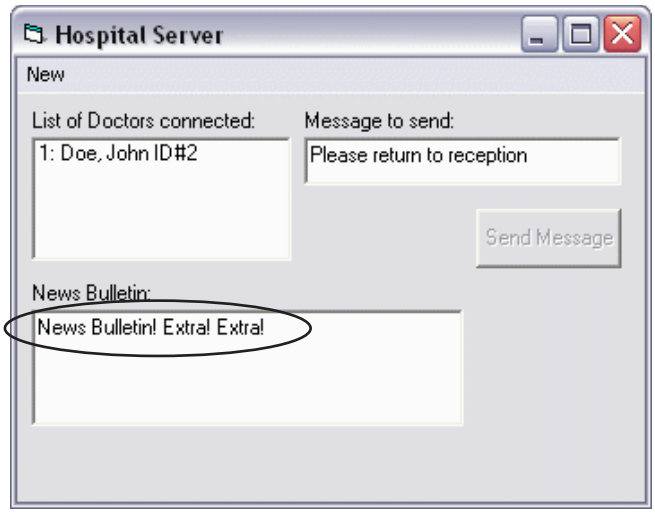


Figure 21. Result of sending the news bulletin

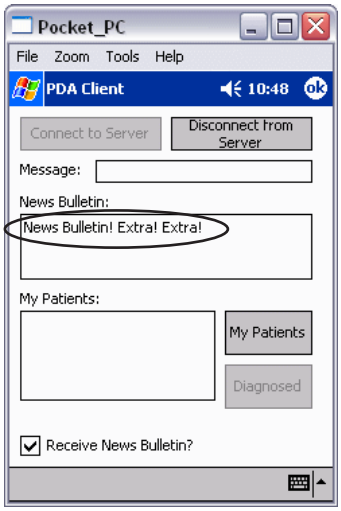


Figure 22. Check box for receiving news bulletin

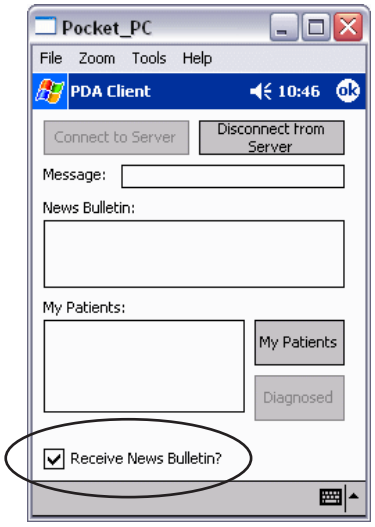


Figure 23. Remote table and the field affected by update

Remote : Table

	r_index	r_remoteip	r_multicast	r_description	d_index
▶	1	192.168.0.2	☑	HP iPaq	2
	2	10.0.0.50	☑	PC	1
	3	10.0.0.100	☑	Laptop	3
*	(AutoNumber)		▩		0

Record: 1 of 3

The client can enable or disable the automatic sending of the news bulletin by selecting or de-selecting the check box marked '*Receive News Bulletin?*" This is shown in Figure 22.

Clicking the '*Receive News Bulletin*' check box triggers send a message to the server application. The server application receives the message from the client application. Once the record is located, the value of `r_multicast` is set to *True* or *False* for *BulletinYes* and *BulletinNo* respectively. Figure 23 highlights the `r_multicast` field that is updated.

CONCLUSIONS AND FUTURE CHALLENGES

The advances of mobile computing and technologies will have to be matched with advanced applications. One of the major applications is related to data-intensive domain. This chapter highlights the importance of mobile data processing, which will play a critical role in our daily lives. This chapter's main focus is to outline a mobile data processing architecture, including the *server strategy* focusing on processing mobile on-demand queries, the *on-air strategy* focusing on data broadcast for disseminating data to mobile users, and the *client strategy* focusing on on-mobile data processing, including caching and other processing which can be done locally in mobile devices.

Mobile data processing is centred around mobile queries. When a mobile user invokes a query, the query may be location-dependent. Therefore, mobile queries are generally location-dependent. The queries become more complex as the queries themselves may move, and therefore, it is important to ensure that the query results are correct when the results finally arrive at the mobile device, which initially sends the query. Not only the queries may move, so do the queried objects themselves. Quering moving objects is one of the main challenges of mobile queries.

In this chapter, we have also demonstrated an application using a simplified e-health (hospital) context to demonstrate some effective uses of pull-based and push-based mechanisms. The hospital scenario relates to doctors as the principal clients to a server application. Furthermore, it demonstrates the usability of wireless networks, and to improve the mobility of doctors through wireless data dissemination.

Data processing in a mobile environment raises interesting challenges, not only in terms of its processing efficiencies, but also its *intelligence* aspects. There have been works in incorporating data mining techniques to analyse mobile users movements (Taniar & Goh, 2007; Goh, Taniar, & Lim, 2006). These researches focus on how movement patterns can be generated by analysing movements of mobile users (Goh & Taniar, 2004a,b,c,d; 2005a,b; 2006 a,b). Future trends would include not only analysing mobile users movements, but also other aspects of mobile users, including communications (e.g. phone calls, text messaging), entertainments (e.g. mobile games) and news (e.g. mobile TV), all of which are related to their locations.

Another important challenge in mobile application involves the context and semantic of mobile query processing. There has been an extensive list of work in mobile *context-aware* (Kottkamp & Zukunft, 1998; Stan & Skadron, 2003). When considering the context or the semantic of mobile queries, *ontology* may also be used. Most of the work on ontology semantic has been in the areas of web and the grid, including web and grid services (Taniar & Rahayu, 2006; Flahive et al, 2004, 2005). It is a challenge to apply the emerging ontology and semantic technology to mobile applications.

Finally, it is undoubted that *XML* technology has been growing rapidly, exist in a wide range of applications, and mobile technology is without exception. Existing work in XML databases include query processing, storage, updates, may be applied to mobile databases (Pardede, Rahayu & Taniar, 2005; Rusu, Rahayu & Taniar, 2004, 2005). By adopting the XML technology in mobile environment, it is expected that the transition from the wired to the wireless technology will be seamless, and information integration among various environments will also be smooth.

REFERENCES

Acharya S., Alonso R., Franklin M. & Zdonk S. (1995). Broadcast Disks: Data Management for Asymmetric Communication Environments, *Proceedings of ACM SIGMOD* (pp.199-210)

Acharya, D., Kumar, V., & Yang, G-C. (2007). DAYS mobile: a location based data broadcast service for mobile users, *Proceedings of the ACM Symposium on Applied Computing* (*ACM SAC*) (pp. 901-905)

Agrawal, D. P. & Zeng, Q.-A. (2006). *Introduction to Wireless and Mobile Systems*, 2nd edition, Thomson

Agrawal, P. & Famolari, D. (1999). Mobile computing in next generation wireless networks, *Proceedings of the 3rd International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, ACM Press (pp. 32–39)

Aksoy D., Altinel M., Bose R., Cetintemel U., Franklin M., Wang J., and Zdonik S. (1999). Research in Data Broadcast and Dissemination, *Proc. of AMCP*, LNCS, 1554:194-207.

Barbara, D. & Imielinski, T. (1994). Sleepers and Workaholics: Caching Strategies in Mobile Environments (Extended Version), *Proceedings of ACM SIGMOD International Conference on Management of Data* (pp. 1-34)

Barbara, D. (1999). Mobile Computing and Databases-A Survey, *IEEE Transactions on Knowledge and Data Engineering*, 11(1), 108-117

Blake, R. (2002), *Electronic Communication Systems*, 2nd Edition, Delmar (div. of Thomson Learning, Inc.), New York, U.S.A

Bluetooth (2008). http://www.bluetooth.com. Last accessed: 02/04/08

Bose, I., et al (2005). Databases for Mobile Applications, *Encyclopedia of Database Technologies and Applications* (pp. 162-169)

Bria A., Gessler F., Queseth O., Stridh R., Unbehaun M., Wu J., & Zander J. (2001). 4 Generation Wireless infrastructure: Scenarios and Research Challenges", *IEEE Personal Communications*, 8(6), 25-31

Cai, Y. & Hua, K.A. (2002). An Adaptive Query Management Technique for Real-Time Monitoring of Spatial Regions in Mobile Database Systems. *Proceedings of the 21st IEEE International Conference on Performance, Computing, and Communications*, IEEE Computer Society (pp. 259-266)

Cao, G. (2003). A Scalable Low-Latency Cache Invalidation Strategy for Mobile", *IEEE Trans. Knowl. Data Eng.* 15(5), 1251-1265

Cao, G. (2002). On Improving the Performance of Cache Invalidation in Mobile Environments", *MONET* 7(4), 291-303

Cao, H., Wang, S., & Li, L. (2003). Location dependent query in a mobile environment", *Inf. Sci.* 154(1-2), 71-83

Chan, B.Y.L., Si, A., & Leong, H.V. (1998). Cache Management for Mobile Databases: Design and Evaluation, *Proceedings of the IEEE International Conference on Data Engineering (ICDE)* (pp: 54-63)

Chang, Y-I. & Yang, C-N. (2002). A complementary approach to data broadcasting in mobile information systems, *Data Knowl. Eng.* 40(2), 181-194

Chang, Y-I. & Chiu, S-Y. (2002). A Hybrid Approach to Query Sets Broadcasting Scheduling for Multiple Channels in Mobile Information Systems. *J. Inf. Sci. Eng.* 18(5), 641-666

Chen, M-S., Wu, K-L. & Yu, P.S. (2003). Optimizing Index Allocation for Sequential Data Broadcasting in Wireless Mobile Computing, *IEEE Trans. Knowl. Data Eng.* 15(1), 161-173

Chiasserini, C.F., Marsan, M.A., Baralis, E., & Garza, P. (2003). Towards feasible topology formation algorithms for Bluetooth-based WPAN's, *Proceedings of the 36th Annual Hawaii International Conference on System Sciences* (p. 313)

Datta, A., VanderMeer, D.E., Celik, A., & Kumar, V. (1999). Broadcast Protocols to Support Efficient Retrieval from Databases by Mobile Users, *ACM Trans. Database Syst.* 24(1), 1-79

DeRose, J.F. (2002). *The Wireless Data Handbook*, 4th edition, Wiley-Interscience, chapter 6

Dixit, S., Guo, Y., & Antoniou, Z. (2001). Resource management and quality of service in third generation wireless network, *IEEE Communication Magazine*, 39(2)

Dulaney, J. (2008). The Evolvement of 3G Mobile: Introduction of Third Generation Cell Phones, http://www.planetomni.com/ARTICLES-The-Evolvement-of-3G-Mobile.shtml

El-Ghazaly, S. & Golio, M. (1996). Challenges in modern wireless personal communications, *Radio Science Conference*, 29, 39-51

Elmagarmid, A.K., et al, (2003). Scalable Cache Invalidation Algorithms for Mobile Data Access, *IEEE Trans. Knowl. Data Eng.* 15(6), 1498-1511

Elmasri R. & Navathe S. B. (2003). *Fundamentals of Database Systems*, 4th edition, Addison Wesley

Englander, E. (2000), *The Architecture of Computer Hardware and Systems Software: An Information Technology Approach*, 2nd Edition, John Wiley & Sons, Inc., U.S.A

Ferro, E., & Potorti, F. (2005). Bluetooth and Wi-Fi wireless protocols: A survey and a comparison, *Wireless Communications*, 12(1), 12-26

Flahive, A., Rahayu, J.W., Taniar, D., & Apduhan, B.O. (2004). A Distributed Ontology Framework for the Grid, *Proceedings of the International Conference on Parallel and Distributed Computing Applications and Technologies* (PDCAT), Lecture Notes in Computer Science volume 3320, Springer, pp. 68-71.

Flahive, A., Rahayu, J.W., Taniar, D., & Apduhan, B.O. (2005). A Distributed Ontology Framework in the Semantic Grid Environment. *Proceedings of the 19th International Conference on Advanced Information Networking and Applications* (AINA), IEEE Computer Society, pp. 193-196.

Gaede, V. & Günther, O. (1998). Multidimensional Access Methods, *ACM Computing Surveys* 30(2), 170-231

Gast, M. (2005). *802.11 Wireless Networks: The Definitive Guide*, 2nd edition, OReilly & Associates, Inc.

Ghavami, M., Michael, L.B., & Kohno, R. (2005). Ultra wideband signals and systems in communications engineering, *Electronics Letters*, 41(25)

Ghosh, A., et al. (2005). Broadband wireless access with WiMax=8O2.16: Current performance benchmarks and future potential, *IEEE Commun. Mag*., 43(2), 129-136

Giuliano, R., & Mazzenga, F. (2005). On the coexistence of power-controlled ultrawideband systems with UMTS, GPS, DCS 1800, and fixed wireless systems, *IEEE Trans. Veh. Technol*., 54(1), 62-81

Goh, J. & Taniar, D. (2006a). MUDSOM: Mobile User Database Static Object Mining", *Proceedings of the 20th International Conference on Advanced Information Networking and Applications (AINA 2006)*, 1 (pp. 528-532)

Goh, J., & Taniar, D. (2006b). On Mining 2 Step Walking Pattern from Mobile Users, *Proceedings of the International Conference on Computational Science and Its Applications - ICCSA 2006*, Part I. *Lecture Notes in Computer Science* 3980, Springer (pp. 1090-1099)

Goh, J. & Taniar, D. (2005a). Mobile User Data Mining: Mining Relationship Patterns, *Proceedings of the International Conference on Embedded and Ubiquitous Computing - EUC 2005, Lecture Notes in Computer Science* 3824 Springer (pp. 735-744)

Goh, J. & Taniar, D. (2005b). Mining Patterns of Mobile Users Through Mobile Devices and the Music They Listens, *Proceedings of the International Conference on Computational Science and Its Applications - ICCSA 2005*, Part IV. *Lecture Notes in Computer Science* 3483 Springer (pp. 1203-1211)

Goh, J. & Taniar, D. (2004a). Mining Physical Parallel Pattern From Mobile Users", *Proceedings of the International Conference on Embedded and Ubiquitous Computing (EUC 2004), Lecture Notes in Computer Science* 3207 Springer (pp. 324-332)

Goh, J. & Taniar, D. (2004b). Mobile Data Mining by Location Dependencies, *Proceedings of the 5th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2004), Lecture Notes in Computer Science* 3177 Springer (pp. 225-231)

Goh, J. & Taniar, D. (2004c). An Efficient Mobile Data Mining Model, *Proceedings of the 2nd International Symposium on Parallel and Distributed Processing and Applications (ISPA 2004), Lecture Notes in Computer Science* 3358 Springer (pp. 54-58)

Goh, J. & Taniar, D. (2004d). Mining Frequency Pattern from Mobile Users", *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2004)*, Part III. *Lecture Notes in Computer Science* 3215 Springer (pp. 795-801)

Goh, J., Taniar, D. & Lim, E-P. (2006). SGPM: Static Group Pattern Mining Using Apriori-Like Sliding Window, *Proceedings of the 10th Pacific-Asia Conference in Knowledge Discovery and Data Mining (PAKDD 2006), Lecture Notes in Computer Science* 3918, Springer (pp. 415-424)

Gunderloy, G. (2002), *ADO and ADO.NET Programming*, SYBEX Inc., Alameda, California, U.S.A.

Hamalainen, M., et al. (2002). On the UWB system coexistence with GSM900, UMTS=WCDMA, and GPS, *IEEE J. Sel. Areas Commun.*, 20(9), 1712-1721

Helal, A., Haskell, B., Carter, J. L., Brice, R., Woelk, D. & Rusinkiewicz, M. (2002). *Introduction to Mobile Computing*, Volume 522, Springer

Hole, K.J., Dyrnes, E., & Thorsheim, P. (2005). Securing Wi-Fi networks, *Computer*, 38(7), 28-34

Hu, Q. & Lee, D.K. (1998). Cache algorithms based on adaptive invalidation reports for mobile environments", *Cluster Computing* 1(1), 39-50

Huang, J-L., Chen, M-S. & Peng, W-C. (2003). Broadcasting Dependent Data for Ordered Queries without Replication in a Multi-Channel Mobile Environment", *Proceedings of the IEEE International Conference on Data Engineering (ICDE)* (pp. 692-694)

Huang, J-L. & Chen, M-S. (2004). Dependent Data Broadcasting for Unordered Queries in a Multiple Channel Mobile Environment", *IEEE Trans. Knowl. Data Eng.* 16(9), 1143-1156

Hung, J-J. & Leu, Y. (2003). An Energy Efficient Data Reaccess Scheme for Data Broadcast in Mobile Computing Environments", *Proceedings of the International Conference on Parallel Processing ICPP* Workshops (pp. 5-12)

IEEE 802.11 (1999). Local and metropolitan area networks: Wireless LAN medium access control (MAC) and physical specifications, ISO/IEC 8802-11

Imielinski T., Viswanathan S. & Badrinath B. R. (1994). Energy Efficient Indexing on Air", *Proceedings of the ACM SIGMOD Conference* (pp.25-36)

Imielinski T. & Badrinath B. (1994). Mobile Wireless Computing: Challenges in Data Management", *Communications of the ACM*, 37(10), 18-28

Imielinski, T. & Badrinath, B. (1992). Querying in Highly Mobile Distributed Environments, *Proceedings of the 18th Very Large Data Bases Conference* (pp. 41- 52)

Jayaputera, J. & Taniar, D. (2005a). Data retrieval for location-dependent queries in a multi-cell wireless environment, *Mobile Information Systems* 1(2), 91-108

Jayaputera, J. & Taniar, D. (2005b). uery Processing Strategies For LocationDependent Information Services", *International Journal of Business Data Communications and Networking*, IGI-Global, 1(2), 17-40

Jayaputera, J. & Taniar, D. (2004a). Defining Scope of Query for Location-Dependent Information Services, *Proceedings of the International Conference on Embedded and Ubiquitous Computing (EUC 2004), Lecture Notes in Computer Science* 3207 Springer (pp. 366-376)

Jayaputera, J. & Taniar, D. (2004b). Location-Dependent Query Results Retrieval in a Multi-cell Wireless Environment, *Proceedings of the 2nd International Symposium on Parallel and Distributed Processing and Applications (ISPA 2004), Lecture Notes in Computer Science* 3358 Springer (pp. 49-53)

Kim, Y. K. & Prasad, R. (2006). *4G Roadmap and Emerging Communication Technologies*, Artech House Publishers

Kottkamp, H-E. & Zukunft, O. (1998). Location-aware query processing in mobile database systems, *Proceedings of the ACM Symposium on Applied Computing* (*ACM SAC*) (pp. 416-423)

Kpper, A. (2005). *Location-Based Services: Fundamentals and Operation*, John Wiley & Sons Ltd.

Lee, D.K., Xu, J., Zheng, B. & Lee, W-C. (2002). Data Management in Location-Dependent Information Services, *IEEE Pervasive Computing*, 2(3), 65-72

Lee, C-H. & Chen, M. (2001). Using Remote Joins for the Processing of Distributed Mobile Queries, *Proceedings of the International Conference on DASFAA* (pp. 226-233)

Lee, K.C.K., Leong, H.V., & Si, A. (2002). Semantic Data Broadcast for a Mobile Environment Based on Dynamic and Adaptive Chunking, *IEEE Trans. Computers* 51(10), 1253-1268

Lim, S.Y, Taniar, D. & Srinivasan, B. (2007a). Mobile Information Processing Involving Multiple Non-Collaborative Sources, *International Journal of Business Data Communications and Networking*, 3(2), 72-93

Lim, S.Y., Taniar, D., & Srinivasan, B. (2007b). Data Caching in a Mobile Database Environment, *Business Data Communications and Networking: A Research Perspective*, Chapter VIII (pp. 187-210)

Lim, S.Y., Taniar, D., & Srinivasan, B. (2006). A Taxonomy of Database Operations on Mobile Devices, *Mobile Multimedia: A Communication Engineering Perspective*, Chapter 10 (pp. 197-215)

Lim, S.Y., Taniar, D. & Srinivasan, B. (2005a). On-Mobile Query Processing Incorporating Multiple Non-Collaborative Servers, *Ingénierie des Systèmes d'Information* 10(5), 9-38

Lim, S.Y., Taniar, D., & Srinivasan, B. (2005b). Mobile Information Processing incorporating Location-Based Services, *Proceedings of the IEEE 3rd International Conference on Industrial Informatics, INDIN 2005*, Perth, Australia, IEEE Computer Society Press (pp. 1-6)

Lo, E., et al (2004). Processing Ad-Hoc Joins on Mobile Devices. *DEXA* 2004 (pp. 611-621)

Lodge, J. H. (1991). Mobile satellite communications systems - toward global personal communications, *IEEE Communications Magazine* 29, 24–30

Malladi R. & Davis, K.C. (2002). Applying Multiple Query Optimization in Mobile Databases, *Proceedings of the 36th Hawaii International Conference on System Sciences* (pp. 294-303)

Myers B.A. & Beigl M. (2003). Handheld Computing, *IEEE Computer Magazine*, 36(9), 27-29

Overview of Wireless Technologies (2008). http://wireless.utk.edu/overview.html. Last accessed: 02/04/08

Ozakar, B., Morvan, F., & Hameurlain, A. (2005). Mobile join operators for restricted sources, *Mobile Information Systems* 1(3), 167-184

Pardede, E., Rahayu, J.W., & Taniar, D. (2005). Preserving Conceptual Constraints During XML Updates. *International Journal of Web Information Systems* 1(2): 65-82.

Paulson L.D. (2003). Will Fuel Cells Replace Batteries in Mobile Devices?, *IEEE Computer Magazine*, 36(11), 10-12

Peng, W-C. & Chen, M.S. (2003). Efficient Channel Allocation Tree Generation for Data Broadcasting in a Mobile Computing Environment, *Wireless Networks* 9(2), 117-129

Pissinou, N., Makki, K., & Campbell, W.J. (1999). On the design of a location and query management strategy for mobile and wireless environments, *Computer Communications* 22(7), 651-666

Pitoura, E. & Samaras, G. (1998). *Data Management for Mobile Computing*, Kluwer Academic Publishers

Porkaew, K., Lazaridis, I. & Mehrotra, S. (2001). Querying Mobile Objects in Spatio-Temporal Databases, *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases* (SSTD '01), Springer-Verlag (pp. 59-78)

Priyantha, N. B., Chakraborty, A. & Balakrishnan, H. (2000). The cricket location-support system", *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking*, ACM (pp. 32-43)

Ren, Q. & Dunham, M. (2000). Using Semantic Caching to Manage Location Dependent Data in Mobile Computing, *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking* (pp. 210-221)

Robertson, M.G., Hansen, S.V., Sorenson, F.E., & Knutson, C.D. (2001). Modeling IrDA performance: The effect of IrLAP negotiation parameters on throughput, *Proceedings of the 10th International Conference on Computer Communications and Networks* (pp. 122-127)

Rusu, L.I., Rahayu, J.W., & Taniar, D. (2004). On Building XML Data Warehouses. *Proceedings of the 5th International Conference on Intelligent Data Engineering and Automated Learning* (IDEAL'2004), Lecture Notes in Computer Science, volume 3177, Springer, pp. 293-299

Rusu, L.I., Rahayu, J.W., & Taniar, D. (2005). Maintaining Versions of Dynamic XML Documents. *Proceedings of the 6th International Conference on Web Information Systems Engineering* (WISE'2005), Lecture Notes in Computer Science, volume 3806, Springer, pp. 536-543

Seydim, A., Dunham, M. & Kumar, V. (2001). Location Dependent Query Processing, *Proceedings of the 2nd ACM International Workshop on Data Engineering for Wireless and Mobile Access*, ACM (pp. 47-53)

Stallings, W. (2001). *Wireless communications and networks*, Prentice Hall

Stamper, D. (2001), *Local Area Networks*, 3rd Edition, Prentice Hall, Inc., New Jersey, U.S.A

Stan, M. C. & Skadron, K. (2003). Power-Aware Computing, *IEEE Computer Magazine*, 36(12), 35-38

Tan, K-L. (2001). Organization of Invalidation Reports for Energy-Efficient Cache Invalidation in Mobile Environments, *MONET* 6(3), 279-290

Taniar, D. & Rahayu, W. (editors) (2006). *Web Semantic and Ontology*, IGI Global.

Taniar, D. & Goh, J. (2007). On Mining Movement Pattern from Mobile Users", *International Journal of Distributed Sensor Networks*, 3(1), 69-86

The IEEE 802.11 Standards (2008). http://standards.ieee.org/getieee802/802.11.html. Last accessed: 02/04/08

Toh, C-K. & Li, V. (1998). Satellite ATM network architectures: An overview, *IEEE Network* 12(5), 61-71

Trajcevski, G., Wolfson, O., Hinrichs, K. & Chamberlain, S. (2004). Managing Uncertainty in Moving Objects Databases, *ACM Trans. Database Syst*. 29(3), 463-507

Trivedi K. S., Dharmaraja S. & Ma X. (2002). Analytic modelling of handoffs in wireless cellular networks", *Information Sciences*, 148, 155-166

Tsalgatidou, A., Veijalainen, J., Markkula, J., Katasonov, A. & Hadjiefthymiades, S. (2003). Mobile e-commerce and location-based services: Technology and requirements, *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Services* (pp. 1-4)

UMTS Forum Report No. 11 (2000b). *Enabling UMTS Third Generation Services and Applications*

UMTS Forum Report No. 9 (2000a). *The UMTS third generation market: Structuring the service revenue opportunities*

Vaughan-Nichols, S.J. (2003). The challenge of Wi-Fi roaming, *Computer*, 36(7), 17-19

Vaughan-Nichols, S.J. (2004). Achieving wireless broadband with Wi-Max, *IEEE Computer*, 37(6), 10-13

Vitsas, V. & Boucouvalas, A.C. (2002). IrDA IrLAP protocol performance and optimum link layer parameters for maximum throughput, *Global Telecommunications Conference*, GLOBECOM'02, Volume 3 (pp. 2270-2275)

Vitsas, V. & Boucouvalas, A.C. (2003). Optimization of IrDA IrLAP link access protocol, *IEEE Transactions on Wireless Communications*, 2(5), 926-938

Waluyo, A.B., Srinivasan, B., & Taniar, D. (2003). Optimal Broadcast Channel for Data Dissemination in Mobile Database Environment, *Proceedings of the 5th International Workshop on Advanced Paral-*

lel Programming Technologies (APPT 2003), Lecture Notes in Computer Science 2834 Springer (pp. 655-664)

Waluyo, A.B., Srinivasan, B. & Taniar, D. (2004a). A Taxonomy of Broadcast Indexing Schemes for Multi Channel Data Dissemination in Mobile Database, *Proceedings of the 18th International Conference on Advanced Information Networking and Applications (AINA 2004),* Volume 1, IEEE Computer Society (pp. 213-218)

Waluyo, A.B., Srinivasan, B. & Taniar, D. (2004b). Location Dependent Queries in Mobile Databases, *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)* (pp. 362-370)

Waluyo, A.B., Srinivasan, B., & Taniar, D. (2004c). Optimizing Query Access Time over Broadcast Channel in a Mobile Computing Environment, *Proceedings of the International Conference on Embedded and Ubiquitous Computing (EUC 2004), Lecture Notes in Computer Science* 3207 Springer (pp. 439-449)

Waluyo, A.B., Srinivasan, B. & Taniar, D. (2004d). Allocation of Data Items for Multi Channel Data Broadcasting in a Mobile Computing Environment, *Proceedings of the International Conference on Embedded and Ubiquitous Computing (EUC 2004), Lecture Notes in Computer Science* 3207 Springer, (pp. 409-418)

Waluyo, A.B., Srinivasan, B., & Taniar, D. (2005a). Data Dissemination in Mobile Databases, *Encyclopedia of Information Science and Technology*, Volume 2 (pp. 691-697)

Waluyo, A.B., Srinivasan, B. & Taniar, D. (2005b). Global Indexing Scheme for Location-Dependent Queries in Multi Channels Mobile Broadcast Environment, *Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA 2005),* Volume 1, IEEE Computer Society (pp. 1011-1016)

Waluyo, A.B., Srinivasan, B. & Taniar, D. (2005c). Research on location-dependent queries in mobile databases", *Comput. Syst. Sci. Eng*. 20(2)

Waluyo, A.B., Srinivasan, B. & Taniar, D. (2005d). Efficient Broadcast Indexing Scheme for Location-dependent Queries in Multi Channels Wireless Environment", *Journal of Interconnection Networks* 6(3), 303-322

Waluyo, A.B., Srinivasan, B. & Taniar, D. (2005e). Research in mobile database query optimization and processing", *Mobile Information Systems* 1(4), 225-252

Waluyo, A.B., Goh, G., Srinivasan, B., & Taniar, D. (2005). On-Building Data Broadcast System in a Wireless Environment, *International Journal of Business Data Communications and Networking*, IGI, 1(4), 15-37

Waluyo, A.B., Goh, G., Taniar, D., & Srinivasan, B. (2005). On Building a Data Broadcasting System for Mobile Databases, *Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA 2005),* Volume 1, IEEE Computer Society (pp. 538-543)

Waluyo, A.B., Zhu, F., Taniar, D., Srinivasan, B. & Rahayu, J.W. (2007). Multiple Entity Types Wireless Broadcast Database System, *Proceedings of the 21st International Conference on Advanced Information Networking and Applications (AINA 2007)* (pp. 179-186)

Waluyo, A.B., Srinivasan, B., Taniar, D., Rahayu, J.W. & Apduhan, B.O. (2006). Performance Analysis of Unified Data Broadcast Model for Multi-channel Wireless Databases, *Proceedings of the 3rd International Conference on Ubiquitous Intelligence and Computing (UIC 2006), Lecture Notes in Computer Science* 4159, Springer (pp. 698-707)

Waluyo, A.B., Srinivasan, B., Taniar, D., & Rahayu, J.W. (2005). Incorporating Global Index with Data Placement Scheme for Multi Channels Mobile Broadcast Environment, *Proceedings of the International Conference on Embedded and Ubiquitous Computing - EUC 2005, Lecture Notes in Computer Science* 3824 Springer (pp. 755-764)

Waluyo, A.B., Hsieh, R., Taniar, D., Rahayu, J.W. & Srinivasan, B. (2004). Utilising Push and Pull Mechanism in Wireless E-Health Environment, *Proceedings of the IEEE International Conference on e-Technology, e-Commerce, and e-Services (EEE 04)*, IEEE Computer Society (pp. 271-274)

Waluyo, A.B., Taniar, D. & Srinivasan, B. (2007). Mobile Information Systems in a Hospital Organization Setting. *Business Data Communication and Networking: A Research Perspective*, Chapter VII, IGI Publishing (pp. 151-186).

Want, R., Schilit, N., Adams, I., Gold, R., Petersen, K., Goldberg, D., Ellis, R. & Weiser, M. (1996). *The ParcTab Ubiquitous Computing Experiment*, Kluwer Academic Publishers, Boston

Ward, A., Jones, A. & Hopper, A. (1997). A new location technique for the active office, *IEEE Journal Personal Communications* 4(5), 42-47

Williams, S. (2000). IrDA: Past, present and future. *Personal Communications*, 7(1), 11-19

Wireless Indoor Positioning System (WIPS) (2007). Technical Documentation. http://www.tslab.ssvl.kth.se/csd/pro jects/0012/technical.pdf. Last accessed: 10/10/2007

Wolfson, O., Xu, B., Yin, H., & Cao, H. (2006). Searching Local Information in Mobile Databases, *Proceedings of the Conference on Data Engineering (ICDE)* (p. 136)

Xu, J., Zheng, B., Zhu, M. & Lee, D.L. (2002). Research Challenges in Information Access and Dissemination in a Mobile Environment, *Proceedings of the PanYellow-Sea International Workshop on Information Technologies for Network Era* (pp. 1-8)

Xu, J., Tang, X., & Lee, D.K. (2003). Performance Analysis of Location-Dependent Cache Invalidation Schemes for Mobile Environments, *IEEE Trans. Knowl. Data Eng.* 15(2), 474-488)

Zak, D. (1999), *Programming with Microsoft Visual Basic 6.0*, Course Technology (div. of International Thomson Publishing), Cambridge, Massachusetts, U.S.A.

Zaslavsky, A. & Tari, Z. (1998). Mobile computing: Overview and current status", *Australian Computer Journal* 30(2), 42-52

Zheng, B., Xu, J., & Lee, D.K. (2002). Cache Invalidation and Replacement Strategies for Location-Dependent Data in Mobile Environments, *IEEE Trans. Computers* 51(10), 1141-1153

About the Editor

David Taniar received a PhD degree in databases from Victoria University, Australia, in 1997. He is now a senior lecturer at Monash University, Australia. He has published more than 100 research articles and edited a number of books in the web technology series. He is in the editorial board of a number of international journals, including *Data Warehousing and Mining*, *Business Intelligence and Data Mining*, *Mobile Information Systems*, *Mobile Multimedia*, *Web Information Systems*, and *Web and Grid Services*. He has been elected as a fellow of the Institute for Management of Information Systems (UK).

Section I
Fundamental Concepts and Theories

This section serves as the foundation for this exhaustive reference tool by addressing crucial theories essential to the understanding of mobile computing. Chapters found within these pages provide an excellent framework in which to position mobile computing within the field of information science and technology. Individual contributions provide overviews of mobile learning, mobile portals, and mobile government, while also exploring critical stumbling blocks of this field. Within this introductory section, the reader can learn and choose from a compendium of expert research on the elemental theories underscoring the research and application of mobile computing.

Chapter 1.1
Ubiquitous Access to Information Through Portable, Mobile and Handheld Devices

Ch. Z. Patrikakis
National Technical University of Athens, Greece

P. Fafali
National Technical University of Athens, Greece

N. Minogiannis
National Technical University of Athens, Greece

N. Kourbelis
National Technical University of Athens, Greece

INTRODUCTION

Use of mobile devices for supporting our everyday communication has become part of our daily routine. Recent statistics illustrate that the penetration of mobile devices in everyday use has reached (and in some cases even surpassed) the penetration of fixed communication devices (ITU, 2004). As a consequence, use of mobile devices for accessing data information also increases, assisted by the rapid development of new technologies especially designed to support multimedia communication. Within the next years, third-generation (3G) wireless services will proliferate, offering multimedia capabilities such as streaming video (BERGINSIGHT, 2005; Raghu, Ramesh, & Whinston, 2002; UMTS forum, 2005). All of these, combined with the establishment of Internet and portal technology as the standard way for information exchange, entertainment, and communication, have created a new scenery that is characterized by access to data "anywhere," "anytime," and by "anyone" (or "any means"). Design issues concerning the particularities of access devices, communication technologies, and volume of information

exchanged are very important in the provision of mobile portal services (Microsoft, 2006).

In this article, we address the issue of providing portal services to users with portable devices such as personal digital assistants (PDAs) or smartphones. We propose a reference architecture for providing mobile portal services, based on the distribution of information between the portal servers and the user devices.

BACKGROUND

The need for mobile portal services lies in the penetration of mobile devices in the global market. However, the services offered today are not widely adopted by the mobile users. Surveys that have been carried out have revealed that cost, both in terms of devices (such as PDAs) and operation/subscriptions, constitutes a prohibitive factor. Furthermore, complexity has been mentioned as another reason for avoiding such services. Many people have also expressed their interest in more personalized content tailored to their profile, or in having the ability to create their favourites and set their preferences. In addition, users consider access speed as a key factor, meaning that they prefer minimum-step navigation, since they are not willing to spend much time and money to reach the information. Last, but not least, the applications that offer mobile services are not offered by the mobile operators or are not preinstalled in the devices, but are sold by third-party vendors. Consequently, many people are not aware of available mobile services.

Despite the aforementioned impediments to the explosion of Web services offered to mobile users, mobile-enabled information and market will define the near future scenery. Besides, this story bears similarity to how mobile phones pierced the whole world. The transition from generic Web portals to mobile portals should not be only associated with the adaptation of the content to the display size of the mobile devices. Mobile services should meet the varying needs of a "moving" user. A mobile user may need immediate access to crucial information, or may be in the process of waiting in a queue or for his flight to take off. Furthermore, mobile portals should focus on supporting concrete services for different target groups.

An attempt to organize mobile portal services into categories, according to global practice (GSA, 2002), leads us to the following categorisation:

- **Information Services:** General news, weather forecasts, financial, and sport news.
- **Food and Lifestyle:** Restaurants, bars, music halls, theater, cinema, events list.
- **Travel Services:** Flight/hotel listings, travel guides, maps, position location, and direction guidance.
- **Entertainment:** Online games, horoscopes, and quizzes.
- **Mobile Commerce (M-Commerce):** With real estate, Web banking, shopping, and auctions.
- **Messaging:** MMS, SMS, Chat, e-mail services.
- **Personal Information Management:** Calendars, contacts, photo albums.

The end-user experience is enhanced by the improved interfaces, use of graphics, touch pads, and technologies, such as VGA screens and cameras built into the devices (*Mobile Tech Review*, 2005). Many mobile portals have been launched combining information from the previously mentioned categories (GSA, 2002).

REQUIREMENTS

The basic idea behind the reference architecture proposed in this article is to overcome the limitations imposed by the handheld devices capabilities (display size, battery) and the cost of

network connectivity into a platform that provides ubiquitous access to a large portfolio of services. Initially, we define the requirements set for the system design.

User Friendly Interface for Users Unacquainted with Computers

Up to now, use of mobile and portable devices in our everyday life for communicating and entertaining ourselves has been a common practice. However, the concept of accessing information through PDAs instead of desktop PCs is quite new and, therefore, special care should be given to the design of applications services and the corresponding user interfaces.

As opposed to the case of voice communication and music entertainment, where the functionality of the device is limited to simple dialling or play-forward-rewind-stop, handling information presents several challenges. The user has to select the information that he needs to access, and then decide whether the result of his/her selection meets his/her demand. Furthermore, links between different types of information have to be specially designed in order to facilitate navigation. The small screens of mobile devices introduce an extra challenge: the "shrinking" of data so that the same level of information fits to much less than a quarter of minimum display of an average desktop computer.

Coherent Site Map to Minimize Navigation and Facilitate Users' Experience while Reducing Network Connectivity Costs

This is actually a requirement for any portal design. However, PDA terminals have special characteristics, that make minimization of navigation steps and connectivity costs very crucial. These characteristics are the low processing power and memory of portable devices, as well as the limitations in network connectivity that is provided over GPRS. Therefore, reaching information with minimum interaction is a key point for successful design of Web pages.

Up-to-Date Content

Ubiquitous access to information places an extra effort for portal designers. If we take into account the nature of information that is expected to be requested from a mobile device (news, weather updates, financial information), then it is obvious that the majority of user requests will be for dynamic content, constantly updated. Therefore, the designers and administrators of mobile portals should focus on data update and back-office mechanisms.

User Notification and Push Content Mechanisms

One major difference between "conventional" portals and mobile portals is the inability of these devices to maintain permanent connections to the portal. Therefore, for example, in a mobile portal that provides information about the stock market, updates on the price of stocks could be provided to desktop users through long last sessions (even for hours). This is not possible in mobile devices, not only due to the nature of the underlying communication infrastructure (GPRS-UMTS), but also due to the fact that deployment of other applications on the device (a phone call) may interrupt the session. Furthermore, the use of the mobile device is not the same as that of a desktop computer that is confined in a certain position on a desk.

For this, special mechanisms for notifications about data updates, and also push content mechanisms should be provided for information that is constantly changing, and this change has to be immediately reported to the user. The case of Blackberry devices (Research In Motion, 2006) and remote management capabilities in Windows mobile 5.0 (Microsoft, 2006) are excellent examples of such mechanisms.

PROPOSED ARCHITECTURE

On the ground of the requirements set, users should have fast response and online feedback on crucial information. An ideal way to achieve both demands is to take advantage of the memory space of the handheld device, and to discriminate content into static and dynamic. The notion is to have locally stored information that need not to be frequently updated, such as travel guides, maps, restaurants' and bars' addresses or description. This kind of data can be preinstalled in the device and can be renewed periodically through a synchronization process, depending on the type of information (i.e., tourist-related information may be updated yearly, while entertainment-related information should be updated more often). The dynamic content can be obtained through direct connection to the mobile portal.

Special provision should be given so that the information provided through the portal is in a form that can be used off-line. This is very crucial for cases where this information regards promotional offers, addresses in terms of phone and fax numbers, and location information. In this way, the user has access to a wide range of services without needing to be always connected to the portal. Especially in cases where use of the mobile device is expected to happen in areas with poor network coverage (i.e., mountain resorts where access to GPRS is not always available), the previous requirement becomes essential.

Another important issue is that of subscription to active information-sources (such as newsfeeds or stock-market) results in periodically updated reports that can be sent by SMS to end-users. Also, users belonging to a specific group (i.e., group of tourists) can be informed by announcements for special events organized. Photos taken during holidays can be uploaded in personal folders hosted under the portal, and can be used for sending e-cards or for creating a photo-album.

There are many issues regarding the frequency with which content should be updated. First of all, most of the online information is provided in the form of RSS-feeds (Loutchko & Birnkraut, 2005), which are information feeds offered by specific content providers. Therefore, there is no burden for the mobile portal administration to update information such as weather forecast, headline news, and so forth. Moreover, weather reports can provide "safe" forecast for a short future period (e.g., for 5 days) so that a user does not have to be connected to the mobile portal on a daily basis. In order to simplify the process of adding offers or dynamic information for the companies that are hosted and promoted by the portal, online tools can be provided for the renewal of the commercial information.

A proposed architecture for an end-to-end implementation of a platform that satisfies these requirements is depicted in Figure 1.

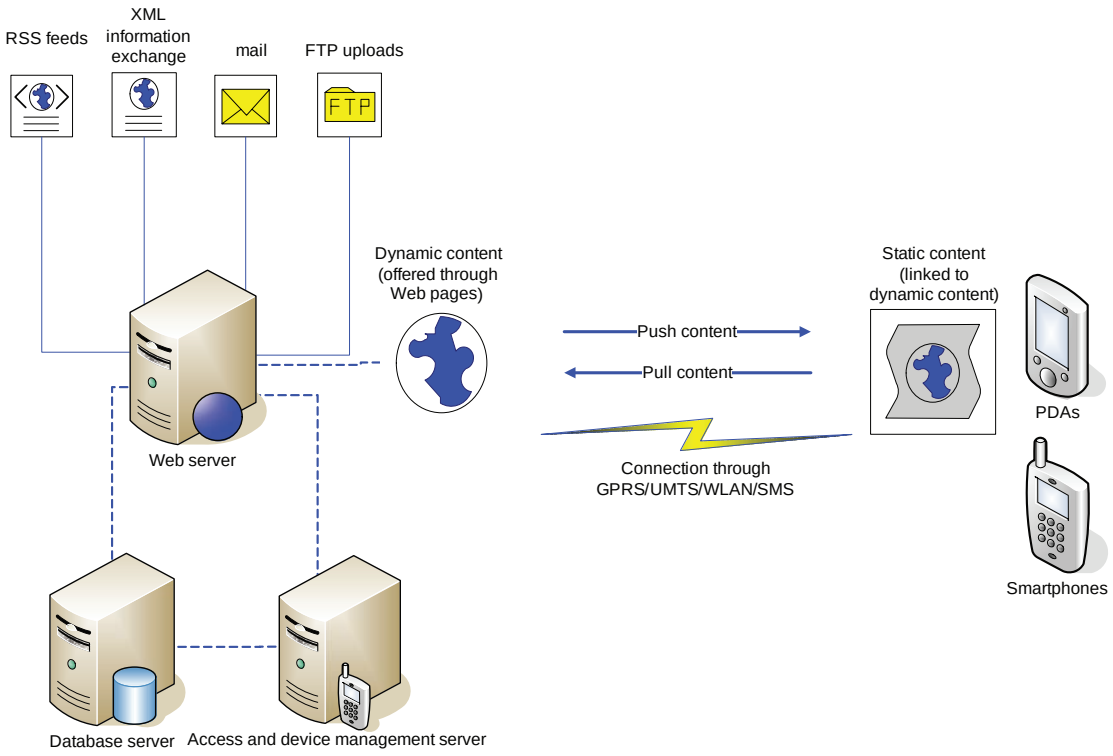
The platform consists of the following components.

Web Server

This constitutes the core component of the architecture. The server is linked to the database server for accessing portal information, while it incorporates interfaces to both end-user devices (PDAs—smartphones) and content providers. For interfacing the end-user equipment, both push and pull technologies are deployed. Thus, it supports access to information over GPRS, UMTS, WLAN, and SMS. Though "pull mode" for content access is easy to understand (as this is the standard way to access information through html), "push mode" is especially applicable in the case of mobile devices. This is offered mainly through the use of SMS for sending information, such as announcements, confirmations, and notifications, without the user having to request it.

Regarding the interface towards the content providers, this is used mainly for the upload of information to the mobile portal. This is achieved through various methods (RSS feeds, XML files, e-mail, and file upload). Information is pass-

Figure 1. Proposed platform implementation



ing from the Web server. As a variation of the architecture at this point, the Web server may be substituted by two components: a Web server that is used solely for hosting the Web pages and acting as the front end of the platform, and an application server that is used for providing the rest of functionality (i.e., access for the content provision mechanism). If we take into account the case of SMS, then a third component (SMS gateway) needs also to be inserted in the platform description. Figure 2 describes the detail breakdown of the Web server into three specialized components:

- Web server (front end);
- Application server (for back office access); and
- SMS gateway (offering SMS interface to the system).

Database Server

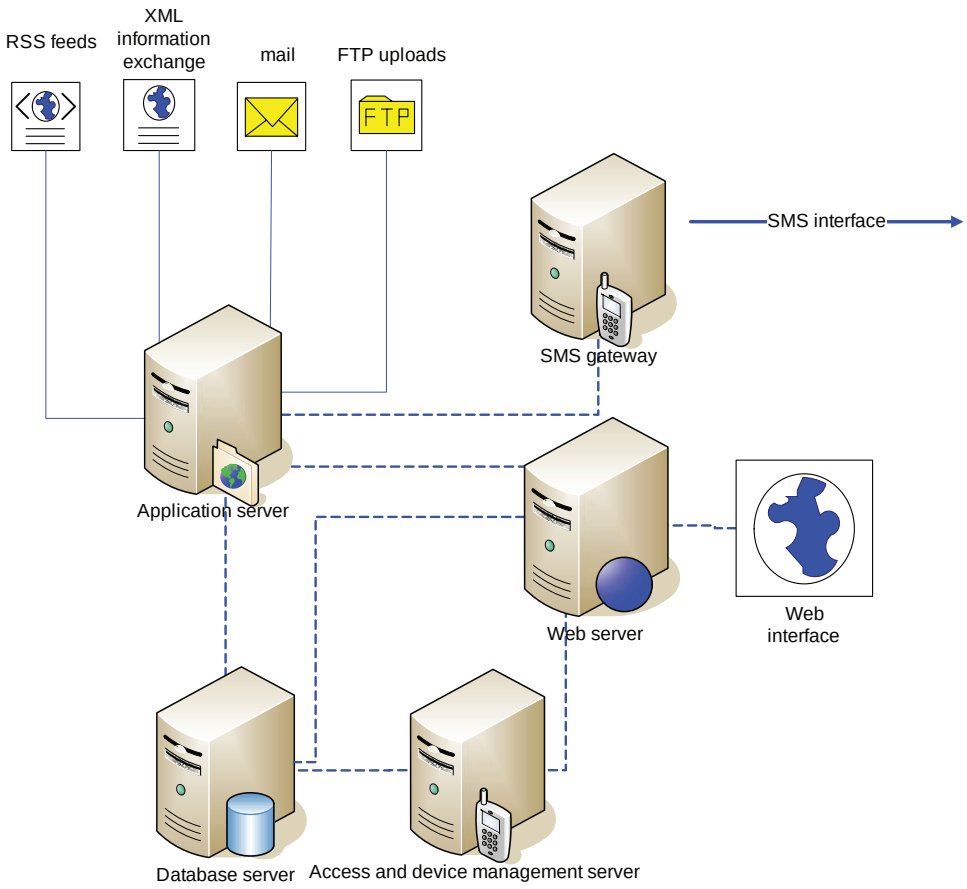
The database server is used to store:

- all the information that is accessed through the Web server; and
- data regarding the devices that have access rights to the information.

For this, apart from the communication towards the Web server, it also incorporates an interface towards the access and device management server, so that the later can control access to the available content and enforce subscription policies. Regarding the interface towards the Web server, this is provided for two reasons:

- for presenting the information to the end user through Web pages (statically or dynamically formed); and

Figure 2. Proposed platform implementation



- for providing access to the content provision mechanisms (through the aforementioned interfaces, deploying either only the Web server as the interfacing point or, alternatively, an application server).

Access and Device Management Server

This component may be optional, in the case where access to the mobile portal is provided without any restriction. However, since access to the information may be offered as a commercial service, this component is necessary to ensure that this access is granted only to registered users. Towards this end, both pull and push mechanisms for content access through the users' devices are being controlled by the access and device management server. Information regarding registered devices and/or users is provided from the database server. An important issue that the access and device management server is called to address is that of activation-deactivation of applications. As it has been mentioned, part of the information is stored to the mobile devices. In the case of a commercial service that is based on subscriptions, access to the information stored on the devices needs to be enabled and disabled, according to the payments status of the user. The access and device management server has to ensure that.

End-User Devices

These are the devices that are used for accessing the mobile portal, and are described as PDAs or smartphones with Web-browsing capabilities. Based on the hardware and firmware capabilities of the devices, we may distinguish two different ways to present Web services to the users:

1. The devices are running a fat client application that is responsible for presenting a comprehensive interface to the user. In this case, the handheld device or mobile phone runs an application (written in a programming language such as Java or C#) that is responsible for supporting the first level of access to information. By this application, the user has the ability to access information stored to his/her device directly, without the need of connecting to the Web portal. Such information, of course, is of the static type, while in the case where updates or access to information that is dynamic (i.e., weather forecasts) is needed, the application connects to the Web portal, accesses this information, and presents it to the user through a native application interface. The advantage of this approach is that the user can be offered comprehensive functionality, surpassing the capabilities of simple Web-based services, while push content mechanisms can be easily implemented, transparently, to the user. However, a drawback of this approach is that it requires the use of sophisticated devices with operating system capabilities, as those of Pocket PCs, while activation and deactivation of the application needs to incorporate a special mechanism (i.e., expiration of licences, SMS, or Web-enabled activation mechanism) while it is vulnerable to cracks and hacks.
2. The devices incorporate a thin client application, such as that of a Web browser. In this case, all the functionality is transferred to the Web server. Of course, in order to reduce the level of interaction, static information is again stored and accessed locally on the user's device, while all dynamic information is located again on the server. The drawback here is that the functionality that is offered to the user is reduced to that supported through the mobile portal Web pages, while in general, push content mechanisms cannot be deployed. On the other hand, management of the information is easier, while access control is simplified (it only requires access control to the Web server).

FURTHER ISSUES

This architecture presents a general approach to the issue of information access through portable, mobile and handheld devices. However, provision of an application or service needs to take into account the particularities of each case, which may introduce differentiations even at architectural level. This will be clear through the example of an application for tourists.

Such an application, apart from the standard functionality for access to information (both static such as hotels, restaurants, museums, and dynamic, such as festivals, theatres), requires extra functionality such as:

- **Translation of Content to Different Languages:** Though this is easy for the static content, in the case of dynamic content that is produced on a daily basis, automatic translation mechanisms need to be incorporated in the system.
- **Location-Based Service Offering:** Location awareness is crucial here. This can be provided through the use of GPS hardware or through location-based services from mobile operators. Correlation of the user's location to the content of the mobile portal is the key point for offering value-added services.

- **Support through a Call Center:** In the case where the service is provided through a "hot line" for assistance to the users, a special mechanism for giving access to the call center empowering this hot line is necessary. This means that the corresponding interfaces and mechanism for data access, customized to the needs of the operators of the call center, needs to be designed and inserted in the architecture.

CONCLUSION

As we see, there is no panacea for the provision of mobile portals. The diversity of user needs, together with the flexibility offered by the ubiquitous computing capabilities of smartphones and PDAs, make each case special. However, the core of requirements, as this is identified in the previous sections of this article, is the first issue that needs to be addressed when designing such services.

Fertile ground for the provision of services through mobile portal access is provided in the areas of:

- Mobile portals at the service of smart home concept (remote monitoring, remote appliance access);
- Digital content access;
- Secure and confidential communications and reliable transactions;
- Web-TV and digital video/audio broadcasting;
- Mobile gaming; and
- Billing.

As the capabilities of mobile devices are increasing in terms of processing power and memory, an increasing number of sophisticated services will appear. Advancements on communications and protocols, on the other hand, will enable the provision of rich audiovisual content that can be streamed to the devices. Up to now, the only burden seems to be the limited life of battery run time for the devices. Depending on the usage pattern of the individual user, the battery can be easily depleted, which constitutes a strong disadvantage when these devices have the role of mobile phones too. Once this final barrier is lifted, the road towards convergence of mobile and traditional portals will open, and the distinction between these two cases introduced by the deployed technology (both for device hardware and communication media) will be eliminated. However, the particularities originating from the user profiles (user mobility, anywhere-anytime access, security needs) will still remain, and be the cornerstone of mobile portal requirements.

REFERENCES

Berginsight. (2005). *Mobile content and entertainment in Western Europe* (pp. 2005-2012).

GSA. (2002). *GSA Quarterly Survey of Mobile Portal Services, 8.*

ITU. (2004). ITU strategy and policy unit news update. *Trends in Mobile Communications, 8.*

Loutchko I., & Birnkraut F. (2005). Mobile knowledge portals: Description schema and development trends. In *Proceedings of I-KNOW '05.*

Mandato D., Kovacs E., Hohl F., & Amir-ALIKHANI H. (2002). Camp: A context aware mobile portal. *IEEE Communications, 40*(1), 90-97.

Microsoft. *Windows mobile Web site*. Retrieved January 22, 2006, from http://www.microsoft.com/windowsmobile/

Mobile Tech Review. *Pocket PC reviews and information: What is a pocket PC (PPC)? What models are out there*? Retrieved January 25, 2005, from http://www.mobiletechreview.com/ppc.htm

Raghu, T. S., Ramesh, R., & Whinston, A. B. (2002). Next steps for mobile entertainment portals. *IEEE Computer, 35*(5), 63-70.

Research in Motion. *Blackberry devices Web site*. Retrieved January 22, 2006, from http://www.discoverblackberry.com/devices/

UMTS forum. (2005) *UMTS towards mobile broadband and personal Internet*, white paper.

KEY TERMS

General Packet Radio Service (GPRS): A technology between the second and third generations of mobile telephony, used to support moderate speed data transfer based on the deployment of unused TDMA channels in the GSM network.

Global Positioning System (GPS): A satellite navigation system that uses broadcasting of precise timing radio signals by satellites for offering accurate positioning of user devices globally.

Multimedia Messaging Service (MMS): A technology used for the exchange of multimedia messages (including images, audio, and video clips) between mobile phones.

Personal Digital Assistant (PDA): A handheld device that offers applications including, address book, task manager, calendar, calculator, and so forth. They may also include mobile phone functionality, word processing, and spreadsheet application capabilities, while newer versions may also support GPS and Wireless LAN access connectivity.

Really Simple Syndication (RSS): A family of Web-feed formats, specified in XML, used in news and Web logs.

Smartphone: A handheld device that combines the functionality of a mobile phone and a PDA. However, the main purpose of the device is to support mobile phone functionality.

Short Message Service (SMS): A service for the exchange of short text-based messages between mobile phones (extended to landline telephones).

Universal Mobile Telecommunications System (UMTS): A third-generation mobile phone technology that is based on the W-CDMA standard and is used in Europe and Japan.

eXtensible Markup Language (XML): A W3C-recommended general-purpose markup language for supporting data sharing across different systems over the Internet.

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 1033-1039,

Chapter 1.2
Mobile Computing and Commerce Framework

Stephanie Teufel
University of Fribourg, Switzerland

Patrick S. Merten
University of Fribourg, Switzerland

Martin Steinert
University of Fribourg, Switzerland

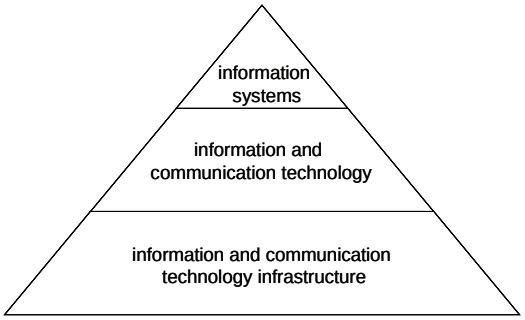
INTRODUCTION

This encyclopedia on mobile computing and commerce spans the entire nexus from mobile technology over commerce to applications and end devices. Due to the complexity of the topic, this chapter provides a structured approach to understand the interrelationship in-between the mobile computing and commerce environment. A framework will be introduced; the approach is based on the Fribourg ICT Management Framework, elaborated at our institute with input from academics and practitioners, which has been tried and tested in papers, books, and lectures on ICT management methods. For published examples, please consult Teufel (2001, 2004), Steinert and Teufel (2002, 2004), or Teufel, Götte, and Steinert (2004).

THE MOBILE CONVERGENCE CHALLENGE

The information revolution has drastically reshaped global society and is pushing the world ever more towards the information-based economy. In this, information has become a commodity good for companies and customers. From an economical perspective, the demand for information at the right time and place, for the right person, and with minimal costs has risen. The transformation towards this information-driven society and economy is based on the developments of modern *information and communication technology (ICT)*. Different industries are able to generate enormous synergy effects from the use of ICT and the *information systems (IS)* building on these technologies, especially the Internet. It

Figure 1. Information and communication technology, infrastructure, and systems



is a possible instrument to chance the structure and processes of entire markets.

As shown in Figure 1, information and communication technology can be differentiated in its infrastructure, the technologies themselves, and the information systems running on these technologies. In general, the infrastructure consists of all hardware- and software-related aspects as well as human resources. Consequently, the technologies themselves enable the collection, storage, administration, and communication of all data. These data can be used to synthesize information in respective systems, supporting the decision process and enabling computer-supported cooperative work.

The term information and communication technology (ICT) appeared in recent years. Due to the harmonization of *information technology (IT)* and the digitalization of the *telecommunications (CT)* infrastructure and the liberalization of the latter business sector, the ICT market established itself (see Figure 3). Consequently, the development and convergence of ICT became increasingly complex. Figure 2 illustrates the associated technology convergence.

Nowadays, a new aspect has entered the arena: mobility. Mobility is perhaps the most important trend on the ICT market. The fundamental

Figure 2. Technology convergence (Teufel, 2004, p. 17)

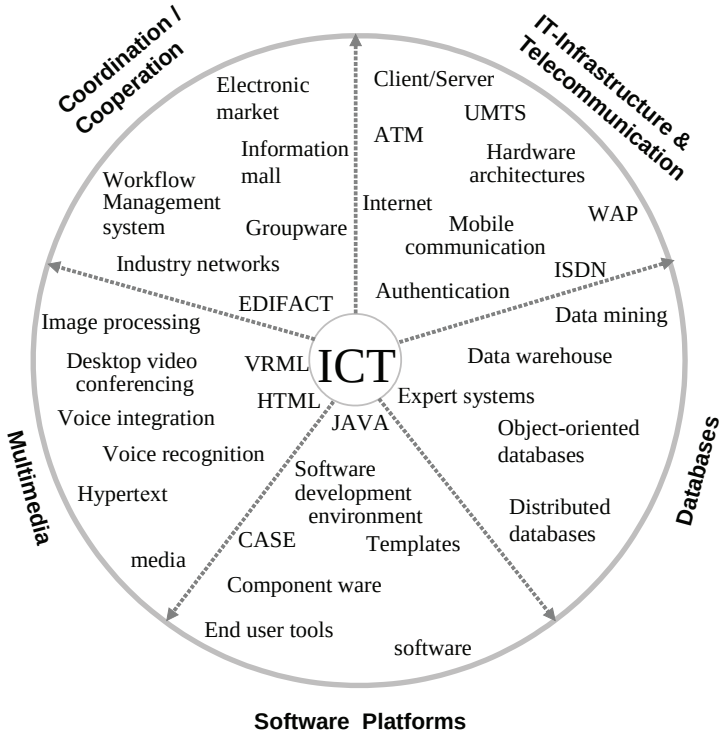
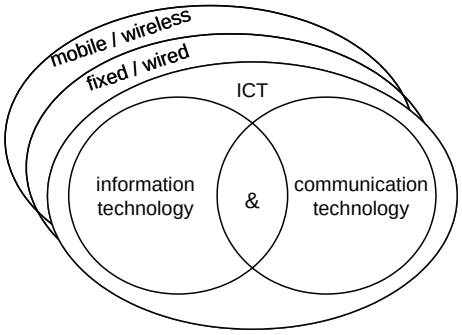


Figure 3. Mobile and fixed-line ICT convergence

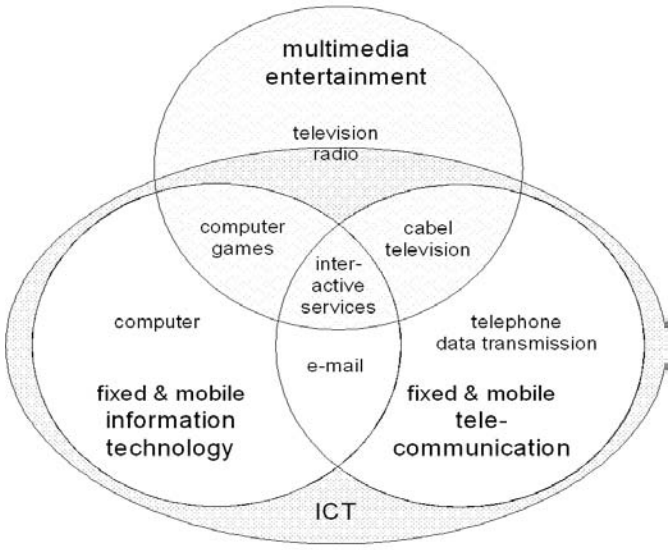


characteristic of mobile technologies is the use of the radio frequency band for (data) communication, which is often referred to as "wireless." The "wireless trend" has influenced not only the telecommunications and IT sector, but also most traditional markets, in the same way wired ICT did before. In addition, a convergence of wired and wireless, respectively fixed and mobile ICT can be observed.

As shown in Figure 3, the convergence of information technology and communication technology to ICT can be seen as the first phase of convergence. This was caused by the digitalization and liberalization in the telecommunications sector. The next phase of convergence was the success of mobile ICT, initializing a competition between wireless and fixed ICT. Meanwhile, information and communication as well as mobile and wired technologies have not only co-existed; they have merged, generating enormous synergy effects for both business and customer. In addition, another not just technological convergence can be observed. The entertainment and multimedia branch has entered the ICT market and vice versa, as illustrated in Figure 4.

The trend shown in Figure 4 becomes obvious when looking at the boom in interactive games or home cinema computerized equipment—again accelerated by the digitalization in a sector, this time the television (DVB) and radio (DAB). Again, the Asian market is leading edge. In South Korea, they are already running a fully functional system, based on the digital mobile broadcasting standard (DMB), bringing video broadcasting directly to the mobile end-device via satellite (tu4u, 2006). Finally, the three dimensions, fixed and mobile ICT convergence plus entertainment/multimedia,

Figure 4. ICT and multimedia entertainment convergence (Teufel, 2004, p. 14)



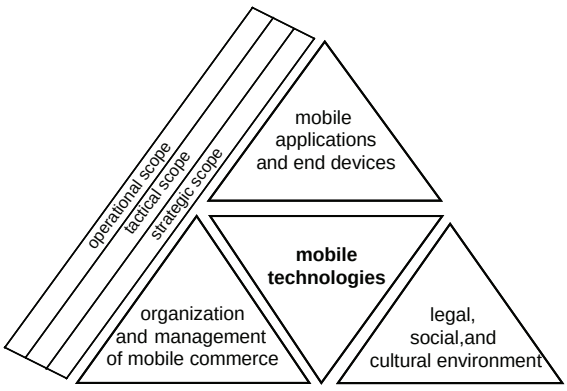
form the core of this encyclopedia's topic: the challenges of mobile computing and commerce.

THE MOBILE COMPUTING AND COMMERCE FRAMEWORK

Mobile computing and commerce comprises all business processes between administration, business, and customer via public or private wireless communication networks and with value creation. To understand the actual trends, recognizing the possibilities and threats and coping with the challenges of mobile computing and commerce are complex tasks. It becomes obvious that mobile computing and commerce consists of multiple dimensions, which are, in addition, interrelated. In order to structure the discussion, a framework for mobile computing and commerce is introduced. Using the classical scientific engineering approach, the framework allows a detailed analysis of single aspects and a reintegration of the diverse solutions in the synthesis. Furthermore, it covers the main issues, controversies, and problems from a market and business perception. Figure 5 features this ,mobile computing and commerce framework.

The four different dimensions of the framework as demonstrated in Figure 5 in addition show a common underlying scope. The strategic scope covers issues of long-term influence (more than five years of impact), as the tactical scope deals with all aspects in a timeframe of one to five years. Finally, all short-term topics are subject of the operational scope and handled within a year's period. The individual four main parts of the framework are examined in the following sections.

Figure 5. Mobile computing and commerce framework



Mobile Technologies

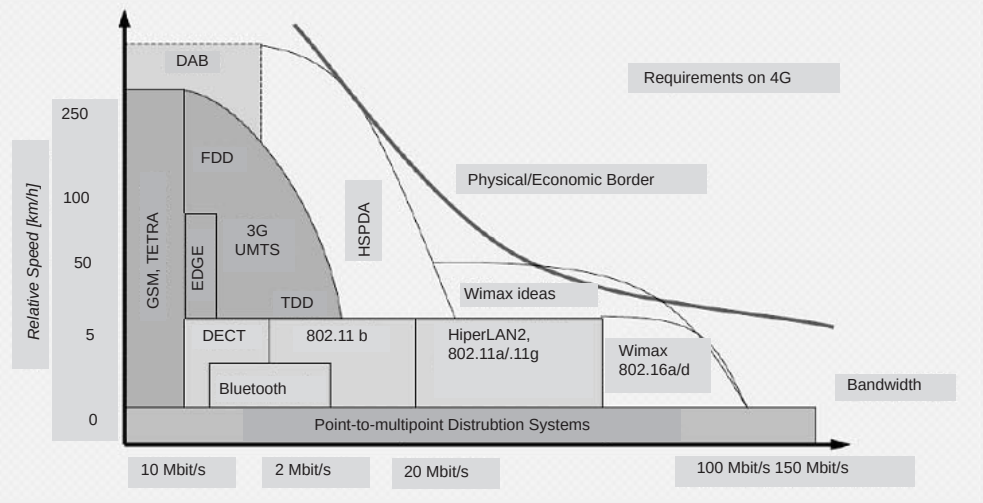
The origin and foundation of every case of mobile computing and commerce are mobile technologies. They are the centerpiece of the framework and comprise the different technological aspects. They are building the foundation for discussing all other aspects of the framework. Mobile technologies have evolved rapidly in the last decade, not only gaining market penetration, but in terms of bandwidth and relative speed. Figure 6 presents today's available wireless access technologies—also introducing a physical and economic border.

As such, this dimension includes aspects which are dealt with in the categories mobile information systems, mobile service technologies, and enabling technologies of this encyclopedia.

Legal, Social, and Cultural Environment

In a mobile environment, corporate social responsibility (CSR) is a fairly new field of increasing attention. It deals with the consequences of globalization, economic and ecological disaster, as well as financial affairs and others. Referring to the Global Compact Program 2000 from the United Nations and the Green Paper on CSR from the European Union, principles and guidelines are available today. These have led to programs that enable companies to continuously analyze and handle the versatile influences and effects on society and vice versa (Teufel et al., 2004).

Figure 6. Wireless access technologies (adapted from Schiller, 2003, p. 450)



Furthermore, the existence, use, and diffusion of mobile technologies are also strongly influenced by environmental aspects, especially from legal, social, and culture sub-environments. Examples are data protection issues, surveillance discussions, and radiation concerns, respectively. Other issues may also include important aspects such as standardization and regulation.

Furthermore, mobile technologies also change the way of living—introducing new concepts like mobile working. Especially the new work-life-(un)balance is subject to heated debates. Mobile technologies, applications, and end devices not only represent new opportunities in a business environment, but also create an interconnected and virtual world. In this, the digital divide more and more becomes a critical threat. Therefore topics of the categories like "mobile enterprise implications for society, business, and security" are to be considered.

Organization and Management of Mobile Commerce

To cope with the business challenges of mobile computing and commerce, all company internal aspects of the organization and the management of mobile commerce form a particular topic space. First of all, the classical roles of the CIO and the CTO have to be re-evaluated, taking mobile ICT into account. Furthermore, mobility also affects a whole set of management issues, which have already been previously influenced by fixed ICT. For example, the information management has to consider the aspect of mobile working when planning information system architectures. This in turn results in an adoption of current business processes and workflow implementations. Especially the procurement and distribution processes go through a fundamental change. In addition, mobile ICT offers new possibilities in the customer relationship management.

This dimension for example includes aspects described in "Mobile Commerce and E-Business."

Mobile Applications and End Devices

The focus point of every examination of mobile computing and commerce is the actual applications and end devices it is running onto. Again

the Asian market can be consulted, to give an example of cutting-edge end device research. NTT DoCoMo is working on a future mobile phone device that uses human fingers as receiver. For this, a wristwatch-like bone conduction terminal is used in contact with the human arm (NTT DoCoMo, 2006). Above all, the Asian market is leading the way towards an all IP-based mobile network environment. Thus this last but probably most important framework dimension features topics such as: mobile to "consumer applications", "mobile applications for the extended enterprise", and "enabling applications."

CONCLUSION

Throughout the previous sections, it has been shown that coping with the challenges of mobile computing and commerce is a complex problem. Therefore, and to structure this encyclopedia, the framework has been introduced. It aims to provide managers, engineers, and practitioners with a profound approach to handle fixed and mobile information and communication technology. In such a mobile computing and commerce environment, the different market players themselves can be subsequently differentiated as shown in Figure 7, following the EITO on their special on "Entering the UMTS era—mobile applications for pocket devices and services."

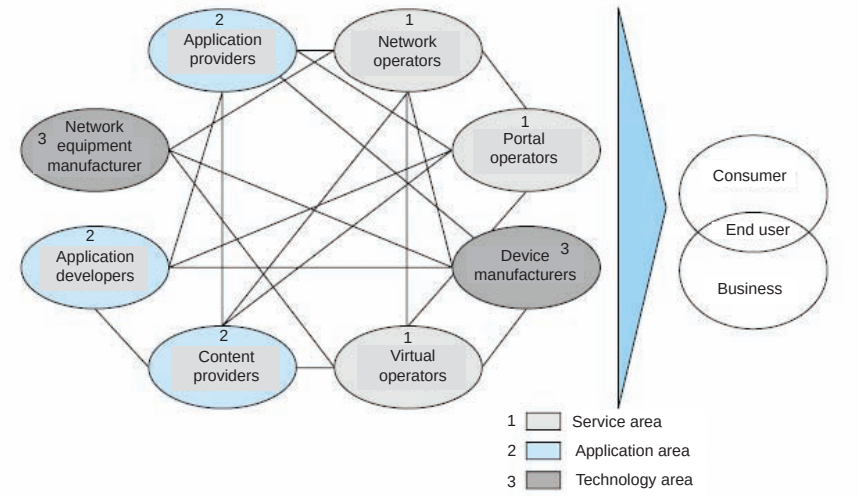
Figure 7 illustrates the mobile data value net. In this interconnected environment, the nine different market players experience an enforced competition, due to the fact that every action influences the entire business network. As a result, a duality of competitive and cooperative business strategies established itself (Steinert & Bult, 2004, p. 31) to generate network effects, introducing the phenomenon of co-opetition (Brandenburger & Nalebuff, 1996). On a more general level, this again shows the complexity of a mobile computing and commerce environment.

REFERENCES

Brandenburger, A., & Nalebuff, B. (1996). *Co-opetition* (1st ed.). New York: Doubleday.

EITO (European Information and Technology Observatory). (2002). *Eito report 2002.*

Figure 7. Mobile data value net (EITO, 2002, p. 205)



NTT DoCoMo. (2006). *R&D.* Retrieved January 12, 2006, from http://www.nttdocomo.com/corebiz/rd/index.html

Schiller, J. (2003). *Mobile communication* (2nd ed.). London: Addison-Wesley.

Steinert, M., & Bult, A. (2004). Strategische unternehmensführung von hightech-unternehmen—insights von swisscom-fixnet. In S. Teufel, S. Götte, & M. Steinert (Eds.), *Managementmethoden für ICT-unternehmen* (p. 12).

Steinert, M., & Teufel, S. (2002). The Asian lesson for mobile provider—An all-out strategic paradigm shift. *Proceedings of ITU Telecom Asia 2002* (pp. 25-44), Hong Kong.

Steinert, M., & Teufel, S. (2004, September 17-19). Beyond e-business—why e-commerce and Web organizations should monitor the mobile dimension. *Proceedings of the 2nd International Conference on Knowledge Economy and Development of Science and Technology (KEST2004)* (pp. 446-454), Beijing, China.

Teufel, S. (2001, August 6-12). ICT-management framework. *Proceedings of the International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet (SSGRR 2001)* (pp. 9-24), L'Aquila, Italy.

Teufel, S. (2004). Managementmethoden für ICT-unternehmen—dargestellt mittels dem Fribourg ICT management framework. In S. Teufel, S. Götte, & M. Steinert (Eds.), *Managementmethoden für ICT-unternehmen*. Zurich: Verlag Industrielle Organisation/Orell Füssli.

Teufel, S., Götte, S., & Steinert, M. (Eds.). (2004). *Managementmethoden für ICT-unternehmen: Aktuelles wissen von forschenden des iimt der Université Fribourg und spezialisten aus der praxis.* Zürich: Verlag. Industrielle Organisation.

tu4u. (2006). *TU media corporation.* Retrieved January 12, 2006, from http://www.tu4u.com/

KEY TERMS

Co-Opetition: Following Brandenburger and Nalebuff (1996), co-opetition is the economic situation between a company and a competing company that provides complementary products and services. Following game theory, a differentiated approach strategic than the generic competitive strategies are necessary (see also *ValueNet*).

Fribourg ICT Management Framework: The framework has been elaborated at the International Institute of Management in Technology (IIMT) of the University of Fribourg (Switzerland) with input of academics and practitioners. It provides an integrated approach to cope with the business challenges of the information-based economy.

Information and Communication Technology (ICT): The result of developments in the fields information technology (IT) and communication technology (CT), and their convergence caused by the digitalization and liberalization in the telecommunication sector.

Legal, Social, and Cultural Environment: This framework dimension covers all aspects and implications of Mobile ICT for Society and Business.

Mobile Application and End Device: Mobile applications running on mobile end devices are the topic of this framework dimension.

Mobile Computing and Commerce Framework: The framework is based upon the Fribourg ICT Management Framework and presents an integrated view on the different fields to be considered, while examining the issues and controversies of mobile computing and commerce.

Mobile ICT Convergence: As ICT can be seen as the first phase of convergence, mobile ICT convergences introduce wireless technologies next to wired ICT.

Mobile Technology: Wireless mobile access technology and the centerpiece of the framework.

Network Effect: Following Katz and Shapiro (1985), each new network participant directly increases the benefit of all other actors in a network, for example, by offering a new communication possibility (primary or direct network effect); an increased size of a network also indirectly increased the value of the entire network indirectly, for example by pushing an industry standard (secondary or indirect network effect).

Organization and Management of Mobile Commerce: All company internal aspects of the organization and the management issues, which are influenced by mobile ICT, also including aspects such as mobile business.

ValueNet or ValueWeb: Instead of a linear value chain, the company, its suppliers, and customers, and also its complementors and competitors, form a ValueNet or ValueWeb. Co-opetition, reciprocal actions, and network effects must be taken into account in the economics of such a value net.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 466-471, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.3
Mobile Electronic Commerce

Samuel Pierre
École Polytechnique de Montréal, Canada

INTRODUCTION

Mobile electronic commerce (or m-commerce) is generally defined as the set of financial transactions that can be carried out over a wireless mobile network (Pierre, 2003; Varshney, 2001; Varshney, Vetter, & Kalakota, 2000). According to this definition, m-commerce constitutes a subset of all electronic commercial transactions (electronic commerce or e-commerce) from business-to-consumer (B2C) or business-to-business (B2B). Thus, short personal messages such as those from short messaging system (SMS) sent between two individuals do not fall within the category of m-commerce, whereas messages from a service provider to a salesperson or a consumer, or vice versa, do fit this very definition. M-commerce appears an emerging manifestation of Internet electronic commerce which meshes together concepts such as the Internet, mobile computing, and wireless telecommunications in order to provide an array of sophisticated services (m-services) to mobile users (Paurobally, Turner, & Jennings, 2003).

Before purchasing a product, clients need services such as those used to search for a product and a merchant who offer the lowest price for this product. Consumers also like to participate in auctions and analyze the quality/price ratio of a product for a certain number of suppliers (Jukic, Sharma, Jukic, & Parameswaran, 2002). Online shopping for a given product is becoming increasingly popular, and electronic purchasing and bargaining consist of looking up and deciphering the contents of electronic catalogues prior to making a decision. To automate this process and to ensure that these documents are comprehensible to computers, they must have a standard format. Such services exist in standard commerce; however, in e-commerce, they require further consideration such as those related to the market dynamics, the variety of platforms, and the languages used by various merchant sites (Itani, & Kayssi, 2003; Lenou, Glitho, & Pierre, 2003).

Just as in standard commerce, e-commerce includes an initial step wherein consumers search for products they wish to purchase by virtually visiting several merchants. Once the product is found, negotiation for this possible transaction can take place between the customer and the merchant. If an agreement is reached, the next step is the payment phase. At each step of the process, a number of problems arise, such as transaction security, confidence in the payment protocol, bandwidth limitations, quality of service, shipping delays, and so forth (Paurobally et al., 2003). The peak withdrawal periods have always presented a major challenge for certain types of distributed applications. The advent of m-commerce further highlights this problem. Indeed, in spite of rather optimistic predictions, m-commerce iss plagued by several handicaps which hinder its commercial development.

This article exposes some basic concepts, technology and applications related to mobile electronic commerce. The background and key technological requirements needed to deploy m-commerce services and applications are discussed, some prominent applications of m-commerce are summarized, future and emerging trends in m-commerce are outlined, and a conclusion of these topics are presented.

BACKGROUND AND RELATED WORK

E-commerce relies upon users' interventions to initiate a transaction and select the main steps of the process. Users' actions are based upon a succession of virtual decisions. Indeed, when shopping with a virtual catalogue, customers can select products that meet their needs, tastes, and respect their price range. Such decisions consistently require the users' input, thus costing them both time and money. These costs are even more exorbitant when a search is launched for an order that includes a variety of products from different providers that have different characteristics (price range, delivery dates, etc.).

Mobile commerce refers to an ability to carry out wireless commercial transactions using mobile applications in mobile devices. M-commerce applications can be as simple as an address-book synchronization or as complex as credit card transactions.

In standard commerce, negotiating a contract or a commercial transaction is a standard practice in purchasing or sales. An agreement between a customer and a merchant can involve various components (price, delivery, warranty, etc.). For example, a volume price can be negotiated (e.g., 20% off the purchase of 100 items or more), price can fluctuate according to the demand (flight and hotel room prices vary according to seasons), and so forth. Once the client has obtained the best offer possible for the product of interest, the negotiation comes to a close. Obviously, the result of such negotiation can vary from one merchant to another. By providing a machine with the appropriate strategies and algorithms, negotiation can be automated and taken over by a computer, hence the concept of electronic negotiation, or e-negotiation.

Significant growth of m-commerce cannot be expected until the required technology (such as SMS services, Bluetooth, WAP, or i-mode) is developed and deployed. Indeed, due to the widely available GSM wireless networks, the SMS service allows GSM users to send short messages of up to 160 characters. These messages are saved and sent within a few seconds, which makes them unsuitable for real-time applications. SMS can become increasingly more important with future improvements once they allow users to send longer messages, multiple messages at once and when they allow users to create mailing lists. Such features will make m-commerce much more accessible.

Bluetooth is a low-powered wireless standard that allows a certain level of communication between many devices. Currently, it is a global

specification for close proximity wireless connections. Given the wide flexibility associated with the variety of terminals it supports, it is expected to play a significant role in m-commerce. It can be deployed on a large scale for short-range m-commerce where terminal proximity is minimal. However, its nonlicensed 2.4 GHz frequency is problematic as it can be encumbered by interference from other devices which use the same frequency.

The design of such applications requires a number of functional components. One of the major components is a mobile terminal that is equipped with sufficient power for its memory, display and communication functionalities. Many of these terminals are currently emerging, such as the Palm Pilot (a PDA with a wireless modem) or the Nokia Communicator (a mobile phone with computer functionalities). These devices offer various capacities involving communication, processor, battery, memory and display. Many of them are actually mobile phones enhanced with laptop features.

Given the enhanced functionalities of the mobile terminal and its improved processing and storage capacities, an operating system to manage the internal resources of the various applications and processes will become an essential requirement. However, operating systems require large storage capacities and they are not adapted to mobile terminals constrained by real-time requirements, limited processing capabilities, mini screen and small memory sizes. Mobile middleware can be defined as a functional layer of software provided by application developers to link their e-commerce applications to an OS and various mobile networks to allow their applications to bypass certain mobility issues.

With the emergence of mobile application environments in the recent years, Europe has focused on WAP technologies, whereas Japan has successfully developed with the i-mode. North American countries use other systems, which can include either of the previous two technologies. Indeed, in order to adapt Web contents to mobile users, Europeans use the Wireless Application Protocol (WAP). The WAP was designed to ensure interoperability amongst various wireless networks, mobile terminals, and applications which use the same type of protocols. It thus allows developers to design e-commerce applications from existing technology, which can function on a large number of mobile terminals.

The i-mode, a proprietary system developed by NTT DoCoMo, has been available in Japan since February 1999. It is a device that allows users to access the Internet from a cellular phone with a color display. It uses the packet switching technique with a bandwidth of 9.6 kbps (CDMA). The i-mode pages must be defined by a tag language called compact HTML (cHTML), which is, actually, a subset of HTML with additional adapted tags. Moreover, instead of paying for the amount of connection time, users pay for the quantity of data transmitted (0.3 penny/packet of 128 bytes).

Java i-mode phones have been available on the Japanese market since the beginning of 2001. These telephones allow users to download Java server applets (called i-appli) for games, agent-type services and other applications. There were nearly 19 million subscribers to the i-mode systems at the beginning of February 2001, and the number is increasing by 1 million every month. This system, which supports 11,000 Web sites and 30 search engines, is completely adapted to m-commerce.

One of the key aspects of m-commerce remains transaction security (Cai et al., 2004; He, & Zhang, 2003; Katsaros, & Honary, 2003; Kim & Chung, 2003). A new protocol for m-commerce was proposed by (Katsaros, & Honary, 2003). Fully applicable to third generation mobile networks, this protocol is characterized by three novel properties, as opposed to the existing methods of m-commerce. In fact, it provides a simplified and secure transaction method, minimizes the number of entities involved in the transaction,

and reduces the source of security threats, thus reducing the risk of fraud.

MOBILE COMMERCE APPLICATIONS

There are a great number of m-commerce applications (see Table 1). According to reliable estimates, in the next few years, over half of European m-commerce will consist of financial services, advertising, and purchasing. Various classes of applications along with their requirements in terms of services, platforms, and networks, are presented here, and four of those classes will be addressed in more detail.

Mobile Financial Applications

Mobile financial applications are likely to become a fertile niche for m-commerce. They include a wide variety of applications, from the banking environment, brokerage firms, mobile money transfers and mobile micro-payments. These mobile financial services can transform a mobile terminal into a business tool that replaces the bank, the ATM and credit cards and allows users to carry out financial transactions with mobile currency. However, to develop these applications, it is necessary to provide the users of these services with better applications and better network infrastructure. Moreover, security issues must be addressed prior to deploying such applications on a large scale.

An interesting mobile financial application is the micropayment, which consists of little purchases involving small transactions. A mobile terminal user can communicate with a sales machine via a wireless local area network (WLAN) to purchase these products.

The micropayment system can be implemented in several different ways (Kim, Lee, Kim, Lee, & Kang, 2002; Renaudin et al., 2004). For example, a user dials a number, and the cost for this call equals the price of the product. Sonera (http// www.sonera.net/asiakaspolvelu/wop. asetukset. html), a wireless service supplier, has tested this approach with a soft drink machine; the soda

Table 1. Applications classes of m-commerce

Application Classes	Type	Examples
Mobile financial applications	B2C, B2B	Banks, brokerage firms, mobile-user fees
Mobile advertising	B2C	Sending custom made advertisements according to user's physical location
Mobile inventory management	B2C, B2B	Finding products and people
Proactive service management	B2C, B2B	Sending information to salespeople regarding age of components (car industry)
Finding products and shopping	B2C, B2B	Locate/order certain products from a mobile terminal
Mobile reengineering	B2C, B2B	Improve quality of service
Mobile auctions	B2C	Customer service to buy or sell certain products
Mobile entertainment services	B2C	Video on demand; other mobile services
Mobile office	B2C	Work from the car, from airports, at conferences
Wireless database	B2C, B2B	Information is downloaded by mobile salespeople or users
Mobile music on demand	B2C	Music is downloaded and listened to while using a mobile service

machine debits a certain amount of money from the user before crediting the same amount to the cola company.

Another way of carrying out these micropayments would consist of using prepaid amounts, bought from, for example, service suppliers, banks, or credit card companies. In order to support the financial transactions, including the micropayments, a mobile-service supplier must play the role of the banker.

Mobile Advertising

Mobile advertising can also constitute a significant part of m-commerce applications. Indeed, using demographic information compiled by mobile-service suppliers and information about the physical location of the user, a highly targeted advertisement can be launched. Advertisements can be tailored to target a given user, according to the information previously provided, during a preliminary stage, or a past shopping expedition. Advertisement can also take advantage of the user's physical location. For example, users could be alerted to sales and feature events occurring in their neighborhood stores and restaurants. This type of advertising functions with a short message service or a pager.

When more wireless bandwidth becomes available, advertisements will become contain more audio, photo, and video content to fit users' specific needs, interests, and habits. Moreover, the network service suppliers will be able to use push-pull methods to make mobile advertising best suited to the user's profile.

The number of advertisements and the level and type of content they include are interesting elements. The number of advertisements must be limited in order to avoid user frustration and network congestion. Wireless networks could consider this type of service as low priority when solving congestion problems that affect the quality of service of the entire network. Because these services require information about the user's physical position, a third module could be used in order to provide localization services. However, this would result in profit sharing among the network service and the position information provider.

Mobile Inventory Management

The mobile inventory management application is used to locate products and, possibly, people. Locating products can help service suppliers specify delivery time to customers, thus improving customer service, a competitive advantage. A very interesting application is the mobile inventory, which could allow a fleet of trucks to transport a significant inventory. As soon as a store requires a certain article or product, the application would locate a truck, preferably one in the area, and obtain just in time delivery of the product. The mobile inventory and delivery applications could significantly reduce inventory cost and space for the store. Moreover, it would also decrease the time span between the moment the merchant sends and receives an order.

The mobile inventory is a B2B type of application, whereas the localization of the products can be considered a B2C application. A wireless network can locate products and services by using a radio/microwave. Since the satellite signal can be disrupted inside a truck, a separate local area network can be deployed for internal communication and to locate products. Determining an appropriate correspondence between the inventory transported in the trucks within a certain geographical area and the requests which vary dynamically remains an interesting challenge to be addressed. Note that road conditions, traffic, and construction in one area can affect just in time delivery to nearby zones.

Prospective customers for mobile inventory management could include shipping companies (UPS, USPS, FedEx, etc.), factories (e.g., automobile, construction), airline companies, the transport industry, and supermarkets. In this context, one of the problems is the integration of

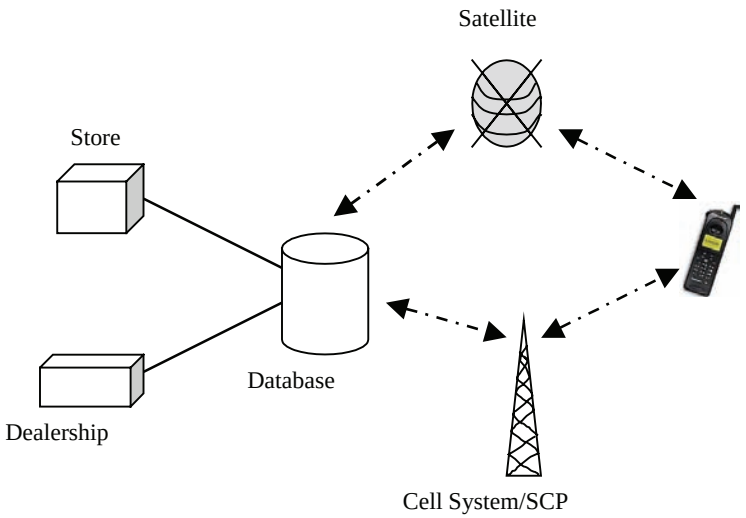
the localization information into a geographical information system (GIS). Progress has been made in this field, which led to the development of products which can find the position of a vehicle and relay this information to a SIG.

Finding Products and Shopping

The finding-products-and-shopping type of application can locate an article within a certain zone or neighborhood. It differs from the previous class because it is focused on finding a specific article or a person who offers a given service within a restricted zone, which can be delimited by the user. A specific article or an equivalent one (if indicated by the user) can be available through many different stores. In standard shopping, for major purchases (such as a new television, a videotape recorder, a car), many people visit several stores to compare various merchants' offers. By using a mobile terminal (e.g., a Palm Pilot, a Nokia Communicator or a Net Phone) and a database that includes information on the products, a user should be able to find the exact address of a store that carries this article. A list of the places as well as the distance from a specific point could be displayed. Then, the user can order this article online using the browser on his mobile terminal. If several stores carry the desired article, they can compete to earn the client's business by offering remote rebates or enticing prices in real time. This type of application can also include other forms of mobile shopping such as retail mobile sales, mobile ticketing or mobile booking.

As shown in Figure 1, a user can send a request to a central location, which can be interfaced with several stores to monitor whether a certain article is available or not, and if so, at what price. Conversely, the stores can connect their inventory systems to this site. Because the inventory systems of various stores generally use different product codes, a uniform product labeling system will be necessary to allow intelligible Web communication. If a database is unused, the mobile user will have to query stores one by one. However, the quantity of wireless traffic can quickly become problematic if the total number of requests per article and person surpasses the capacity of the wireless network. In order to avoid bottlenecks, it is preferable to use codes rather than specific data to refer to the articles. Thus, two factors must be considered:

Figure 1. Locating products and shopping



- How will the database invoice users?
- How can one verify the database or Web site reliability concerning the availability and pricing of the goods and services?

The use of mobile agents can be very efficient for these types of applications (Lenou et al., 2003). Thus, many cooperating and negotiating agents can be deployed to carry out transactions in various places.

FUTURE TRENDS

Some important issues and concerns must be addressed and solved in order to embrace and deploy mobile commerce. Future and emerging trends include three main challenges: security (Cai et al., 2004; Itani, & Kayssi, 2003; Renaudin et al., 2004), service discovery and transaction management (Veijalainen, Terziyan, & Tirri, 2003; Younas, Chao, & Anane, 2003).

Mobile commerce offers an exciting new set of capabilities that service providers can leverage to increment their revenue base while attracting new services that enhance the end-user's experience. With these new opportunities, the risk of new security threats also arises (Cai et al., 2004). New mobile devices such as PDAs and GSM/UMTS terminals enable easy access to the Internet and strongly contribute to the development of e-commerce and m-commerce services, whereas Smartcard platforms will enable operators and service providers to design and deploy new m-commerce services. This development can only be achieved if a customer's information and transactions are guaranteed to be protected by a high level of security (Renaudin et al., 2004). Thus, establishing security mechanisms which allow diverse mobile devices to support a secure m-commerce environment in wireless Internet is critical (Kim et al., 2003).

Providing security provisions for the m-commerce community is also challenging due to the insecure air interface of wireless access networks, limited computational capability of mobile devices, and users' mobility (He & Zhang, 2003). The limited equipment resources in terms of equipment require the e-payment protocol in the wireless Internet environment to be designed in consideration of the efficiency of the computing functions and the storage device.

Until now, much of the research on m-commerce has focused on the problem of service discovery. However, once a service is discovered, it needs to be provisioned according to the goals and constraints of the service provider and consumer. In this context, automated negotiation protocols and strategies that are applicable in m-commerce environments must be proposed (Paurobally et al., 2003). Specifically, time-constrained bilateral negotiation algorithms that allow software agents to adapt to the quality of the network and/or their experience with similar interactions must be developed and evaluated.

Finally, transaction management is a major issue in m-commerce. It enables people to order goods and access information anywhere, anytime. Given the nature of mobile computing, there is a need for a generic approach that adapts to the needs of m-commerce applications (Younas et al., 2003).

CONCLUSION

This article exposed some basic concepts, technology, and applications related to mobile electronic commerce. After having presented the background and related work, it summarized some prominent applications of m-commerce and outlined future and emerging trends in m-commerce.

End-to-end security remains a fundamental priority for large-scale deployment of m-commerce applications. It is also important to provide mobile terminals with a generic cryptographic functionality in its own right, which is accessible from the application layer. Moreover, because m-

commerce transactions imply sharing confidential information such as credit card numbers, it is important that mobile terminals be equipped with a safe storage unit for data as well as mechanisms for authentication and access control. In addition, an infrastructure equipped with a public key is essential to authenticate both actors and ensure secure transactions. Finally, m-commerce applications should provide a consistent user interface for easy and intuitive access to security functionalities.

Because people are becoming increasingly more nomadic, many interesting services can be offered through mobile terminals and mobile networks, such as buying and selling shares on demand or simply entertainment or information services. Such services could also include mobile games and mobile music.

Finally, remember that the user plays the key role in accepting and deploying m-commerce applications. According to studies and market analysis carried out in this field, this technology still remains somewhat immature and large-scale deployment cannot occur until security problems are solved. However, long-term projections offer a very promising future for m-commerce, which is currently gaining ground with the younger generation.

REFERENCES

Cai, Y., Kozik, J., Raether, H. L., Reid, J. B., Starner, G. H., Thadani, S., et al. (2004). Authorization mechanisms for mobile commerce implementations in enhanced prepaid solutions. *Bell Labs Technical Journal, 8*(4), 121-131.

He, L. S., & Zhang, N. (2003). An asymmetric authentication protocol for m-commerce applications. *Proceedings of the Eighth IEEE Symposium on Computers and Communications, 1,* 244-250.

Itani, W., & Kayssi, A. I. (2003, March 16-20). J2ME end-to-end security for m-commerce. *Proceedings of the IEEE International Conference on Wireless Communications and Networking, 3,* 2015-2020.

Jukic, N., Sharma, A., Jukic, B., & Parameswaran, M. (2002, May 19-22). *3-m-commerce: Analysis of impact on marketing orientation.* International Conference on Issues and Trends of Information Technology Management in Contemporary Organizations, Seattle, WA, (vol. 1, pp. 305-307).

Katsaros, I., & Honary, B. (2003, June 25-27). *Novel m-commerce security protocol for third generation mobile networks.* Fourth International Conference on 3G Mobile Communication Technologies (3G 2003), London, (pp. 23-26).

Kim, M., Kim, H., & Chung, M. (2003). Design of a secure e/m-commerce application which integrates wired and wireless environments. *Proceedings of the Third IASTED International Conference on Wireless and Optical Communications* (pp. 259-264).

Kim, M. A., Lee, H. K., Kim, S. W., Lee, W. H., & Kang, E. K. (2002, June 29-July 1). Implementation of anonymity-based e-payment system for m-commerce. *IEEE 2002 International Conference on Communications, Circuits and Systems, 1,* 363-366.

Lenou, B. E., Glitho, R., & Pierre, S. (2003). A mobile agent-based advanced service architecture for internet telephony: Implementation and evaluation. *IEEE Transactions on Computers, 52*(6), 690-705.

Paurobally, S., Turner, P. J., & Jennings, N. R. (2003, November). Automating negotiation for m-services. *IEEE Transactions on Systems, Man and Cybernetics, Part A, 33*(6), 709-724.

Pierre, S. (2003). *Réseaux et systèmes informatiques mobiles.* Montréal, Québec, Canada: Presses Internationales Polytechnique.

Renaudin, M., Bouesse, F., Proust, P., Tual, J. P., Sourgen, L., & Germain, F. (2004, February 16-20). High security smartcards. *Proceedings of Europe Conference on Design, Automation and Exhibition, 1,* 228-232.

Varshney, U. (2001). Addressing location issues in mobile commerce local computer networks. *Proceedings LCN 2001, 26th Annual IEEE Conference* (pp. 184 -192).

Varshney, U., Vetter, R. J., & Kalakota, R. (2000, October). Mobile commerce: A new frontier. *Computer, 33*(10), 32-38.

Veijalainen, J., Terziyan, V., & Tirri, H. (2003, January 6-9). Transaction management for m-commerce at a mobile terminal. *Proceedings of the 36th Hawaii International Conference on Systems Sciences,* Big Island.

Younas, M., Chao, K. M., & Anane, R. (2003). M-commerce transaction management with multi-agent support. *Proceedings of 17th International Conference on Advanced Information Networking and Applications* (pp. 284-287).

Zhang, J. J., Yuan, Y., & Archer, N. (2002). Driving forces for m-commerce success. *Journal of Internet Commerce, 1*(3), 81-105.

KEY TERMS

Business-To-Business Transaction (B2B): Electronic commercial transaction from business to business.

Business-To-Consumer Transaction (B2C): Electronic commercial transaction from business to consumer.

Electronic Commerce (E-Commerce): A set of financial transactions that can be carried out over a network. E-commerce relies upon users' interventions to initiate a transaction and select the main steps of the process.

Electronic Negotiation (E-Negotiation): Standard practice in purchasing or sales consisting of using a networked environment to negotiate in order to reach an agreement (price, delivery, warranty, etc.) between a customer and a merchant.

Mobile Commerce (M-Commerce): A set of financial transactions that can be carried out over a wireless mobile network.

Mobile Middleware: A functional layer of software provided by application developers to link their e-commerce applications to an OS and various mobile networks to allow their applications to bypass certain mobility issues.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 786-791, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.4
Mobile Communications and Mobile Commerce:
Conceptual Frames to Grasp the Global Tectonic Shifts

Nikhilesh Dholakia
University of Rhode Island, USA

Morten Rask
Aarhus School of Business, Denmark

Ruby Roy Dholakia
University of Rhode Island, USA

ABSTRACT

In this keynote chapter, we provide an overview of the emerging global landscape of mobile communications and mobile commerce, circa 2005. We introduce the four core CLIP functionalities—communications (C), locatability (L), information (I), exchange and payment (P) facilitation—on which mobile commerce systems and services are based. We then explore the various requirements for creating successful mobile commerce portals, or m-portals, using the CLIP functionalities as well as ways for personalization, permission and specification of service formats and content.

TECTONIC SHIFTS IN GLOBAL MOBILE COMMUNICATIONS

In 2004, the nation of China was adding five million new mobile telephone customers every month. That is the equivalent of adding the whole nation of Denmark, or Finland, *every month* to the mobile user base of the world's most populous country. India, the world's second most populous nation, was far behind China, but its mobile user base was also galloping ahead at a phenomenal pace. By 2005, India had, by some estimates, over 79 million users and various observers expected the number to double in 12-18 months.

While emerging nations such as China, India, Vietnam and South Africa were adding mobile telecom users at a phenomenal rate, in the advanced countries with very high mobile penetration rates, the race was on to promote new patterns of life based on mobile technologies. Take the example of the United States. Although the U.S. was slower than most European nations and the leading Asian nations in terms of mobile technology penetration and mobile data applications, by the mid-2000s a distinct pattern of making mobile communications and applications ubiquitous was becoming evident in many American cities (see Box 1 "The Race to Ubiquitous Mobile Connectivity").

Mobile commerce, or m-commerce, refers to monetary transactions conducted via a mobile telecommunications network using devices such as mobile phones, personal digital assistants (PDAs), enhanced alphanumeric handheld gadgets and so on. The global wireless mobile networks of various kinds, and the user bases of such networks, constitute the bedrock infrastructure of mobile commerce. The growing variety of terminal devices and services are the facilitative and revenue-producing tentacles of the mobile telecommunications networks. Together, the network, the devices and the services constitute the growing, globalizing and ever morphing "mobile ecosystem."

As we survey the mobile ecosystem circa 2005, tectonic shifts are occurring in it. Such shifts will continue into the foreseeable future. Including the explosive growth in Asia's mobile user base, the following represent the main tectonic shifts expected to shape the mobile commerce landscape for decades:

- Emergence of China as the world's biggest mobile communications market and the likely impact of this on everything from services to technical standards.

Box 1. The Race to Ubiquitous Mobile Connectivity

Towards the end of 2005, many cities in the United States started receiving proposals from a variety of information technology companies to blanket the entire city with Wi-Fi mobile connectivity. For example:

- Google proposed to make the entire city of San Francisco into a large, urban Wi-Fi network. Users would of course be able to take their laptops and be connected to the Internet. With the newly launched "Google Talk" service, using VOIP technology, users would also be able to make
- Earthlink, a major Internet Service Provider, similarly offered to blanket the city of Philadelphia with a ubiquitous Wi-Fi network, and to offer highly discounted services to Earthlink users while on the move anywhere in the city.
- Intel, the maker of the Centrino and other mobile data communications chips, launched programs for Wi-Fi blanketing not only in the United States but also in a dozen cities across the world.

These offers of "Wi-Fi blanketing" were of course made because of the obvious commercial benefits to the firms making these offers. While these developments of creating ubiquitous urban mobile networks were going on, the venture capital firms in the United States were bank rolling a large number of startup companies developing mobile applications.

Of course, looking into the future, many uncertainties and glitches remain. But it is almost certain that certain areas in the United States would become so saturated with free or nearly free mobile networks that people would begin to reorient their lifestyles – carrying a single mobile device of some type that would be phone, a wallet, and a browser all rolled into one.

Source: Authors' research.

- Emergence and rapid growth of a massive mobile user base in the low-income economy of India, paving the way for super-discounted services.
- Launching of third generation (3G) services in Europe, Asia, North America and elsewhere. The advanced 3G and the soon-to-follow 4G networks provide super-fast data speeds capable of opening the gates for new classes of mobile commerce offerings.
- Increasing degrees of "convergence" across various mobile communications formats (cellular, Wi-Fi, WLAN, RFID, Bluetooth and satellite-aided) and between mobile media and other communications media (landlines, cable TV, broadcast and satellite TV) and other information technologies (the Internet and computers). Each instance of convergence opens new product and service possibilities.
- Complex interplay of "standards" (CDMA, TDMA and GSM — just to name some of the cellular standards) and "generations" (2G, 2.5G, 3G, 4G, etc.). Incompatibilities of standards create barriers, but they also present opportunities for multi-format and integrative devices and services.

Because of these ongoing shifts and complexities, it is difficult to fit neat conceptual frameworks on the patterns of evolution of mobile services and mobile commerce. Nonetheless, conceptual structures are necessary for strategic purposes as well as for the practical need of training hundreds of thousands of people to work effectively in the mobile sector. In this keynote chapter, we present some basic precepts, cutting across global regions and technology formats, to help tame the complexities of mobile commerce services.

Structure of This Chapter

We begin by providing a simple definition of mobile commerce, and then review the four core ingredients — Communications (C), Locatability (L), Information (I) provision and Payment (P) facilitation — that underlie most mobile commerce applications.

Next, we introduce the idea of a mobile commerce portal, or m-portal, the electronic window through which users become aware of and deal with m-services. We outline success requirements for building versatile and appealing m-portals, based on thorough integration of functionalities. Additional factors that lead to mobile commerce success — personalization of content, permission seeking and specification of formats and content — are discussed. Finally, we reflect on elements that service partners of m-portals should pay attention to, and provide conclusions and an overview of the country perspectives that constitute the remaining chapters of this book.

M-COMMERCE: THE CORE "CLIP" INGREDIENTS

M-commerce refers to monetary transactions conducted via a mobile telecommunications network by employing devices such as mobile phones or palmtop units. For mobile commerce to happen, at the minimum the device and the network should be configured to enable communications (C), information (I) exchange and payments (P). Adding the additional geographical dimension of "locatability" (L) creates the CLIP — Communications, Location, Information and Payment — framework. CLIP functionalities are very useful for designing mobile portals (or m-portals) and providing mobile services.

The winners of the battle for leadership will be the m-portals that can utilize the key success factors for m-commerce — mobility and locatability — with a high degree of integration. Even though we are in the initial stage of m-commerce, where locatability is not fully implemented, effective business strategies for mobile commerce in the future hinge on locatability in addition to

the other core functions: communication, information and payment.

The communication (C) applications include the basic offerings of Internet service providers (ISPs), fixed-line service providers (FSPs) and wireless service providers (WSPs). Regarding voice, most mobile phones handle calls supplied by both WSPs and FSPs. Text messages come in multiple flavors like e-mail, fax, SMS (Short Message Services) and MMS (Multimedia Message Services). In many countries it is possible to route calls from the FSP to the WSP and to get e-mail messages forwarded to mobile phones or other handheld devices. Sometimes, to accomplish such integration, users have to buy a mobile phone that is WAP-enabled or can handle POP3 protocol and therefore accept e-mail. In order words, if the user chooses the right service provider and buys the right mobile device, the communication (C) functionalities could be fully integrated.

The state-of-the-art regarding information (I) delivery circa 2005 relied on SMS, MMS, WAP and Web. In the simplest versions, text-based data can be accessed. Many m-portals team up with content providers to deliver news and entertainment, and some also give access to the employing company's information system and/or private information stored in a personal information manager such as MS Outlook. Multimedia and streaming video content are gradually becoming available via network enhancements (transition to 3G networks) and device enhancements (phones with cameras).

When it comes to payment (P) functions requiring efficient and secure exchange of financial data, the methods of integration are still evolving. We are not aware of any portal that can handle stock trades, m-banking, e-wallet and billing at the same time. Terminals and services offering separate applications of these types do exist in some countries. Billing information flows directly form the WSP to the mobile end user as SMS. Most of the larger Scandinavian banks offer m-banking solutions with stock trades included. E-wallet trials are also available in Scandinavia, focused on payment in supermarkets, for parking and paying highway tolls.

Locatability (L) functionalities are also still evolving. These are based upon geo-coded data. Aggregation and integration of such data are still at low levels. We are aware of mapmakers making it possible to download maps to palmtop units that also can be equipped with a GPS receiver. Also, some trials are underway where users can get the location of the nearest restaurant, bar or convenience store based upon geographical position determined by the WSP. Most geo-coded datasets, however, are not yet available in ways that m-portals can use to personalize CLIP functions.

M-PORTAL: USER'S WINDOW TO M-COMMERCE

Users interact with mobile communications and mobile commerce systems through the small screen-based interface on their handheld devices. To the users, this small screen opens up an electronic window to the world. Besides text, tones and icons, the mobile device can also potentially offer music, photos, video, animation and other types of content. Of course, both the network and the device have to be advanced for such multimedia content to flow to the user.

To the users, the handheld device — and especially the small screen — represent the mobile portal, or m-portal, the gateway to mobile services. For service providers and device makers, the challenge is to make the m-portal versatile and capable in CLIP terms, so as to seamlessly and easily deliver a range of services to the users.

Profiling a Versatile M-Portal

To illustrate, consider the case of Angela, a sales engineer traveling from Stockholm's Arlanda airport to Tokyo's Narita airport. Upon arrival

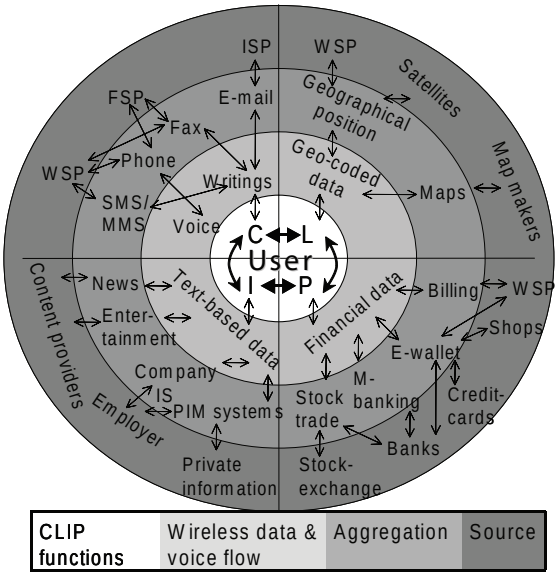
at Narita airport, her CLIP device automatically shifts from her Stockholm portal to the Tokyo portal and only shows the links relevant to Angela and Tokyo. When she enters a convenience store at the Narita airport terminal, the m-portal lists goods offered in that store based on her previous purchase history — even pointing out the shelf location of *Financial Times*, her favorite newspaper, and her preferred flavor of Altoid mints. After some personal shopping while on the hour-long Narita Express train ride to downtown Tokyo and the New Otani hotel, Angela checks her CLIP device for new e-mail messages. In one of the e-mails, a new purchasing officer at Fuji Xerox, the Japanese client firm she has come to visit, introduces himself and explains that he will be at Angela's impending sales presentation. At the Wi-Fi enabled New Otani hotel, Angela uses her mobile device as a remote control, and on the flat plasma TV screen checks out the profile of the purchasing officer on the client company's WAP site. After a quick shower and change, as she heads to the client's offices in a taxi, Angela adjusts two slides of her presentation located on her own company's Intranet, and leaves the taxi, paying with the e-wallet equipped mobile CLIP device.

The m-portal is an individual-specific portal tailored for both personal and professional tasks. In addition to the personalization features evident in the Tokyo trip illustration, the m-portal is PIM-based.[1] It can draw on all of Angela's contact, schedule and task information and use such information to automatically generate the contents of the portal. The success of the m-portal depends on a continuous-loop personalization. Such a personalization makes it very difficult to maintain the distinction between Angela's private and professional lives. It is entirely possible, for example, that a musical greeting card sent by her 7-year old son back home in Stockholm could pop-up on the handheld screen in the middle of the sales meeting in Tokyo. Angela needs to ensure that her handheld device is programmed to keep a discrete line of separation between personal and professional content.

Figure 1 shows the need for integration, which is the primary key success factor for the m-portal. All the four CLIP functionalities need to work well within each individual function (e.g., integrating various payment types [P], such as from e-wallet and credit cards) and across the CLIP functions (e.g., Angela's locatability [L] at Narita triggering information [I] about the availability of her favorite flavor of Altoid mint, and enabling her to pay [P] for the purchase). Figure 1 illustrates the business opportunities for the m-portal, where the arrows symbolize the needed integration. The first-level integration of the communication, location, information and payment functions happens in the CLIP device. The figure also shows that the m-portal owner has to integrate already existing offerings or build applications that integrate the possible wireless data flows, aggregations and sources. Seamless and smooth-functioning partnerships with shared revenue are needed for effective integration of sources and services.

In the initial phase of the evolution of m-com-

Figure 1. CLIP integration requirements for an m-portal



merce, for some of the larger players, the key strategic goal will be to attain a leadership position in the m-services space, i.e., to become an m-portal. For other firms, and for the firms that fail to become m-portals, strategies will have to evolve in terms of becoming effective m-portal service partners. While it is too early to predict what the competitive field of m-commerce will look like in various global regions, we can utilize Figure 1 to delineate some of the success requirements. We can do so for three situations: the global battles for leadership in the m-commerce space, the strategic requirements for the leading successful m-portals and the strategic requirements for m-portal service partners.

DYNAMICS OF LEADERSHIP IN M-COMMERCE SPACE

It is evident from Figure 1 that the Wireless Service Providers (WSPs) are well positioned for attaining leadership positions in the m-commerce business space. Besides being in charge of the wireless data and voice flow to and from the CLIP device, WSPs also have access to sources that provide the value-adding communication, location, information and payment features. Additionally, some WSPs are also building applications that access the information systems of the users' employers. For example, the Danish WSP Sonofon has teamed up with HP to create access to the company's Intranet (Hewlett-Packard, 2000). In a report, the consultant firm Strategis Group Europe (2000a, 2000b) concludes that "wireless portals will provide operators with key competitive edge in Europe" and the WSP and the device manufacturers have core competencies in creating m-portals. Durlacher, another European consultancy, suggests that WSPs team up with traditional Internet portals because they have complementary strengths. WSPs bring experiences with mobile communications, billing and location information to the table. These elements represent the weaknesses of the traditional Internet portals that, in return, have strengths in portal configuration, content creation and presentation, application and partnering experiences (Müller-Veerse, 1999). Partnering will be a key success factor for m-portals, a theme that we will visit later in this chapter.

With the exception of Japan's NTT DoCoMo (see Bradley & Sandoval, 2002; see also the Japan chapter in this volume), WSPs did not have a very good start in the m-portal business.[2] There have also been a lot of teething troubles with the first version of the preferred WAP protocol. [3] Many WSPs bet on the previously used "walled garden" content model, which restricts subscribers' access to third party portals. That approach — limiting of services to a set controlled and promoted by the WSP —– had no success at all (see the Denmark chapter in this volume).[4]

The key success requirement for mobile commerce is the seamless integration of mobility and locatability. In the case of Angela presented earlier in this chapter, her m-portal configured itself as she moved, first internationally from Stockholm to Tokyo, then locally in the Narita convenience store and then to her Wi-Fi enabled hotel room. Leading mobile commerce portals would be those that offer high degrees of integration of CLIP services, based on location and context, for users on the move. At the current stage of mobile commerce, the geo-capabilities of locatability are not fully implemented yet. Future business strategies for mobile commerce — especially in those global regions where 3G networks exist — must be based on locatability being a key feature of the mobile commerce network.

ADDITIONAL ELEMENTS FOR M-PORTAL SUCCESS

To be effective and appealing, m-portals must blend elements of personalization, permission and specification of the CLIP features in m-commerce

services. From the prior experience of landline and desktop terminal-based e-commerce, we know that the e-commerce players that survived and thrived did a very good job of *personalizing* the content to the users, carefully seeking the users' *permission* for various types of communications and services, and allowing the users to *specify* how the services and content should be presented to them. As Table 1 shows, for m-commerce these three elements — personalization, permission and specification — take on an even stronger role than for e-commerce.

The essential task of the m-portal is to be an intermediary between service provider and user, and a mediator between multiple media and service formats. In principle, the m-portal can be a database permit, specify and personalize the communication, provide information, enable payment and provide location functions, where the primary mobile communications provider delivers all the data and voice. This is illustrated in Table 2.

The m-portal will handle the permission element by giving the user certain rights to define the types of communication, information and payment features. The m-portal will also offer one-button (or voice activated) disabling functions so that pre-set permissions — such as determining and communicating the user's location — can be suspended temporarily. Successful m-portals would have to allow users to become partly or totally invisible to the commercial side of the network for those situations and contexts when the users desire privacy.

Table 1. Personalization, permission, and specification in m-commerce and e-commerce

Dimension	E-Commerce	M-Commerce
PERSONALIZATION		
User-centric database	**Slow Evolution:** Evolves from navigation and transaction behavior of the user	**Fast Evolution:** Evolves from daily communications and linking of multiple databases
Tailoring of services and content	**Somewhat Limited:** Depends on inferences about user's preferences and roles	**Possibly Extreme:** User revealed preferences, inferred roles and preferences and location factors can be used to tailor offerings
Learning and intelligence	**Limited:** Based on collaborative filtering and profiling	**Extensive:** Based on collaborative filtering and profiling applied to multiple databases
PERMISSION		
Scope of permission	**Relatively Narrow:** Merchant-specific, defined in user agreement	**Relatively Broad:** Often unspecific and location based
Depth of permission	**Relatively Shallow:** Very specific transactions and charges are permitted	**Relatively Deep:** Extensive range of transactions and payments are permitted
SPECIFICATION		
Role demarcation	**Sharp:** Especially in firewalled work environments	**Blurred:** Difficult to tell whether user is on or off duty
Nature of role specification	**Static:** Determined by the location of the client terminal	**Dynamic:** Depends on user preferences, merchant preferences and geographic location
Service or content specification	**Somewhat Configurable:** Depends on client terminal IP address and revealed user identity	**Evolving and Dynamic:** Depends on user preferences, merchant capabilities and location characteristics

Table 2. Contents of an effective business strategy for m-portals

	Communication	Location	Information	Payment
Permission	Types of communication and senders can be permitted or forbidden	Types of information and senders can be permitted or forbidden	Payment features can be enabled or disabled, individually or collectively	Locatability and geo-positioning features can be enabled or disabled
Specification	Off/on duty "button," preferences, current time of the day and location of the user specify which messages go through	Off/on duty "button," preferences, current time of the day and location of the user specify types of information	Off/on duty "button,", user and merchant preferences, current time of the day and location of the user specify types of transactions	Geographical position feeds CLIP specification features
Personalization	Dynamic unified inbox	"Me & My": Personalized information portal for news, travel information, PIM, company information and entertainment	Personal e-wallet, stock portfolio and phone bills	Dedicated maps

SUCCESS REQUIREMENTS FOR M-COMMERCE SERVICE PARTNERS

Firms with diverse interests and competences have to cooperate to create and operate effective and attractive m-portals. Three vital groups of partners are central for mobile commerce: Device Manufacturers, Infrastructure Enablers and Content Providers. Table 3 shows these types of firms according to the communication, information, payment, and location features in mobile commerce.

Mobile phone manufacturers such as Nokia, Sony-Ericsson, Motorola, Samsung and Kyocera are working intensely to create standard devices for mobile commerce communications and transactions. Firms such as Palm, Psion, Handspring and Microsoft have wireless strategies allowing the use of handheld Personal Digital Assistants (PDAs) as the main m-commerce device. These firms will attempt to drive mobile commerce in the directions they think are most profitable. Some outsiders, however, will also be in the game. These include Pager firms such as Research in Motion (maker of BlackBerry devices), Glenayre and Tandy Radio Shack; and GPS receiver makers such as Garmin, Lowrance and Magellan. No de facto standards and protocols have emerged yet, so it is too early to describe the general interface between the devices and the m-portal. It seems reasonable, however, to focus strongly on the mobile phone producers because they already have developed solutions for all four primary CLIP functions — communication, location, information and payment — of the m-portal.

When it comes to infrastructure enablers, the most important partners for the m-portal are the wireless service providers (WSPs). Other contenders include Internet service providers (ISPs), fixed-line service providers (FSPs), GPS network providers, content aggregators, Internet portals, banks and credit card firms. In some cases, electric utility companies, transportation firms and television firms may play important roles in enabling mobile commerce. In the initial years of mobile commerce, WSPs hold an advantage — they are already positioned as the mobile com-

Table 3. M-portal partner products and services

	Communication	Location	Information	Payment
Device Manufacturers	Mobile Phones, PDAs and Pagers	Mobile Phones and GPS receivers	Mobile Phones, PDAs and Pagers	Mobile Phones, PDAs
Infrastructure Enablers	ISP, FSP and WSP	WSP and GPS Networks	Content aggregators and Internet portals	WSP, Banks and Credit card firms
Content Providers	WSP	Map makers	News agencies. Travel firms, Entertainment firms, PIM firms and Employers	Banks, Exchanges, WSP and Virtual and Physical businesses

munications companies that have either their own or strongly dedicated infrastructures. Over time, however, other firms could make inroads — just as they did in the fixed line and ISP businesses in the recent past.

Perhaps the most important group of partners is the content providers. Each of the CLIP elements has specialized partners. The appropriate partner to handle the communication part will be the WSP. For location-related services, mapmakers are best positioned to provide location-related content. Many third-party information and entertainment providers could be partners in providing information. In order to handle the payment function, banks, exchanges, WSPs and virtual and physical stores are likely to provide content. A big challenge for all will be the need to supply geo-coded information so that specification of the m-portal services can be appropriate to the location-role of the user. Another and even bigger challenge is to integrate multiple and often competing technologies such as the following:

- Network Technologies (GSM, HSCSD, GPRS, EDGE and 3G) and short-distance wireless technologies (such as Wi-Fi and Bluetooth)
- Service Technologies (SMS, MMS, USSD, Cell Broadcast, SIM Application Toolkit, WAP, Web Clipping and MexE)
- Mobile Middleware (Mobile Portal Platforms, Mobile Commerce Platforms, Mobile Payment Platforms and Mobile Banking Platforms)
- Mobile Commerce Terminals (Operating Systems, Physical Terminals, Microbrowser, Bluetooth, Smartcards, PKI and Synchronization)
- Mobile Location Technologies (GPS, TOA, E-OTD, COO and LFS Independent)
- Transportation modes (bicycles, cars, buses, trains, airplanes and boats)
- Mobile Personalization Technologies
- Content Delivery And Format (XML, WML, VXML and cHTML)

While the competitive picture at this stage is emergent and blurred — with many potential m-portal partners — it is a good strategic stance to have a strong focus on the specific business strengths of each potential m-portal partner.

QUESTIONS FOR DISCUSSION

1. Which of the CLIP dimensions have proved to be most challenging in the global m-commerce context so far? Why?
2. Why are the concepts of "Personalization," "Permission" and "Specification" important

for the success for m-commerce services? Illustrate using the examples of at least two mobile commerce services.

3. Discuss how partnerships among various service provider and technology developer organizations help in the creation and promotion of m-commerce offerings.

REFERENCES AND ADDITIONAL READINGS

Bradley, S.P., & Sandoval, M. (2002). Case study: NTT DoCoMo – The future of the wireless Internet? *Journal of Interactive Marketing, 16*(Spring), 80-96.

Dholakia, N., Dholakia, R. R., Lehrer, M., & Kshetri, N. (2004). Global heterogeneity in the emerging m-commerce landscape. In Nan Si Shi (Ed.), *Wireless communications and mobile commerce* (pp. 1-22). Hershey, PA: Idea Group Publishing.

Funk, J. L. (2001). *Global competition between and within standards — The case of mobile phones.* New York; London: Palgrave.

Godin, S. (1999). *Permission marketing: Turning strangers into friends, and friends into customers.* New York: Simon and Schuster.

Hamilton, E. (2000). *Japan mobile Internet case study: NTT DoCoMo i-Mode* (Presentation). Washington, DC: The Strategis Group. Retrieved from http://www.strategisgroup.com/press/pubs/docomo.pdf

Hewlett-Packard. (2000). *Intranet åbnes for WAP.* Retrieved August 24, 2000, from http://www.hp.dk/firma_information/presse/2000/000313.html

May, P. (2001). *Mobile commerce: Opportunities, applications, and technologies of wireless business.* Cambridge UK: Cambridge University Press.

Müller-Veerse, F. (1999). *Mobile commerce report.* London: Durlacher. Retrieved November 19, 1999, from http://www.durlacher.com/research/resrep-detail20.asp

Pelkonen, T., & Dholakia, N. (2004). Understanding emergent m-commerce services by using business network analysis: The case of Finland. In Nan Si Shi (Ed.), *Wireless communications and mobile commerce* (pp. 105-31). Hershey, PA: Idea Group Publishing.

Peppers, D., & Rogers, M. (1993). *The one-to-one future: Building relationships one customer at a time*. New York: Currency/Doubleday.

Rask, M., & Dholakia, N. (2004). Configuring m-Commerce portals for business success. In Nan Si Shi (Ed.), *Mobile commerce application* (pp. 76-94) Hershey PA: Idea Group Publishing.

Samuelsson, M., & Dholakia, N. (2004). Assessing the market potential of network-enabled 3G m-business services. In Nan Si Shi (Ed.), *Wireless communications and mobile commerce* (pp. 23-48). Hershey, PA: IGP.

Strategis Group Europe. (2000a). *European wireless portals: Strategies & market positioning* (Presentation). London: The Strategis Group. Retrieved July 13, 2000, from http://www.strategisgroup.com/press/pubs/ewireless.pdf

Strategis Group Europe. (2000b). *Wireless portals will provide operators with key competitive edge in Europe* (Press Releases). Strategis Group. Retrieved July 5, 2005, from http://www.strategisgroup.com/pr

ENDNOTES

[1] PIM stands for Personal Information Manager. It refers to software, devices and databases that keep track of personal calendars, addresses, notes, etc.

[2] See Brandt, 2000; Economist, 1999c; Economist, 2000c; Hamilton, 2000; Hara, 1999; Hoffman, 2000; and Kunii, 2000.

[3] See Baker, Gross, Kunii, & Crockett, 2000; E-business Forum, 2000; Economist, 2000b; Financial Times, 2000g; Hara, 1999; Müller-Veerse, 1999; Nielsen, 2000a; Nielsen, 2000b; and Young, 2000.

[4] See Baker, 2000; Economist, 2000a; Economist, 2000b; Nielsen, 2000b; Smith, 2000; Strategis Group Europe, 2000b; and Young, 2000.

This work was previously published in M-Commerce: Global Experiences and Perspectives, edited by N. Dholakia, M. Rask, and R. Dholakia, pp. 1-14, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 1.5
Adoption and Diffusion of M-Commerce

Ranjan B. Kini
Indiana University Northwest, USA

Subir K. Bandyopadhyay
Indiana University Northwest, USA

INTRODUCTION

Mobile commerce (or in short, m-commerce) is currently at the stage where e-commerce was a decade ago. Many of the concerns consumers had regarding e-commerce (such as security, confidentiality, and reliability) are now directed towards m-commerce. To complicate the matter further, the lack of a standardized technology has made m-commerce grow in multiple directions in different parts of the world. Thus, the popularity of m-commerce-based services varies by country, by culture, and by individual user. For example, in Europe the most popular application is SMS (short message service) or text messaging, in Japan interactive games and picture exchange via NTT DoCoMo i-mode, and in North America e-mail via interactive pagers (such as RIM BlackBerry) and wireless application protocol-based (WAP-based) wireless data portals providing news, stock quotes, and weather information. It is safe to predict that these applications will take on different forms as the technologies mature, devices become more capable in form and functionality, and service providers become more innovative in their business models.

It is true that m-commerce has witnessed spectacular growth across the globe. It is also encouraging that several factors are expected to accelerate the pace of adoption of m-commerce. Notable among these drivers is convergence in the voice/data industry, leaping improvements in related technology and standards, adoptive technology culture in many parts of the world, and governmental and regulatory initiatives.

Despite the undisputed promise of m-commerce, there are several barriers that are slowing the pace of adoption of m-commerce. The major barriers include: (a) lack of good business models to generate revenues, (b) perception of lack of

security, (c) short product lifecycle due to rapidly changing technology, (d) non-convergence of standards, (e) usability of devices, (f) limitation of bandwidth, and (g) cost.

Many of the aforesaid were common to e-commerce also at its introduction and growth stage. We strongly believe it is worthwhile to investigate how e-commerce has been able to overcome these barriers so that we can incorporate some of the successful strategies to m-commerce. In our study, we will first compare and contrast e-commerce and m-commerce with respect to a set of common criteria such as: (1) hardware requirement, (2) software requirement, (3) connection or access, and (4) content. In the process, we will identify the principal barriers to the development of m-commerce as outlined in the above list.

The Growth in E-Commerce

Electronic commerce or e-commerce is the mode of commerce wherein the communication and transactions related to marketing, distributing, billing, communicating, and payment related to exchange of goods or services is conducted through the Internet, communication networks, and computers. Since the Department of Defense opened up the Internet for the public to access in 1991, there has been exponential growth in the number of Web sites, users on the Web, commerce through the Web, and now change of lifestyle through the Web (Pew, 2006).

The chronology of events shows that as the Internet became easier and cheaper to use, and as the applications (such as e-mail and Web interaction) became necessary or useful to have, the rate of adoption of the Internet accelerated. In fact, the rate of adoption of the Internet surpassed all projections that were made based on the traditional technology adoption rates that were documented for electricity, automobile, radio, telephone, and television (Pew, 2006). Unfortunately, the over-enthusiastic media hyped up the growth rate to an unsustainable level, leading to unprecedented growth of investment in the Internet technologies and followed by a melt-down in the stock market. This shattered the confidence in Internet technologies in the investment market. Although there was a significant deceleration in IT investment, e-commerce has rebounded to a large extent since the dot.com bust. It has been growing at about 30% compound rate per year (Pew, 2006).

In the last 10 years, the adoption of e-commerce has been extensively studied both by academicians as well as practitioners. During this period e-commerce and the scope of its definition also went through various iterations. For example, people may not buy a car on the Internet, but it is documented that 65% of car buyers have done extensive research on the Web about the car they eventually buy. Is this e-commerce? Should we restrict the e-commerce definition to financial exchange for goods or services? We have various such examples in the marketplace where extensive research about the product or service is conducted on the Internet, but the final purchase is made in the physical environment. Hence, although the number of consumer financial transactions has not grown to the level industry projected initially, there has been a significantly high rate of adoption of the activities supporting e-commerce.

In addition, there has been a very high rate of adoption of business-to-business (B2B) commerce both in terms of financial and supporting transactions. In this article, we are interested in business-to-consumer (B2C) commerce. Hence, the comparison and contrast is made between e-commerce and m-commerce. All our discussion henceforth will be on B2C commerce using desktop and/or mobile technologies.

The Growth Potential of M-Commerce

Mobile commerce is the model of commerce that performs transactions using a wireless device and data connection that result in the transfer of value in exchange for information, services, or

goods. Mobile commerce is facilitated generally by mobile phones and newly developed handheld devices. It includes services such as banking, payment, ticketing, and other related services (DEVX, 2006; Kini & Thanarithiporn, 2005).

Currently, most m-commerce activity is performed using mobile phones or handsets. This type of commerce is common in Asian countries led by Japan and South Korea. Industry observers are expecting that the United States will catch up soon, with mobile phones replacing existing devices such as ExxonMobil's Speedpass (eMarketer, 2005; Kini & Thanarithiporn, 2005).

Although the U.S. is lagging behind many countries in Asia and Europe in m-commerce, a UK-based research firm projects North American m-commerce users to total 12 million by 2009, with two-thirds of them using the devices to buy external items such as tickets and goods, and a third of them using it to make smaller transactions through vending machines (eMarketer, 2005). The firm also notes that there is a large potential number of the 95 million current American teens who are already making purchases on the Web that will adopt m-commerce. However, the study also remarks that generating widespread user interest in m-commerce and addressing security fears of mobile payment technologies and m-commerce services are critical in achieving a high level of adoption (eMarketer, 2005).

While the Asia Pacific Research Group (APRG, 2006) projected in 2002 that global m-commerce would reach US$10 billion 2005, Juniper Research currently projects that the global mobile commerce market, comprising mobile entertainment downloads, ticket purchases, and point-of-sale (POS) transactions, will grow to $88 billion by 2009, largely on the strength of micro-payments (e.g., vending machine type purchases). See eMarketer (2005) for more details.

Today, a large percentage of mobile phone users use mobile phones to download ring tones and play games; hence content-based m-commerce is expected to make up a small percentage of m-commerce. One recent study, however, projects that in the future mobile phone users will move up the value chain from purchases that are used and enjoyed on the mobile phone to external items such as tickets, snacks, public transportation, newspapers, and magazines (eMarketer, 2005).

Diffusion Models of Technology Adoption

There are many models that have been formulated and studied with regard to technology adoption, acceptance, diffusion, and continued adoption. These theories identify factors that are necessary to support different levels of adoption of information and communication technologies (ICTs). Notable among these models are the innovation-diffusion theory (Roger, 1995), technology acceptance model (or TAM) based on the theory of reasoned action (Davis, 1989; Fishbein & Ajzen, 1975), extended TAM2 model that incorporates social factors (Venkatesh & Davis, 2000), technology adoption model based on the theory of planned behavior (Ajzen & Fishbein, 1980), post acceptance model based on marketing and advertising concepts (Bhattacherjee, 2001), and SERVQUAL (Parasuraman, Berry, & Zeithaml, 1988) for service quality. These models have been extensively used to predict and evaluate online retail shopping and continued acceptance of ICTs. In addition, varieties of integrated models have been developed to measure the success of information systems, ICT, and Internet adoption and diffusion. Currently, many of these models are being tested in the context of mobile technology (primarily mobile phone services).

The integration models mentioned above have been empirically tested in the e-commerce area. The models have been authenticated and proven to be extremely useful in predicting behavior of users of ICT and e-commerce. In the case of m-commerce, the results have been slightly inconsistent. Primarily these inconsistencies have been found because of the differing market maturity levels or

the usage pattern of mobile devices. For example, in a South Korean study where mobile phones have been in use for quite some time, the results of testing an integrative m-commerce adoption model yielded different results for actual use than in a similar study conducted in Thailand where mobiles devices were introduced much later in the market. South Koreans were not influenced much by advertising, unlike Thai people in the initial adoption phase of m-commerce. Conversely, Thai people were not influenced by word-of-mouth to the extent South Koreans were influenced in the initial adoption (Thanarithiporn, 2005). According to Thanarithiporn (2005), this is due to the fact South Koreans are at a more advanced level of adoption for ICTs. Furthermore, Thanarithiporn (2005) found that, unlike in South Korea where content availability had no influence in the continued use of mobile phones, it had a strong influence in Thailand on mobile usage rate. Also, in both countries self-efficacy had no influence one way or the other in the initial adoption of the mobile phone.

Key Factors that Affect the Adoption and Diffusion of E-Commerce and M-Commerce

As expected, many factors influence the rate of adoption and diffusion of technological innovations. We reviewed the extant literature, as outlined above, to identify those factors. In particular, we were interested in a set of factors that have significant influence in the adoption and diffusion of both e-commerce and m-commerce. These include: (a) hardware requirement, (b) software requirement, (c) connection or accessibility, and (d) content. In the following paragraphs, we will outline how these factors have influenced the development of e-commerce, and are currently influencing the adoption and diffusion of m-commerce.

Hardware Requirement

E-Commerce

Computer users were used to the QWERTY keyboard (of typewriters), thus they easily adapted to the standardized desktop of the first personal computers (PCs) in the 1980s. The development of graphical user interface (GUI), mice, and various other multimedia-related accessories has made PCs and variations thereof easy to use. With the introduction of open architecture, the adoption and diffusion of PCs proliferated. The introduction of the Internet to the common public, and the introduction of the GUI browser immediately thereafter, allowed PC users to quickly adopt the Web browsers and demand applications in a hurry. The limitation of hardware at the user level was only restricted by the inherent rendering capability of a model based on the processors, configuration, and accessories that supported them. Since the Web and e-commerce server technologies that serve Internet documents or Web pages are also based on open architecture, limitations were similar to that of desktops.

M-Commerce

The hardware used for mobile devices are complex. The evolution of the hardware technology used in mobile devices is diverse because of the diversity in fundamental architecture. These architectures are based on diverse technology standards such as TDMA, CDMA, GPRS, GSM, CDMA/2000, WCDMA, and i-mode. In addition, these architectures have gone through multiple generations of technology such as 1G (first generation – analog technology); 2G (second generation – digital technology, including 2.5G and 2.75G); and 3G, to meet the demands of customers in terms of bandwidth speed, network capabilities, application base, and corresponding price structures. The lack of uniform global standards and varied sizes and user interfaces to operate the devices has further disrupted the smoother adoption process.

While the U.S. still suffers from a lack of uniform standard, Europe is moving towards uniformity through some variation of TDMA technology, and China is modifying CDMA technology to develop its own standard. Other countries are currently working towards a uniform standard based on a variation of base TDMA or CDMA technology (Keen & Mackintosh, 2001).

The innovation in the changing standards, devices, applications, and cultural temperament have constantly maintained a turbulent environment in the adoption and diffusion of commerce through mobile devices. For example, if the device is WAP-enabled, then Web services can be delivered using standardized WML, CHTML, or J2ME development tools. But the WAP enabling has not given scale advantages because hardware standards have not converged, at least not in the U.S. where consumers use a multitude of devices such as Palm, different Web-enabled phones, and different pocket phones.

Software Requirement

E-Commerce

The standardization and open architecture of PCs, along with the high degree of penetration of PCs in the office and home environment, allowed for standardization of client devices. This allowed for the development of text browsers, and subsequently the development of the graphical interface through Web browsers. Apples, PCs, and other UNIX-based workstations were able to use the device-independent Web browsers, thus leading to rapid adoption and expansion in the usage of Web browsers. The low price of earlier browsers such as Mosaic and Netscape, and the distribution of Internet Explorer with the Windows Operating System by Microsoft allowed the diffusion of the browsing capability in almost every client in the market.

Standardized browser software and interface, along with market dominant operating systems such as the Windows family of desktop operating systems and server platforms, facilitated the exponential growth of Internet users and applications. The availability, integration, and interoperability of application development tools, and the reliance on open systems concept and architecture, fueled further changes in the interactivity of the Web and indirectly boosted the commerce on the Web. The development of hardware-independent Java (by Sun Microsystems) and similarly featured tools allowed growth in the interactivity of the Web and application integration both at the front end and backend of the Web. The interoperability of Web applications to communicate with a wide variety of organizational systems initiated a concern for security of the data while in transit and storage. In the early stages of e-commerce, major credit card companies did not trust the methodologies that were used, although they allowed the transactions. Beginning in 1999, they started protecting the online customers just as they protected off-line customers (namely, a customer is only responsible for $50 if she reports the card stolen within 24 hours). The technology companies and financial service organizations collaboratively created and standardized methodologies for online secure transactions, and originated the concept of third-party certification of authority. This certification practice further strengthened the security of online commerce and established a strong basis for consumers to trust and online commerce to grow.

M-Commerce

Software for mobile technologies is dependent on the technology standard used and type of applications suitable for the mobile device. In most nations, like in the U.S., the use of mobile devices started with the use of analog cellular phones. These required proprietary software and proprietary networks. The digitization of handheld devices started with personal digital assistants (PDAs) for personal information management. The transformation of the PDA as a digital communication tool was made possible

by private networks, operating systems, and applications developed by companies such as Palm. However, as Microsoft's Windows CE (Compact Edition) and BlackBerry started offering e-mail, information management tools, and Web surfing using micro-browsers, the growth in the use of handheld devices for Web applications started growing. The handheld industry responded with a variety of applications and made WAP a standard for applications development.

Concurrently, the telecom industry brought out digital phones and devices that could offer voice, personal information management (PIM), and data applications. However, until now, operating systems, servers, and Web applications are not standardized in the handheld market. The diversity of server software and client operating systems, and the availability of applications have not made these devices interoperable. In addition, with each player offering its own network and original content or converted content (i.e., content originally developed for the desktop computers), the interest in commerce using mobile devices has not been too enthusiastic. Furthermore, the lack of common security standards has made mobile commerce adoption very slow.

Connection or Access

E-Commerce

In the United States, where telephone wire lines have been in existence for over 100 years, it was natural for the telecom companies to focus on offering Internet connectivity through the existing telephone network. In the early stages of pubic offering of the Internet, it was easy for people to adopt the Internet using their modem from a private network. As the Internet evolved into the World Wide Web, and innovation brought faster modems to the market, more Internet service providers (ISPs) started providing ramps to the Internet. When the Windows98 Operating System with its integrated Internet Explorer was introduced to the marketplace, the Internet adoption was growing in triple digits per year. The major infrastructural components were already in place. The telecom sector invested heavily into building the bandwidth and router network to meet the insatiable demand for Web surfing. Worldwide Internet adoption and use was growing exponentially. The ICT industry responded with innovative technologies, software and services using standardized PCs, modems, support for (Internet protocol suite) TCP/IP protocol of Internet, and highly competitive pricing. The e-tail industry subsequently started growing rapidly, and the financial service industry introduced innovative products and services while collaboratively designing secure electronic payment mechanisms with ICT industry players.

The drop in pricing, availability of bandwidth, security, and quality of products and services bolstered the commerce activity on the Internet until the 'dot.com bust' of May 2000. Although the bust slowed the growth rate of e-commerce, in reality e-commerce continuously grew despite the bust. Support for e-commerce from the U.S. government to fuel the e-commerce growth through moratorium on taxes by two administrations considerably helped the diffusion of e-commerce. The concern about the security in e-commerce shown by laggards was eased by a variety of security and encryption tools, and the creation of the certification of authority concept by strong security services offered by companies such as Verisign, TRUSTe, and others.

Lately, the demand for highly competitive broadband service availability, and the availability and delivery of media-rich content, has brought media and entertainment industry to the Web with greater force. These technological advances in the e-commerce sector have received increased attention, thus ensuring a strong global growth rate in e-commerce.

M-Commerce

In the mobile arena, customers may have been using analog cellular phones (1G) for a long of

time. During the era of analog cellular phones, the common mobile commerce activity was the downloading of ring tones. This type of commerce activity is still quite prevalent in developing nations. In addition to this type of commerce, other types of commerce conducted using these devices are the same as the ones that can be performed using a standard desk phone, such as ordering tickets for an event, ordering catalog items, and similar tasks.

With the introduction of digital devices (2G), mobile phones quite suddenly have become the lifeline for many transactions, such as e-mail, voicemail, and text messaging. With 2.5G, 2.75G, and now with 3G devices, more varied and complex applications such as photo transfers, interactive games, and videos have become the norm. The capabilities of these devices are determined by technical ability of the devices and the support of terrestrial tower structures by the vendors offering these services. In addition, the content availability and their desirability by the customers also determine the adoption of such services. The technology, standards, and competition have left U.S. vendors in the distance in rolling out new technology and services. While Asia's (South Korea, Japan, and China) mobile penetration growth is three times that of the United States, Europe is closely behind Asia, with England (87%) and Finland (75%) achieving very high penetration rates (Shim, 2005). In the U.S., the major players in the telecom industry are collaborating to achieve the 3G-standard Universal Mobile Telecommunication System (UMTS) to provide penetration and support rollout of new technology and services. Several countries including South Korea were planning to offer a more advanced technology called the Digital multimedia broadband (DMB) or wire broadband (WiBro) by the end of 2006 (Shim, 2005). According to Shim (2005), it will take a while to obtain DMB cellular phone services in the U.S., since technical standards and logistical barriers will have to be overcome first.

The private networks built by the wireless service providers through the customized devices will determine the access and speed available in the future in the United States. The investment in the network, along with the rollout of new technology and methods used to price the services, will be strong factors in building the capacity. Government policies are also vital in this respect. According to Shim (2005), the government commitment and push for IT strategy and long-term goals are among the most important factors to advance a country's cellular mobile business, particularly for less-developed countries.

Content

E-Commerce

Identifying the most preferred method for delivery of any content has always been a thorny issue. In electronic commerce, the complete digital conversion of all media into technology mandated by the FCC by 2008 would be much easier (FCC, 2006). Voice, as well as radio and television signals, will be broadcast digitally. The Internet has built capacity to deliver rich media content at high speed using the fiber network in the U.S. The convergence of devices such as TV monitors and PC monitors has already brought down the prices for such devices due to scale effects. The stumbling blocks to achieve a greater level of broadband adoption (from the current 53% in the U.S.) are pricing and quality of content (Pew, 2006). In e-commerce, content can be provided by anyone using standardized development tools and can be served on the standardized server software since most desktops can handle all the content delivered through the Web. The diffusion of such innovations is constrained by the pricing and the investment made by consumers at the client level. The industry has converged in standardizing hardware, software, and protocols. Globally as well as in the U.S., there is a clear trend to make the technology affordable throughout the world through the open systems concept. This has helped

tremendously, especially in developing countries, in the adoption and diffusion of the Internet and generalized applications.

M-Commerce

In mobile commerce, the content such as data, text, audio, video, and video streaming can be delivered through the devices provided by service providers through their network infrastructure. As the service providers rollout new network technologies with greater capabilities to adapt to the new generation of hardware and software technologies, consumers can expect more media-rich content. Any content that is available in the e-commerce world will be specially modified for mobile delivery using specific development tools for WAP-enabled devices such as WML, CHTML, and J2ME.

Depending on the type of device, the content will have to be delivered in device-specific configuration—for example, the content has to be delivered differently to a PocketPC, WAP-enabled mobile phone, and WAP-enabled PDAs. This type of dynamic configuration in the content delivery requires investment from service providers and/or value-added intermediaries. The special intermediaries provide enormous value-added services in converting the e-commerce content for different mobile devices and become consolidators of content and applications and essentially become data portals for mobile devices. The diversity of devices available in the market will require a significant amount of investments in the U.S. to offer it nationwide, unless it focuses only on high-population density regions to maximize returns.

CONCLUSION

Based on the foregoing discussion, we can say that the introduction of e-commerce has been comparatively smoother than m-commerce. The development of the hardware capability (from PC to GUI to other multimedia-related accessories such as printers, camera, etc.), the software capability (such as browsers, open operating systems, payment schemes, secure systems, etc.), better accessibility (such as phone lines, cables, etc.), and more varied content (such as voice, radio, and television signals) ensured a fast adoption and diffusion of e-commerce throughout the world.

It is true that m-commerce also enjoys many advantages similar to e-commerce. For example, the mobile phone—the principal mode of m-commerce—is witnessing a spectacular growth throughout the world. Unfortunately, unlike e-commerce, m-commerce does not enjoy an open architecture that can accommodate varied standards in hardware, software, connection technology, and the content. Several countries (such as Japan and South Korea) are further ahead of the U.S. in solving this issue of incompatible technologies. It is heartening to see a sincere effort in many countries, including the U.S., to achieve convergence in technologies so that m-commerce is able to grow true to its full potential.

REFERENCES

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice Hall.

APRG. (2006). Retrieved from http://www.aprg.com

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly, 25*(3), 351-370.

Cassidy, J. (2002). *dot.con: The greatest story every sold.* New York: Harper Collins.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quarterly, 13*(2), 319-339.

DEVX. (2006). Retrieved from http://www.devx.com/wireless/Door/11297

eMarketer. (2005). Mobile marketing and m-commerce: Global spending and trends. *eMarketer,* (February 1).

FCC. (2006). Retrieved from http://www.fcc.gov/cgb/consumerfacts/digitaltv.html

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Keen, P., & Mackintosh, R. (2001). *The freedom economy: Gaining the m-commerce edge in the era of the wireless Internet.* Berkeley, CA: Osborne/McGraw-Hill.

Kini, R. B., & Thanarithiporn, S. (2004). M-commerce and e-commerce in Thailand—A value space analysis. *International Journal of Mobile Communications, 2*(1), 22-37.

Parasuraman, A., Berry, L. L., & Zeithaml, V. A. (1988). SERVQUAL: A multiple-item scale for measuring customer perceptions of service quality. *Journal of Retailing, 64*(1), 12-40.

Pew. (2006). Retrieved from http://www.pewinternet.org

Rogers, E. M. (1995). *Diffusion of innovations.* New York: The Free Press.

Schifter, D. E., & Ajzen, I. (1985). Intention, perceived control, and weight loss: An application of the theory of planned behavior. *Journal of Personality and Social Psychology, 49*(3), 843-851.

Shim, J. P. (2005). Korea's lead in mobile cellular and DMB phone services. *Communications of the Association for Information Systems, 15,* 555-566.

Thanarithiporn, S. (2004). *A modified technology acceptance model for analyzing the determinants affecting initial and post intention to adopt mobile technology in Thailand*. Unpublished dissertation, Bangkok University, Thailand.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186-204.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 32-37, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.6
Evolution of Mobile Commerce Applications

George K. Lalopoulos
Hellenic Telecommunications Organization S.A. (OTE), Greece

Ioannis P. Chochliouros
Hellenic Telecommunications Organization S.A. (OTE), Greece

Anastasia S. Spiliopoulou-Chochliourou
Hellenic Telecommunications Organization S.A. (OTE), Greece

INTRODUCTION

The tremendous growth in mobile communications has affected our lives significantly. The mobile phone is now pervasive and used in virtually every sector of human activity—private, business, and government. Its usage is not restricted to making basic phone calls; instead, digital content, products, and services are offered. Among them, mobile commerce (m-commerce) holds a very important and promising position.

M-commerce can be defined as: using mobile technology to access the Internet through a wireless device such as a cell phone or a PDA (Personal Digital Assistant), in order to sell or buy items (products or services), conduct a transaction, and perform supply-chain or demand-chain functions (Adams, 2001).

Within the context of the present study, we shall examine widespread used and emerging m-commerce services, from early ones (i.e., SMS [Short Message Service]) to innovative (i.e., mobile banking and specific products offered by known suppliers). We shall also investigate some important factors for the development of m-commerce, as well as some existing risks. Particular emphasis is given to the issue of collaboration among the key-players for developing standardization, interoperability, and security, and for obtaining market penetration.

M-COMMERCE SERVICES AND COMMERCIAL PRODUCTS

M-commerce products and services involve a range of main players, including Telcos (telecommunications service providers), mobile operators, mobile handset manufacturers, financial institutions, suppliers, payment service providers, and customers. Each party has its own interests (e.g., Telcos and mobile operators are interested not only in selling network airtime, but also in becoming value-added services providers offering additional functionality; banks consider the adaptation of their financial services to mobile distribution channels). However, successful cooperation of the involved parties is the key to the development of m-commerce.

Today's most profitable m-commerce applications concern entertainment (e.g., SMS, EMS [Enhanced Message Service], MMS [Multimedia Message Service], ring tones, games, wallpaper, voting, gambling, etc.). However, new interactive applications such as mobile shopping, reservations and bookings, ticket purchases via mobile phones (for train and bus travel, cinemas and theaters, car parking, etc.), m-cash micro purchases (for vending machines, tollbooths, etc.), mobile generation, assignment and tracking of orders, mobile banking, brokering, insurance, and business applications (e.g., accessing corporate data) have emerged and are expected to evolve and achieve significant market penetration in the future. In addition, future m-commerce users are likely to view certain goods and services not only as m-commerce products, but also in terms of situations such as being lost or having a car break down, where they will be willing to pay more for specific services (e.g., location awareness, etc.).

Mobile banking (m-banking) is the implementation of banking and trading transactions using an Internet-enabled wireless device (e.g., mobile phones, PDAs, handheld computers, etc.) without physical presentation at a bank branch. It includes services such as balance inquiry, bill payment, transfer of funds, statement request, and so forth. However, there are some problems regarding future development and evolution of mobile banking services. Many consumers consider those services difficult to use and are not convinced about their safety, while financial institutions are probably waiting for a payoff from their earlier efforts to get people to bank using their personal computers and Internet connections (Charny, 2001). As a consequence, the growth of mobile banking has been relatively slow since the launch of the first m-banking products by European players in 1999 and 2000. Currently, the main objective of mobile banking is to be an additional channel with a marginal role in a broader multi-channel strategy. Nevertheless, these strategic purposes are expected to change with the development of new applications of the wireless communication market, especially in the financial sector.

Now we will examine some characteristic m-commerce products. Japan's NTT DoCoMo was the first mobile telephone service provider to offer m-commerce services by launching the i-mode service in 1999 (NTT DoCoMo, 2004, Ryan, 2000). Key i-mode features include always-on packet connections, NTT's billing users for microcharges on behalf of content providers, and user's open access to independent content sites.

T-Mobile has developed a suite of applications called Mobile Wallet and Ticketing in the City Guide (T-Mobile, 2003). The first is a mobile payment system designed for secure and comfortable shopping. T-Mobile customers in Germany already use this system via WAP (Wireless Application Protocol). The highlight of the service is that customers do not have to provide any sensitive data like payment or credit card information when they make mobile purchases. Instead, after logging-in using personal data such as name, address, and credit card or bank details, they receive a personal identification number (PIN). By entering this PIN, a user can make a purchase from participating retailers.

With the Ticketing in the City Guide application, T-Mobile demonstrates a special future mobile commerce scenario. Here, entrance tickets for events such as concerts or sporting events can be ordered using a UMTS (Universal Mobile Telecommunications System) handset and paid for via Mobile Wallet. The tickets are sent to the mobile telephone by SMS in the form of barcodes. The barcodes can be read using a scanner at the venue of the event and checked to confirm their validity; subsequently, a paper ticket can be printed using a connected printer.

Nokia offers mobile commerce solutions such as the Nokia Payment Solution and the Wallet applications (Nokia Press Releases, 2001). The first one networks consumers, merchants, financial institutions, content/service providers, and various clearing channels in order to enable the exchange of funds among these parties and to allow users to make online payments for digital content, goods, and services via the Internet, WAP, or SMS. It collects, manages, and clears payments initiated from mobile phones and other Web-enabled terminals through various payment methods like credit and debit cards, operator's pre-paid or post-paid systems, and a virtual purse, which is an integrated pre-paid account of Nokia's Payment Solution that can be used with specific applications (e.g., mobile games). The solution enables remote payments from mobile terminals (e.g., electronic bill payment and shopping, mobile games, ticketing, auctioning, music downloading, etc.) and local payments (e.g., vending machines, parking fees, etc.).

Wallet is a password-protected area in the phone where users can store personal information such as payment card details, user names, and passwords, and easily retrieve it to automatically fill in required fields while browsing on a mobile site.

FACTORS AND RISKS

The development of advanced m-commerce applications, in combination with the evolution of key infrastructure components such as always-on high-speed wireless data networks (e.g., 2.5G, 3G, etc.) and mobile phones with multi-functionality (e.g., built-in-camera, music player, etc.) is stimulating the growth of m-commerce. Other key drivers of m-commerce are ease-of-use, convenience, and anytime-anywhere availability. On the other hand, a customer's fear of fraud is a major barrier. The nature of m-commerce requires a degree of trust and cooperation among member nodes in networks that can be exploited by malicious entities to deny service, as well as collect confidential information and disseminate false information. Another obvious risk is loss or theft of mobile devices. Security, therefore, is absolutely necessary for the spreading of m-commerce transactions with two main enablers:

- A payment authentication to verify that the authorized user is making the transaction; and
- Wireless payment-processing systems that make it possible to use wireless phones as point-of-sale terminals.

These elements of security are fundamental in order to gain consumer trust.

Mobile phones can implement payment authentication through different solutions: single chip (authentication functionality and communication functionality integrated in one chip—SIM [Subscriber Identification Module]); dual chip (separate chips for authentication and communication); and dual slot (authentication function is built in a carrier card separate from the mobile device, and an external or internal card reader intermediates the communication of the card and the mobile device) (Zika, 2004).

Furthermore, several industry standards have been developed: WAP, WTLS (Wireless Transport Layer Security), WIM (Wireless Identity Module), and so forth. In particular, as far as authentication is concerned, many security companies have increased their development efforts in wireless security solutions such as Public Key Infrastructure (PKI), security software (Mobile PKI), digital signatures, digital certificates, and smart-card technology (Centeno, 2002). PKI works the same way in a wireless environment as it does in the wireline world, with more efficient usage of available resources (especially bandwidth and processing power) due to existing limitations of wireless technology. Smart-card technology allows network administrators to identify users positively and confirm a user's network access and privileges. Today, mobile consumers are using smart cards for a variety of activities ranging from buying groceries to purchasing movie tickets. These cards have made it easier for consumers to store information securely, and they are now being used in mobile banking, health care, telecommuting, and corporate network security. An example of a security mechanism is the Mobile 3-D Secure Specification developed by Visa International (Cellular Online, Visa Mobile, 2004; Visa International, 2003).

New advanced mobile devices have tracking abilities that can be used to deliver location-specific targeted advertisements or advanced services (e.g., directions for traveling, information about location of the nearest store, etc.). This additional convenience, however, has its risks due to its intrusive nature, since tracking technology may be seen as an invasion of privacy and a hindrance to an individual's ability to move freely (the "Big Brother" syndrome).

The existence of many different solutions for m-commerce leads to a need for standardization, which can result from market-based competition, voluntary cooperation, and coercive regulation.

Voluntary Cooperation

Some significant forums for the development of m-commerce are the following:

- **Mobile Payment Forum (http://www.mobilepaymentforum.org/):** A global, cross-industry organization aiming to develop a framework for secure, standardized, and authenticated mobile payment that encompasses remote and proximity transactions, as well as micro-payments. It also is taking a comprehensive approach to the mobile payments process and creating standards and best practice for every phase of a payment transaction, including the setup and configuration of the mobile payment devices, payment initiation, authentication, and completion of a transaction. Members include American Express, Master Card, Visa, Japan Card Bureau, Nokia, TIM, and so forth.
- **MeT—Mobile Electronic Transaction (http://www.mobiletransaction.org/):** It was founded to establish a common technology framework for secure mobile transactions, ensuring a consistent user experience independent of device, service, and networks, and building on existing industry security standards such as evolving WAP, WTLS, and local connectivity standards such as Bluetooth. Members include Ericsson, Motorola, Nokia, Siemens, Sony, Wells Fargo Bank, Verisign, Telia, and so forth.
- **Mobey Forum (http://www.mobeyforum.org/):** A financial industry-driven forum, whose mission is to encourage the use of mobile technology in financial services. Activities include consolidation of business and security requirements, evaluation of potential business models, technical solutions, and recommendations to key-players in order to speed up the implementation of

solutions. Members include ABN AMRO Bank, Deutsche Bank, Ericsson, Nokia, Siemens, Accenture, NCR, and so forth.

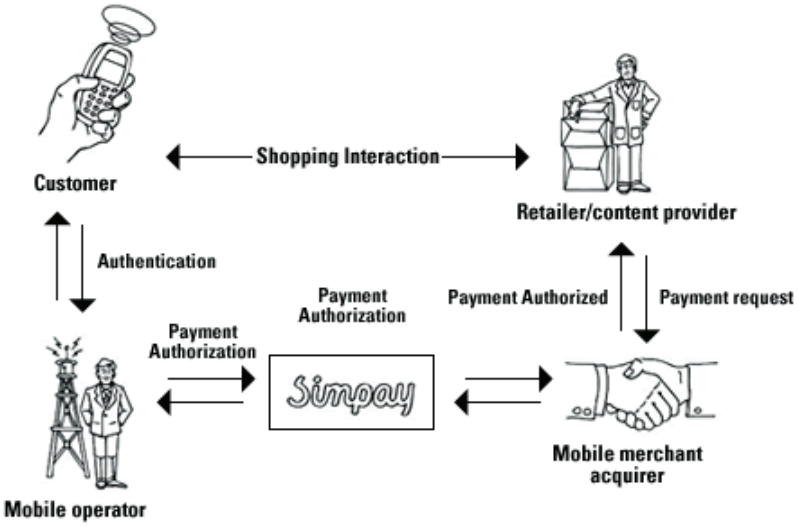
- **Open Mobile Alliance (OMA) (http://www.openmobilealliance.org/):** The mission of OMA is to deliver high-quality, open technical specifications based upon market requirements that drive modularity, extensibility, and consistency among enablers, in order to guide industry implementation efforts and provide interoperability across different devices, geographies, service providers, operators, and networks. Members include Bell Canada, British Telecommunications, Cisco Systems, NTT DoCoMo, Orange, Lucent Technologies, Microsoft Corporation, Nokia, and so forth.
- **Simpay (http://www.simpay.com/):** In order to facilitate mobile payments and deal with the lack of a single technical standard open to all carriers, four incumbent carriers (Orange, Telefonica Moviles, T-Mobile, and Vodafone) founded a consortium called Simpay (formerly known as Mobile Services Payment Association [MPSA]). Simpay was created to drive m-commerce through the development of an open and interoperable mobile payment solution, providing clearance and settlement services and a payment scheme that allow customers to make purchases through mobile-operator-managed accounts (see Figure 1).

The mobile merchant acquirer (MA), after signing an agreement with Simpay, aggregates merchants (e-commerce sites that sell goods or services to the customer [in Figure1, retailers/content providers]) by signing them up and integrating them with the scheme. Any industry player (i.e., mobile operators, financial institutions, portals, etc.) can become an MA, provided that they have passed the certification and agree on the terms and conditions contractually defined by Simpay.

Membership in Simpay includes mobile operators and other issuers of SIM cards such as service providers and Mobile Virtual Network Operators (MVNOs).

When the customer clicks the option to pay with Simpay, the mobile operator provides details of the transaction to the customer's mobile phone screen. The customer clicks to send confirmation. Simpay then routes the payment details (the payment request and the authorization) between the mobile operator (a Simpay member) and the

Figure 1. Simpay's mobile payment solution



merchant acquirer who, in turn, interacts with the merchant. Purchases will be charged to the customer's mobile phone bill or to a pre-paid account with the customer's particular operator.

The technical launch for Simpay was expected at the end of 2004 and the commercial one early in 2005 (Cellular Online, Simpay Mobile, 2004). At launch, Simpay would focus on micropayments of under 10 euros for digital content (e.g., java games, ringtones, logos, video clips, and MP3 files). Higher-priced items such as flights and cinema tickets with billing to credit or debit cards will follow.

- **Wireless Advertising Association (http://www.waaglobal.org/):** An independent body that evaluates and recommends standards for mobile marketing and advertising, documents advertising effectiveness, and educates the industry on effective and responsible methods. Members include AT&T Wireless, Terra Lycos, Nokia, AOL Mobile, and so forth.

Regulation

Directives from the authorities can boost consumer trust in m-commerce. This is the case in Japan, where regulators have set up standards for operators who wish to offer m-payment facilities to their users. The system also requires companies who allow for mobile payments to be registered with government regulations, so that consumers know they can get a refund if a service is not delivered as promised (Clark, 2003).

EU Directives

The European Commission has proposed some directives in an effort to harmonize regulatory practices of member countries. In September 2000, two directives on e-money were adopted: the ELMI Directive (Directive 46/EC, 2000) of the European Parliament and the Council of 18 September 2000 on the taking up, pursuit of, and prudential supervision of the business of electronic money institutions; and Directive 28/EC, 2000 of the European Parliament and the Council of 18 September 2000, amending Directive 12/EC, 2000 relating to the taking up and pursuit of the business of credit institutions.

The e-money Directives introduced a set of harmonized prudential rules that should be adopted by national regulators. By implementing these requirements, the national regulators would be allowed to authorize and supervise e-money issuers that could enter the whole market of the EU without the necessity of authorization in other countries (Zika, 2004). This strategy, however, might create some problems due to the wide disparity in implementation from country to country in the EU (e.g., e-money issuers in Italy have strict regulatory demands compared to the relatively laissez-faire attitude toward regulation of mobile transactions in Finland). Consequently, some EU members' mobile payments and related content services infrastructure could develop much more quickly than others, based solely on a country's legislative approach to implementation of supposedly standard Europe-wide legislation. Therefore, a balanced approach is needed in order to facilitate competition and to develop mobile business throughout Europe, toward smoothing the existing differences between different countries in the EU (EU Information Society Portal, 2003).

Moreover, under the umbrella of the e-Europe 2005 Action Plan, which is part of the strategy set out at the Lisbon European Council to modernize the European economy and to build a knowledge-based economy in Europe, a blueprint on mobile payments is under development (working document). This blueprint aims at providing a broadly supported approach that could give new momentum to industry-led initiatives and accelerate the large-scale deployment of sustainable mobile payment services, including pre-paid, post-paid, and online services, as well as payments at the point-of-sale (e-Europe Smart Card, 2003).

The EU Blueprint formally supports two objectives of the Action Plan eEurope 2005, which sets the scene for a coordinated European policy approach on information society issues:

- Interoperability
- Reduce barriers to broadband deployment (including 3G communications)

Issues like security and risk management, technical infrastructure, regulation and oversight of payment services provision, stimulation and protection of investments, and independence of mobile services providers from mobile networks are examined within the scope of the blueprint, which is expected to be endorsed by the main stakeholders (i.e., critical mass of market actors in both the financial and telecommunications sectors, as well as the relevant public authorities) by the end of 2005.

Regulation in the U.S.

The U.S. approach, in contrast to that of the EU, is based on a more relaxed view of e-money. From the beginning, the Federal Reserve (Fed) pointed out that early regulation might suppress innovation. This does not imply, however, that the regulatory interventions in the U.S. are minimal compared to the EU. In fact, besides the great number of regulatory and supervisory agencies applying a broad range of very confined rules, there also are many regulators at the state and federal level. Among them, the Uniform Money Services Act (UMSA) aims at creating a uniform legal framework in order to give non-banks the opportunity to comply with the various state laws when conducting business on a nationwide level. UMSA covers a wide range of financial (payment) services, not just e-money activities (Zika, 2004).

CONCLUSION

Mobile commerce (m-commerce) is seen as a means to facilitate certain human activities (i.e., entertainment, messaging, advertising, marketing, shopping, information acquisition, ticket purchasing, mobile banking, etc.), while offering new revenue opportunities to involved parties in the related supply chain (i.e., mobile operators, merchants/retailers, service providers, mobile handset manufacturers, financial institutions, etc.).

However, there are some barriers preventing m-commerce from taking off. They include lack of user trust in m-commerce technology, doubts about m-commerce security, and lack of widely accepted standards. As a consequence, the main income source for today's m-commerce services is the entertainment sector with low-price applications such as ringtones, wallpapers, games, lottery, horoscopes, and so forth.

With the advent of high-speed wireless networks (e.g., 2.5G, 3G, etc.) and the development of advanced applications such as mobile shopping, mobile ticketing, mobile banking, and so forth, m-commerce is expected to take off within the next three to five years.

The worldwide acceptance and use of standards such as Japan's i-mode and Europe's WAP, in combination with the work performed by market-based competition, collaboration of key-players, and regulations imposed by regulation authorities, are expected to boost consumer trust in m-commerce and strengthen its potential and perspectives.

REFERENCES

Adams, C. (2001). Mobile electronic payment systems: Main technologies and options. Retrieved August 9, 2004, from *http://www.bcs.org.uk/branches/hampshire/docs/mcommerce.ppt*

Cellular On-line. (2004). SIMPAY mobile payment platform announces first product. Retrieved August 11,2004, from *http://www.cellular.co.za/news_2004/feb/022704-simpay_mobile_payment_platform_a.htm*

Cellular On-line. (2004). Visa mobile 3D secure specification for m-commerce security. Retrieved August 10, 2004, from *http://www.cellular.co.za/technologies/mobile-3d/visa_mobile-3d.htm*

Centeno, C. (2002). Securing Internet payments: The potential of public key cryptography, public key infrastructure and digital signatures [*e*PSO background paper no.6]. Retrieved August 9, 2004, from *http://epso.jrc.es/backgrnd.html*

Charny, B. (2001). Nokia banks on mobile banking. *CNET News*. Retrieved August 9, 2004, from *http://news.com.com/2100-1033-276400.html?legacy=cnet&tag=mn_hd*

Clark, M. (2003). Government must regulate m-commerce. *Electric News Net*. Retrieved August 11, 2004, from *http://www.enn.ie/frontpage.news-9375556.html*

e-Europe Smart Card. (2003). Open smart card infrastructure for Europe, v2 , part 2-2: ePayments: Blueprint on mobile payments.*TB5 e/m Payment*. Retrieved August 12, 2004, from *http://www.eeurope-smartcards.org/Download/01-2-2.PDF*

EU Information Society Portal. (2004). e-Europe 2005, e-business. Retrieved August 12, 2004, from *http://europa.eu.int/information_society/eeurope/2005/all_about/mid_term_review/ebusiness/index_en.htm*

European Parliament (EP). (2000). Directive 2000/12/EC of the European Parliament and of the Council of 20 March 2000 relating to the taking up and pursuit of the business of credit institutions. *Official Journal, L 126*. Retrieved August 11, 2004, from *http://europa.eu.int/eur-lex/en*

European Parliament. (2000). Directive 2000/28/EC of the European Parliament and of the Council of 18 September 2000 amending Directive 2000/12/EC relating to the taking up and pursuit of the business of credit institutions. *Official Journal, L 275*. Retrieved August 11, 2004, from *http://europa.eu.int/eur-lex/en*

European Parliament. (2000). Directive 2000/46/EC of the European Parliament and of the Council of 18 September 2000 on the taking up, pursuit and prudential supervision of the business of electronic money institutions. *Official Journal, L 275*. Retrieved August 11,2004, from *http://europa.eu.int/eur-lex/en*

Nokia Press Releases. (2001). Nokia payment sSolution enables mobile e-commerce services with multiple payment methods and enhanced security. Retrieved August 11,2004, from *http://press.nokia.com/PR/200102/809553_5.html*

NTT DoCoMo Web Site. (2004). I-mode. Retrieved August 10, 2004, from *http://www.nttdocomo.com/corebiz/imode*

Ryan, O. (2000). Japan's m-commerce boom. *BBC NEWS*. Retrieved August 10, 2004, from *http://news.bbc.co.uk/1/business/945051.stm*

T-Mobile Web Site. (2003). T-Mobile with CeBIT showcases on the subject of mobile commerce. Retrieved August 10, 2004, from *http://www.t-mobile-international.com/CDA/*

T-mobile_deutschland_newsdetails,1705,0,newsid-1787-yearid-1699-monthid-1755,en.html?w=736&h=435

Visa International Web Site. (2003). 3-D secure: System overview V.1.0.2 70015-01external version. Retrieved August 10, 2004, from *http://www.international.visa.com/fb/paytech/secure/pdfs/3DS_70015-01_System_Ove rview_external01_System_Overvie w_external_v1.0.2_May_2003.pdf*

Zika, J. (2004). Retail electronic money and prepaid payment instruments, thesis, Draft 1.4. Retrieved August 10, 2004, from *http://www.pay.czweb.org/en/Payment V1_4.pdf*

KEY TERMS

Bluetooth: A short-range radio technology aimed at simplifying communications among Internet devices and between devices and the Internet. It also aims to simplify data synchronization between Internet devices and other computers.

EMS: Enhanced Messaging Service. An application-level extension to SMS for cellular phones available on GSM, TDMA, and CDMA networks. An EMS-enabled mobile phone can send and receive messages that have special text formatting (i.e., bold or italic), animations, pictures, icons, sound effects, and special ringtones.

I-Mode: A wireless Internet service for mobile phones using HTTP, popular in Japan and increasingly elsewhere (i.e., USA, Germany, Belgium, France, Spain, Italy, Greece, Taiwan, etc.). It was inspired by WAP, which was developed in the U.S., and it was launched in 1999 in Japan. It became a runaway success because of its well-designed services and business model.

M-Commerce: Mobile commerce. Using mobile technology to access the Internet through a wireless device, such as a cell phone or a PDA, in order to sell or buy items (i.e., products or services), conduct a transaction, or perform supply chain or demand chain functions.

MMS: Multimedia Message Service. A store-and-forward method of transmitting graphics, video clips, sound files, and short text messages over wireless networks using the WAP protocol. It is based on multimedia messaging and is widely used in communication between mobile phones. It supports e-mail addressing without attachments.

MVNO: Mobile Virtual Network Operator. A company that does not own or control radio spectrum or associated radio infrastructure, but it does own and control its own subscriber base with the freedom to set tariffs and to provide enhanced value added services under its own brand.

PKI: Public Key Infrastructure. A system of digital certificates, certified authorities, and other registration authorities that verify and authenticate the validity of each party involved in an Internet transaction.

WAP: Wireless Application Protocol. A secure specification that allows users to access information instantly via handheld devices such as mobile phones, pagers, two-way radios, and so forth. It is supported by most wireless networks (i.e., GSM, CDMA, TETRA, etc.). WAP supports HTML and XML.

This work was previously published in Electronic Commerce: Concepts, Methodologies, Tools, and Applications, edited by A. Becker, pp. 808-816, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.7
Context-Awareness in Mobile Commerce

Jun Sun
Texas A&M University, USA

Marshall Scott Poole
Texas A&M University, USA

INTRODUCTION

Advances in wireless network and multimedia technologies enable mobile commerce (m-commerce) information service providers to know the location and surroundings of mobile consumers through GPS-enabled and camera-embedded cell phones. Context awareness has great potential for creating new service modes and improving service quality in m-commerce. To develop and implement successful context-aware applications in m-commerce, it is critical to understand the concept of the "context" of mobile consumers and how to access and utilize contextual information in an appropriate way. This article dissects the context construct along both the behavioral and physical dimensions from the perspective of mobile consumers, developing a classification scheme for various types of consumer contexts. Based on this classification scheme, it discusses three types of context-aware applications—non-interactive mode, interactive mode and community mode—and describes newly proposed applications as examples of each.

UTILIZING CONSUMER CONTEXT: OPPORTUNITY AND CHALLENGE

M-commerce gets its name from consumers' usage of wireless handheld devices, such as cell phones or PDAs, rather than PCs as in traditional e-commerce (Mennecke & Strader, 2003). Unlike e-commerce users, m-commerce users enjoy a pervasive and ubiquitous computing environment (Lyttinen & Yoo, 2002), and therefore can be called "mobile consumers."

A new generation of wireless handheld devices is embedded or can be connected with GPS receivers, digital cameras and other wearable sensors. Through wireless networks, mobile consumers can share information about their location, surroundings and physiological conditions with m-commerce service providers. Such

information is useful in context-aware computing, which employs the collection and utilization of user context information to provide appropriate services to users (Dey, 2001; Moran & Dourish, 2001). The new multimedia framework standard, MPEG-21, describes how to adapt such digital items as user and environmental characteristics for universal multimedia access (MPEG Requirements Group, 2002). Wireless technology and multimedia standards give m-commerce great potential for creating new context-aware applications in m-commerce.

However, user context is a dynamic construct, and any given context has different meanings for different users (Greenberg, 2001). In m-commerce as well, consumer context takes on unique characteristics, due to the involvement of mobile consumers. To design and implement context-aware applications in m-commerce, it is critical to understand the nature of consumer context and the appropriate means of accessing and utilizing different types of contextual information. Also, such an understanding is essential for the identification and adaptation of context-related multimedia digital items in m-commerce.

CONSUMER CONTEXT AND ITS CLASSIFICATION

Dey, Abowd and Salber (2001) defined "context" in context-aware computing as "any information that can be used to characterize the situation of entities (i.e., whether a person, place or object) that are considered relevant to the interaction between a user and an application …" (p. 106). This definition makes it clear that context can be "any information," but it limits context to those things relevant to the behavior of users in interacting with applications.

Most well-known context-relevant theories, such as Situated Action Theory (Suchman, 1987) and Activity Theory (Nardi, 1997), agree that "user context" is a concept inseparable from the goals or motivations implicit in user behavior. For specific users, interacting with applications is the means to their goals rather than an end in itself. User context, therefore, should be defined based on typical user behavior that is identifiable with its motivation.

According to the Merriam-Webster Collegiate Dictionary, the basic meaning of context is "a setting in which something exists or occurs." Because the typical behavior of mobile consumers is consumer behavior, the user context in m-commerce, which we will term *consumer context*, is a setting in which various types of consumer behavior occur.

Need Context and Supply Context

Generally speaking, consumer behavior refers to how consumers acquire and consume goods and services (both informational and non-informational) to satisfy their needs (e.g., Soloman, 2002). Therefore, consumer behavior is, to a large extent, shaped by two basic factors: consumer needs and what is available to meet such needs. Correspondingly, consumer context can be classified conceptually into "need context" and "supply context." A *need context* is composed of stimuli that can potentially arouse a consumer's needs. A *supply context* is composed of resources that can potentially meet a consumer's needs.

This behavioral classification of consumer context is based on perceptions rather than actual physical states, because the same physical context can have different meanings for different consumers. Moreover, a contextual element can be in a consumer's need and supply contexts simultaneously. For example, the smell or sight of a restaurant may arouse a consumer's need for a meal, while the restaurant is part of the supply context. However, it is improper to infer what a consumer needs based on his or her supply context (see below). Therefore, this conceptual differentiation of consumer contexts is important for the implementation of context-aware applications in

m-commerce, which should be either need context-oriented or supply context-oriented.

The needs of a consumer at any moment are essential for determining how a context is relevant to the consumer. However, "consumer need" is both a multi-level construct and a personal issue. According to Maslow (1954), human need is a psychological construct composed of five levels: physiological, safety, social, ego and self-actualization. While it is feasible to infer some of the more basic needs of mobile consumers, including physiological and safety needs, based on relevant context information, it is almost impossible to infer other higher-level needs. Moreover, consumer need is a personal issue involving privacy concerns. Because context-aware computing should not violate the personal privacy of users by depriving them of control over their needs and priorities (Ackerman, Darrell & Weitzner, 2001), it is improper to infer a consumer's needs solely based on his or her supply context and provide services accordingly. It is for this reason that pushing supply context information to mobile consumers based on where they are is generally unacceptable to users.

When consumers experience emergency conditions, including medical emergencies and disastrous events, they typically need help from others. Necessary services are usually acceptable to consumers when their urgent "physiological" and "safety" needs can be correctly inferred based on relevant context information. Context-aware applications can stand alert for such need contexts of consumers and provide necessary services as soon as possible when any emergencies occur. Such context-awareness in m-commerce can be denoted as *need-context-awareness*.

Under normal conditions, context-aware applications should let consumers determine their own needs and how certain supply contexts are relevant. The elements of supply contexts, including various sites, facilities and events, usually locate or occur in certain functionally defined areas, such as shopping plazas, tourist parks, traffic systems, sports fields and so on. Information about such contextual elements in certain areas can be gathered from suppliers and/or consumers and stored in databases. *Supply-context-awareness*, therefore, concerns how to select, organize and deliver such information to mobile consumers based on their locations and needs.

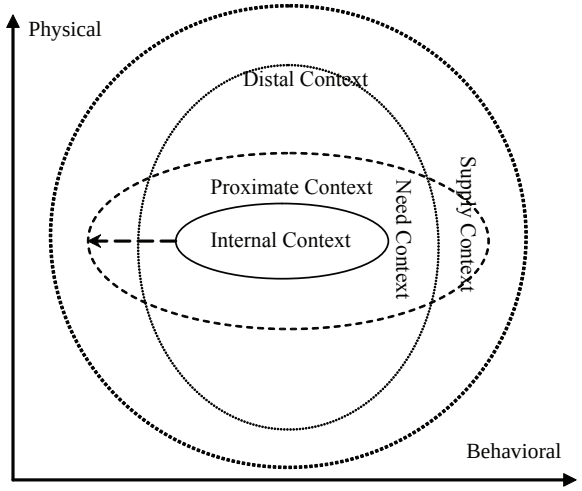
Internal Context, Proximate Context and Distal Context

Besides the behavioral classification, contextual elements can also be classified based on their physical locus. According to whether the contextual elements are within or outside the body of a consumer, a consumer context can be divided into internal and external contexts. An *internal context* is comprised of sensible body conditions that may influence a consumer's needs. By definition, internal context is part of need context. An *external context*, however, can refer to both the supply context and part of the need context that is outside of a consumer.

According to whether the contextual elements can be directly perceived by a consumer, his or her external context can be divided into "proximate context" and "distal context." A *proximate context* is that part of external context close enough to be directly perceivable to a consumer. A *distal context* is that part of external context outside the direct perception of a consumer. Mobile consumers do not need to be informed of their proximate context, but may be interested in information about their distal context. Context-aware information systems, which are able to retrieve the location-specific context information, can be a source of distal context information for mobile consumers. Besides, consumers can describe or even record information about their proximate context and share it with others through wireless network. To those who are not near the same locations, the information pertains to their distal contexts.

Figure 1 illustrates a classification scheme that combines two dimensions of consumer context,

Figure 1. A classification of consumer context



Note: ←– – indicates the range of direct perceivability.

physical and behavioral. The need context covers all the internal context and part of the external context. A subset of need context that can be utilized by need context-aware applications is *emergency context*; includes *internal emergency context,* which comprises urgent physiological conditions (e.g., abnormal heart rate, blood pressure and body temperature); and *external emergency context,* which emerges at the occurrence of natural and human disasters (e.g., tornado, fire and terrorist attacks). The supply context, however, is relatively more stable or predictable, and always external to a consumer. Supply context-aware applications mainly help mobile consumers obtain and share desirable supply context information. This classification scheme provides a guideline for the identification and adaptation of context-related multimedia digital items in m-commerce.

CONTEXT-AWARE APPLICATIONS IN M-COMMERCE

Context-aware applications in m-commerce are applications that obtain, utilize and/or exchange context information to provide informational and/or non-informational services to mobile consumers. They can be designed and implemented in various ways according to their orientation towards either need or supply context, and ways of collecting, handling and delivering context information.

It is generally agreed that location information of users is essential for context-aware computing (e.g., Grudin, 2001). Similarly, context-aware applications in m-commerce need the location information of mobile consumers to determine their external contexts and provide location-related services. Today's GPS receivers can be made very small, and they can be plugged or embedded into wireless handheld devices. Therefore, it is technically feasible for context-aware applications to acquire the location information of mobile consumers. However, it is not ethically appropriate to keep track of the location of consumers all of the time because of privacy concerns. Rather, consumers should be able to determine whether and/or when to release their location information except in emergency conditions.

There can be transmission of contextual information in either direction over wireless networks between the handheld devices of mobile consumers and information systems that host context-aware applications. For applications oriented towards the internal need context, there is at least the flow of physiological and location information from the consumer to the systems. Other context-aware applications typically intend to help mobile consumers get information about their distal contexts and usually involve information flow in both directions.

In this sense, mobile consumers who use context-aware applications are communicating with either information systems or other persons (usually users) through the mediation of systems. For user-system communications, it is commonly believed that the interactivity of applications is largely about whether they empower users to exert control on the content of information they can get from the systems (e.g., Jensen, 1998). Therefore,

the communications between a consumer and a context-aware system can be either non-interactive or interactive, depending on whether the consumer can actively specify and choose what context-related information they want to obtain. Accordingly, there are two modes of context-aware applications that involve communication between mobile consumers and information systems: the non-interactive mode and the interactive mode. For user-user communications, context-aware applications mediate the exchange of contextual information among mobile consumers. This represents a third mode: the community mode. This classification of context-aware applications into non-interactive, interactive and community modes is consistent with Bellotti and Edwards' (2001) classification of context awareness into responsiveness to environment, responsiveness to people and responsiveness to the interpersonal. Below, we will discuss these modes and give an example application for each.

Non-Interactive Mode

Successful context-aware applications in m-commerce must cater to the actual needs of mobile consumers. The non-interactive mode of context-aware applications in m-commerce is oriented toward the need context of consumers: It makes assumptions about the needs that mobile consumers have in certain contexts and provides services accordingly. As mentioned above, the only contexts in which it is appropriate to assess consumer needs are certain emergency conditions. We can call non-interactive context-aware applications that provide necessary services in response to emergency contexts Wireless Emergency Services (WES). Corresponding to the internal and external emergency contexts of mobile consumers, there are two types of WES: Personal WES and Public WES.

Personal WES are applications that provide emergency services (usually medical) in response to the internal emergency contexts of mobile consumers. Such applications use bodily attached sensors (e.g., wristwatch-like sensors) to keep track of certain physiological conditions of service subscribers. Whenever a sensor detects anything abnormal, such as a seriously irregular heart rate, it will trigger the wearer's GPS-embedded cell phone to send both location information and relevant physiological information to a relevant emergency service. The emergency service will then send an ambulance to the location and medical personnel can prepare to administer first-aid procedure based on the physiological information and medical history of the patient. The connection between the sensor and cell phone can be established through some short-distance wireless data-communication technology, such as Bluetooth.

Public WES are applications that provide necessary services (mainly informational services) to mobile consumers in response to their external emergency contexts. Such applications stand on alert for any disastrous events in the coverage areas and detect external context information through various fixed or remote sensors or reports by people in affected areas. When a disaster occurs (e.g., tornado), the Public WES systems gather the location information from the GPS-embedded cell phones of those nearby through the local transceivers. Based on user location and disaster information, the systems then give alarms to those involved (e.g., "There are tornado activities within one mile!") and display detailed self-help information, such as evacuation routes and nearby shelters, on their cell phones.

Interactive Mode

The interactive mode of context-aware applications in m-commerce does not infer consumer needs based on contextual information, but lets consumers express their particular information requirements regarding what they need. Therefore, the interactive mode is not oriented towards the need contexts of consumers, but their supply

contexts. The Information Requirement Elicitation (IRE) proposed by Sun (2003) is such an interactive context-aware application.

In the IRE approach, mobile consumers can express their needs by clicking the links on their wireless handheld devices, such as "restaurants" and "directions," that they have pre-selected from a services inventory. Based on such requests, IRE-enabled systems obtain the relevant supply context information of the consumers, and elicit their information requirements with adaptive choice prompts (e.g., food types and transportation modes available). A choice prompt is generated based on the need expressed by a consumer, the supply context and the choice the consumer has made for the previous prompt. When the information requirements of mobile consumers are elicited to the level of specific suppliers they prefer, IRE-enabled systems give detailed supplier information, such as directions and order forms.

The IRE approach allows the consumers to specify which part of their distal supply context they want to know in detail through their interactions with information systems. It attempts to solve the problem of inconvenience in information search for mobile consumers, a key bottle neck in m-commerce. However, it requires consumers to have a clear notion of what they want.

Community Mode

The community mode of context-aware applications in m-commerce mediates contextual information exchange among a group of mobile consumers. Consumers can only share information about what is directly perceivable to them, their proximate contexts. However, the information shared about the proximate context may be interesting distal context information for others if it is relevant to their consumption needs or other interests. A group of mobile consumers in a functionally defined business area have a common supply context, and they may learn about it through sharing context information with each other. Some applications in DoCoMo in Japan have the potential to operate in the community mode.

Wireless Local Community (WLC) is an approach to facilitate the exchange of context information for a group of mobile consumers in a common supply context, such as a shopping plaza, tourist park or sports field (Sun & Poole, working paper). In such an area, mobile consumers with certain needs or interests can join a WLC to share information about their proximate supply contexts with each other (e.g., seeing a bear in a national park). Because the information shared by different consumers is about different parts of the bigger common supply context, the complementary contributions are likely to achieve an "informational synergy." Compared with the IRE approach, the WLC approach allows mobile consumers to obtain potentially useful or interesting context information without indicating what they want.

Table 1 illustrates the primary context orientations of three modes of context-aware applications. The need context-aware applications are usually non-interactive. Personal WES applications are

Table 1. Primary context orientations of context-aware applications

Physical / Behavioral	Internal Context	Proximate Context	Distal Context
Need Context	(Personal WES)	←Non-Interactive→	(Public WES)
Supply Context	N/A	Community (WLC)	Interactive (IRE)

oriented towards the internal need context of mobile consumers, while Public WES applications are oriented towards the external (especially distal) need context of mobile consumers. The supply context-aware applications should be either of the interactive mode or community mode. As an example of interactive mode applications, IRE systems help mobile consumers know the part of their distal supply context they are interested in through choice prompts. As an example of community mode applications, WLC enables mobile consumers to share their proximate supply context with each others.

CONCLUSION

The advance in multimedia standards and network technology endows m-commerce great potential in providing mobile consumers context-aware applications. An understanding of consumer context is necessary for the development of various context-aware applications, as well as the identification and adaptation of context-related multimedia digital items. This article defines dimensions of consumer context and differentiates three modes of context-aware applications in m-commerce: the non-interactive, interactive and community modes. While applications for the interactive and community modes are in rather short supply at present, all indications are that they will burgeon as m-commerce continues to develop. Example applications are given to stimulate the thoughts on developing new applications.

Further technical and behavioral issues must be addressed before the design, implementation and operation of context-aware applications in m-commerce. Such issues may include: network bandwidth and connection, digital elements compatibility, content presentation, privacy protection, interface design, service sustainability and so on. We hope that this article can enhance further discussions in this area.

REFERENCES

Ackerman, M., Darrell, T., & Weitzner, D.J. (2001). Privacy in context. *Human-Computer Interaction, 16,* 167-176.

Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: Human considerations in context-aware systems. *Human-Computer Interaction, 16,* 193-212.

Dey, A.K. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal,* (1), 4-7.

Dey, A.K., Abowd, G.D., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction, 16,* 97-166.

Greenberg, S. (2001). Context as a dynamic construct. *Human-Computer Interaction,* 16, 257-268.

Grudin, J. (2001). Desituating action: Digital representation of context. *Human-Computer Interaction, 16,* 269-286.

Jensen, J.F. (1998). Interactivity: tracing a new concept in media and communication studies. *Nordicom Review,* (1), 185-204.

Lyttinen, K., & Yoo, Y. (2002). Issues and challenges in ubiquitous computing. *Communication of the ACM,* (12), 63-65.

Maslow, A.H. (1954). *Motivation and personality.* New York: Harper & Row.

Mennecke, B.E., & Strader, T.J. (2002). *Mobile commerce: Technology, theory and applications.* Hershey, PA: Idea Group Publishing.

Moran, T.P., & Dourish, P. (2001). Introduction to this special issue on context-aware computing. *Human-Computer Interaction, 16,* 87-95.

MPEG Requirements Group. (2002). *MPEG-21 Overview.* ISO/MPEG N5231.

Nardi, B. (1997). *Context and consciousness: Activity theory and human computer interaction.* Cambridge. MA: MIT Press.

Solomon, M.R. (2002). *Consumer behaviour: buying, having, and being* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication.* Cambridge: University Press.

Sun, J. (2003). Information requirement elicitation in mobile commerce. *Communications of the ACM, 46*(12), 45-47.

Sun, J. & Poole, M.S. (working paper). Wireless local community in mobile commerce. Information & Operations Management, Texas A&M University.

KEY TERMS

Consumer Context: The setting in which certain consumer behaviour occurs. It can be can be classified conceptually into "need context" and "supply context," and physically into "internal context," "proximate context" and "distal context."

Distal Context: The physical scope of a consumer context that is outside the direct perception of the consumer. Most context-aware applications intend to help mobile consumers obtain useful and interesting information about their distal context.

Information Requirement Elicitation (IRE): An interactive mode of context-aware application that helps consumers specify their information requirements with adaptive choice prompts in order to obtain desired supply context information.

Internal Context: The physical scope of a consumer context comprised of sensible body conditions that may influence the consumer's physiological needs. Certain context-aware applications can use bodily-attached sensors to keep track of the internal context information of mobile consumers.

Need Context: The conceptual part of a consumer context composed of stimuli that can influence the consumer's needs. A subset of need context that can be utilized by need context-aware applications is emergency context, from which the applications can infer the physiological and safety needs of consumers and provide services accordingly.

Proximate Context: The physical scope of a consumer context that is external to the body of consumer but close enough to be directly sensible to the consumer. Mobile consumers can describe and even record the information about their proximate contexts and share it with others.

Supply Context: The conceptual part of a consumer context composed of resources that can potentially supply what the consumer needs. Supply context-aware applications mainly help consumers obtain interesting and useful supply context information regarding their consumption needs.

Wireless Emergency Service (WES): A non-interactive mode of context-aware applications that provide necessary services in response to emergency contexts. Corresponding to the internal and external need contexts of mobile consumers, there are two types of WES: personal WES and public WES.

Wireless Local Community (WLC): A community mode of context-aware applications that facilitate the exchange of context information for a group of mobile consumers in a common supply context.

This work was previously published in the Encyclopedia of Multimedia Technology and Networking, edited by M. Pagani, pp. 123-129, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.8
Context as a Necessity in Mobile Applications

Eleni Christopoulou
University of Patras & Ionian University, Greece

ABSTRACT

This chapter presents how the use of context can support user interaction in mobile applications. It argues that context in mobile applications can be used not only for locating users and providing them with suitable information, but also for supporting the system's selection of appropriate interaction techniques and providing users with a tool necessary for composing and creating their own mobile applications. Thus, the target of this chapter is to demonstrate that the use of context in mobile applications is a necessity. It will focus on the current trend of modeling devices, services and context in a formal way, like ontologies, and will present an ontology-based context model.

INTRODUCTION

The future of computer science was marked by Weiser's vision (Weiser, 1991), who introduced the term ubiquitous computing (ubicomp) by defining a technology that can be seamlessly integrated into the everyday environment and aid people in their everyday activities. A few years later, the European Union, aiming to promote "human-centered computing," presented the concept of ambient intelligence (AmI) (ISTAG, 2001), which involves a seamless environment of computing, advanced networking technology and specific interfaces. So, technology becomes embedded in everyday objects such as furniture, clothes, vehicles, roads, and smart materials, providing people with the tools and processes that are necessary in order to achieve a more relaxing interaction with their environment.

Several industry leaders, like Philips and Microsoft, have turned to the design of ubicomp applications with a focus on smart home applications. However, people nowadays are constantly on the move, travel a lot, and choose to live in remote or mobile environments. In the near future, each person will be "continually interacting with hundreds of nearby wirelessly connected computers" (Weiser, 1993). Therefore, the need for mobile applications is now more evident than ever.

Recent years have seen a great breakthrough occur in the appearance of mobile phones. Initially they were used as simple telephone devices. Today,

mobiles have evolved into much more than that. Although the majority of people still use mobile phones as communication devices, an increasing number of users have begun to appreciate their potential as information devices. People use their smart mobile phones to view their e-mails, watch the news, browse the Web, and so forth. Eventually, mobile phones and other mobile handheld devices became an integral part of our daily routine.

Both scientists and designers of ubicomp applications have realized that the mobile phone could be considered as one of the first AmI artefacts to appear. As mobile phones are becoming more powerful and smarter this fact is increasingly proven true. Thus, scientists wanting to take advantage of the emerging technology have implemented a great number of mobile applications that enable human-computer interaction through the use of handheld devices like mobile phones or personal digital assistants (PDAs). Such applications include visitor guides for cities and museums, car navigation systems, assistant systems for conference participants, shopping assistants and even wearable applications.

A closer examination of mobile applications shows that most of them are location-aware systems. Specifically, tourist guides are based on users' location in order to supply more information on the city attraction closer to them or the museum exhibit they are seeing. Nevertheless, recent years have seen many mobile applications trying to exploit information that characterizes the current situation of users, places and objects in order to improve the services provided. Thus, context-aware mobile applications have come to light.

Even though significant efforts have been devoted to research methods and models for capturing, representing, interpreting, and exploiting context information, we are still not close to enabling an implicit and intuitive awareness of context, nor efficient adaptation to behavior at the standards of human communication practice. Most of the current context-aware systems have been built in an ad-hoc approach, deeply affected by the underlying technology infrastructure utilized to capture the context (Dey, 2001). To ease the development of context-aware ubicomp and mobile applications it is necessary to provide universal models and mechanisms to manage context.

Designing interactions among users and devices, as well as among devices themselves, is critical in mobile applications. Multiplicity of devices and services calls for systems that can provide various interaction techniques and the ability to switch to the most suitable one according to the user's needs and desires. Context information can be a decisive factor in mobile applications in terms of selecting the appropriate interaction technique.

Another inadequacy of current mobile systems is that they are not efficiently adaptable to the user's needs. The majority of ubicomp and mobile applications try to incorporate the users' profile and desires into the system's infrastructure either manually or automatically observing their habits and history. According to our perspective, the key point is to give them the ability to create their own mobile applications instead of just customizing the ones provided.

The target of this chapter is to present the use of context in context-aware ubicomp and mobile applications and to focus on the current trend of modeling devices, services and context in a formal way (like ontologies). Our main objective is to show that context in mobile applications can be used not only for locating users and providing them with suitable information, but also for supporting the system's selection of appropriate interaction techniques and for providing them with a tool necessary for composing and creating their own mobile applications.

In the background section, which follows, we define the term context and present how context is modeled and used in various mobile applications focusing on ontology-based context models. In the subsequent sections we present our perspective of context, an ontology-based context model for

mobile applications as well as the way in which human-computer interaction can be supported by the use of context. The Future section embraces our ideas of what the future of human-computer interaction in mobile applications can bring by taking context into account. Finally we conclude with some prominent remarks.

BACKGROUND

What is Context

The term "context-aware" was first introduced by Schilit and Theimer (1994), who defined context as "the location and identities of nearby people and objects, and changes to those objects." Schilit, Adams, and Want (1994) defined context as "the constantly changing execution environment" and they classified context into computing environment, user environment, and physical environment. Schmidt (2000) also considered situational context, such as the location or the state of a device, and defined context as knowledge about the state of the user and device, including surroundings, situation and tasks and pointing out the fact that context is more than location.

An interesting theoretical framework has been proposed by Dix et al. (2000), regarding the notions of space and location as constituent aspects of context. According to this framework context is decomposed into four dimensions, which complement and interact with each other. These dimensions are: system, infrastructure, domain, and physical context.

One of the most complete definitions for context was given by Dey and Abowd (2000); according to them context is "any information that can be used to characterize the situation of an entity. An entity should be treated as anything relevant to the interaction between a user and an application, such as a person, a place, or an object, including the user and the application themselves."

When studying the evolution of the term "context" one notices that the meaning of the term has changed following the advances in context-aware applications and the accumulation of experience in them. Initially the term "context" was equivalent to the location and identity of users and objects. Very soon, though, the term expanded to include a more refined view of the environment assuming either three major components; computing, user and physical environment, or four major dimensions; system, infrastructure, domain, and physical context. The term did not include the concept of interaction between a user and an application until Dey and Abowd (2000). This definition is probably at present the most dominant one in the area.

Context Modeling in Context-Aware Applications

A number of informal and formal context models have been proposed in various systems; the survey of context models presented in Strang and Linnhoff-Popien (2004) classifies them by the scheme of data structures. In Partridge, Begole and Bellotti (2005) the three types of contextual models, which are evaluated, are environmental, personal, and group contextual model.

Among systems with informal context models, Context Toolkit (Dey, Salber & Abowd, 2001) represents context in the form of attribute-value tuples, and Cooltown (Kindberg et al., 2002) proposed a Web-based model for context in which each object has a corresponding Web description. Both ER and UML models are used for the representation of formal context models in Henricksen, Indulska, and Rakotonirainy (2002).The context modeling language is used in Henricksen and Indulska (2006) in order to capture user activities, associations between users and communication channels and devices and locations of users and devices.

Truong, Abowd and Brotherton (2001) point out that the minimal set of issues required to be

addressed when designing and using applications are: who the users are, what is captured and accessed, when and where it occurs, and how this is performed. Designers of mobile applications should also take these issues into account. Similar to this approach Jang, Ko and Woo (2005) proposed a unified model in XML that represents user-centric contextual information in terms of 5W1H (who, what, where, when, how, and why) and can enable sensor, user, and service to differently generate or exploit a defined 5W1H-semantic structure.

Given that ontologies are a promising instrument to specify concepts and their interrelations (Gruber, 1993; Uschold & Gruninger1996), they can provide a uniform way for specifying a context model's core concepts as well as an arbitrary amount of subconcepts and facts, altogether enabling contextual knowledge sharing and reuse in a Ubicomp system (De Bruijn, 2003). Ontologies are developed to provide a machine-processable semantics of information sources that can be communicated between different agents (software and humans). A commonly accepted definition of the term ontology was presented by Gruber (1993) and stated that "an ontology is a formal, explicit specification of a shared conceptualization." A "conceptualization" refers to an abstract model of some phenomenon in the world which identifies the relevant concepts of that phenomenon; "explicit" means that the type of concepts used and the constraints on their use are explicitly defined and "formal" refers to the fact that the ontology should be machine readable. Several research groups have presented ontology-based models of context and used them in ubicomp and mobile applications. We will proceed to briefly describe the most representative ones.

In the Smart Spaces framework GAIA (Ranganathan & Campbell, 2003) an infrastructure that supports the gathering of context information from different sensors and the delivery of appropriate context information to ubicomp applications is presented; context is represented as first-order predicates written in DAML+OIL. The context ontology language (Strang, Linnhoff-Popien & Frank, 2003) is based on the aspect-scale-context information model. Context information is attached to a particular aspect and scale and quality metadata are associated with information via quality properties. This contextual knowledge is evaluated using ontology reasoners, like F-Logic and OntoBroker.

Wang, Gu, Zhang et al. (2004) created an upper ontology, the CONON context ontology, which captures general features of basic contextual entities, a collection of domain specific ontologies and their features in each subdomain. An emerging and promising context modeling approach based on ontologies is the COBRA-ONT (Chen, Finin & Joshi, 2004). The CoBrA system provides a set of OWL ontologies developed for modeling physical locations, devices, temporal concepts, privacy requirements and several other kinds of objects within ubicomp environments.

Korpipää, Häkkilä, Kela et al. (2004) present a context ontology that consists of two parts: structures and vocabularies. Context ontology, with the enhanced vocabulary model, is utilized to offer scalable representation and easy navigation of context as well as action information in the user interface. A rule model is also used to allow systematic management and presentation of context-action rules in the user interface. The objective of this work is to achieve personalization in mobile device applications based on this context ontology.

Although each research group follows a different approach for using ontologies in modeling and managing context in ubicomp and mobile applications, it has been acknowledged by the majority of researchers (Biegel & Cahill, 2004; Dey et al., 2001; Ranganathan & Campbell, 2003) that it is a necessity to decouple the process of context acquisition and interpretation from its actual use, by introducing a consistent, reliable and secure context framework which can facilitate the development of context-aware applications.

Context Utilisation in Mobile Applications

In context-aware mobile applications location is the most commonly used variable in context recognition as it is relatively easy to detect. Thus, a lot of location-aware mobile systems have been designed, such as shopping assistants (Bohnengerger, Jameson, Kruger et al., 2002) and guides in a city (Davies, Cheverst, Mitchell et al., 2001) or campus area (Burrell, Gay, Kubo et al., 2002). Many location-aware mobile applications are used in museum environments; a survey is presented in (Raptis, Tselios & Avouris, 2005). In the survey of Chen and Kotz (2000) it is evident that most of the context-aware mobile systems are based on location, although some other variables of context like time, user's activity and proximity to other objects or users are taken into consideration.

User activity is much more difficult to identify than location, but some aspects of this activity can be detected by placing sensors in the environment. Advanced context-aware applications using activity context information have been put into practice for a specific smart environment (Abowd, Bobick, Essa et el., 2002). The concept of activity zones (Koile, Tollmar, Demirdjian et al., 2003) focuses on location, defines regions in which similar daily human activities take place, and attempts to extract users' activity information from their location.

Sensor data can be used to recognize the usage situation based on illumination, temperature, noise level, and device movements, as described for mobile phones in Gellersen, Schmidt and Beigl (2002) and PDA in Hinkley, Pierce, Sinclair et al.(2000), where it is suggested that contextual information can be used for ring tone settings and screen layout adaptation. The mobile device can observe the user's behavior and learn to adapt to a manner that is perceived to be useful at a certain location as was the case with the comMotion system (Marmasse & Schmandt, 2000).

Sadi and Maes (2005) propose a system that can make adaptive decisions based on the context of interaction in order to modulate the information presented to the user or to carry out semantic transformation on the data, like converting text to speech for an audio device. CASIS (Leong, Kobayashi, Koshizuka et al., 2005) is a natural language interface for controlling devices in intelligent environments that uses context in order to deal with ambiguity in speech recognition systems. In Häkkilä and Mäntyjärvi (2005) context information is used in order to improve collaboration in mobile communication by supplying relevant information to the cooperating parties, one being a mobile terminal user and the other either another person, group of people, or a mobile service provider.

Perils of Context-Awareness

The promise and purpose of context-awareness is to allow computing systems to take action autonomously; enable systems to sense the situation and act appropriately. Many researchers, though, are skeptical and concerned because of the problems that emerge from context-awareness.

A main issue regarding context-aware computing is the fear that control may be taken away from the user (Barkhuus & Dey 2003). Experience has shown that users are still hesitant to adopt context-aware systems, as their proactiveness is not always desired. Another aspect of this problem is that users often have difficulties when presented with adaptive interfaces.

Apart from control issues, privacy and security issues arise. The main parameters of context are user location and activity, which users consider as part of their privacy. Users are especially reluctant to exploit context-aware systems, when they know that private information may be disclosed to others (Christensen et al., 2006).

Even recent research projects suffer from difficulties in automated context fetching; in order to overcome this, the user is asked to provide context

manually. Studies have shown that users are not willing to do much in order to provide context and context that depends on manual user actions is probably unreliable (Christensen et al., 2006). Additionally, systems that ask from users to supply context fail, as this affects the user's experience and diminishes his benefit from the system.

Practice has shown that there is a gap between how people understand context and what systems consider as context. The environment in which people live and work is very complex; the ability to recognize the context and determine the appropriate action requires considerable intelligence. Skeptics (Erickson, 2002) believe that a context-aware system is not possible to decide with certainty which actions the user may want to be executed; as the human context is inaccessible to sensors, we cannot model it with certainty. They, also, argue whether a context-aware system can be developed to be so robust that it will rarely fail, as ambiguous and uncertain scenarios will always occur and even for simple operations exceptions may exist. A commonly applied solution is to add more and more rules to support the decision making process; unfortunately this may lead to large and complex systems that are difficult to understand and use.

An issue that several researchers bring forward (Bardram, Hansen, Mogensen et al., 2006) is that context-aware applications are based on context information that may be imperfect. The ambiguity over the context soundness arises due to the speed at which the context information changes and the accuracy and reliability of the producers of the context, like sensors.

It is a challenge for context-aware systems to handle context, that may be non accurate or ambiguous, in an appropriate manner. As Moran and Dourish (2001) stated, more information is not necessarily more helpful; context information is useful only when it can be usefully interpreted.

WHAT IS CONTEXT FOR MOBILE APPLICATIONS?

Considering the use of context in the mobile applications discussed in the background section, we may conclude that, for these applications, context is almost synonymous to location and, specifically, to user location. However, context is quite more than just that. In this section, we will present our perspective on the parameters of context that are necessary for mobile applications. In order to figure out these parameters we have to identify the concepts that constitute the environment in which mobile applications exist. The primary concepts are indubitably people, places, time, objects and physical environment.

A mobile application is context-aware if it uses context to provide relevant information to users or to enable services for them; relevancy depends on a user's current task and profile. The user context issue has been addressed by many researchers of context-awareness (Crowley, Coutaz, Rey et al., 2002; Schimdt, 2002). However, the key for context-aware mobile applications is to capture user activity and preferences. Apart from knowing who the users are and where they are, we need to identify what they are doing, when they are doing it, and which object they focus on. In the background section we mentioned that, until now, most mobile applications determine user activity by their location; it is apparent, however, that a more elaborate model is necessary for representing this activity. Stahl (2006) proposes a model that represents a user's goals, activities and actions; he suggests that the distinction between an activity and an action lies in the fact that an activity takes a time span, while actions occur instantaneously. The system can define user activity by taking into account various sensed parameters like location, time, and the object that they use. For example, when a user opens the front door he is thought to be either entering or leaving the house, when the bed is occupied

and the television is turned on he is watching a movie, but when the television is turned off he is probably sleeping. User preferences are also very important for context-aware mobile applications, but it is difficult for the system to define them. Users have to incorporate their preferences into the application on their own, although the system can also gather information from the interaction with them in order to acquire experience based on history. By exploiting system experience the application may also infer a user's mood, a factor that cannot be measured by any sensor.

In order to identify user location various technologies are being used. In outdoors applications, and depending on the mobile devices that are used, satellite supported technologies, like GPS, or network supported cell information, like GSM, IMTS, WLAN, are applied. Indoors applications use RFID, IrDA and Bluetooth technologies in order to estimate the users' position in space. Although location is the determining factor in identifying where users are, orientation is also a very important parameter; the system has to know what users are looking at or where they are going to. However, in order to efficiently exploit the information on user location and orientation, the mobile application needs to have a representation of the layout of the place in which users are. Spaces can be classified into the following types: public, private, an area in which restrictions may apply, transient, places where people do not congregate easily or frequently, like hallways and corridors, social, public places where people arrange to meet, like coffee shops, informative, places that are used for public announcements (Mitchell, Race & Suggitt, 2006). Additionally a space can also be divided into districts, for example a home may have a living room, kitchen and bedroom, while a museum could have ancient Greek, paintings and modern art sections, as well into zones, such as lower left, upper left, and so forth.

Time is another significant parameter of context as it can play an important role in order to extract information on user activity; for example if it is early in the morning and the front door is opening the user is probably leaving the house, not entering it. Time can be used in various forms such as hour (daytime), night, day, weekday, week, month, season and year.

The objects that are used in mobile applications are the most crucial context sources. In mobile applications the user can use mobile devices, like mobile phones and PDAs and objects that are enhanced with computing and communication abilities (AmI artefacts). Sensors attached to artefacts provide applications with information about what the user is utilizing. However, this is not the most important parameter of context sensed by the artefacts. In order to present the user with the requested information in the best possible form, the system has to know the physical properties of the artefact that will be used, for example the display size of the artefact is determinant for the modulation of information. Additionally, the types of interaction interfaces that an artefact provides to the user need to be modeled; the system has to know if an artefact can be handled by both speech and touch techniques or if a mobile phone can vibrate. Apart from the physical properties of an artefact, the system must know how it is designed. A table with only one weight sensor in the centre cannot provide to the application information on whether an object is at its edge; thus the system has to know the number of each artefact's sensors and their position in order to gradate context information with a level of certainty. Based on information on the artefact's physical properties and capabilities, the system can extract information on the services that they can provide to the user; this is considered to be the most crucial context information related to artefacts. The application has to know if a printer can print both black-and-white and color text or if it can supply free maps and guidelines to a user that is close enough to a city's info center.

Finally, context from the physical environment may include current weather conditions, illumination, noise level, overcrowding. Taking into

account the illumination of a room the application may decide to turn on an additional light when a user is reading a book or, if a user is in a noisy public space, the system may decide to vibrate his mobile phone when he has a call.

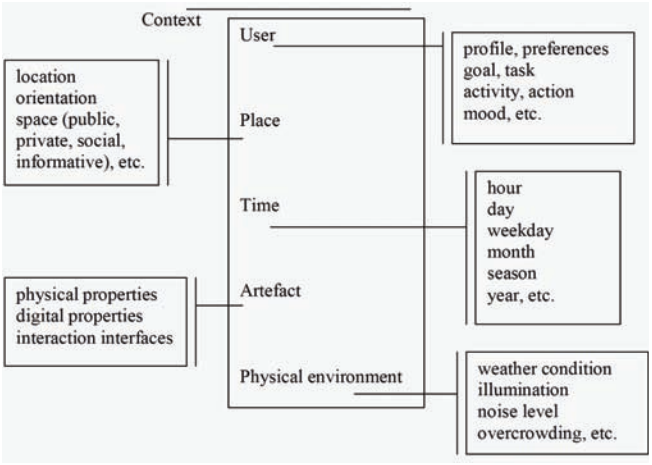
We selected to model the parameters of context illustrated in Figure 1 creating an ontology and taking into account the acknowledgement, shared by the majority of researchers (Biegel & Cahill, 2004; Dey et al., 2001; Ranganathan & Campbell, 2003), that it is a necessity to decouple the process of context acquisition and interpretation from its actual use. In the next section the details of this ontology-based context model are discussed.

AN ONTOLOGY-BASED CONTEXT MODEL FOR MOBILE APPLICATIONS

The key idea behind the proposed context model is that artefacts of AmI environments can be treated as components of a context-aware mobile application and users can compose such applications by creating associations between these components. In the proposed system, artefacts are considered as context providers. They allow users to access context in a high-level abstracted form and they inform other application's artefacts so that context can be used according to the application needs. Users are able to establish associations between the artefacts based on the context that they provide; keep in mind that services enabled by artefacts are provided as context. Thus defining the behavior of the application that they create, they can also denote their preferences, needs and desires to the system.

The set of sensors attached to an artefact measure various parameters such as location, time, temperature, proximity, motion, and so forth; the raw data given by its sensors is the artefact's low level context. As the output of different sensors that measure the same artefact parameter may differ, for example sensors may use different metric system, it is necessary to interpret the sensors' output into higher level context information. Aggregation of context is also possible meaning that semantically richer information may be derived based on the fusion of several measurements that come from different homogeneous or heterogeneous sensors. Thus, an artefact based on its own experience and use has two different levels of context; the low level which represents information acquired from its own sensors and the

Figure 1. Context in mobile applications



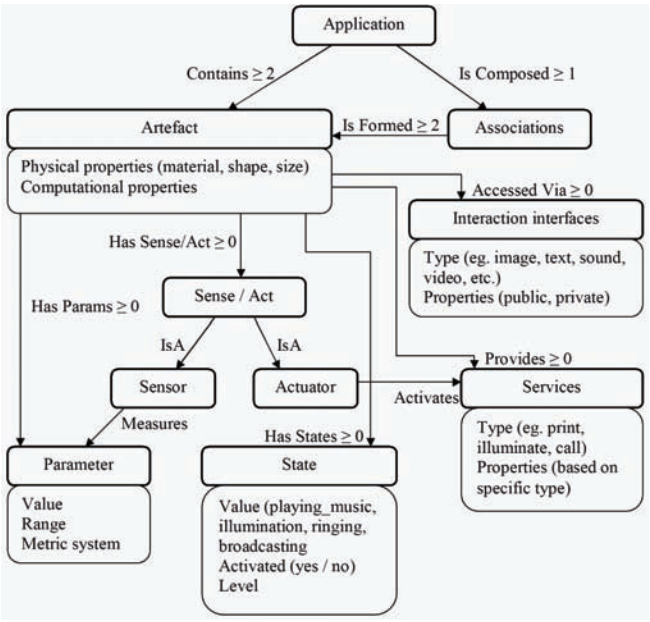
high level that is an interpretation of its low level context information. Additionally, an artefact can get context information from the other artefacts; this context can be considered as information from a "third-person experience."

When a user interacts and uses an artefact it affects its state; for example turning on the television sets it in a different state. An artefact may decide to activate a response based on both a user's desires and these states; for example when the user's PDA perceives that it is close to a specific painting in a museum, it will seek information about this painting. Such decisions may be based on the artefact's local context or may require context from other artefacts. The low and high level context, their interpretation and the local and global decision-making rules can be encoded in an ontology.

The ontology that we propose to represent the context of mobile applications is based on the GAS Ontology (Christopoulou & Kameas, 2005). This ontology is divided into two layers: a common one that contains the description of the basic concepts of context-aware applications and their inter-relations representing the common language among artefacts and a private one that represents an artefact's own description as well as the new "knowledge or experience" acquired from its use.

The common ontology, depicted in Figure 2, defines the basic concepts of a context-aware application; such an application consists of a number of artefacts and their associations. The concept of artefact is described by its physical properties and its communication and computational capabilities; the fact that an artefact has a number of sensors and actuators attached is also defined in our ontology. Through the sensors an artefact can perceive a set of parameters based on which the state of the artefact is defined; an artefact may also need these parameters in order to sense its interactions with other artefacts as well as with the user. Artefacts may provide various services to the environment, for example a printer provides the print service, a lamp provides illumination and a phone the call service; these services are activated

Figure 2. The common ontology



either by the user or by other artefacts using the actuators attached to artefacts. The interaction interfaces via which artefacts may be accessed are also defined in our ontology in order to enable the selection of the appropriate one.

We have decided that each parameter of context in our context-aware mobile applications, for example user, space, time and physical environment, is represented as an application's artefact. For instance, the notion of time is integrated into such applications only if a watch or a clock may provide this context as a service. The necessary information about the users that interact with such applications may be provided by the users' mobile phone or PDA. The services provided by such artefacts may be regarded as context; for instance the information that a thermometer provides is context related to the weather and we consider that the thermometer provides a temperature service. So, based on the concepts of context and their subcategories as presented in Figure 1, we have designed a service classification.

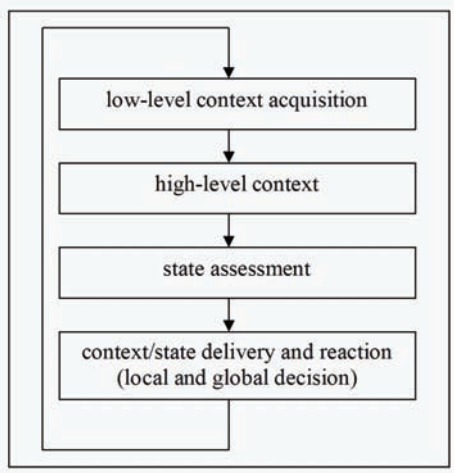
The common ontology represents an abstract form of the concepts represented, especially of the context parameters, as more detailed descriptions are stored into each artefact's private ontology. For instance, the private ontology of an artefact that represents a house contains a full description of the different areas in a house as well as their types and their relations.

The question that arises is where should these ontologies be stored? The system's infrastructure is responsible for answering this question. For a centralized system the common ontology as well as all the artefacts' ontologies can be stored in a central base. However, the majority of context-aware mobile applications are based on ad-hoc or p2p systems. Therefore, we propose that each artefact should store the common ontology and its private one itself; although when an artefact has limited memory resources its private ontology could be stored somewhere else. Another issue is where should place, time, environment and user ontologies be stored? The artefact that measures time, for example a clock, is responsible to store the time ontology; similarly there is an artefact for the environmental context. The place ontology can be stored either in a specific artefact that represents the space, for example an info kiosk in the entrance of a museum, or in the digital representation of the space managed by the application, for example the context from the sensors located in a room should be handled by the system through the use of the e-room and stored in an artefact with sufficient memory and computational capabilities. The user ontology in a similar way to the place ontology can be stored either in the user's mobile phone or in a digital self. These ontologies could also be stored in a web server in order to be accessible from artefacts.

The basic goal of the proposed ontology-based context model is to support a context management process, presented in Figure 3, based on a set of rules that determine the way in which a decision is made and are applied to existing knowledge represented by this ontology. The rules that can be applied during such a process belong to the following categories: rules for an artefact's state assessment that define the artefact's state based on its low and high level context, rules for local decisions which exploit an artefact's knowledge only in order to decide the artefact's reaction (like the request or the provision of a service) and finally rules for global decisions that take into account various artefacts' states and their possible reactions in order to preserve a global state defined by the user (Christopoulou, Goumopoulos & Kameas, 2005).

The ontology that is the core of the described context-management process was initially developed in the extrovert-Gadgets (eGadgets) project (http://www.extrovert-gadgets.net). In the e-Gadgets project our target was to design and develop an architectural framework (the Gadgetware Architectural Style—GAS) that would support the composition of ubicomp applications from everyday physical objects enhanced with sensing, acting, processing and communication

Figure 3. Context-management process



abilities. In this project we implemented the GAS Ontology (Christopoulou & Kameas, 2005), which served the purpose of describing the semantics of the basic concepts of a ubicomp environment and defining their inter-relations. The basic goal of this ontology was to provide a common language for the communication and collaboration among the heterogeneous devices that constitute these environments; it also supported a service discovery mechanism necessary for that ubicomp environment. Already, at this early stage, we had decided on issues like how this ontology would be stored in each artefact, by dividing it into two layers, and a module had been implemented, which was responsible for managing and updating this ontology.

This work evolved in the PLANTS project (http://plants.edenproject.com) that aimed to enable the development of synergistic, scalable mixed communities of communicating artefacts and plants (Goumopoulos, Christopoulou, Drossos et al., 2004). In this project we extended the concept of "context" in order to allow for the inclusion of plants as components of our ubicomp applications, by attaching sensors to them that provided information regarding the plants' state. The ontology that was inherited from the e-Gadgets project was extended and refined in order to include all the parameters of context that were identified as necessary for our applications. The ontology-based context model and the context-management process, presented in Figure 3, were defined at that stage. Experience showed that our system managed to decouple the process of context acquisition and interpretation from its actual use. Our context-management process is based on a set of rules that define the state of each artefact or plant in an application. Based on these rules and their state, each artefact determines its local decisions; the set of rules on various artefacts determine global decisions made by the whole application. These rules are defined by the users themselves via a graphical user interface. Each artefact stores its ontology as well as its rule base as defined by the user; the decision-making process, part of the context-management process, is supported by an inference engine. Experience has shown that users could easily define their own applications, denoting the rules that govern both each artefact and the whole application; the fact that the reasoning process permits user-defined rules that can be dynamically updated was another positive point. A drawback of our system is that the inference engine, which was used

required significant memory that was not always available; a workaround to this problem was to host the inference engine in an artefact with the required capabilities. Details on the design and implementation of this system as well as a case study of an application in the e-health domain and an evaluation of the outcome are presented in Christopoulou et al. (2005).

HOW CONTEXT CAN SUPPORT USER INTERACTION IN MOBILE APPLICATIONS

Recalling the use of context in mobile applications presented in the background section, we reach the conclusion that context has not been adequately exploited so far in order to support human-computer interaction. In this section we will present how our ontology-based context model enables the use of context in order to assist human-computer interaction in mobile applications and to achieve the selection of the appropriate interaction technique.

The goal of context in computing environments is to improve interaction between users and applications. This can be achieved by exploiting context, which works like implicit commands and enables applications to react to users or surroundings without the users' explicit commands (Schmidt, 2000). Context can also be used to interpret explicit acts, making interaction much more efficient. Thus, context-aware computing completely redefines the basic notions of interface and interaction.

The future of human computer interaction is going further than WIMP (Windows Icons Menus Pointing) interfaces. Jones and Marsden (2005) present various mobile interaction techniques that are trying to better exploit a user's capabilities like auditory (hearing) and haptic (touch and movement sensing) abilities as well as gestural skills, such as the expressive movements users can make with their hands or heads. More senses (vision, hearing, touch) and more means of expression (gestures, facial expression, eye movement and speech) are involved in human-computer interaction. A comparable analysis of mobile interaction techniques is presented in (Ballagas, Borchers, Rohs et al., 2006).

Rukzio et al. (2006) conclude from their experimental comparison of touching, pointing and scanning interaction techniques that users tend to switch to a specific physical mobile interaction technique dependent on location, activity and motivation; for example when a user is close enough to an artefact he prefers to touch it, otherwise he has no motivation for any physical effort. Thus, mobile systems have to provide multi-modal interfaces so that users can select the most suitable technique based on their context.

The ontology-based context model that we presented in the previous section captures the various interfaces provided by the application's artefacts in order to support and enable such selections. The application based on context can adapt to the information provided to the user; for example if a user tries to hear a message sent by his child on the mobile phone in a noisy environment the application may adjust the volume.

Similarly the context can determine the most appropriate interface when a service is enabled. Imagine that a user is in a meeting and an SMS is received by his mobile phone; even though he may have forgotten to enable the phone's silent profile, the application can select to enable the vibration interface instead of the auditory one based on the context about place and activity. Another example is the following: a user is with his children in a museum and he receives a high priority e-mail and the display of his PDA is too small for him to read the whole document that a colleague sent him; the application tries to identify a larger display to present the document based on proximate artefacts' context and taking into account environmental parameters, like whether there are other users close to it, and issues of

privacy and security, like whether the document is confidential.

This infrastructure could also be useful for people with special needs. Consider how useful a museum guide application could be if it can provide more auditory information or even a model that the user can touch when it identifies a user with impaired vision entering a gallery.

Another aspect of mobile applications is that they are used simultaneously by several users. The mobile application has to consider the number of users and their preferences and attempt to form groups of people with similar profiles and interests. The application can base its decisions on place context when many users exploit it. In a museum guide, it is easier to form groups of people with similar interests than in city guides. People in social places are more willing to share artefacts and services than in private spaces. In a home application the system can give priority to a father to print his last version of a work instead of first printing a child's painting, whereas in a work environment application it is arguable whether the boss should have greater priority.

An important issue in mobile applications is system failure because of device unavailability; a mobile phone may run out of battery or be out of range. The service classification represented in the proposed context-ontology can handle such situations, as it merely needs to identify another artefact that provides the same or similar services, therefore is abstracting the user from such problems.

Ubiquitous and mobile interfaces must be proactive in anticipating needs, while at the same time working as a spatial and contextual filter for information so that the user is not inundated with requests for attention (Brumitt, Meyers, Krumm et al., 2000). At the same time, ubiquitous interfaces must allow the user control over the interface (Abowd & Mynatt, 2000). Barkhuus and Dey (2003) presented an interesting case study on some hypothetical mobile phone services and have shown that users prefer proactive services to personalized ones. Providing proactive context aware services based on perceived user context is one of the major focuses of mobile and ubiquitous computing. However, proactive systems involving multiple smart artefacts often create complex problems if their behavior is not inline with user preferences and implicit understandings.

The ontology-based context model that we propose empowers users to compose their own personal mobile applications. In order to compose their applications they first have to select the artefacts that will participate and establish their associations. They set their own preferences by associating artefacts, denoting the sources of context that artefacts can exploit and defining the interpretation of this context trough rules in order to enable various services. As the context acquisition process is decoupled from the context management process, users are able to create their own mobile applications avoiding the problems emerging from the adaptation and customisation of applications like disorientation and system failures. A similar approach is presented in Zhang and Bruegge (2004).

Finally context can also assist designers to develop mobile applications and manage various interfaces and interaction techniques. Easiness is an important requirement for mobile applications; by using context according to our approach, designers are abstracted from the difficult task of context acquisition and have merely to define how context is exploited from various artefacts by defining simple rules. Our approach presents an infrastructure capable of handling, substituting and combining complex interfaces when necessary. The rules applied to the application's context and the reasoning process support the application's adaptation. The presented ontology-based context model is easily extended; new devices, new interfaces as well as novel interaction techniques can be exploited into a mobile application by simply defining their descriptions in the ontology.

FUTURE TRENDS

A crucial question that emerges is what the future of user interaction techniques and interfaces in mobile and ubicomp applications is. Aarts (2004) presented that the ultimate goal of user interaction in such applications is realizing "magic." Watching the movie Matilda (DeVito, 1996), a number of interaction techniques that designers try to integrate into mobile applications are presented as magic; eyes blinking can lead to opening or closing of the blinds, simple gesture movements may open or close the windows and pointing at specific devices switches them on and off.

However, can ubiquitous and mobile computing enable forms of magic? The answer is yes. As Scott (2005) mentions "by embedding computing, sensing and actuation into everyday objects and environments, it becomes feasible to provide new abilities to users, allowing them to exert levels of control and sensing in the physical world that were not previously possible." All superhuman or magic powers related to mobile applications are closely connected with context as defined in the previous sections. When users establish associations among artefacts define how artefacts should react on various context changes; a form of telekinesis is implemented as devices are ubiquitously controlled. Teleathesia can also be implemented using context; having associated their mobile phone with their house, users can be informed via their phone if someone is entering or leaving house by merging place's and family members' context. When a user drives back to home, this context information about the user's activity can be presented via a toy's display to his child who is playing waiting to go to the zoo; thus telepresence is enabled by context. Precognition and postcognition abilities can also be supported by exploiting context; from system experience and artefacts's knowledge important results from the past can be concluded, whereas precognition is also feasible if users have particularly incorporated information into the applications about future meetings, appointments, and so forth.

Magic is not applicable only to user interaction and interfaces in mobile applications. The artefacts that will be created may embody forms of magic. Consider the Weasley's clock in the Harry Potter book series (Rowling), it presents information about each member to the family based on their current activity and state. Context could enable the design and development of such artefacts.

Ontologies will play an important role in context representation for mobile applications as well as rule-based infrastructures and inference engines will be exploited for context reasoning in such applications. However a number of critical questions arise. For example, the location where ontologies are stored is still in dispute. Various infrastructures propose general ontologies centrally stored, whereas others prefer smaller and application-specific ontologies stored in distributed locations. Concerning the context-reasoning based on rule-based infrastructures, the issue that emerges is whether existing inference engines are suitable for mobile applications or need we turn our focus on different, more light-weight systems.

A research opportunity within the domain of this topic is how various interaction techniques and interfaces can be classified and represented into the ontology-based context model in order to provide a more effective selection of interaction techniques. During the previous years a number of markup languages were created in order to represent and describe interfaces; we believe that ontologies are the most suitable formal model for representing interfaces for mobile applications. Additionally, a formal model of interfaces described by an ontology may also assist the evaluation of interfaces used in mobile applications.

It is evident that the progress made in the last decade in the field of context-awareness in mobile systems is significant; however, certain critical issues remain open. Proactive mobile applications need to be certain for the context information

based on which they decide their reaction in order to be trusted by the users; furthermore, mobile applications are usually multi-user so privacy and security are crucial.

CONCLUSION

The objective of this chapter was to present how context can support user interaction in mobile applications. Context-aware applications exploit location information in order to deliver location-aware services; when a user is identified by the system, personalized and adaptive services are provided. Whenever the user activity can be determined, the infrastructure provides the user with a proactive system that transforms his environment to a smart one; when the environmental parameters can be exploited along with the activity the system can best adapt the conditions or select the most suitable interaction method and interface. More advanced scenarios of proactive systems can even accommodate for the failures of particular system components.

However, users are still hesitant to adopt context-aware systems. The major reason for this is the fear that control may be taken away from them (Barkhuus & Dey 2003). Also, the gap between human expectations and the abilities of context-aware systems is sometimes big, especially when systems must handle ambiguous and uncertain scenarios or when the context on which decisions are based is imperfect.

The ontology-based context model that we presented in a previous section offers the benefits that were described above. Additionally, it allows users to setup their own context-aware applications and define the way that artefacts react to changes, giving them at the same time the sense of retaining control over the system. The context-management process assesses the state of an artefact in a two step process; the low-level context may contain impure information that is refined in order to produce the high-level context. In our system the user is able to dynamically update the rules that define the environment; so he is capable of foreseeing possible exceptions.

ACKNOWLEDGMENT

I would like to deeply thank the various people who, during the several months in which this endeavor lasted, provided me with useful and helpful assistance.

As part of the research described in this chapter carried out in the e-Gadgets and PLANTS projects. I would like to thank all my fellow researchers in these projects; especially thank Achilles Kameas, Christos Goumopoulos, Irene Mavrommati, and all my colleagues in the DAISy team of the Research Unit 3 of the Research Academic Computer Technology Institute for their encouragement and patience throughout the duration of these projects.

I would like to thank the anonymous reviewers, who read an early (and rather preliminary) proposal of this chapter and provided me with helpful feedback and invaluable insights, as well as Joanna Lumsden, the editor of this book, for her personal invitation to me to contribute to this book and her support.

I would like to commend the interest and great job done by Dimitris Dadiotis and Ourania Stathopoulou, who reviewed and proofed this chapter.

Most important, to Dimitris, who put up with lost weekends and odd working hours.

REFERENCES

Aarts, E. (2004). Keynote speak. *Adaptive Hypermedia Conference 2004*, Eindhoven, Netherlands.

Abowd, G., & Mynatt, E. (2000). Charting past, present, and future research in ubiquitous com-

puting. *ACM Transactions on Computer-Human Interaction, 7*(1), 29-58.

Abowd, G., Bobick, A., Essa, I., Mynatt, E., & Rogers, W. (2002). The aware home: Developing technologies for successful aging. *Workshop held in conjunction with American Association of Artificial Intelligence (AAAI) Conference, Alberta, Canada*.

Ballagas, R., Borchers, J., Rohs, M., & Sheridan, J. G. (2006). The smart phone: A ubiquitous input device. *IEEE Pervasive Computing, 5(1),* 70-77.

Bardram, J., Hansen, T., Mogensen, M., & Soegaard, M. (2006). Experiences from real-world deployment of context-aware technologies in a hospital environment. *In proceedings of Ubicomp 2006* (pp. 369-386). Orange County, CA.

Barkhuus, L., & Dey, A. K. (2003). Is context-aware computing taking control away from the user? Three levels of interactivity examined. In *Proceedings of UbiComp 2003* (pp. 150-156). Springer.

Biegel, G., & Cahill, V. (2004, March 14-17). *A framework for developing mobile, context aware applications.* In 2nd IEEE Conference on Pervasive Computing and Communications. Orlando, FL

Bohnengerger, T., Jameson, A., Kruger, A., & Butz, A. (2002). User acceptance of a decision-theoretic location-aware shopping guide. In *Proceedings of the Intelligent User Interface 2002* (pp. 178-179). San Francisco: ACM Press

Brumitt, B., Meyers, B., Krumm, J., Kern, A., & Shafer, S. A. (2000). EasyLiving: Technologies for intelligent environments. *In proceedings of the 2nd international symposium on Handheld and Ubiquitous Computing* (pp.12-29). Bristol, UK.

Burrell, J., Gay, G. K., Kubo, K., & Farina, N. (2002). Context-aware computing: A test case. In *Proceedings of Ubicomp 2002* (pp. 1-15).

Chen, G., & Kotz, D. (2000). *A survey of context-aware mobile computing research*. (Tech. Rep. TR2000-381). Department of Computer Science, Dartmouth College.

Chen, H., Finin, T., & Joshi, A. (2004). An ontology for context aware pervasive computing environments. *Knowledge Engineering Review—Special Issue on Ontologies for Distributed Systems*. Cambridge: Cambridge University Press.

Christensen, J., Sussman, J, Levy, S., Bennett, W. E., Wolf, T. V., & Kellogg, W. A. (2006). Too much information. *ACM Queue, 4*(6).

Christopoulou, E., & Kameas, A. (2005). GAS Ontology: An ontology for collaboration among ubiquitous computing devices. *International Journal of Human-Computer Studies, 62*(5), 664-685.

Christopoulou, E., Goumopoulos, C., & Kameas, A. (2005). An ontology-based context management and reasoning process for UbiComp applications. *In proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies* (pp. 265-270). Grenoble, France.

Crowley, J. L., Coutaz, J., Rey, G., & Reignier, P. (2002). Perceptual Components for Context Aware Computing. In *the proceedings of UbiComp 2002.*

Davies, N., Cheverst, K., Mitchell, K., & Efrat, A. (2001). Using and determining location in a context-sensitive tour guide. *In IEEE Computer, 34*(8), 35-41.

De Bruijn, J. (2003). *Using ontologies — Enabling knowledge sharing and reuse on the semantic Web.* (Tech. Rep. DERI-2003-10-29). Digital Enterprise Research Institute (DERI), Austria.

DeVito, D. (Director). (1996). Matilda. *Sony and TriStar Pictures.*

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing, Special issue on Situated Interaction and Ubiquitous Computing, 5*(1), 4-7.

Dey, A.K., & Abowd, G.D. (2000). Towards a better understanding of context and context-awareness. *CHI 2000, Workshop on The What, Who, Where, When, Why and How of Context-awareness* (pp.1-6). ACM Press

Dey, A. K., Salber, D., & Abowd, G. D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction Journal, 16*(2-4), 97-166.

Dix, A., Rodden, T., Davies, N., Trevor, J., Friday, A., & Palfreyman, K. (2000). Exploiting space and location as a design framework for interactive mobile systems. *ACM Transactions on Computer-Human Interaction, 7*(3), 285-321.

Erickson, T. (2002). Some problems with the notion of context-aware computing. *Communications of the ACM, 45*(2), 102-104.

Gellersen, H.W., Schmidt, A., & Beigl, M. (2002). Multi-sensor context-awareness in mobile devices and smart artefacts. *Mobile Networks and Applications, 7,* 341-351.

Goumopoulos, C., Christopoulou, E., Drossos, N. & Kameas, A. (2004). The PLANTS System: Enabling Mixed Societies of Communicating Plants and Artefacts. In *proceedings of the 2nd European Symposium on Ambient Intelligence* (pp. 184-195). Eindhoven, the Netherlands.

Gruber, T. G. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5*(2), 199–220.

Häkkilä, J., & Mäntyjärvi, J. (2005). Collaboration in context-aware mobile phone applications. In *Proceedings of the 38th International Conference on System Sciences.* Hawaii

Henricksen, K., & Indulska, J. (2006). Developing context-aware pervasive computing applications: Models and approach. *Journal of Pervasive and Mobile Computing,* 2(1), 37-64.

Henricksen, K., Indulska, J., & Rakotonirainy, A. (2002). Modeling context information in pervasive computing systems. In F. Mattern & M. Naghshineh (Eds.), *Pervasive 2002* (pp. 167–180). Berlin: Springer Verlag.

Hinkley, K., Pierce, J., Sinclair, M., & Horvitz, E. (2000). Sensing techniques for mobile interaction. *In CHI Letters, 2*(2), 91-100.

IST Advisory Group (ISTAG). (2001). *Scenarios for Ambient Intelligence in 2010-full*. http://www.cordis.lu/ist/istag-reports.htm

Jang, S., Ko, E. J., & Woo, W. (2005). Unified context representing user-centric context: Who, where, when, what, how and why. In *proceedings of International Workshop ubiPCMM05.* Tokyo, Japan.

Jones, M., & Marsden, G. (2005). *Mobile interaction design.* John Wiley & Sons.

Kindberg, T., Barton, J., Morgan, J., Becker, G., Caswell, D., Debaty, P., Gopal, G., Frid, M., Krishnan, V., Morris, H., Schettino, J., Serra, B., & Spasojevic M. (2002). People, places, things: Web presence for the real world. *Mobile Networks and Applications, 7*(5), 365–376.

Korpipää, P., Häkkilä, J., Kela, J., Ronkainen, S., & Känsälä, I. (2004). Utilising context ontology in mobile device application personalisation. In *proceedings of the 3rd international conference on Mobile and ubiquitous multimedia* (pp.133–140).

Koile, K., Tollmar, K., Demirdjian, D., Shrobe, H., & Darrell, T. (2003). Activity zones for context-aware computing. In *proceedings of UbiComp 2003 conference* (pp. 90-106). Seattle, WA.

Leong, L. H., Kobayashi, S., Koshizuka, N., & Sakamura, K. (2005). CASIS: A context-aware speech interface system. In *Proceedings of the 10th international conference on Intelligent user interfaces* (pp.231-238). San Diego, CA.

Marmasse, N., & Schmandt, C. (2000). Location-aware information delivering with comMotion. In *Proceedings of HUC 2000* (pp.157-171). Springer-Verlag.

Mitchell, K., Race, N. J.P., & Suggitt, M. (2006). iCapture: Facilitating spontaneous user-interaction with pervasive displays using smart sevices. In *PERMID workshop at the Pervasive 2006*. Dublin, Ireland.

Moran, T. P., & Dourish, P. (2001). Introduction to this special issue on Context-Aware Computing. *Human Computer Interaction 16*(2-4), 1-8.

Partridge, K., Begole, J., & Bellotti, V. (2005, September 11). Evaluation of contextual models. In *Proceedings of the First Internaltional Workshop on Personalized Context Modeling and Management for UbiComp Applications*. Tokyo, Japan.

Ranganathan, A., & Campbell, R. (2003). An infrastructure for context-awareness based on first order logic. *Personal and Ubiquitous Computing*, 7(6), 353–364.

Raptis, D., Tselios, N. & Avouris, N. (2005). Context-based design of mobile applications for museums: a survey of existing practices. In *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices &Amp; Services. MobileHCI '05, 111* (pp. 153-160).. *ACM Press.*

Rowling, J. K. Harry Potter book series. *Bloomsbury Publishing Plc.*

Rukzio, E., Leichtenstern, K., Callaghan, V., Holleis, P., Schmidt, A., & Chin, J. (2006). An experimental comparison of physical mobile interaction techniques: Touching, pointing and scanning. In *proceedings of the 8th International Conference UbiComp 2006*, Orange County, CA.

Sadi, S. H., & Maes, P. (2005). xLink: Context management solution for commodity ubiquitous computing environments. In *proceedings of International Workshop ubiPCMM05.* Tokyo, Japan.

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. In *proceedings of the IEEE Workshop on Mobile Computing Systems and Applications* (pp.85-90). Santa Cruz, CA.

Schilit, B., & Theimer, M. (1994). Disseminating active map information to mobile hosts. *IEEE Network, 8*, 22-32.

Schmidt, A. (2000). Implicit human computer interaction through context. *Personal Technologies, 4*(2-3), 191-199.

Schmidt, A. (2002). *Ubiquitous computing—Computing in context.* Unpublished Ph.D. thesis, Department of Computer Science, Lancaster University, UK.

Scott, J. (2005). UbiComp: Becoming superhuman. *In the UbiPhysics 2005 workshop, Designing for physically integrated interaction.* Tokyo, Japan.

Stahl, C. (2006). Towards a notation for the modeling of user activities and interactions within intelligent environments. In *proceedings of the 3rd International Workshop on the Tangible Space Initiative (TSI 2006).* In Thomas Strang, Vinny Cahill, Aaron Quigley (Eds.), *Pervasive 2006 Workshop Proceedings* (pp. 441-452).

Strang, T., & Linnhoff-Popien, L. (2004). A context modeling survey. In *proceedings of the 1st International Workshop on Advanced Context Modelling, Reasoning And Management* (pp. 33-40). Nottingham, UK.

Strang, T., Linnhoff-Popien, L., & Frank, K. (2003). CoOL: A context ontology language to enable contextual interoperability. In *LNCS*

2893 Proceedings of 4th IFIP WG 6.1 International Conference on Distributed Applications and Interoperable Systems (pp. 236–247). Paris, France.

Truong, K. N., Abowd, G. D., & Brotherton, J. A. (2001). Who, what, when, where, how: Design issues of capture & access applications. In *proceedings of the International Conference: Ubiquitous Computing (UbiComp 2001)* (pp. 209-224). Atlanta, GA.

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods, and applications. *Knowledge Engineering Review, 11*(2), 93–155.

Wang, X. H., Gu, T., Zhang, D. Q., & Pung, H. K. (2004). Ontology based context modeling and reasoning using OWL. *Workshop on Context Modeling and Reasoning at IEEE International Conference on Pervasive Computing and Communication.* Orlando, FL.

Weiser, M. (1991). The computing for the 21st century. *Scientific American, 265*(3), 94-104.

Weiser, M. (1993). Some computer science issues in ubiquitous computing. *Communications of the ACM, 36*(7), 75-84.

Zhang, T., & Bruegge, B. (2004, August). *Empowering the user to build smart home applications.* Second International Conference on Smart homes and health Telematics. Singapore.

KEY TERMS

Ambient Intelligence (AmI): Implies that technology will become invisible, embedded in our natural surroundings, present whenever we need it, enabled by simple and effortless interactions, accessed through multimodal interfaces, adaptive to users and context and proactively acting.

Context: Any information that can be used to characterize the situation of entities (i.e., whether a person, place or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves.

Context-Aware Application: An application based on an infrastructure that captures context and on a set of rules that govern how the application should respond to context changes.

Mobile Computing: The ability to use technology in remote or mobile (non static) environments. This technology is based on the use of battery powered, portable, and wireless computing and communication devices, like smart mobile phones, wearable computers and personal digital assistants (PDAs).

Ontology: A formal, explicit specification of a shared conceptualisation. A tool that can conceptualise a world view by capturing general knowledge and providing basic notions and concepts for basic terms and their interrelations.

Ubiquitous Computing (Ubicomp): Technology that is seamlessly integrated into the environment and aids human in their everyday activities. The embedding computation into the environment and everyday objects will enable people to interact with information-processing devices more naturally and casually than they currently do, and in whatever locations or circumstances they find themselves.

This work was previously published in the Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 187-204, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.9
A Proposed Framework for Mobile Services Adoption:
A Review of Existing Theories, Extensions, and Future Research Directions

Indrit Troshani
University of Adelaide, Australia

Sally Rao Hill
University of Adelaide, Australia

ABSTRACT

Mobile services are touted to create a significant spectrum of business opportunities. Acceptance of these services by users is, therefore, of paramount importance. Consequently, a deeper insight is required to better understand the underlying motivations leading users to adopting mobile services. Further, enhanced understanding would also help designing service improvements and appropriate adoption strategies. Most of the existing theoretical acceptance models available originate from organisational contexts. As mobile services bring additional functional dimensions, such as hedonic or experiential aspects, using extant models for predicting mobile services acceptance by individuals may be inadequate. The aim of this chapter is to explore and critically assess the use of existing acceptance theories in the light of evolving mobile services. Constructs affecting adoption behaviour are discussed and relevant extensions are made which culminate with a framework for mobile services adoption. Managerial implications are explored and future research directions are also identified.

INTRODUCTION

Mobile technologies and services are touted to create a significant spectrum of business opportunities. According to the International Telecommunications Union (ITU) mobile phone penetration rates have increased significantly in many coun-

tries in Northern Europe (e.g., Sweden—98.05%, Denmark—88.72%, Norway—90.89%) (Knutsen, Constantiou, & Damsgaard, 2005). Similarly, Japan and Korea have consistently experienced very high diffusion rates of mobile devices and services (Carlsson, Hyvonen, Repo, & Walden, 2005; Funk, 2005). While experts predict that by 2010 online access via mobile channels is expected to reach 24% of homes in North America, 27% in Eastern Europe, and 33% in North-Western Europe (Hammond, 2001), the current penetration rate in many countries in the Western hemisphere and Asia-Pacific, including the U.S. and Australia lags behind the forerunners (Funk, 2005; Ishii, 2004; Massey, Khatri, & Ramesh, 2005). Given the difference between rapid growth rates in the adoption of mobile technologies and associated services in some countries and the relatively slow growth rates in others (Bina & Giaglis, 2005; Knutsen et al., 2005), it is important to identify the factors and predictors of further adoption and integrate them into a consolidated framework.

Mobile technology is enabled by the collective use of various communication infrastructure technologies and portable battery-powered devices. Examples of mobile devices include notebook computers, personal digital assistants (PDAs) and PocketPCs, mobile, "smart" and Web-enabled phones, and global positioning system (GPS) devices (Elliot & Phillips, 2004). There is a variety of communication infrastructure technologies that can enable these devices. Data networking technologies, such as GSM, GPRS, and 3G, are typically used for connecting mobile phones. WiFi (wireless fidelity) is used for connecting devices in a local area network (LAN). Mobile devices can be connected wirelessly to peripherals such as printers and headsets via the Bluetooth technology and virtual private networks (VPNs) enable secure access to private networks (Elliot & Phillips, 2004). Mobile devices are powered by mobile applications which deliver various services while enhancing flexibility, mobility, and efficiency for users within business and life domains. Despite the availability of technologically advanced mobile devices there is evidence that advanced mobile services which run on these have not been widely adopted (Carlsson et al., 2005; Khalifa & Cheng, 2002). The adoption of advanced mobile services is important for the mobile telecommunications industry because mobile services associated with technologically advanced devices constitute a massive source of potential revenue growth (Alahuhta, Ahola, & Hakala, 2005; Massey et al., 2005).

The adoption of advanced mobile technologies and services requires further research as most of the current technology acceptance models are based on research conducted in organisational contexts (Carlsson et al., 2005), and there has been only limited research from consumers' perspective (Lee, McGoldrick, Keeling, & Doherty, 2003). The features of mobile technologies and services, such as short message service (SMS), multimedia messaging service (MMS), e-mail, map, and location services, allow for single wireless devices, such as mobile phones, to be used seamlessly and pervasively across traditionally distinct spheres of life, such as work, home, or leisure, and with various levels of time commitment and self-ascribed roles (Dholakia & Dholakia, 2004). The interactions of these aspects are more intense than ever before (Knutsen et al., 2005). As mobile technologies and services add other functional dimensions, such as hedonic and/or experiential aspects (Kleijen, Wetzels, & de Ruyter, 2004; Mathwick, Malhotra, & Rigdon, 2001), applying extant theories outright to determine the acceptance and adoption by individual users may be questionable and inadequate (Knutsen et al., 2005).

Moreover, more research is called for in the adoption of mobile technologies because of the levels of complexity and diversity that may be encountered during their adoption. A number of factors contribute to this level of complexity and diversity. First, there is a strong relationship between the mobile devices and their users

because the former always carries the identity of the latter (Chae & Kim, 2003). As a result, spatial positioning and identification of users is easier in the mobile context than in the traditional innovation adoption (Figge, 2004). Second, most mobile devices have limited available resources including memory, processing power, and user interface, which have the potential to offset ubiquity benefits (Chae & Kim, 2003; Figge, 2004). Third, the lifecycle of mobile technologies is usually short, which increases adoption risks because new technologies become rapidly obsolete and may, therefore, need to be replaced by newer ones. During this process, a certain amount of consumer learning might be required before adopters can be confident and satisfied in using the mobile devices and services (Saaksjarvi, 2003). Again, this supports the argument that current models of technology acceptance may not be applied directly in predicting mobile adoption behaviour because they do not reflect the levels of complexity and diversity in the adoption of mobile technologies.

This chapter focuses on mobile phones and the associated services. Examples of mobile services include mobile e-mail, commercial SMS, and MMS services, downloads to portable devices, access to news through a mobile phone, mobile ticket reservations, mobile stock trading, as well other customised services which may be made available by mobile phone operators (Bina & Giaglis, 2005). Research shows that ownership of technologically advanced mobile phones is a main driver for advanced mobile services (Carlsson et al., 2005). Therefore, the adoption of mobile services should also be considered in the context and the technologies which enable them.

The aim of this chapter is to extend the existing models and to propose an integrated conceptual and parsimonious framework which explains adoption behaviour of users of mobile technologies and services. To accomplish this, we first provide an overview of recent developments of mobile technologies and services. Then, a critical assessment of existing acceptance models is made. Next, acceptance constructs and their relevance to mobile technologies and services are discussed. These constructs are then integrated into a new framework about mobile services adoption. In the last section, the implications of this model and future research directions are also discussed.

OVERVIEW OF MOBILE TECHNOLOGY EVOLUTION

In this section, an overview of the evolution of mobile phone technologies is provided. The recognition the evolution of these technologies is important because it puts the adoption constructs discussed later in the appropriate context. The diagram in Figure 1 summarises the evolution of the technologies.

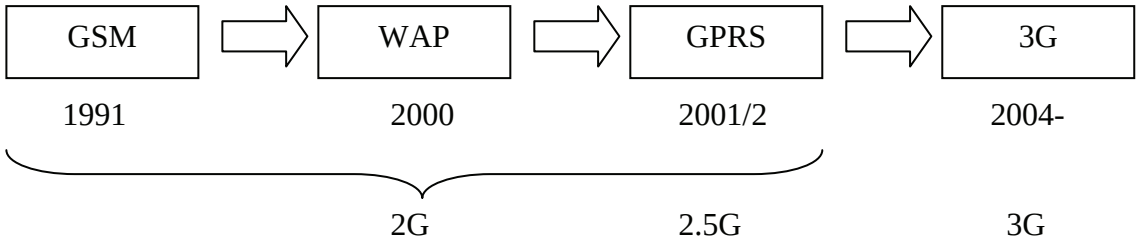
Second Generation Wireless Devices

The second generation of wireless devices (2G) introduced the digitisation of mobile communication and encompasses several standards which incrementally introduced new services and improved existing ones. It was a big leap forward from the first generation wireless communication (1G) which used analog standards and was characterised by poor quality and narrow bandwidth which resulted in limited adoption by both businesses and individuals (Elliot & Phillips, 2004). The commonly used standards by 2G are the Global System for Mobile Communications, the Wireless Applications Protocol, and the General Packet Radio Service. These are explained in more detail in the following sections.

Global System for Mobile Communication

Launched in the early 1990s, the Global System for Mobile Communications (GSM) constitutes the world's fastest growing and most popular mobile

Figure 1. Evolution of wireless technologies (Source: Carlsson et al., 2005)



telephony. Available through over 500 networks and serving almost a billion customers in 195 countries, GSM has been expanding exponentially (UMTS, 2003). Most countries, including underdeveloped and developing or even countries with a very low population density have at least two GSM network operators (Rossotto, Kerf, & Rohlfs, 2000). This has increased product and service offerings and competitiveness which has boosted GSM popularity even further.

One of the key advantages of the GSM technology is that it unified a range of different mobile communication standards into a single standard which constitutes a complete and open network architecture. This allows GSM-compatible mobile devices to be connected to any GSM network, therefore, enhancing interoperability. Further, GSM uses digital encoding which encrypts communications between a mobile phone and its base station, which makes interception more difficult. This results in improved security (Elliot & Phillips, 2004).

Another feature of the GSM technology is the automatic country-to-country communication, also known as global roaming. Because international travel for both business and pleasure has increased in recent years, roaming between mobile networks has become valuable as it generates as much as 15% of mobile operator's average revenue per user (ARPU) (UMTS, 2003). The subscriber identity module (SIM) card is another aspect of the GSM technology which is central to its popularity. The SIM card allows operators to manage information about their customers, including customer profile and billing, security access and authentication, virus intrusion and downloading capabilities (UMTS, 2003).

In addition to features such as caller ID, call forwarding, and call waiting, SMS emerged as the first unique mobile service and became the most popular mobile service after 1995 when adopters began using mobiles to send and receive limited amount of data in the form of short messages (Carlsson et al., 2005). While SMS later became the foundational platform for a variety of other services, it is considered to be cumbersome by many users because in addition to memorise service codes, users are also required to type text using the keypad of the mobile device (Carlsson et al., 2005).

Wireless Application Protocol

The Wireless Application Protocol (WAP) was introduced with the aim of providing advanced telephony and data access from the Internet using mobile terminals such as mobile phones, PDAs, smart phone, and other portable handheld devices (van Steenderen, 2002). With WAP, mobile devices can access Web sites specifically designed and built for them. WAP was, therefore, expected to provide the opportunity for connecting two of the fastest growing sectors of the telecommunications industry, namely, the Internet and mobile communications. As a matter of fact, the hype generated by WAP reached such dizzying heights

that from January to August 2000, the number of WAP-compatible Web pages increased from almost zero to 4.4 million (Teo & Pok, 2003).

The benefits of WAP were supposed to extend to several industries ranging from mobile operators, developers of WAP applications, manufacturers of mobile devices, and consumers in terms of various services, including banking, ticket reservations, entertainments, voice and fax mail notifications (Klasen, 2002; van Steenderen, 2002). Nevertheless, except for NTT DoCoMo's i-mode successful demonstration of mobile Internet, WAP has turned out to be a major disappointment with early adopters and other enthusiasts experiencing cognitive dissonance due to the relative oversell (Carlsson et al., 2005; Ratliff, 2002; Teo & Pok, 2003; Xylomenos & Polyzos, 2001).

Other challenges have also adversely affected widespread diffusion of WAP. Narrow bandwidth, low storage memory, and small screen limitations have resulted in slow communications; abridged Internet access has resulted in mediocre interfaces and almost no graphics. This has considerably limited Web site effectiveness (Klasen, 2002). By the end of 2000, only 12 million Europeans had WAP-compatible mobile devices, and of these, only 6% regularly used WAP functionality (Robins, 2003). Worldwide, only 10-15% of whom own WAP-compatible mobile devices would ever use WAP services, suggesting that "WAP had no future" (Klasen, 2002, p. 196). Nonetheless, the introduction of WAP constitutes a major step forward as it showed that Internet browsing is possible in mobile devices in general and phones in particular (Carlsson et al., 2005).

General Packet Radio Service

Simply known as the GPRS, the General Packet Radio Service constitutes an improvement over the GSM technology. GPRS uses packet-based data transfer mechanisms to provide continuous Internet accessibility (Elliot & Phillips, 2004; Hart & Hannan, 2004). With GPRS, users are not required to stay connected all the time in order to use a service. As a result, they are not charged on the basis of the connection time, rather, on the basis of the amount of downloaded data (Carlsson et al., 2005). Overall, GPRS is more efficient and cheaper than GSM, and yet, less widespread among users. Advancements associated with GPRS include the introduction of cameras, colour screens, multimedia messaging service (MMS), and video streaming (Carlsson et al., 2005). Because GPRS enhances 2G services, it is often referred to as the 2.5G technology (Elliot & Phillips, 2004).

Third Generation Wireless Devices

3G represents the next generation of mobile communication technologies, and it makes considerable improvements over its predecessors. These improvements include broad bandwidth which results in higher connection speeds, variety of multimedia capabilities and improved screen display, enhanced security features, and increased storage capacity (Elliot & Phillips, 2004). These enhancements enable users to receive digital photographs, moving video images, high quality sound in their mobile devices, and full unabridged e-mail and Internet access (Elliot & Phillips, 2004). Corporate users are also able to connect remotely to office computers and networks in order to access and download files quickly and easily (Robins, 2003). Because 3G technology mainly improves and enhances many existing services, it is considered to be an evolution rather than a revolution over the previous generation (Carlsson et al., 2005).

3G ensures that anybody, anywhere can access the same services (Grundström & Wilkinson, 2004). Further, 3G aims at integrating both the business and the social domains of the user's life which is the reason why 3G terminals are also referred to as "lifestyle portals" (Elliot & Phillips, 2004, p. 7). Another feature of the 3G technology is its capability to provide location-based services

(LBS) which could support health, transport, entertainment, data mining, and so forth (Casal, Burgelman, & Bohlin, 2004; UMTS, 2003). There is evidence that there is demand for such services which constitutes the main economic incentive for the development of the 3G technology (Alahuhta et al., 2005; Repo, Hyvonen, Pantzar, & Timonen, 2004).

A CRITICAL REVIEW OF THEORETICAL MODELS OF TECHNOLOGY ACCEPTANCE

A review of technology acceptance literature revealed many competing theoretical models, each with different focus and tested in different contexts. A significant amount of research effort has been put into building theories to examine how and why individuals adopt new information technologies and predict their level of adoption and acceptance. While one stream of research focuses on individual acceptance of technology (Compeau & Higgins, 1995; Davis, Bagozzi, & Warshaw, 1989), other streams have focused on implementation success at the organizational level (Leonard-Barton & Deschamps, 1988).

Many of the previously empirically researched models have been drawn from social psychology, for example, theory of reasoned action (TRA), motivational model, theory of planned behaviour (TPB), and sociology, for example, social cognitive theory (SCT) and innovation diffusion theory (IDT). Others specifically apply to technology adoption, for example, technology acceptance model (TAM). While each of these models made unique contributions to the literature on technology acceptance and adoption, most of these theoretical models theorise behaviour intention and/or usage as the key dependent variable in explaining acceptance of information technology because behavioural intentions are motivational factors that capture how hard people are willing to try to perform a behaviour (Ajzen, 1991). For example, TPB suggests that behavioural intention is the most influential predictor of behaviour; after all, a person does what s/he intends to do. In a meta-analysis of 87 studies, an average correlation of 0.53 was reported between intentions and behaviour (Sheppard, Hartwick, & Warshaw, 1988). As mobile services and underlying technologies are emerging information technologies, it is appropriate to consider this as the point of departure and use it to form the basis of a theoretical framework in mobile services and technology acceptance and adoption. The models that have been most frequently quoted in the technology acceptance and adoption literature are discussed next.

Theory of Reasoned Action (TRA)

Theory of reasoned action models are considered to be the most systematic and extensively applied approaches to attitude and behaviour research. According to TRA, the proximal determinant of a behaviour is a behavioural intention, which, in turn, is determined by attitude. These models propose that an individual's actual behaviour is determined by the person's intention to perform the behaviour, and this intention is influenced jointly by the individual's attitude and subjective norm. Attitude is defined as "a learned predisposition to respond in a consistently favourable or unfavourable manner with respect to a given object" (Fishbein & Ajzen, 1975, p. 6). A person's attitude towards a behaviour is largely determined by salient beliefs about the consequences of that behaviour and the evaluation of the desirability of the consequences (Fishbein & Ajzen, 1975). Subjective norm is defined as "the person's perception that most people who are important to him think he should or should not perform the behaviour in question" (Dillon & Morris, 1996). In brief, TRA asserts that attitude and subjective norm and their relative weights directly influence behavioural intention.

Theory of Planned Behaviour (TPB) and Decomposed Theory of Planned Behaviour

TPB, which generalizes TRA by adding a third construct—perceived behavioural control (Ajzen, 1991)—has been one of the most influential theories in explaining and predicting behaviour, and it has been shown to predict a wide range of behaviours (Sheppard et al., 1988). TPB asserts that the actual behaviour is determined directly both by behavioural intention and perceived behavioural control. Behavioural intention is formed by one's attitude, subjective norm, and perceived behavioural control (Ajzen, 1991). Further, a decomposed TPB includes constructs such as relative advantage, compatibility, influence of significant others, and risk from the innovation diffusion literature, and decomposing the three perceptions in TPB into a variety of specific belief dimensions. This model offers several advantages over TPB and is considered more complete and management-relevant by focusing on specific factors that may influence adoption and usage (Teo & Pok, 2003).

Technology Acceptance Model (TAM)

TAM can be seen as an adaptation of the theory of reasoned action (TRA) and was developed to predict and explain individual system use in the workplace (Davis, 1989). This model further suggests that two beliefs—perceived usefulness and perceived ease of use—are instrumental in explaining the user's intentions of using a system. Perceived usefulness refers to the degree to which "a person believes that use of the system will enhance his or her performance" whereas perceived ease of use is the degree to which "a person believes that using the system will be free of effort". Simply put, a technology that is easy to use and is useful will lead to a positive attitude and intention towards using the technology.

The main advantage of this model over others is that the two related beliefs can generalize across different settings. Thus, some argue that it is the most robust, parsimonious, and influential model in explaining information technology adoption behaviour (Elliot & Loebbecke, 2000; Teo & Pok, 2003; Venkatesh, Morris, Davis, & Davis, 2003). Indeed, since its development, it has received extensive empirical support through validations, applications, and replications for its prediction power (Taylor & Todd, 1995, 1995a; Venkatesh & Morris, 2000a). A number of modified TAM models were proposed to suit new technologies including Internet and intranet (Agarwal & Prasad, 1998; Chau, 1996; Chau & Hu, 2001; Horton, Buck, Waterson, & Clegg, 2001). For example, TAM has been used to predict Internet purchasing behaviour (Gefen, Karahanna, & Straub, 2003; Kaufaris, 2002).

A major theoretical limitation of TAM is the "exclusion of the possibility of influence from institutional, social, and personal control factors" (Elliot & Loebbecke, 2000, p. 49). Thus the suitability of the model for predicting general individual acceptance needs to be re-assessed as the main TAM constructs do not fully reflect the specific influences of technological and usage-context factors that may alter user acceptance (King, Gurbaxani, Kraemer, McFarlan, Raman, & Yap, 1994; Taylor & Todd, 1995). In response to this, a number of modifications and changes to the original TAM models have been made. The most prominent of these is the unified theory of acceptance and use of technology (UTAUT), a unified model that integrates constructs across eight models (Venkatesh et al., 2003). UTAUT provides a refined view of how the determinants of intention and behaviour evolve over time and assumes that there are three direct determinants of intention to use (performance expectancy, effort expectancy, and social influence) and two direct determinants of usage behaviour (intention and facilitating conditions). However, both TAM and UTAUT have received criticisms with

the fundamental one being about the problems in applying these beyond the workplace and/or organisation for which originally created (Carlsson et al., 2005).

Motivational Theories

Motivation theories are rooted in psychological research to understand individuals' acceptance of information technology (Davis, Bagozzi, & Warshaw, 1992; Igbaria, Parasuraman, & Baroudi, 1996). These theories often distinguished extrinsic and intrinsic motivation. While extrinsic motivation refers to the performance of an activity in helping achieve valued outcomes, intrinsic motivation puts emphasis on the process of performing an activity (Calder & Staw, 1975; Deci & Ryan, 1985). For example, perceived usefulness is an extrinsic source of motivation (Davis et al., 1992) while perceived enjoyment (Davis et al., 1992), perceived fun (Igbaria et al., 1996), and perceived playfulness (Moon & Kim, 2001) can be considered intrinsic sources of motivation. Both sources of motivation affect usage intention and actual usage. Therefore, in addition to ease of use and usefulness, intrinsic motivators, such as playfulness, will also play an important role in increasing usability in a usage environment in which information technology applications are both used for work and play (Moon & Kim, 2001).

Innovation Diffusion Theory

The innovation diffusion theory is concerned with how innovations spread and consists of two closely related processes: the diffusion process and adoption process (Rogers, 1995). Diffusion is a macro process concerned with the spread of an innovation from its source to the public whereas the adoption process is a micro process that is focused on the stages individuals go through when deciding to accept or reject an innovation. Key elements in the entire process are the innovation's perceived characteristics, the individual's attitude and beliefs, and the communication received by individuals from their social environment. In relation to the factors pertaining to innovation, factors such as, relative advantage, complexity, trialability, observability, and compatibility, were considered important in influencing individual's acceptance of the innovation (Rogers, 1995).

TOWARDS AN ACCEPTANCE MODEL FOR MOBILE SERVICES

This section develops an acceptance model for mobile technology and services that may be empirically tested. This development begins with identifying the latent constructs in extant technology adoption literature. However, mobile services differ from traditional systems in that mobile services are ubiquitous, portable, and can be used to receive and disseminate personalised and localised information (Siau, Lim, & Shen, 2001; Teo & Pok, 2003). Thus, the models examined in the previous section and the constructs included in these models may not be applicable to mobile services adoption. In particular, we discuss the various antecedents of attitude towards mobile services and develop a new model based on the widely used TAM model to predict adoption of new mobile services.

User Predisposition

User predisposition refers to the internal factors of an individual user of mobile services. Personal differences strongly influence adoption. There is evidence that successful acceptance of innovations depends as much on individual adopter differences as on the innovation itself. Indeed, individual differences help identify segments of adopters who are more likely to adopt technology innovations than others, which in turn, helps providers address adopter needs more closely (Massey et al., 2005). Diffusion resources can also be used

more effectively and efficiently (Agarwal & Prasad, 1998). Early adopters, for example, can then act as opinion leaders or change agents to facilitate the diffusion of the technology further (Rogers, 1995). There are several dimensions used to capture individual differences, including personal innovativeness, perceived costs, demographic factors, psychographic profiles, and personality traits (Dabholkar & Bagozzi, 2002). In this chapter, we define user predisposition as the collection of factors including the individual's prior knowledge and experience of existing mobile services, compatibility, behavioural control, personal innovativeness, perceived enjoyment, and price sensitivity.

First, *prior knowledge* is essential for the comprehension of the technology and related services. According to Rogers (1995), knowledge occurs when a potential adopter learns about the existence of an innovation and gains some understanding concerning its functionality. Like other technologies, the mobile technology is comprised of both the hardware (i.e., the physical mobile device) and software domains (i.e., the applications consisting of the instructions to use the hardware as well as other information aspects) (Rogers, 1995). Thus knowledge from both hardware and software domains might be required for complete comprehension (Moreau, Lehmann, & Markman, 2001; Saaksjarvi, 2003). Prior knowledge consists of two components, namely, familiarity and expertise. For instance, the former constitutes the number of mobile services-related experiences accumulated by consumers over time, which includes exposure to advertising, information search, interaction with salespersons, and so on. The latter represents the ability to use the mobile services, and it includes beliefs about service attributes (i.e., cognitive structures) as well as decision rules for acting on those beliefs (i.e., cognitive processes) (Alba & Hutchinson, 1987). In any case, familiarity alone cannot capture the complexity of consumer knowledge (Alba & Hutchinson, 1987), which suggests the learning is required (Saaksjarvi, 2003).

With learning, consumers use the "familiar" component of existing knowledge as a means to understand and comprehend new phenomena in the innovation which is being adopted (Roehm & Sternthal, 2001). Specifically, existing knowledge in general and analogical learning in particular have been shown to be powerful and highly persuasive communication devices in acquiring in-depth understanding of innovation benefits and functionality (Moreau et al., 2001; Roehm & Sternthal, 2001; Yamauchi & Markman, 2000). An analogy compares and contrasts a known base innovation to an unknown target innovation. The base and the target share structural attributes, but are different in terms of surface attributes. A cellular phone versus a personal digital assistant (PDA) versus a "smart phone" are good examples. Research shows that "a message containing an analogy is better comprehended and is more persuasive when the recipient has expertise with regard to the base product [innovation]." (Roehm & Sternthal, 2001, p. 269). However, expertise alone is insufficient to ensure analogy persuasiveness. Substantial resources, training/usage instructions, and a positive mood are also required to facilitate learning (Roehm & Sternthal, 2001). However, while knowledge is important, by itself, it has limited usefulness, and therefore, "knowledge alone cannot determine the basis for adoption" (Rogers, 1995, p. 167) of a technology or service.

Adopters' previous positive or negative *experiences* with a technology or service can have a significant impact on their perceptions and attitudes towards that technology (Lee et al., 2003; Taylor & Todd, 1995a). Specifically, experience may influence adopters in forming positive or negative evaluations concerning innovations, which can boost or impair adoption of mobile technologies and services. Because of their greater clarity and certainty, direct prior experiences are likely to have a stronger impact

on perceptions and attitudes towards usage than indirect or incomplete evidence (i.e., pre-trial) (Knutsen et al., 2005; Lee et al., 2003).

The second variable within the user predisposition construct is *compatibility*. Rogers (1995) defines compatibility as the degree to which an innovation is perceived to be consistent with existing values of potential adopters. In general, high incompatibility will adversely affect potential adopters of an innovation, which decreases the likelihood of adoption (Saaksjarvi, 2003). In contrast, high compatibility is likely to increase adoption propensity. In the context of wireless devices, lifestyle compatibility is the extent to which adopters believe mobile devices and services can be integrated into their daily lives. For example, adopters' lifestyle in terms of degree of mobility is likely to have a strong impact on their decision to adopt the technology (Pagani, 2004; Teo & Pok, 2003). For example, a person who leads a busy lifestyle, and is employed in an information-intensive job, and is always on the move is more likely to adopt a wireless device and its associated services compared to a person who leads a sedentary lifestyle.

Third, perceived *behavioural control*, a dynamic and socio-cognitive concept, has attracted a lot of attention in adoption literature. Earlier work by Ajzen (1991) considered it as a uni-dimensional variable. More recent empirical findings suggest that perceived behavioural control has two distinct components: self-efficacy, which is an individual's judgement of their capability to perform a behaviour, and controllability, which constitutes an individual's beliefs if they have the necessary resources and opportunities to adopt the innovation. It denotes a subjective judgment of the degree of control over the performance of a behaviour not the perceived likelihood that performing the behaviour will produce a given outcome (Ajzen, 1991). In the context of mobile service adoption, perceived behavioural control refers to the individual perception of how easy or difficult it is to get mobile services.

Fourth, *personal innovativeness* is the willingness of an individual to try out and embrace new technologies and their related services for accomplishing specific goals. Also known as technology readiness, personal innovativeness embodies the risk-taking propensity which exists in certain individuals and not in others (Agarwal & Prasad, 1998; Massey et al., 2005; Parasuraman, 2000). This definition helps segment potential adopters into what Rogers (1995) characterises as innovators, early adopters, early and late majority adopters, and laggards. Personal innovativeness represents a confluence of technology-related beliefs which jointly determine an individual's predisposition to adopt mobile devices and related services. The adoption of any innovation in general, and of innovative mobile phones and services in particular is inherently associated with greater risk (Kirton, 1976). Therefore, given the same level of beliefs and perceptions about an innovation, individuals with higher personal innovativeness are more likely to develop positive attitudes towards adopting it than less innovative individuals (Agarwal & Prasad, 1998).

Fifth, *perceived enjoyment* refers to the degree to which using an innovation is perceived to be enjoyable in its own right and is considered to be an intrinsic source of motivation (Al-Gahtani & King, 1999). Because the market for mobile innovations and services is comprised of both corporate users and consumers, factors focusing on perceived enjoyment constitute an important consideration (Carlsson et al., 2005; Pagani, 2004). That is, adopters use an innovation for the pleasure or enjoyment its adoption might bring and, therefore, serve as an end unto itself. Further, intrinsic enjoyment operates outside valued outcomes or immediate material needs (i.e., extrinsic motivations), such as enhanced job performance, increased pay, and so forth (Mathwick et al., 2001; Moon & Kim, 2001). Most research on enjoyment is based on the "flow theory" according to which flow represents "the holistic sensation that people feel when they act with total involvement" (Csikszentmihalyi,

1975). In a "flow state" individuals interact more voluntarily with innovations within their specific context, which determines their subjective experiences (Csikszentmihalyi, 1975). Consequently, individuals who have a more positive enjoyment experience with an innovation are likely to have stronger adoption intentions than those who do not (Moon & Kim, 2001).

That is, intrinsic enjoyment can positively affect the adoption and use of innovative mobile services, and is therefore, a significant determinant of intention and attitude towards adoption (Kaufaris, 2002; Novak, Hoffman, & Yung, 2000). Further, upon adoption, individuals are more likely to use the mobile services that offer enjoyment more extensively than those which do not. As a consequence, perceived enjoyment is also seen to have a significant effect beyond perceived usefulness (Davis et al., 1989a). However, the complexity of a mobile innovation or service has a negative effect on perceived enjoyment, suggesting that the potential impact of enjoyment may not be fully realised (Igbaria et al., 1996).

The final variable that needs to be added to the existing technology adoption models is *price sensitivity*. In the original technology acceptance models, the costs of adopting an innovation were not considered to be a relevant construct because the actual users did not have to pay for the technology. In an organisational setting, the cost would be incurred by the organisation. However, in the context of individual private adoption, cost becomes a relevant factor. There is evidence showing that perceived financial resources required to adopt mobile technologies and services constitute a significant determinant of behavioural intention (Kleijen et al., 2004; Lin & Wang, 2005). However, evidence also shows that adopters of mobile devices and services also attempt to assess the value of adoption by comparing perceived costs against the benefits (Pagani, 2004). Perceived costs are directly related to income and socioeconomic status of potential adopters which are recognised to have a strong impact on technology adoption and diffusion (Lu, Yu, Liu, & Yao, 2003). For example, in Europe individuals earning income beyond certain levels were found to have a high propensity to embrace mobile technologies, such as WAP mobile phones, handheld computers, and so forth (Crawford, 2002). Similarly, there's evidence that in fast growing economies, individuals with higher income spend more on mobile devices (Lu et al., 2003).

Perceived Usefulness

Perceived usefulness is "the degree to which a person believes that using a particular system would enhance his or her job performance" (Davis, 1989, p. 320). Perceived usefulness is also known as performance expectancy (Venkatesh et al., 2003). An innovation is believed to be of high usefulness when a potential adopter believes that there is a direct relationship between use on the one hand and productivity, performance, effectiveness, or satisfaction on the other (Lu et al., 2003).

Usefulness recognition is important because it has been found to have a strong direct effect on the intention of adopters to use the innovation (Adams, Nelson, & Todd, 1992; Davis, 1989). In addition, potential adopters assess the consequences of their adoption behaviour and innovation usage in terms of the ongoing desirability of usefulness (Chau, 1996; Venkatesh & Davis, 2000). Although an innovation might provide at least some degree of usefulness, a potential reason not to adopt exists when adopters fail to see the "need" to adopt (Zeithaml & Gilly, 1987). Adopters may not be able to recognise their needs until they become aware of the innovation or its consequences (Rogers, 1995). Need recognition is, therefore, likely to drive potential adopters to educate themselves in order to be able to utilise the innovation fully before being able to recognise its usefulness. This in turn is likely to lead to a faster rate of adoption (Rogers, 1995; Saaksjarvi, 2003).

Perceived usefulness can be split into two parts. Near-term usefulness is perceived to have

an impact on the near-term job fit, such as job performance or satisfaction (Thompson, Higgins, & Howell, 1994). Long-term usefulness is perceived to enhance the future consequences of adoption including career prospects, opportunity for preferred job assignments, or social status of adopters (Chau, 1996; Thompson et al., 1994). Evidence shows that even though perceived near-term usefulness has the most significant impact on the behavioural intention to adopt an innovation, perceived long-term usefulness also exerts a positive, yet lesser impact (Chau, 1996; Jiang, Hsu, Klein, & Lin, 2000).

In the case of mobile technology and services, perceived usefulness is defined as the degree to which the mobile technology and services provide benefits to individuals in every day situations (Knutsen et al., 2005). The range and type of service offerings as well as the compatibility of the user's existing computing devices influence perceived usefulness (Pagani, 2004). In addition, Pagani (2004) also finds that usefulness emerges as the strongest determinant in the adoption of three generation mobile services which is consistent finds of research concerning the adoption of other innovations (Venkatesh et al., 2003).

Perceived Ease of Use

Perceived ease of use is the "degree to which a person believes that using a particular system would be free of effort" (Davis, 1989, p. 320). Other constructs that capture the notion of perceived ease of use are complexity and effort expectancy (Rogers, 1995; Venkatesh et al., 2003). Perceived ease of use may contribute towards performance, and therefore, near-term perceived usefulness. In addition, lack of it can cause frustration, and therefore, impair adoption of innovations. Nevertheless, "no amount of EOU [ease of use] will compensate for low usefulness" (Keil, Beranek, & Konsynski, 1995, p. 89).

In the mobile setting, perceived ease of use represents the degree to which individuals associate freedom of difficulty with the use of mobile technology and services in everyday usage (Knutsen et al., 2005). For example, there is evidence in the media that using certain services on a mobile device can be quite tedious, especially when browsing Internet-like interfaces on mobile devices is required (Teo & Pok, 2003). Together with relatively small screen sizes and associated miniaturized keypads, the overall usage experience may be adversely affected. This suggests that input and output devices are likely to influence perceived ease of use (Pagani, 2004). In addition, user-friendly and usable intuitive man-machine interfaces, including clear and visible steps, suitable content and graphical layouts, help functions, clear commands, symbols, and meaningful error messages are likely to influence adoption as well (Condos, James, Every, & Simpson, 2002). Further, Pagani (2004) argues the mobile system response time affects perceived ease of use suggesting that mobile bandwidth is important as well.

Social Influences

Social influence constitutes the degree to which individuals perceive that important or significant others believe they should use an innovation (Venkatesh et al., 2003). Venkatesh et al. (2003) believe that the social influence constructs may only become significant drivers on intention to adopt when users adopt an innovation in order to comply mandatory requirements. In these circumstances, social influence seems to be significant in the early phases of adoption and its effect decreases with sustained usage (Venkatesh & Davis, 2000). Conversely, in voluntary settings, social influence appears to have an impact on perceptions about the innovation (Venkatesh et al., 2003). Social influence is related to three similar constructs, namely, subjective and social norms, and image.

In Taylor and Todd's study (1995), subjective norms are defined to include the influence of other

people's opinions otherwise known as reference groups. These include peers, friends, superiors, computer, and technology experts. Subjective norms have a greater impact during the initial adoption phase when potential adopters have little or no experience or when the adoption behaviour is new (Thompson et al., 1994). Research shows that pressure from reference groups to adopt an innovation is effective because it contributes to reducing perceived risk associated with adoption (Teo & Pok, 2003).

Social factors constitute another construct of social influence. Social factors represent cues individuals receive from members of their social structure which prompt them to behave in certain ways (Thompson, Higgins, & Howell, 1991). For example, in Japan, teenagers regard smart phones as fashion items (Lu et al., 2003). Further, there is evidence that unique communications patterns determined by key social and cultural factors, such as group-oriented nationality, have positively affected adoption practices of using the Internet via mobile phones in East Asia (Ishii, 2004).

A third critical construct related to social influence is image. The adoption of an innovation can be seen to enhance one's status or image in their social system. For certain adopters a mobile device may be more of a lifestyle than a necessity (Bina & Giaglis, 2005; Teo & Pok, 2003). For example, early adopters of mobile computing devices might be image-conscious users who wish to be seen as trend-setters or technology savvy enthusiasts.

Facilitating Conditions

Facilitating conditions refer to external controls and catalysts in the adoption environment which aim at facilitating adoption and diffusion of new technologies (Terry, 1993). Facilitating conditions are important because they are considered to be direct usage antecedents, and are therefore, likely to make adoption behaviour less difficult by removing any obstacles to adoption and sustained usage (Thompson et al., 1994; Venkatesh et al., 2003). These conditions can be provided by both governments and mobile operators. For example, governments or the representative agencies can act as facilitators by bringing together the telecommunication industry, academia, and research community. Government agencies can also set up protocol standardization policies and regulations favouring the future growth of mobile communication systems (Lu et al., 2003). Likewise, mobile operators can encourage adoption by mass advertising campaigns and active promotion aimed at increasing awareness about mobile devices and related services (Teo & Pok, 2003). Further, promoting and enforcing appropriate interconnection agreements and adequate regulatory mechanisms among mobile operators help adopters of mobile devices take advantage of roaming services and consequently be conducive to adoption (Rossotto et al., 2000).

Facilitating conditions also capture the existence of a trusting environment that is external to the mobile operator's control. A trusting environment constitutes an important factor in the adoption of mobile technologies and services. It determines the user's expectations from the relationship with their service providers, and it increases their perceived certainty concerning the provider's expected behaviour. Generally, trust is essential in all economic activities where undesirable opportunistic behaviour is likely to occur (Gefen et al., 2003). However, trust becomes vital in a mobile environment, where situational factors such as uncertainty or risk and information asymmetry are present (Ba & Pavlou, 2002). On the one hand, adopters of mobile technology are unable to judge the trustworthiness of service providers, and on the other, the latter can also easily take advantage of the former by engaging in harmful opportunistic behaviours. For example, service providers can sell or share the transactional information of its users or their personal information.

There are two key elements in a trusting environment, namely, security and privacy (Lu

et al., 2003). In a wireless environment, security encompasses confidentiality, authentication, and message integrity. Because mobile devices have limited computing resources and wireless transmissions are more susceptible to hacker attacks, security vulnerabilities can have serious consequences (Galanxhi-Janaqi & Nah, 2004; Lu et al., 2003). There are several remedies against the dangers of insecurity, for example, public key infrastructure and certificate authority which use public key cryptography to encrypt and decrypt mobile transmissions and authenticate users.

Ironically, the same information practices which provide value to both users and providers of mobile technology and services also cause privacy concerns. Some of these concerns include: the type of information that can be collected about users and the ways in which it will be protected; the entities that can access this information and their accountability; and the ways in which the information will be used (Galanxhi-Janaqi & Nah, 2004). In mobile adoption research the trust environment has been encapsulated in a construct called perceived credibility (Lin & Wang, 2005; Wang, Wang, Lin, & Tang, 2003). Evidence shows that there is a "significant direct relationship between perceived credibility and behavioural intention" (Lin & Wang, 2005, p. 410) to use mobile services.

Moderating Variables

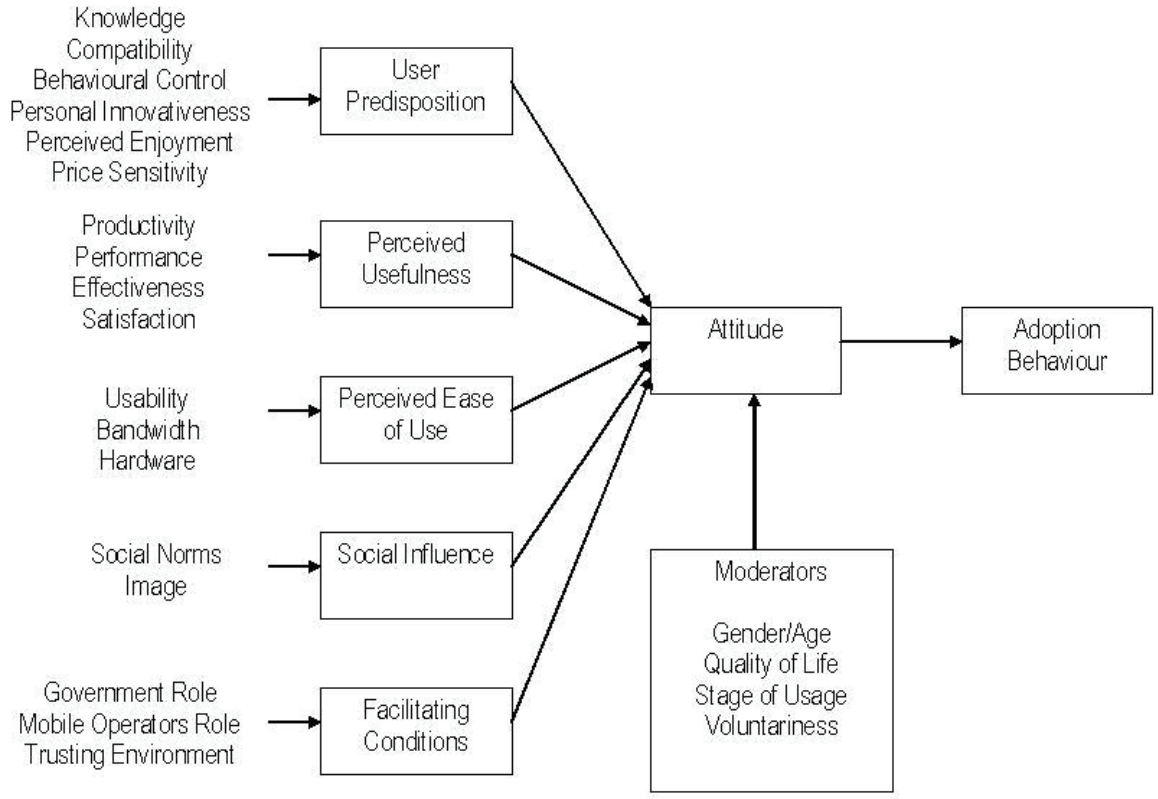
Evidence shows that gender and age might influence the adoption of technology and related services due to their moderating effects on other constructs (Venkatesh et al., 2003). In general, men tend to exhibit task-oriented attitudes suggesting that usefulness expectations might be more accentuated in men than women (Minton & Scheneider, 1980). This is particularly the case for younger men (Venkatesh & Morris, 2000a). On the other hand, ease of use expectations are more salient for women and older adopters (Bozionelos, 1996). Further, women are predisposed to be more sensitive to the opinions of members of their social structure. As a result women are more likely to be affected by social influence factors when deciding to adopt new mobile technologies and services (Venkatesh & Morris, 2000a). Similarly, because affiliation needs increase with age (Rhodes, 1983), older adopters are more likely to be affected by social influence.

Quality of life of potential adopters is another moderating variable which is likely to affect the adoption of mobile devices and services. "Quality of life" is an established social sciences notion which represents "a global assessment of a person's life satisfaction according to his chosen criteria" (Diener & Suh, 1997; Shin & Johnson, 1978). There is evidence which indicates that mobile technology and services have enhanced the perceived quality of social and work life of adopters (Jarvenpaa, Lang, Takeda, & Tuunainen, 2003). Bina and Giaglis (2005) present evidence that the reverse is also true. They indicate that adopters who are satisfied with specific life domains exhibit favourable attitudes towards the adoption of specific mobile services (Bina & Giaglis, 2005).

Further, evidence also shows that stage of use and voluntariness of usage have moderating effects on adoption attitudes through various constructs. For example, perceived ease of use is significant during the initial period of usage when process-related issues constitute obstacles to be overcome (Venkatesh et al., 2003). Perceived ease of use, however, becomes insignificant during periods of extended usage (Agarwal & Prasad, 1998; Davis et al., 1989a; Thompson et al., 1994). Similarly, ease of use has a significant positive effect on attitude toward use in both voluntary and mandatory usage contexts (Al-Gahtani & King, 1999; Keil et al., 1995; Venkatesh & Davis, 2000).

To summarise the constructs discussed in this section, Figure 2 portrays a proposed model of acceptance of mobile services. The implications of this model are discussed next.

Figure 2. Proposed model of acceptance of mobile services



PROPOSED METHOD

In order to validate the model discussed in the previous section, we propose a two-stage research design, consisting of both qualitative and quantitative approaches.

Qualitative Stage

This stage is the first round of the fieldwork data collection. Data collection at this stage would involve conducting face-to-face in-depth interviews in order to study the perceptions of all stakeholders who contribute directly or indirectly to providing mobile services. Stakeholder targets include mobile operators or carriers, industry and government associations, user groups, mobile application developers, content providers, aggregators, as well as manufacturers of mobile devices. The aim of these interviews is to gather an in-depth understanding of the perceptions and perspectives of all stakeholders involved in the adoption and diffusion of mobile services.

To ensure validity in this research design, several tactics may be used. Construct validity would be addressed by using the multiple sources of evidence as noted previously. The issue of internal validity would be considered by using the techniques of pattern matching the data to a predicted pattern of variables, and formulating rival explanations. In addition, an interview protocol should be developed to guide the data collection. A single pilot case study is recommended to be used in order to refine data collection procedures and improve conceptualisation of the model prior to finalising

the set of theoretical propositions developed from the literature.

The second round of the qualitative stage would involve focus groups with mobile users. Because the research phenomenon is contemporary and little prior research has been conducted, focus groups would be appropriate for generating ideas and obtaining insights from existing mobile service users and potential users (Carson, Gilmore, Gronhaug, & Perry, 2001). Focus groups are useful when investigating complex behaviour and motivations. By comparing the different points of view that participants exchange during the interactions in focus groups, researchers can examine motivation with a degree of complexity that is usually not available with other methods (Morgan & Krueger, 1993). The use of a focus group is more valuable many times over compared with any representative sample for situations requiring the investigation of complex decision-making processes, as is the case for this research. Based on demographic characteristics, we propose setting up homogeneous groups because discussions within homogeneous groups produce more in-depth information than discussions within heterogenous groups (Bellenger, Bernhardt, & Goldtucker, 1989). These groups would be selected based on main moderating variables identified in the literature, such as, gender/age, quality of life, stage of usage, and voluntariness.

We believe that at least two investigators should conduct all interviews and moderate the focus groups (Denzin, 1989; Patton, 1990). This kind of triangulation reduces the potential bias which is commonly cited as a limitation of interviews and focus gruops (Frankfort-Nachmias & Nachmias, 1996; Yin, 1994).

Quantitative Stage

The last stage of this project would involve an online survey. The collected data would help understand and confirm the determinants and the adoption intentions of the consumers of mobile services. Random sampling should be used to select the sample. We propose that two types of data analysis should be performed on the survey data: descriptive analysis and inferential analysis. Descriptive analysis should be carried out for transformation of raw data into a form that would provide information to describe a set of factors in a situation (Sekaran, 2000). For the inferential analysis, a structural equation model (SEM) should be used to test the refined model.

MANAGERIAL IMPLICATIONS

Mobile technologies and the associated services integrate both the business and social domains of the user's life (Elliot & Phillips, 2004; Knutsen et al., 2005). 3G services in general and location-based services in particular can provide anytime-anyplace tracking of adopters (UMTS, 2003). This creates the opportunity for developing accurate adopter profiles both in their work- and leisure-related domains. In addition, live video and location-based information can also be gathered (Robins, 2003). While such information can help address the needs of adopters better, it can also be misused by businesses for unethical direct business-to-consumer marketing (Casal et al., 2004), raising privacy concerns, overcontrol and overwork of individual adopters (Yen & Chou, 2000). For example, by reducing space and time constraints, mobile communications provide an immensely flexible work environment for some individuals while bringing about overwork or intrusion problems for others (Gerstheimer & Lupp, 2004). As a result, existing privacy protection policies and regulations about employees and consumers should reflect these new conditions. These policies should also account for overcontrol prevention that is likely to result from organisations' attempts to monitor individual performance (Yen & Chou, 2000).

Designing content suitable for mobile phones constitutes an important issue that affects the

adoption and diffusion of mobile technology and associated services. This has implications for service providers, developers, policymakers, and academics. Content providers must design content "for value-contexts specific for mobile use which provide users freedom from complicated configuration procedures, and ubiquitously serve and support current day-to-day individual social practices" (Knutsen et al., 2005a, p. 7). Developers of mobile applications need to recognise that mobile applications are quite different from PC applications (Funk, 2005). Developers should use established standards, such as HTML and Java. More importantly, the usage contexts, the adopters, and their evolving behaviour should be important considerations. Further, because "made-for-the-medium" content type and design may be required (Massey et al., 2005), the available technologies which determine screen size, display quality and processing speeds should be taken into consideration as well (Funk, 2005). The combined effect of these factors on navigation patterns, adopters' cognitive overload, and subjective perceptions about the usability and ease of use of mobile applications can have a critical impact on uptake (Chae & Kim, 2004).

Segmentation of mobile service adopters must not only be based on adopter type (e.g., pioneers, early adopters, majority adopters, and laggards) but also on individual differences. The basis of segmentation should constitute the foundation in developing marketing strategies. For example, individuals with high personal innovativeness or novelty seekers are likely to be willing to experiment with new mobile devices and services, in which case these should be marketed as technological innovations. For individuals who are reluctant to use the same devices and services and are likely to feel discomfort and insecurity while using them, lifestyle promotions may be more appropriate (Dabholkar & Bagozzi, 2002; Teo & Pok, 2003). In addition, endorsements by peers, famous celebrities, or other referent groups may be adequate if these individuals appreciate social norms and image (Hung, Ku, & Chang, 2003; Teo & Pok, 2003). Marketing mobile applications for adopters in one category is likely to frustrate adopters in the other. Therefore, developers and marketers should be prudent in recognising that the confluence of various individual characteristics with varying levels of prior experience, perceptions, and learning predispositions are all likely to influence adoption and retention patterns (Card, Moran, & Newell, 1983; Hung et al., 2003; Massey et al., 2005).

Further, the interface design of mobile applications should encompass both intrinsic and extrinsic motivation dimensions (Moon & Kim, 2001). Based on the proposed model, marketers should promote attributes such as usefulness, ease of use, and enjoyment as important aspects when attempting to persuade potential users in adopting specific mobile phones and services as well as to increase their loyalty and retention (Dabholkar & Bagozzi, 2002; Hung et al., 2003; Lin & Wang, 2005). In particular, personalisation is a well-suited and an achievable goal as mobile phones are identifiable. 3G phones also enable identification of the location of individual handsets, making location-specific marketing possible. Messages promoting the services of businesses, such as restaurants, hotels, grocery stores, and so forth, can be transmitted when users are detected within range (Robins, 2003). Evidence shows that despite privacy concerns, many users of mobile devices are happy to receive unsolicited promotional messages provided that such messages are relevant and personalised (Robins, 2003).

Governments and mobile operators should design appropriate and dedicated strategies to promote the relative advantages of mobile phones and services. Such promotion strategies are important because of their impact on the perceptions of potential adopters (Knutsen et al., 2005). Moreover, the development of wireless communication infrastructures and the provision of incentives are likely to contribute towards the minimisation of the digital divide which results from demographic factors such as varying income levels, education

and experience, gender, and age (Lin & Wang, 2005). The digital divide not only prevents the exploitation of the full market potential, but it also adversely impacts the maximization of benefits for current adopters due to limited network externalities effects (Katz & Shapiro, 1986).

CONCLUSION AND FUTURE RESEARCH

User acceptance of mobile technology and related services is of paramount importance. Consequently, a deeper insight into theory-based research is required to better understand the underlying motivations and barriers that will lead users to inhibit them from adopting these technologies and services. This in turn will also help designing technology and service improvements as well as appropriate adoption and diffusion strategies. There are several theoretical models in the literature which attempt to determine acceptance and adoption of new technologies. However, most of these models originate from organisational contexts. As mobile technologies and services add other functional dimensions such as hedonic or experiential aspects, applying extant theories outright to determine the acceptance and adoption of mobile services may be questionable and inadequate.

In this chapter, we have explored and critically reviewed existing technology acceptance theories. Relevant constructs of extant models were discussed in the light of evolving mobile technologies and services and then incorporated into a synthesised acceptance model of mobile services. The proposed model attempts to view acceptance of mobile services beyond traditional organisational borders and permeate everyday social life practices. The proposed model which can be tested empirically provides the foundation to guide further validation and future research in the area of mobile services adoption.

In addition, a plethora of mobile services have become available recently (Alahuhta et al., 2005). Because all services would be available to adopters through a single user interface of the current technology, the appropriation of these services by users may be interconnected and at different stages of maturity (Knutsen et al., 2005). These interconnections are temporal and are also likely to have mutually enhancing, suppressing, or compensating effects on each other (Black & Boal, 1994). This adds dynamism and complexity to acceptance and, therefore, cannot be explained by simply considering factors impacting individual or aggregate adoption at single points in time (Knutsen et al., 2005; Pagani, 2004). Consequently, future research should develop and test dynamism-compatible acceptance models because these models may provide a deeper understanding and help in explaining how and why technology acceptance perceptions change as the appropriation process progresses. Further, with a wide variety of mobile devices and services available and their applicability in distinct spheres of life, the definition of a unit of analysis in mobile services adoption has become a challenging task (Knutsen et al., 2005). Additional research in this aspect is also needed.

REFERENCES

Adams, D. A., Nelson, R. R., & Todd, P. A. (1992). Perceived usefulness, ease of use, and usage of information technology: A replication. *MIS Quarterly, 16*(2), 227-247.

Agarwal, R., & Prasad, J. (1998). A conceptual and operational definition of personal innovativeness in the domain of information technology. *Information Systems Research, 9*(2), 204-215.

Ajzen, I. (1991). The theory of planned behavior. *Organisational Behavior and Human Decision Process, 52*(2), 179-211.

Alahuhta, P., Ahola, J., & Hakala, H. (2005). *Mobilising business applications: a survey about the opportunities and challenges of mobile business applications and services in Finland* (Technology Review No. 167/2005). Helsinki: Tekes.

Alba, J. W., & Hutchinson, J. W. (1987). Dimensions of consumer expertise. *Journal of Consumer Research, 13*(3), 411-454.

Al-Gahtani, S. S., & King, M. (1999). Attitudes, satisfaction and usage: Factors contributing to each in the acceptance of information technology. *Behaviour & Information Technology, 18*(4), 277-297.

Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: price premiums and buyer behavior. *MIS Quarterly, 26*(3), 243-268.

Bellenger, D. N., Bernhardt, K. L., & Goldtucker, J. L. (1989). Qualitative research techniques: Focus group interviews. In T. J. Hayes, & C. B. Tathum (Eds.), *Focus group interviews: A reader* (pp. 7-28). Chicago: American Marketing Association.

Bina, M., & Giaglis, G. M. (2005). *Exploring early usage patterns of mobile data services.* Paper presented at the International Conference on Mobile Business, Sydney, Australia, July 11-13.

Black, J. A., & Boal, K. B. (1994). Strategic resources: Traits, configurations, and paths to sustainable competitive advantage. *Strategic Management Journal, 15,* 131-148.

Bozionelos, N. (1996). Psychology of computer use: Prevalence of computer anxiety in British managers and professionals. *Psychological Reports, 78*(3), 995-1002.

Calder, B. J., & Staw, B. M. (1975). Self-perception of intrinsic and extrinsic motivation. *Journal of Personality and Social Psychology, 31*(4), 599-605.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Carlsson, C., Hyvonen, K., Repo, P., & Walden, P. (2005). *Adoption of mobile services across different platforms.* Paper presented at the 18th Bled eCommerce Conference, Bled, Slovenia, June 6-8.

Carson, D., Gilmore, A., Gronhaug, K., & Perry, C. (2001). *Qualitative research in marketing.* London: Sage.

Casal, C. R., Burgelman, J. C., & Bohlin, E. (2004). Propects beyond 3G. *Info, 6*(6), 359-362.

Chae, M., & Kim, J. (2003). What's so different about the mobile Internet? *Communications of the ACM, 46*(12), 240-247.

Chae, M., & Kim, J. (2004). Do size and structure matter to mobile users? An empirical study of the effects of screen size, information structure, and task complexity on user activities with standard web phones. *Behaviour & Information Technology, 23*(3), 165-181.

Chau, P. Y. K. (1996). An empirical assessment of a modified technology acceptance model. *Journal of Management Information Systems, 13*(2), 185-204.

Chau, P. Y. K., & Hu, P. J.-H. (2001). Information technology acceptance by individual professionals: a model comparison approach. *Decision Science, 32*(4), 699-719.

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly, 23*(2), 189-211.

Condos, C., James, A., Every, P., & Simpson, T. (2002). Ten usability principles for the development of effective WAP and m-commerce services. *Aslib Proceedings, 54*(6), 345-355.

Crawford, A. M. (2002). International media habits on the rise. *Ad Age Global, 2*(11). Retrieved

from http://web.ebscohost.com/ehost/detail?vid=3&hid=101&sid=ff86c2ae-e7f7-4388-96b4-7da9c1bc4eb3%40sessionmgr106

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety.* San Francisco: Jossey-Bass.

Dabholkar, P. A., & Bagozzi, R. P. (2002). An attitudinal model of technology-based self-service: Moderating effects of consumer traits and situational factors. *Journal of Academy of Marketing Science, 30*(3), 184-201.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance in information technology. *MIS Quarterly, 13*(3), 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1002.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology, 22*, 1111-1132.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum Press.

Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods (3rd ed.).* Englewood Cliffs, N. J.: Prentice Hall.

Dholakia, R. R., & Dholakia, N. (2004). Mobility and markets: Emerging outlines for m-commerce. *Journal of Business Research, 57*(12), 1391-1396.

Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social and subjective indicators. *Social Indicators Research, 40*(1-2), 189-216.

Dillon, A., & Morris, M. (1996). User acceptance of information technology: theories and models. *Journal of American Society for Information Science, 31*, 3-32.

Elliot, G., & Phillips, N. (2004). *Mobile commerce and wireless computing systems.* Harlow: Pearson Education Limited.

Elliot, S., & Loebbecke, C. (2000). Interactive, inter-organizational innovations in electronic commerce. *Information Technology & People, 13*(1), 46-66.

Figge, S. (2004). Situation-dependent services: A challenge for mobile operators. *Journal of Business Research, 57*(12), 1416-1422.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behaviour: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Frankfort-Nachmias, C., & Nachmias, D. (1996). *Research methods in the social sciences (5th ed.).* New York: St. Martin's Press.

Funk, J. L. (2005). The future of the mobile phone Internet: An analysis of technological trajectories and lead users in the Japanese market. *Technology in Society, 27*(1), 69-83.

Galanxhi-Janaqi, H., & Nah, F. F.-H. (2004). U-commerce: Emerging trends and research issues. *Industrial Management & Data Systems, 104*(9), 744-755.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. *MIS Quarterly, 27*(1), 51-90.

Gerstheimer, O., & Lupp, C. (2004). Needs versus technology: The challenge to design third-generation mobile applications. *Journal of Business Research, 57*(12), 1409-1415.

Grundström, C., & Wilkinson, I. F. (2004). The role of personal networks in the development of industry standards: A case study of 3G mobile telephony. *Journal of Business and Industrial Marketing, 19*(4), 283-293.

Hammond, K. (2001). B2C e-commerce 2000-2010: What experts predict. *Business Strategy Review, 12*(1), 43-50.

Hart, J., & Hannan, M. (2004). The future of mobile technology and mobile wireless computing. *Campus-Wide Information Systems, 21*(5), 201-204.

Horton, R. P., Buck, T., Waterson, P. E., & Clegg, C. W. (2001). Explaining intranet use with the technology acceptance model. *Journal of Information Technology, 16*, 237-249.

Hung, S.-Y., Ku, C.-Y., & Chang, C.-M. (2003). Critical factors of WAP services adoption: An empirical study. *Electronic Commerce Research and Applications, 2*(1), 42-60.

Igbaria, M., Parasuraman, S., & Baroudi, J. J. (1996). A motivational model of microcomputer usage. *Journal of Management Information Systems, 13*(1), 127-143.

Ishii, K. (2004). Internet use via mobile phone in Japan. *Telecommunications Policy, 28*(1), 43-58.

Jarvenpaa, S. L., Lang, K. R., Takeda, Y., & Tuunainen, V. K. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41-44.

Jiang, J. J., Hsu, M. K., Klein, G., & Lin, B. (2000). E-commerce user behaviour model: An empirical study. *Human Systems Management, 19*(4), 265-276.

Katz, M. L., & Shapiro, C. (1986). Technology adoption in the presence of network externalities. *Journal of Political Economy, 94*(4), 822-841.

Kaufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behaviour. *Information Systems Research, 13*(2), 205-223.

Keil, M., Beranek, P. M., & Konsynski, B. R. (1995). Usefulness and ease of use: Field study evidence regarding task considerations. *Decision Support Systems, 13*(1), 75-91.

Khalifa, M., & Cheng, S. K. N. (2002). *Adoption of mobile commerce: Role of exposure.* Paper presented at the 35th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Hawaii, January 7-10 (pp. 46-52). IEEE Computer Society.

King, J. L., Gurbaxani, V., Kraemer, K. L., McFarlan, F. W., Raman, K. S., & Yap, C. S. (1994). Institutional factors in information technology innovation. *Information Systems Research, 5*(2), 139-169.

Kirton, M. (1976). Adopters and innovators: a description and measure. *Journal of Applied Psychology, 61*(5), 622-629.

Klasen, L. (2002). Migrating an online service to WAP: Case study. *The Electronic Library, 20*(3), 195-201.

Kleijen, M., Wetzels, M., & de Ruyter, K. (2004). Consumer acceptance of wireless finance. *Journal of Financial Services Marketing, 8*(3), 206-217.

Knutsen, L., Constantiou, I. D., & Damsgaard, J. (2005). *Acceptance and perceptions of advanced mobile services: Alterations during a field study.* Paper presented at the International Conference on Mobile Business, Sydney, Australia, July 11-13.

Lee, M. S. Y., McGoldrick, P. J., Keeling, K. A., & Doherty, J. (2003). Using ZMET to explore barriers to the adoption of 3G mobile banking services. *International Journal of Retail & Distribution Management, 31*(6), 340-348.

Leonard-Barton, D., & Deschamps, I. (1988). Managerial influence in the implementation of new technology. *Management Science, 34*(10), 1252-1265.

Lin, H., & Wang, Y. (2005). *Predicting consumer intention to use mobile commerce in Taiwan.* Paper presented at the International Conference on Mobile Business, Sydney, Australia, July 11-13.

Lu, J., Yu, C., Liu, C., & Yao, J. E. (2003). Technology acceptance model for wireless Internet. *Internet Research: Electronic Networking Applications and Policy, 13*(3), 206-222.

Massey, A. P., Khatri, V., & Ramesh, V. (2005). *From the Web to the wireless Web: Technology readiness and usability.* Paper presented at the 38th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Hawaii, January 3-6 (p. 32b). IEEE Computer Society.

Mathwick, C., Malhotra, N., & Rigdon, E. (2001). Experiental value: Conceptualization, measurement and application in the catalog and Internet shopping environment. *Journal of Retailing, 77*(1), 39-56.

Minton, G. C., & Scheneider, F. W. (1980). *Differential psychology.* Prospect Heights, IL: Waveland Press.

Moon, J.-W., & Kim, Y.-G. (2001). Extending the TAM for a World-Wide-Web context. *Information & Management, 38*(4), 217-230.

Moreau, C. P., Lehmann, D. R., & Markman, A. B. (2001). Entrenched knowledge structures and consumer response to new products. *Journal of Marketing Research, 38*(1), 14-29.

Morgan, D. L., & Krueger, R. A. (1993). When to use focus groups and why. In D. L. Morgan (Ed.), *Successful focus groups* (pp. 1-19). London: Sage Publications.

Novak, T. P., Hoffman, D. L., & Yung, Y. (2000). Measuring the customer experience in online environments: A structural modeling approach. *Marketing Science, 19*(1), 22-42.

Pagani, M. (2004). Determinants of adoption of third generation mobile multimedia services. *Journal of Interactive Marketing, 18*(3), 46-59.

Parasuraman, A. (2000). Technology readiness index: A multiple item scale to measure readiness to embrace new technologies. *Journal of Service Research, 2*(4), 307-320.

Patton, M. Q. (1990). *Qualitative evaluation and research methods (2nd ed.).* London: Sage Publications.

Ratliff, J. M. (2002). NTT DoCoMo and its i-mode success: Origins and implications. *California Management Review, 44*(3), 55-71.

Repo, P., Hyvonen, K., Pantzar, M., & Timonen, P. (2004). *Users intenting ways to enjoy new mobile services: The case of watching mobile videos.* Paper presented at the 37th Hawaii International Conference on System Sciences, Hawaii, January 5-8 (p. 40096.3). IEEE Computer Society.

Rhodes, S. R. (1983). Age-related differences in work attitudes and behavior: A review of conceptual analysis. *Psychological Bulletin, 93*(2), 328-367.

Robins, F. (2003). The marketing of 3G. *Marketing Intelligence & Planning, 21*(6), 370-378.

Roehm, M. L., & Sternthal, B. (2001). The moderating effect of knowledge and resources on the persuasive impact of analogies. *Journal of Consumer Research, 28*(2), 257-272.

Rogers, E. M. (1995). *Diffusion of innovations.* New York: Free Press.

Rossotto, C. M., Kerf, M., & Rohlfs, J. (2000). Competition in mobile telecommunications: Sector growth, benefits for the incumbent and policy trends. *Info, 2*(1), 67-73.

Saaksjarvi, M. (2003). Consumer adoption of technological innovations. *European Journal of Innovation Management, 6*(2), 90-100.

Sekaran, U. (2000). *Research methods for business: A skill building approach*. New York: John Wiley and Sons.

Sheppard, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research, 15*(3), 325-343.

Shin, C. C., & Johnson, D. M. (1978). Avowed happiness as an overall assessment of quality of life. *Social Indicators Research, 5,* 475-492.

Siau, K., Lim, E. P., & Shen, Z. (2001). Mobile commerce: Promises, challenges, and research agenda. *Journal of Databases Management, 12*(2), 4-13.

Taylor, S., & Todd, P. A. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research, 6*(2), 144-176.

Taylor, S., & Todd, P. A. (1995a). Assessing IT usage: The role of prior experience. *MIS Quarterly, 19*(4), 561-570.

Teo, T. S. H., & Pok, S. H. (2003). Adoption of WAP-enabled mobile phones among Internet users. *Omega: The International Journal of Management Science, 31*(6), 483-498.

Terry, D. J. (1993). Self-efficacy expectancies and the theory of reasoned action. In D. C. Terry, C. Gallois, & M. McCamish (Eds.), *The theory of reasoned action: Its application to AIDS-preventive behaviour* (pp. 135-152). Oxford: Pergamon.

Thompson, R., Higgins, C., & Howell, J. (1994). Influence of experience on personal computer utilization: Testing a conceptual model. *Journal of Management Information Systems, 11*(1), 167-187.

Thompson, R. L., Higgins, C. A., & Howell, J. M. (1991). Personal computing: Toward a conceptual model of utilization. *MIS Quarterly, 15*(1), 125-143.

UMTS. (2003). *Mobile evolution: Shaping the future.* Retrieved August 28, 2005, from http://www.umts-forum.org/servlet/dycon/ztumts/umts/Live/en/umts/Resources_Papers_index

van Steenderen, M. (2002). Business applications of WAP. *The Electronic Library, 20*(3), 215-223.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science, 46*(2), 186-204.

Venkatesh, V., & Morris, M. G. (2000a). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly, 24*(1), 115-139.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Wang, Y.-S., Wang, Y.-M., Lin, H.-H., & Tang, T.-I. (2003). Determinants of user acceptance of Internet banking: An empirical study. *International Journal of Service Industry Management, 14*(5), 501-519.

Xylomenos, G., & Polyzos, G. C. (2001). Quality and service support over multi-service wireless Internet links. *Computer Networks, 37*(5), 601-615.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 776-795.

Yen, D. C., & Chou, D. C. (2000). Wireless communications: Applications and managerial issues. *Industrial Management & Data Systems, 100*(9), 436-443.

Yin, R. K. (1994). *Case study research: Design and methods.* Beverley Hills: Sage.

Zeithaml, V. A., & Gilly, M. C. (1987). Characteristics affecting the acceptance of retailing technologies: A comparison of elderly and nonelderly consumers. *Journal of Retailing, 63*(1), 49-68.

This work was previously published in Mobile Multimedia Communications: Concepts, Applications, and Challenges, edited by G. Karmakar and L. Dooley, pp. 85-108, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.10
Mobile Learning:
Learning on the Go

Steve Chi-Yin Yuen
The Univeristy of Southern Mississippi, USA

Patrivan K. Yuen
William Carey University, USA

INTRODUCTION

The mobile revolution is finally here. The evidence of mobile penetration and adoption is irrefutable: smartphones, personal digital assistants (PDAs), portable game devices, portable media players, MP3 and MP4 players, tablet PCs, and laptops abound and can be found everywhere. Also, the increasing availability of high-bandwidth network infrastructures and advances in wireless technologies have opened up new accessibility opportunities (Kinshuk, 2003). No demographic is immune from this phenomenon. People from all walks of life and in all age groups are increasingly connected and communicate electronically with each other nearly everywhere they go (Wagner, 2005). The development of and adoption rate of mobile technologies are advancing rapidly on a global scale (Brown, 2005). Since 2000, there is considerable interest from educators and technical developers in exploiting the universal appeal and unique capabilities of mobile technologies for the use in education and training settings (Naismith, Lonsdale, Vavoula, & Sharples, 2004).

The use of mobile technologies to support, enhance, and improve access to learning is a relatively new idea and many learners are quite comfortable with various mobile devices. M-learning (mobile learning) is consequently an emerging concept as educators are beginning to explore more with mobile technologies in teaching and learning environments. Already, there are numerous applications for mobile technologies in education—from the ability to transmit learning modules and administrative data wirelessly, to enabling learners to communicate with instructors and peers "on-the-go" (Brown, 2005).

Still in its early stages, m-learning is comparable to where e-learning was a few years ago. M-learning is at the point by which mobile computing and e-learning intersect to produce an anytime, anywhere learning experience. Advances in mobile technologies have enhanced m-learning tools at just the right moment to meet the need for more cost-effective just-in-time training op-

tions—Learning on the Go. Today, the evidence is overwhelming that m-learning is beginning to take hold:

- The population of mobile and remote access workers in the United States alone will grow to 55.4 million by 2004 (Shepherd, 2001).
- Over 50% of all employees spend up to half of their time outside the office.
- The average employee had less than three days of training in 2003.
- There will be more than 1 billion wireless Internet subscribers worldwide by 2005.
- Multipurpose handheld devices (PDAs and telephones) will outsell laptop/desktop computers combined by 2005.
- Most major U.S. companies will either switch to or adopt wireless networks by 2008 (Ellis, 2003).
- More than 1.5 billion mobile phones are used in the world today. This is more than three times the number of personal computers, and today's sophisticated phones have the processing power of a mid-1990s personal computer (Attewell, 2005; Prensky, 2004).
- Smartphones rose by 17% year-on-year in the first part of 2005 in Europe and the Middle East. In contrast, standard mobile phones rose by only 11% (Canalys, 2005).
- Global sales of smart phones will reach 170 million in 4 to 5 years, compared slightly more than 20 million in 2004 (Attewell, 2005).
- More than 16 million 3G phones were sold worldwide in the beginning of 2005, compared to only 10 million 3G handsets sold in September 2004.
- Total U.S. spending on wireless communications will grow 9.3% in 2005, to $158.6 billion.
- The wireless market will grow at 10% compound annual growth rate through 2008 (Wagner, 2005).

While mobile devices are approaching ubiquity today, the industry is still in its infancy. Fusing mobile technology and e-learning is very natural. Mobile devices are a natural extension of e-learning because mobile devices have the power to make learning even more widely available and accessible. Imagine the power of learning that is truly "just-in-time," where learners could actually access training at the precise place and time on the job when needed (Kossen, 2001).

BACKGROUND

Conventional e-learning, delivered to a desktop computer, is leaving a large part of the learners out in the cold. As Elliott Masie (Shepherd, 2001, p. 1) points out:

The assumption here is to dramatically expand the accessibility of learning beyond the physical footprint of the PC. If we remember that over 50% of the workforce does not sit at a desk, but instead is standing, walking or moving around a factory, we see the potential of breaking the tether of the Ethernet wire.

M-learning is designed to fit with the unique work-style requirements of the mobile workforce, linked to their office by mobile devices.

Vavoula and Sharples (2002) suggest three ways in which learning can be considered mobile: (a) learning is mobile in terms of space, (b) learning is mobile in different areas of life, and (c) learning is mobile with respect to time. Their definition suggests that m-learning systems are capable of delivering educational content anywhere and anytime the learners need it.

According to Quinn (2000), m-learning is the intersection of mobile computing and e-learning. M-learning includes anytime, anywhere resources, strong search capabilities, rich interaction, powerful support for effective learning, and performance-based assessment. Chabra and

Figueiredo defined m-learning as "the ability to receive learning anytime, anywhere and on any device," while Harris referred m-learning to "the point at which mobile computing and eLearning intersect to produce an anytime, anywhere learning experience" (Dye, K'Odingo, & Solstad, 2003, p. 6).

Commonly, m-learning refers to learning opportunities through the use of mobile solutions and handheld devices (i.e., mobile phones, smartphones, and PDAs) which are connected to information networks. Mobile implies movement and mobility. Likewise, m-learning implies the opportunity to learn "on the go" (Vanska, 2004). M-learning can be an educational environment in which wireless technology is used to assist students in their studies—both inside and outside the classroom. In a mobile learning scenario, students can access their learning materials from anywhere: on the bus, at the cafeteria, or waiting in line. Also, students can easily contact fellow students, check e-mail, or get feedback from their instructors. Unlike being limited to working online in a computer lab, the library, or at home, students can access online materials regardless of their location. M-learning translates to flexibility in accessing course materials, fellow students, and their instructor anytime, anywhere.

Evans (2005), at the Think-Tank Day for the UK mobile learning community, identified several unique features of mobile devices which could enhance the learning experience:

- **Privacy:** The small size of mobile devices makes it possible to learn "unobtrusively" whenever the learner is located.
- **Support for learning styles:** The mobile devices have potential to support learners with preferences for textual, audio and video presentation of material.
- **Immersive:** The richness and diversity of both content and activity can immerse the learners in their experience.
- **Capture of data:** The mobile devices allow the capture of data anywhere and analyze later.
- **Context:** The ability to automatically receive relevant information.
- **User control:** Learners have more control over when and where they choose to study, and over their interaction with other learners.

In his book *The Future of Learning: From E-Learning to M-Learning*, Keegan (2002) discusses the progression of types of learning from distance, to electronic, to mobile. He indicates that the logical extension of PC-based distance learning is mobile learning. He analyzes about 30 global m-learning initiatives regarding to the experimental use of wireless technologies (including wireless Internet environments and wireless classrooms) and various mobile devices for teaching and learning. He concludes regarding the emergence and growing importance of m-learning. M-Learning is the logical extension of asynchronous learning, available not only anytime, but also anywhere.

Many educators and trainers are optimistic about the potentials of m-learning. Wagner (2005) believes that m-learning represents the next step in a long tradition of technology-mediated learning. M-learning will employ new learning strategies, practices, tools, applications, and resources to realize the promise of ubiquitous, pervasive, personal, and connected learning. M-learning connects formal education experience (i.e., taking a class, attending a workshop or seminar, or participating a training session) with informal, situated learning experience (i.e., learning on the go while riding the bus, waiting for a flight in an airport, or receiving performance support while on the job). Wagner further states that m-learning will be built upon the foundations of previous educational technology frameworks (i.e., distance learning, e-learning, flexible learning,

modular instructional design, learning and content management), and thus can take full advantage of the experiences, empirical evidence, and effective practice guidelines derived by researchers and practitioners from the preceding technology revolutions in education (Wagner, 2005).

MOBILE TECHNOLOGIES AND NEW LEARNING PARADIGMS

There is no theory of mobile learning. However, m-learning supports a new dimension in the educational process. In the review of new learning and teaching practices, Sharples (2003) concludes the following:

- Learning involves constructing understanding. Learners use their knowledge to construct new knowledge.
- Learning takes place within a community of practice and not only in the classroom or in form of the computer.
- Learning starts from conversation—with oneself and with others. Learning is part of collaborative processes in professional, educational, and daily-life settings.
- Problems provide resources for learning.
- Learning is part of daily living. Learning is dependent on the situation—physical as well as emotional—that it takes place in.
- Learning is lifelong. It takes place over a long period of time and beyond formal education.

Similarly, Ferscha (2002) summarized the new learning paradigms as: (a) individual/ learner centered, (b) collaborative learning, (c) situated learning, (d) contextual learning, (e) ubiquitous, and (f) lifelong. In the review of literature concerning new learning and teaching practices and mobile technologies, Naismith et al. (2004) reveal six learning theories and areas of learning relevant to mobile technologies:

1. **Behaviorist learning:** Learning activities that promote learning as a change in observable actions. Mobile technologies provide the ideal opportunity to present content, gather responses, and provide appropriate feedback.
2. **Constructivist learning:** Learning activities in which learners actively construct new ideas or concepts based on both their previous and current knowledge. Mobile devices provide unique opportunities to transform learners from passive recipients of information to active constructors of knowledge.
3. **Situated learning:** Learning activities that promote learning within an authentic context and culture. The portability of mobile devices allows the learning environment to be extended beyond the classroom into authentic and appropriate contexts of use.
4. **Collaborative learning:** Learning activities that promote learning through social interaction. Mobile devices enable learners to share data, files, and messages and provide means of coordination without attempting to replace human-human interactions.
5. **Informal and lifelong:** Learning activities that support learning outside a dedicated learning environment and formal curriculum. Mobile devices with small size and ease of use make them well suited for learning applications outside of formal education.
6. **Learning and teaching support:** Activities that assist in the coordination of learners and resources for learning activities. Mobile devices can be used to support learning-related activities for students, teachers, and administrators (Naismith et al., 2004).

WHY MOBILE LEARNING?

According to Brown (2003), m-learning is a natural extension of e-learning. It has the potential to further expand where, how, and when we learn

and perform in all the aspects of our life. One of the key benefits of m-learning is its potential for increasing productivity by making learning available anywhere and anytime, allowing learners to participate in educational activities without the restrictions of time and place. Mobile technologies have the power to make learning even more widely available and accessible than we are used to in existing e-learning environments. M-learning could be the first step towards learning that is truly just-in-time where learners could actually access education and training at the place and time that they need it. Brown (2003) further states that integrating electronic performance support systems (EPSS) into the mobile environment will take m-learning even further: m-learning with on-demand access to information, tools, learning feedback, advice, support, learning materials, and so forth.

Mobile technologies can support and monitor student learning activity in real time outside the traditional classroom and promote a learning community. They can help students access learning records, register attendance, access media rich learning materials, collaborate with other learners, and keep in touch with teachers and mentors. Furthermore, mobile devices can support and facilitate learning assessment and the creation of portfolios (Evans, 2005). In addition, there are many other benefits of m-learning:

- Inform learners that training is available—just in time vs. just in case.
- Minimize barriers that prevent people from accessing training when they need it.
- Enable organizations to be more responsive to changes in the environment.
- Provide compelling, personalized, on-demand learning.
- Provide real world skills.
- Offer just-in-time learning/reference tool for quick access to data in the field.
- Provide rich interaction with others.
- Offer increased opportunities for students to research by accessing electronic resources (Evans, 2005).

The findings of the m-learning project, funded by the European Commission's Information Society Technologies (IST) initiative, indicate that mobile devices can be used successfully to involve some of the hardest to reach and most disadvantaged young adults in learning. M-learning has the potential to help these youngsters improve both their skills and their self-confidence (Attewell, 2005). Furthermore, Attewell concludes that the use of m-learning may have a positive contribution in the following areas:

- M-learning helps learners improve their basic skills.
- M-learning can be used to encourage both independent and collaborative learning experiences.
- M-learning helps learners identify areas where they need assistance and support.
- M-learning helps bridge the gap between mobile phone literacy and information and communication technology (ICT) literacy.
- M-learning helps learners engage in learning and maintain their interest levels.
- M-learning helps learners remain more focused for longer periods.
- M-learning helps raise learners' self-esteem and self-confidence (Attewell, 2005).

In a survey of expert expectations about m-learning conducted in Germany, Switzerland, and Austria in 2005, Kuszpa finds that a time and place-independent learning alternative is the greatest advantage of m-learning. Also, a learner can individually control his/her speed of learning during the use of mobile devices is considered a strong advantage. However, the greatest disadvantage of m-learning is seen in the need for a higher self-discipline when learning on mobile devices. Furthermore, the majority of experts

participated in the survey feel m-learning is an impersonal way of learning. They criticize the small displays and limited input possibilities on mobile devices that give little space for a good presentation of the learning content (Kuszpa, 2005). Another problem for m-learning is the lack of a standardized platform. The current mobile devices utilize a variety operating environments, display and sound characteristics, and input devices, making it difficult to develop educational content that will work anywhere for every mobile device (Shepherd, 2001).

Mobile devices are getting smaller and more powerful. They have the ability to deliver learning objects and provide access to online systems and services. However, network infrastructure has not quite kept up with the development of mobile hardware. As a result, bandwidth is not yet sufficient for substantial m-learning and coverage, and signal problems are still barriers in many areas when traveling. Attewell suggests a mixture of online learning and learning using materials downloaded onto mobile devices for use off-line is necessary. In addition, due to the immature mobile standards, it is a challenge for educators to develop and implement mobile learning projects. It is almost impossible to develop one generic version of mobile applications to run on all mobile platforms. As a result, educators often develop several versions of learning materials specifically for particular platforms (Attewell, 2005). To support flexible learning requirements of m-learning, solutions are needed that not only support m-learning but also develop frameworks that support automatic adaptation of educational content to suit various mobile devices and individual preferences of the learners using those devices (Kinshuk, 2003).

FUTURE TRENDS

According to Wagner (2005), current trends suggest that educational games, language instruction, and performance-support and decision-support tools are likely to lead the mobile movement in the next few years. Particularly, wireless games have taken the world by storm. There are 170 million wireless games worldwide. Eighteen million Americans play wireless games and 6 million users download games to their mobile device each month in the U.S. It is very possible that educational games will provide m-learning with its first success in wide-spread adoption in education.

The future mobile devices will be even more embedded, ubiquitous, and networked than those available today. The capabilities of mobile phones, PDAs, games consoles, and cameras will likely merge within the next few years to provide a networked, personal, portable, and multimedia device that is always with the user. According to Naismith et al. (2004), future mobile technologies will have a greater impact on learning. Learning will move more and more outside of the classroom and into both real and virtual learner's environments. Learning will involve making rich connections within these environments to both resources and to other people (Naismith et al., 2004). In addition to accessing Internet resources on the move, learners will be able to manage their learning through consultations with their personal diaries and institution-based virtual learning environments. The ability to instantly publish learners' observations and reflections as digital media will empower them to be researchers. Context-aware applications will enable learners to easily capture and record events in their lives to both assist later recall and share their experiences for collaborative reflection. Opportunities for distributed collaboration and mobile team working will be greatly enhanced (Naismith, et al., 2004).

CONCLUSION

With the immense penetration and the continuously increasing capabilities of mobile devices,

there is a great potential of m-learning in education and training. Learning everywhere and anytime can be a valuable complement, but definitely is not a replacement for traditional learning methods (Kuszpa, 2005). M-learning just provides another way of learning using new mobile technology. As educators, we should embrace the rich learning enhancing possibilities that m-learning already provides and will provide even more so in the future. M-learning fulfils the growing demands for life-long learning opportunities that enable learners to "learn while you are on the go" (Brown, 2005).

M-learning allows truly anywhere, anytime, personalized learning. It can also be used to enrich, enliven, or add variety to conventional lessons or courses. However, the challenge of m-learning is to take advantage of the special needs of mobile learners and the unique characteristics of the mobile devices they use, and to provide an improved m-learning service along with other learning systems (Shepherd, 2001). The success of m-learning does not solely depend on the technological developments and the possibilities they provide. Effective m-learning programs will require digital communication skills, new pedagogies, and new learning strategies and practices (Wagner, 2005). The ability of educators and instructional designers to develop m-learning activities that provide rich, collaborative, and conversational learning experience is imperative. Also, it is important to identify those applications of mobile technologies that contribute to the optimizing of teaching and learning in the new learning environments.

REFERENCES

Attewell, J. (2005). *Mobile technologies and learning.* London, UK: Learning and Skills Development Agency.

Brown, T.H. (2003). *The role of m-learning in the future of e-learning in Africa?* Retrieved September 22, 2007, from http://www.tml.hut.fi/Opinnot/T-110.556/2004/Materiaali/brown03.pdf

Brown, T. (2005). *Towards a model for m-learning in Africa. International Journal on E-Learning, 4*(3), 299-315. Norfolk, VA: AACE.

Canalys (2005). *Changing times in the smart mobile deice market* (Canalys press release). Retrieved September 22, 2007, from http://www.canalys.com/pr/2005/r2005094.htm

Dye, A., K'Odingo, J.A., & Solstad, B. (2003). *Mobile education - a glance at the future.* Retrieved September 22, 2007, from http://www.dye.no/articles/a_glance_at_the_future/

Ellis, K. (2003). *Moving into m-learning. Training, 40*(10), 56-59.

Evans, D. (2005). *Potential uses of wireless and mobile learning.* Retrieved September 22, 2007, from http://www.jisc.ac.uk/elearning_innovation.html

Ferscha, A. (2002). *Wireless learning network.* Grundlagenkonferenz e-elearning, Wien.

Keegan, D. (2002). *The future of learning: From e-learning to m-learning.* Hagen, Germany: Zentrales Institut für Fernstudienforschung. (ERIC Document Reproduction Service No. ED472435)

Kinshuk. (2003). *Adaptive mobile learning technologies.* Retrieved September 22, 2007, from http://www.globaled.com/articles/Kinshuk2003.pdf

Kossen, J.S. (2001). *When e-learning becomes m-learning.* Retrieved September 22, 2007, from http://www.palmpowerenterprise.com/issues/issue200106/elearning001.html

Kuszpa, M. (2005). The future of mobile learning – a survey of expert expectations about learning on mobile phones. In *Online Educa 2005, Book of Abstracts,* Berlin, Germany.

Naismith, L., Lonsdale, P., Vavoula, G., & Sharples, M. (2004). *NESTA futurelab series, report 11: Literature review in mobile technologies and learning.* Bristol, UK: Nesta Futurelab.

Prensky, M. (2004). *What can you learn from a cell phone? -- almost anything.* Retrieved September 22, 2007, from http://www.marcprensky.com/writing/Prensky-What_Can_You_Learn_From_a_Cell_Phone-FINAL.pdf

Quin, C. (2000). *M-Learning: Mobile, wireless, in-your-pocket learning.* Retrieved September 22, 2007, from http://www.linezine.com/2.1/features/cqmmwiyp.htm

Sharples, M. (2003, June). *Portable and mobile educational technology research.* Paper presented a the BECTA Expert Technology Seminar, London.

Shepherd, C. (2001). *M is for maybe*. Retrieved September 22, 2007, from http://www.fastrak-consulting.co.uk/tactix/features/mlearning.htm

Vanska, R. K. (2004). *Mobile learning in Europe: A multidisciplinary approach.* Retrieved September 22, 2007, from http://db.kmkg.de/cgi-bin/congress/course.pl?language=1&eve_id=25&cou_id=1754

Vavoula, G., & Sharples, M. (2002, August 29-30). KLeOS: A person, mobile, knowledge and learning organization system. In M. Mirad, U. Hoppe & Kinshuk (Eds.), *Proceedings of the IEEE International Workshop on Mobile and Wireless Technologies in Education [WMTE 2002],* Vaxjo, Sweden, (pp. 152-156).

Wagner, E.D. (2005). *Enabling mobile learning.* Retrieved September 22, 2007, from http://www.educause.edu/er/erm05/erm0532.asp

KEY TERMS

3G: Third-generation mobile telephone technology. The 3G services provide the ability to transfer both voice data and non-voice data (music, videos, e-mail, and instant messaging) at the speed of up to two megabits per second.

4G: Fourth-generation mobile telephone technology. It is not yet available. 4G will be the successor to 3G and will feature high-speed mobile wireless access with a data transmission speed of up to 100 megabits per second.

Enhanced Data Rates for GSM Evolution (EDGE): EDGE is an upgrade of GRS system for data transfer in GSM networks. EDGE increases the capacity and quality and allows the use of advanced services over the existing GSM network.

General Packet Radio Service (GPRS): A mobile data service available to users of GSM mobile phones. It is often described as "2.5G," a technology between the second generation (2G) and third generation (3G) of mobile telephony. It provides moderate speed data transfer, "always on" data connections that are much faster than the traditional 9600 bps, by using unused TDMA channels in the GSM network.

M-Learning: A term that refers to the delivery of learning content via mobile devices including PDAs, cell phones, or other handheld devices. It allows users to learn what they want, where they want, and when they want.

Multimedia Messaging System (MMS): The successor to SMS. MMS allows subscribers to send multimedia (digital photos, audio, and video) material along with their messages.

Short Message Service (SMS): A digital mobile phone service that allows single short messages of up to 160 characters to be passed between mobile phones, fax machines, or e-mail addresses.

Smartphone: Smartphones are a hybrid of the functionality of PDAs and mobile phones. They usually provide a means of connecting to a desktop

or laptop to perform the same functions as a PDA docking and synchronization cradle.

WiMax: A standard based on IEEE 802.16. WiMax offers mobile devices with a wireless, direct connection to the Internet at the speeds of up to 75 megabits per second and over distances of several kilometers.

This work was previously published in the Encyclopedia of Information Technology Curriculum Integration, edited by L. Tomei, pp. 580-586, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.11
Environments for Mobile Learning

Han-Chieh Chao
National Dong Hwa University, Taiwan, R.O.C.

Tin-Yu Wu
National Dong Hwa University, Taiwan, R.O.C.

Michelle T.C. Kao
National Dong Hwa University, Taiwan, R.O.C.

INTRODUCTION

As we enter the electronic age, technologies enabling e-learning have increased flexibility of learning location. Wireless communication technologies further increase the options for learning location (Johnson & Maltz, 1996; Wu, Huang & Chao, 2004). Advances in wireless communication technologies have recently provided the opportunity for educators to create new educational models. With the aid of wireless communication technology, educational practice can be embedded into mobile life without wired-based communication. With the trend of the educational media becoming more mobilized, portable and individualized, the learning form is being modified spectacularly. The mobile learning environment possesses many unique characteristics (Chen, Kao & Sheu, 2003):

- **Urgency of learning need:** The wireless applications can be used for an urgent matter of learning, such as linking problem solving and knowledge. Otherwise, the learner may record the questions and look for the answer later in the library, on houseline with a computer or from the experts.
- **Initiative of knowledge acquisition:** Frequently, information provided by wireless applications are based on the learners' requests; that is, information on demand. Being based on the learners' requests, together with the help of current state-of-the-art I/O devices, such as Radio Frequency

Identification (RFID), Voice Extensible Markup Language (VXML) and so forth (Page, 1993; Andersson, 2001), interactive personal information can be communicated between learners and the databases so that the wireless application can provide closely related information in time and in need.

- **Mobility of learning setting:** Wireless devices are developed to be more and more portable. Therefore, the educational practice can be performed at any time and any place and always on, such as on a tour bus, camping area, exhibit room, and so forth. All kinds of field trip situations can be facilitated. This kind of learning setting can be preplanned or be opportunistic in nature.
- **Interactivity of the learning process:** Through the interfaces of voices, pointing, mails, icons, even videos, the learner can communicate with experts, peers or other materials effectively in the form of synchronous or asynchronous communication. Hence, the expert is more reachable and the knowledge is more available.
- **Situating of instructional activity:** Via wireless applications, the learning could be embedded in daily life. The problems encountered, as well as the knowledge required, are all presented in authentic context, which helps learners notice the features of problem situations that make particular actions relevant.
- **Integration of instructional content:** The wireless learning environment integrates many information resources, and supports learners to do non-linear, multidimensional and flexible learning and thinking. It especially facilitates complex and ill-structured learning content, such as cross-subject, theme-based learning activities.

WIRELESS TECHNOLOGY

Next-generation wireless networks (2.5G, 3G, B3G, 4G) offer the promise of high-speed access to mobile hosts along with IP-based data services, the General Packet Radio Service (GPRS) communication network that can transmit data and speech sounds at the same time with limited bandwidth and third generation of mobility communication network (Khan, 2001). The powerful third-generation mobility network, 3G, has much larger wireless bandwidth capabilities and more multi-media services than the Global System for Mobile Communications (GSM)/GPRS cellphone system. 3G features a bandwidth of 2M bits/second when users are motionless, a bandwidth of 384k bits/second when users move in a low speed, and a bandwidth of 144K bits/second when users move at a high speed (Andersson, 2001). More than that, GSM can combine Wireless Local Area Network (WLAN) to accomplish a double network with WLAN and a cellular network (Wang, 2001). The bandwidth offered by a double network with WLAN and a cellular network enables learners to enjoy all kinds of service on the Internet, anytime, anywhere.

While these technologies are enabling mobile e-learning options, there are problems, including bandwidth, Internet Protocol (IP) and roaming limitations. Bandwidth problems can be solved simply by Internet Service Providers (ISPs) developing the backbone of broader bandwidth.

NEXT GENERATION INTERNET PROTOCOL—IPv6

Providing enough IP addresses for worldwide use is presenting challenges, given the limitations of the current Internet Protocol, version 4 (IPv4). With universal access and use of the Internet, IP has to offer the capability for worldwide use of Internet resources. In the early 1990s, the

Internet Engineering Task Force (IETF) had already identified difficulties with maintaining the Internet with IPv4. With a global population of more than six billion, the 42 million possible IP addresses are insufficient to meet the needs of users needing one or more IP addresses. The next generation of IP, Internet Protocol version 6 (IPv6) allows 5.4×10^{28}, more than enough for everyone (Deering & Hinden, 1995). While IPv6 has been around for a number of years, it is not yet widely adopted. In the meantime, dynamic allocation of IPv4 addresses for mobile users increases the number of possible addresses and may be sufficient for the near future. However, not only have Europe and Asia Pacific put great attention on IPv6, but the United States (U.S.) Department of Defense (DoD) announced in June 2003 that it will convert all of its systems, networks and applications to IPv6 by 2008 (French, 2003; DoD CIO memo, 2003). This action raises the interest of the U.S. Department of Commerce (DoC) and Department of Homeland Security, which led to the DoC Request for Comments on Deployment of Internet Protocol, version 6 in January 2004. U.S. commercial companies foresaw potentially billions of dollars in upgrades and have started to bring out all kinds of IPv6 products, such as, routers, switches and so forth.

Besides the abundance of addresses, IPv6 carries some other advantages: IPv4 can not enable the speed and efficient application processing required for e-learning. These problems are being solved by Mobile IPv6 (Chao & Chu, 2001; Chao & Huang, 2003; Chao & Chu, 2003). IPv6 uses a new method for transmission, Anycast Protocol (Doi, 2004), which differs from the current Unicast and Multicast Protocol because it transmits packets using sophisticated metrics for finding the least-delay time path, route with the lowest price or less routing hop. At the network level, Anycast determines load balance and uses techniques related to the Domain Name Server (DNS) to transmit information. Adopting Anycast will enable learners to get connected to the Internet and utilize e-learning applications in a much faster and more efficient way.

Internet Protocol Security (IPSec) is widely used in IPv6, and those features are embedded in IPv6's extension headers so that end-to-end security can be achieved without interruption. IPv6 resolves some of the Internet security problems that can detract from e-learning applications (Huang, 2000; Arkko, 2004). IPSec utilizes two security protocol, Encapsulating Security Payload (ESP) and Authentication Header (AH) to enable confidentiality, information integrity, identity of package sources, access control, protection of replay and traffic flow confidentiality. The Encryption Algorithms and Conformance Requirements of Hash Algorithms method, Internet Key Exchange (IKE), Security Association (SA) and others enable learner privacy. Moreover, the providers of mobile learning can use the identification offered by IPSec and the Authentication Authorization Accounting (AAA) function to authenticate users and allocate functions (Wang, Chen, & Chao, 2004).

CONCLUSION

Mobile computing using IPv6 allows applications such as Next Generation Learning Environment (NeGL) to set up learning systems, identification, personal learning record, personal preference record and personal learning information management seamlessly. It offers learners the opportunities to use all kinds of Mobile Node or anything that can connect to an Internet learning equipment system to be accessed by using ALL-IP communication networks. Besides, for providers of digital content, they can follow the Shareable Content Object Reference Model (SCORM) to compose information (Bohl, 2002). In the future, problems like the compatibility between digital content and related facilities will no longer occur.

As you can imagine, the condition of the learning mode in the future will be like an international, immediate and interactive classroom that enables learners to learn and interact.

REFERENCES

Andersson, C. (2001). *GPRS and 3G wireless applications: Professional developer's guide*. John Wiley & Sons.

Andersson, E. (2001). VoiceXML: Letting people talk to your HTTP server through the telephone. *ArsDigita Systems Journal.* Retrieved March 5, 2001 from *www.eveandersson.com/arsdigita/asj/vxml*

Arkko, J., Devarapalli, V., & Dupont, F. (2004). *Using IPsec to protect mobile Ipv6 signaling between mobile nodes and home agent.* RFC 3776.

Bohl, O., Scheuhase, J., Sengler, R., & Winand, U. (2002). The sharable content object reference model (SCORM)—a critical review. *International Conference on Computers in Education, 2*, 950-951.

Chao, H.C., & Chu, Y.M. (2001). Seamless supports for the mobile Internet Protocol based cellular environments. *International Journal of Wireless Information Networks, 8*(3), 133-153.

Chao, H.C., & Huang, C.Y. (2003). A micro mobility mechanism for smooth handoffs in an integrated ad-hoc and cellular IPv6 network under high speed movement. *IEEE Transactions on Vehicular Technology, 52*(6), 1576-1593.

Chao, H.C., & Chu, Y.M. (2003). An architecture and communication protocol for IPv6 packet-based picocellular networks. *Journal on Special Topics in Mobile Networking and Applications, 8*(6), 663-674.

Chen, Y.S., Kao, T.C., & Sheu, J.P. (2003). A mobile learning system for scaffolding bird Watching learning. Journal of Computer Assisted Learning *(special issue on Wireless and Mobile Technologies in Education), 19*(3), 347-359.

Deering, S., & Hinden, J. (1995). *Internet Protocol, version 6 (IPv6) specification.* RFC 1883.

DOD CIO Memo (2003). Internet Protocol version 6 (IPv6) dtd, June 9, 2003, *Memorandum by John P. Stenbit.*

Doi, S., Ata, S., Kitamura, H., & Murata, M. (2004). IPv6 Anycast for simple and effective service—oriented communications. *IEEE Communication Magazine, 42*(5), 163-171.

French, M. (2003). Military sees network benefits from IPv6. Retrieved December 9, 2003 from *FCW.COM*

Huang, H., & Ma, J. (2000). IPv6—future approval networking. *International Conference on Communication Technology Proceedings, 2*(21-25), 1734-1739.

Johnson, D.B., & Maltz, D.A. (1996). Dynamic source routing in ad hoc wireless networks. In T. Imielinski & H. Korth (Eds.), *Mobile Computing* (pp. 81-153). Norwell, MA: Kluwer Academic Publishers.

Khan, J. (2001). Introduction to 3G/4G wireless network architectures. *IEEE International Symposium on Circuits and Systems Tutorial guide*, 7.1.1-7.1.13.

Page, R. (1993). A low power RFID transponder. *RF Design*, 31-34.

Wang, J. (2001). *Broadband wireless communications: 3G, 4G and wireless LAN.* Norwell: Kluwer Academic Publishers.

Wang, R.C., Chen, R.Y., & Chao, H.C. (2004). AAA architecture for mobile IPv6 based on WLAN. *International Journal of Network Management, 14*(5), 305-313.

Wu, T.Y., Chao, H.C., & Huang, C.Y. (2004). A survey of mobile IP in cellular and mobile ad-hoc network environments. To appear in the *Ad Hoc Networks Journal*.

This work was previously published in the Encyclopedia of Distance Learning, Vol. 2, edited by C. Howard, J. Boettcher, L. Justice, K. Schenk, P. Rogers, and G. Berg, pp. 853-856, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.12
Mobile Education:
Lessons Learned

Holger Nösekabel
University of Passau, Germany

ABSTRACT

Mobile education, comprising learning, teaching, and education related administrative services delivered via mobile technologies, has incited several projects and discussion in the last years. When reviewing these projects, it becomes apparent that most of them are technology driven, and only a few were formally evaluated at the end. However, certain lessons, chances and obstacles can be identified which may be helpful for further development in this sector. One critical issue is the distribution of costs for mobile services. As both educational institutions and students act on a limited budget, it is necessary to choose an infrastructure which meets the requirements of the users and addresses all relevant obstacles. Consequently, there is no single ideal technological alternative, but each project needs to make a situational choice.

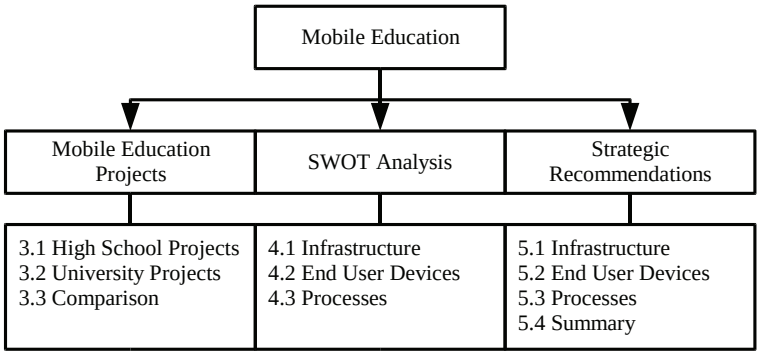
INTRODUCTION

Technological progress continually creates new opportunities for creating, storing, and disseminating knowledge. One aspect is the utilization of new technologies for learning and teaching: e-learning. Recent endorsements in this sector were *mobile devices*, which can increase mobility, flexibility, and personalization compared to traditional, PC-based approaches. The term "e-learning" was thus extended to "*m-learning*," or "*mobile education*."

Mobile education covers three distinct but interconnected areas in which mobile devices may be implemented: learning, teaching, and administration. A major focus in the past was placed on learning activities, mobile learning or m-learning, as the term itself was derived from e-learning. Teaching and administrative tasks were either omitted or understood as learning tasks. Consequently, m-learning can either be understood as a subclass of e-learning or as a distinct area of research (Nösekabel, 2005).

After establishing a framework by clarifying what will be termed as mobile education in this chapter, a survey of m-education projects establishes the state of the art. Selected projects are grouped into high school and university projects, as the didactic requirements for these

Figure 1. Chapter overview



educational institutions are different. Universities, for example, allow their students a higher degree of self-determination and self-direction in learning.

The results of this comparison are compiled into a *SWOT analysis*, which is used to point out experiences, obstacles, and chances for existing and future mobile education projects. Both the analysis and the strategic recommendation focus on mobile infrastructures, end user devices, and educational processes.

MOBILE EDUCATION

Defining *m-education* is the focus of the following discussion, which helps identify relevant projects which are then analysed. First, a restriction should be placed on the devices used for educational purposes. Devices need to be mobile, as stated by Lyytinen and Yoo (2002), which means they must have a high degree of mobility but only a low degree of embeddedness. This would include mobile phones, personal digital assistants, and other devices (e.g., MP3 players), but excludes laptops, as laptops are only portable and cannot be used easily while in motion. Another factor is that laptops do not have the same technical restrictions as mobile devices; thus, services and experiences from e-learning are mostly applicable and do not require a new view on these issues. The restriction also excludes pervasive and ubiquitous devices (Dourish, 2001), which are both highly embedded and could be subject to research in "pervasive" or "ubiquitous education."

Second, m-education addresses—as already mentioned—learning, teaching, and administration, affecting not only students, but lecturers and possibly administrative staff alike. One result for the selection of projects is that so called "*classroom applications*" (Myers, 2001) using mobile devices are also included in the survey. These "*classroom applications*" run on mobile devices, often in combination with a non-mobile PC or laptop. They foster interaction between students and teachers, for example, by offering the ability to conduct polls or to remotely annotate presentation slides as a group.

Third, a network connection is not permanently required when using mobile education. This allows the inclusion of applications where data are transmitted to a mobile device via a stationary PC, for example, during synchronisation. Further included are Java applets (J2ME MIDlets) on mobile devices, which possibly make use of a network connection only during installation over the air or during data transmissions.

These various aspects are covered by several definitions, even though most authors define "Mobile Learning," not "Mobile Education." Nyiri,

for example, reasons that the primary purpose for mobile devices is interpersonal communication, and, therefore m-learning is learning *"[...] as it arises in the course of person-to-person mobile communication"* (2002, p. 123). Communication aspects are also mentioned by Hummel and Hlavacs (2003). Constructivistic learning theories emphasise that communication is a decisive element in the learning process, supporting this kind of definition. However, there are mobile devices which are designed for *personal information management* (*personal digital assistants – PDAs*) and coordination. Clarke and Flaherty (2002) therefore include the aspect of collaboration, defining m-learning as *"[...] an approach to teaching and learning that utilizes wireless technologies to communicate and collaborate in an educational context"* (p. 68). Sharma and Kitchens (2004) additionally include the notion of context, stating that learning content should be specifically prepared for a learner in a given situation.

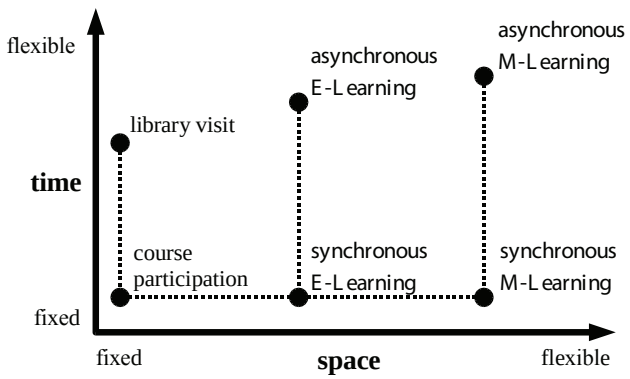
Lehner, Nösekabel, and Lehmann follow a more technical approach and define mobile education as *"[...] any service or facility that supplies a learner with general electronic information and educational content that aids in the acquisition of knowledge, regardless of location and time"* (2004, p. 24).

Introducing mobile technologies into learning-related processes primarily results in an increased flexibility for both learners and instructors. Depending on the underlying learning theories, support for a learning process can be assumed. Mobile behaviouristic learning, for example, allows the learning loop of information presentation, question, answer, analysis, and feedback to be initiated at any time and in any place. Constructivistic approaches, on the other hand, benefit from increased spatial flexibility, which allows teachers to employ real-life locations and scenarios in the learning process.

The effect of technology on learning with regard to flexibility is shown in Figure 2. In a space-time matrix, where the amount of flexibility is differentiated, physical participation in a course or class always requires a person to be at a specific place at a specific time. With synchronous learning, for example by transmitting the course via video, students no longer are required to be at a certain place. Thus, spatial flexibility is gained. However, non-mobile computing still has requirements regarding network connectivity, power supply, and physical transportation of hardware equipment. As a result, synchronous e-learning is not as flexible as synchronous m-learning.

Learning can also take place outside predetermined times, for example when students visit the library. In this case, the place is fixed (the

Figure 2. Comparison of non-electronic, electronic, and mobile learning



physical location of the library), but the time is more flexible compared to a course participation. Restrictions, such as opening times, may still exist, prohibiting greater flexibility. Electronic resources, such as e-books, recorded lectures, or other learning materials, can be accessed with a PC without these restrictions. Again, however, a non-mobile PC must be present, which is not always the case. Mobile technologies, on the other hand, provide greater time and space flexibility, as these devices can be carried around and are readily available.

To demonstrate the use of mobile technologies in educational processes, a set of exemplary *educational tasks* will serve as a background for discussion, and key elements will be extended in the SWOT analysis. Table 1 lists positive and negative aspects when supporting certain tasks with mobile technologies. In order to overcome limitations or problems, enabling factors may provide solutions.

One typical learning task is using various learning materials: presentation slides, audio or video recordings, images, or text. With mobile technologies, these materials can be accessed without spatial or temporal restrictions, but currently mobile devices often lack the ability to display more complex materials satisfactorily. A short term solution to this problem would be to create learning materials suited for mobile device. In the long run, hardware advancements could lift these limitations. A second learning task is planning educational activities, where mobile device offer *personal information management* (PIM) services to manage contacts and appointments. Exchanging this kind of information, however, is sometimes hindered by proprietary

Table 1. Effect of mobile technologies for educational tasks

Educational task	Positive aspects	Negative aspects	Enabling factors
Access to learning material	Ubiquitous access to content anywhere, anytime; utilization of short free time periods	Device limitations (display size, computing power); bandwidth or memory requirements	Mobile hardware advancements; fast and inexpensive network access; dedicated mobile content
Planning of educational activities	Integrated *Personal Information Management* (PIM); group coordination and collaboration	Support for common data exchange format required; network connectivity	Conformity to standards; inexpensive network access
Giving lessons	New teaching methods possible (classroom applications); increased cooperation	Device limitations (display size, input methods); change to teaching process	Mobile hardware advancements; didactic integration of mobile services
Communication	Direct and immediate communication; ability to communicate when necessary	Many messages to few experts; information overload; additional communication channel	Communication etiquette; unified messaging to combine different communication channels
Activities regarding curriculum	Real-time access to information; individuality through personal device	Security risks when device is stolen or damaged; device limitation (display size, input methods)	Improved security technologies; mobile hardware advancements
Management of resources	*M-Payment* could allow monetary transactions; flexible initiation of transaction	Integration of backend management system required; security risks	Interoperability through open interfaces, data exchange, or middleware; *M-Payment*

implementations or interpretations of standardized data formats.

Giving lessons is a teaching task which can be supported with mobile technologies, for example by implementing *classroom applications* as described before. Since this requires changes to the teaching process, it could be met with resistance. Consequently, a didactic integration into existing processes is necessary. Another teaching task regards the communication between students and teachers. With mobile technologies, communication is direct and immediate (via SMS, MMS, mobile e-mail, instant messaging, etc.) and popular recipients, for example experts and teachers, may face a high number of incoming messages. Unified messaging and the adherence to communication etiquette can help to channel or reduce the information load.

Administrative tasks include, among others, activities concerning student curricula and resource management. With mobile technologies, a real-time, individualized access to information and a flexible initiation of transactions is possible. Both require a technical integration of backend systems and high standards regarding privacy and data security. Improvements in mobile security technologies and data exchange/middleware are thus enabling factors in this area.

Curtis, Williams, Norris, O'Leary, and Soloway (2003) list several positive effects which were observed in projects after introducing handheld computers in classrooms. Students had equitable access to IT, motivation increased, collaboration among students was easier, and learning environments could be individualized and organized.

In summary, mobile technologies can provide additional flexibility and increased motivation, although it should be noted that these effects are not compulsory and the results depend on the integration of mobile devices into existing activities. The technologies also offer an opportunity to enhance or alter educational processes. In the following section, projects in the area of mobile education will be presented to give an overview of similarities, differences, and trends.

MOBILE EDUCATION PROJECTS

For the survey, a total of 30 *mobile education projects* were analyzed regarding infrastructure, devices, and educational processes (Nösekabel, 2005; Lehner, Nösekabel, & Bremen, 2004; Lehner, Nösekabel, & Lehmann, 2004). Ten of these projects were initiated at high schools, and 20 of them at universities. For further information on these projects, consult the corresponding URL provided in Table 2.

A first differentiation separates the projects based on the implemented infrastructure: local transmission of files, for example via synchronization, wireless local access via WLAN, or wireless remote access via a cellular network. File transmissions include sending and receiving data and information as well as applications, which are installed on the device. These transmissions may occur with the help of a data cable, which connects the mobile device to a stationary PC, or with a wireless connection (an infrared interface or Bluetooth). WLAN, typically used to describe wireless networks based on IEEE 802.11 standards, may also include other local wireless technologies like Hiperlan, although IEEE 802.11b and 802.11g networks were implemented most often. Cellular networks encompass all digital mobile phone networks, for example GSM, HSCSD, GPRS, and IMT-2000.

A second differentiation determines whether the projects used PDAs, mobile phones or other mobile devices. PDAs, or handheld computers, are operated with a stylus on a touch sensitive display, or with a small hardware keyboard (Hansmann, Merk, Nicklous, & Stober, 2003). They can be equipped with WLAN and Bluetooth access. Mobile phones, on the other hand, have smaller displays, which are not touch sensitive, and some models are equipped with an alphanumerical

Table 2. Project information URLs

Project	URL for information
Ballard High School	http://ballard.seattleschool.org/
Consolidated Highschool District 230	http://www.d230.org/Handheld/default.htm
Hartland Farms Intermediate	http://www.oetc.org/handhelds.html
King Middle School	http://www.pcmag.com/article2/0,4149,15154,00.asp
Gymnasium Landau a. d. Isar	http://www.gymnasium-landau.de/wissen/journada/jo-ziele.htm
MobiSkoolz	http://www.mobiwave.com/
Palm Education Pioneers	http://www.palmgrants.sri.com/
RAFT Project	http://www.raft-project.net/
Anglia Polytechnical University	http://www.ultralab.ac.uk/
Berlin Univ. der Künste	http://www.campus-mobile.de/
Wirtschaftsuniv. Wien	http://nm.wu-wien.ac.at/palm.shtml
Carnegie Mellon University	http://www.cmu.edu/computing/handheld/index.html
Cornell University	http://mobile.mannlib.cornell.edu/
East Carolina University	http://www.ecu.edu/handheld/
Harvard Medical School	http://www.theanswerpage.com/
Kentucky Migrant Project	http://www.migrant.org/
Purdue University	http://www.purdue.edu/UNS/html4ever/010724.Chan.pda.html
Stanford School of Medicine	http://mednews.stanford.edu/stanmed/2001fall/mobilemed.html
Stanford University	http://palm.stanford.edu/
University of North Carolina	http://aa.uncwil.edu/numina/
University of Michigan	http://hi-ce.org/
University of South Dakota	http://www.usd.edu/pda/
University Twente	http://usa.nfia.nl/publish/su2001news.pdf
Wake Forest University	http://www.palm.com/us/enterprise/studies/study9.html

hardware keyboard. They connect to cellular networks, and an increasing number features Bluetooth connectivity. Smart phones combine mobile phones and PDAs into one device. They offer a touch sensitive display with cellular network connectivity, sometimes extended with WLAN. For the following analysis, smart phones are either included in the "PDA" or in the "Mobile" section. Other mobile devices include appliances with no or small displays, and some can connect to a WLAN network. Examples include MP3 players, digital cameras, or portable game consoles.

A third differentiation examines which of the three educational processes (learning, teaching, and administration) are covered by the project implementation. Learning processes include all activities pursued by a learner to gain knowledge or skills, whereas teaching processes include all activities by a lecturer to aid this acquisition process. Both processes interact with one another, and both depend on their underlying didactic model. Administrative processes focus on activities concerning the curriculum of a learner, for example registering for tests or enrolling in courses.

Table 3. Infrastructure, end user devices and processes in high school projects

	Infrastructure			End User Device			Process		
Project	**File**	**WLAN**	**Cell**	**PDA**	**Mobile**	**Other**	**Learn**	**Teach**	**Admin**
Ballard High School	x			x			x		x
Cons. HS District 230	x	x		x			x		
Hartland Farms Inter.	x			x			x		
King Middle School	x			x			x		
Gym. Landau a. d. Isar	x	x		x			x		x
Lessenger Elementary	x			x			x		
Mead Elementary	x			x			x		
MobiSkoolz	x		x	x			x	x	
Palm Educ. Pioneers	x			x			x		x
RAFT Project		x	x	x	x		x	x	
Total	**9**	**3**	**2**	**10**	**1**	**0**	**10**	**2**	**3**

High School Projects

Table 3 summarizes the findings for high school projects. File transfers are used by all but one project, whereas wireless transmissions via WLAN or cellular networks were implemented in only four projects. It can also be noted that wireless technologies were used by only one project exclusively.

A similar result is apparent with regard to the mobile devices used. All projects targeted PDAs as an end user device; again only one project used mobile phones. Considering that these projects piloted a new form of education, most of them focused on the learning process, where immediate benefits could be expected. Two projects additionally used mobile devices to support teaching activities, and three other projects allowed users to perform administrative tasks with their end device.

University Projects

In the next table, the same analysis is presented for university projects. Even though the results are more diverse compared to high school projects, a few observations can be made. First, once again most projects preferred file transfer to wireless transmission. In contrast to high school projects, those institutions that chose to implement wireless networks did not offer file transfer as an additional option—except for one case. Second, PDAs were a common end device in university projects, although a few projects supported the use of mobile phones. Third, almost all projects tried to improve learning and administration processes, while only three targeted the teaching process.

Comparison

When contrasting *m-education projects* at high schools and universities, three differences can be seen:

Table 4. Infrastructure, end user devices and processes in university projects

	Infrastructure			End User Device			Process		
Project	**File**	**WLAN**	**Cell**	**PDA**	**Mobile**	**Other**	**Learn**	**Teach**	**Admin**
Anglia Polytechnical Univ.			x	x	x		x		
Berlin Univ. der Künste			x		x				x
Wirtschaftsuniv. Wien	x			x			x		
Carnegie Mellon University		x		x			x	x	
Columbia Sch. of Nursing	x			x			x		x
Cornell University	x			x					x
East Carolina University	x			x			x		
Harvard Medical School	x			x			x		
Kentucky Migrant Project			x	x			x		
Purdue University		x		x					x
Stanford Sch. of Medicine		x		x			x	x	
Stanford University	x			x					x
University of Michigan	x			x	x	x	x		
Univ. of North Carolina	x	x		x			x	x	
University of Buffalo	x			x					x
University of Minnesota	x			x			x		
University of South Dakota	x			x			x		x
University Twente			x		x		x		x
Wake Forest University	x			x			x		x
Total	**12**	**4**	**4**	**17**	**4**	**1**	**14**	**3**	**9**

1. Unlike high schools, universities install and use wireless networks. This is motivated by the facts that university students spend more time between classes on a campus, and university campuses tend to be larger than school campuses, with higher computer literacy and usage at universities.
2. High schools focus on PDAs; universities tend to support mobile phones. PDAs are easier to use and are able to hold larger amounts of data than mobile phones. They are therefore better suited for learning purposes. University projects, on the other hand, strive not only to support learning, but also to target a larger and readily available hardware base.
3. Administrative and organisational information is more common at university projects. Students at universities have more freedom in learning compared to students at high schools. Furthermore, they need to acquire all information themselves, which can be a time consuming task if there is no central information repository. To remedy these factors, support for administrative and organisational tasks becomes more relevant at universities than at high schools.

It should be noted that some projects were initiated in cooperation with PDA device manufacturers, which additionally explains the high number of projects using PDAs. Taking the implemented functions into consideration, it is possible to identify those which were available in most projects:

- **Access to learning material:** Theoretically, this function is relatively easy to implement if there is already an e-learning platform: access to learning materials can be achieved by transferring the electronic documents to a mobile devices, for example by providing a URL. The problem is that mobile devices often lack the software to present certain media formats, and small display sizes limit legibility. As a result, learning material should be designed with a possible mobile use in mind.
- **Carrying out knowledge assessments:** Knowledge assessments can be carried out as part of a classroom application, or as a method for self assessment. The latter can be implemented easily with various technologies (J2ME stand-alone quiz applications, WAP-based questionnaire) if multiple-choice answers are allowed. Development of a *classroom application* is more complex and depends on the available devices and network infrastructures in the classroom.
- **Communication between students and lecturers:** Communication is the main function of mobile devices. One way to enable synchronous communication is to insert a WTAI (Wireless Telephony Application Interface) link in WAP pages, which, when clicked, dial the provided phone number (Larsson, 2000). J2ME application can also access mobile device functions (Schmatz, 2004). Depending on the number of supported communication protocols and their complexity (e.g., AIM/ICQ, E-mail, voice calls, SMS, MMS, fax, newsgroups) communication services are moderately hard to implement.
- **Information about courses and events:** Lecturers can inform students via push or pull methods. Push methods actively inform users about new information, while pull methods require the user to query this information regularly. Pull services are relatively easy to implement, as course information and events are usually published in text form which can be presented without problems. Push services are more complex, as they could require additional technologies, such as SMS gateways.
- **Enroll for courses and exams:** With these functions, students can use mobile devices to enrol for courses or exams. Technically, such services are not complex. They need, however, a connection to administrative backend systems. This creates potential security and legal risks that must be addressed.
- **Querying dates and examination results:** Similarly, students could be informed with push or pull methods about examination results. Again, this requires a connection to systems containing sensitive personal data, creating security and legal issues.
- **Reservation of resources (e.g., lab places, books, PCs):** An educational institution possesses a number of resources where access to them needs to be managed. Specific books or technical equipment are not available in unlimited numbers, and thus students and lecturers must make reservations for them in advance. If those reservations are already stored in a database, mobile access to these is easily granted. Users can then make reservations or check their status.
- **Information about external data (public services, events):** It is also possible to include external data and services in a m-education system. Depending on the complexity and openness of the service, such an inclusion

can be difficult. For example, if an external service is Web-based and financed with online advertising, the provider is not interested in allowing direct access to his database. In this case, one common solution is to pass the input of the mobile interface to the Web-based service and parse the resulting HTML page. This technique, however, is technically and legally problematic.

Some of these services were mobile extensions to existing functions accessible via a PC. Other functions were created with both wireless and wired usage in mind. Only a few services, primarily communication and collaboration functions, were usable exclusively with mobile devices. As the discussion shows, most functions can be implemented with medium effort (depending on the complexity approximately 3-6 person months per function) but their integration into existing technical and organizational frameworks needs additional attention. The following SWOT analysis provides decision support by presenting strength and weaknesses of various directions when planning and designing mobile education services.

SWOT ANALYSIS

The *SWOT analysis* was developed to identify strengths, weaknesses, opportunities, and threats (SWOT) as internal and external factors, and to help create adequate strategies for an organization. Even as it was originally intended for use in a business environment, it has also been applied in educational contexts (Balamuralikrishna, & Dugger, 1995, Gorski, 1991; Rosenberg, 2001). Criticism has resulted in further developments (Novicevic, Harvey, Autry, & Bond, 2004; Valentin, 2001); however, a SWOT analysis can provide initial insights, which can then be refined. The following SWOT analysis examines mobile education in general, not a specific project or institution. Additional factors must thus be taken into account when planning to implement a mobile education system. Again, the SWOT analysis will be structured along the three key factors from the presented projects:

1. Infrastructure
2. End user devices
3. Processes

Each of these factors will be analysed regarding internal, external, favourable, and unfavourable effects. The result is then assigned to the appropriate SWOT quadrant.

Infrastructure

A favourable, internal factor arising for mobile education from the *infrastructure* is flexibility. Learners and lecturers have the ability to access information and carry out educational processes regardless of location and time. The gained flexibility is even higher than with traditional, PC-based e-learning, as this requires a stationary desktop computer.

While flexibility is a favourable factor, interoperability between different technical infrastructures currently is unfavourable. Content and information can be distributed over multiple infrastructure channels, for example a video lecture can be broadcast via WLAN or via a cellular network, and course information are available via file transfer, cellular push (messaging) or WLAN. But switching between infrastructure technologies with a single device is problematic, as technical details of data transmissions are different. This is also due to interface limitations in end user devices (only smart phones can be used for file transfers, WLAN, and cellular networks).

External factors, which are determined by network providers, contain market demand as a favourable factor, and costs as an unfavourable factor. Market demand results from both users and infrastructure providers looking for ways to

utilize mobile networks. Furthermore, increased workforce mobility and life long learning can be seen as drivers for mobile learning.

Costs for infrastructures cover several elements. First, the m-education provider (the educational institution offering mobile services) needs to install required hard- and software, resulting in initial costs. Second, the provider has to cover recurring costs caused by network data traffic and maintenance. Third, m-education users also need to purchase dedicated hardware in order to connect to the network, and, fourth, bear recurring costs for data traffic. Not all of these costs occur simultaneously, but rather depending on the chosen technology, and not all costs are distributed evenly among the participants.

Table 5 summarizes the three infrastructure technologies with regard to range, and costs for both provider and user. Cellular connections are divided into pull services (when a user actively requests data) and push services (when a user is informed by an m-education provider, for example with a SMS) as their cost structures vary. It should be noted that the range applies during an exchange of data. File transfers require the mobile device to be physically close to the transmitting source device, and after the transfer the data can be used on the mobile device without any further restrictions. However, the data can only be as current as the most recent transfer, decreasing timeliness of information (Lehner, Nösekabel, & Lehmann, 2004).

File transfers incur very few costs, as they are based on wired or wireless personal area networks. m-education providers only need to implement a distribution channel for information and content, which can then be transferred by the users to their already available devices. Costs for data transfer are, in this case, negligible.

WLAN requires an initial installation of access points (based on either IEEE 802.11 or Bluetooth technology), and users need to be equipped with hardware enabling them to utilize the wireless network. Costs for data transfer are, again, negligible, although there are costs for maintaining the hardware.

In order to use cellular networks, the provider has to create an appropriate service, for example a WAP site, or a SMS gateway. As this kind of infrastructure primarily targets mobile phones as an end user device, it can be assumed that most users already possess an adequate device. By differentiating between pull and push technologies, the resulting cost structures for a provider and a user can be taken into account. When users pull desired information (e.g., from a WAP site), they are charged for the data connection. If the information is pushed to the users (e.g., with a SMS), the sender is charged for each data transmission.

End User Devices

A favourable, internal factor of mobile education is the availability of appropriate *mobile devices* (Le Bodic, 2003), although PDAs are not as common as mobile phones. Therefore, most projects using PDAs as the end user device made arrangements for lending or sponsoring, reducing the initial costs

Table 5. Range and cost structures for infrastructure technologies

infrastruct. technology	typical end user device	range	costs for provider		costs for user	
			initial	recurring	initial	recurring
File	PDA	short	low	none	none	none
WLAN	PDA	medium	high	low	medium	none
Cell. (pull)	mob. phone	high	low	none	none	high
Cell. (push)	mob. phone	high	low	high	none	none

for users. Other mobile devices, such as portable multimedia players, can be either popular (MP3 player) or uncommon (DVD player).

Unfavourable factors are technical limitations imposed by mobile devices. Even as technological progress—an external, favourable factor—increases computing power and battery life with each device generation, physical limitations, such as screen size and cumbersome input methods, remain. Not all of these limitations can be overcome, though. The screen size, for example, is dependent on the overall size of the device, which is limited due to mobility considerations.

One drawback of short development cycles in device development is a lack of standard conformity, an external, unfavourable factor. As a result, interoperability can be hindered, if implementations are not compatible to standards. Such an incompatibility can occur when either a standard is not yet defined, is unspecific, or when a manufacturer explicitly decides not to follow it. One example is the case of "Smart Messaging," a system developed by Nokia in the late 1990s to extend the capabilities of SMS. With "Smart Messaging," messages could contain multimedia content and text formatting. In 2001, the official 3GPP "Enhanced Messages Service" (EMS) was formally standardized and Nokia abandoned "Smart Messages" in favour to EMS (Dornan, 2002).

Processes

As mobile devices are personal devices, *educational processes* can address each user individually, making this a favourable, internal factor. Individual data for a learner encompasses—among others—learning progress, courses taken, personal preferences, and administrative data. It should be noted that such a system requires a high level of technical integration and security, on a level not always achievable with wireless networks (Sikora, 2001). An unfavourable, internal factor is caused by the already mentioned limitations of mobile devices, and extra effort has to be taken to produce content which is usable on a wide range of these devices. For example, long text, detailed graphics, or complex animations are difficult to read on small screens. Multimedia files take up too much memory, although in these cases streaming technologies can provide a solution.

Opportunities arise when mobile education services are didactically integrated into teaching and learning processes, resulting in added flexibility for both the learner and the lecturer. However, such integration builds on the willingness of all users to adapt such services, and resistance to change poses a threat.

STRATEGIC RECOMMENDATIONS

Table 6 summarizes the aforementioned factors, contrasting strengths, weaknesses, opportunities, and threats. Based on these factors, a number of strategies can be deducted for implementing a mobile education service. Again, the discussion will revolve around feasible choices for an infrastructure and for end user devices, and which processes should be integrated.

Before engaging in a mobile education project, a decision on whether to implement a mobile education system or not has to be made. This decision should be part of an e-learning strategy (Back, Bendel, & Stoller-Shai, 2001; Rosenberg, 2001) for several reasons. First, mobile services should extend an e-learning system, which provides a basic data and content infrastructure. Non-mobile e-learning systems also offer easier and faster access for lecturer and administrative staff, who are usually working at a stationary PC. Second, an e-learning strategy defines the framework for mobile services, for example which functions should be implemented, or who should be responsible—organizationally and financially. Third, it is possible that the strategy already contains statements regarding the choices for an infrastructure, for end user devices, or for

Table 6. SWOT factors for mobile education

	favourable factors	unfavourable factors
internal factors	Strengths	Weaknesses
	1. flexibility	1. interoperability
	2. availability	2. technical limitations
	3. individuality	3. effort
external factors	Opportunities	Threats
	1. market demand	1. cost
	2. technological progress	2. standard conformity
	3. didactic integration	3. resistance to change

didactic processes (e.g., when the support for lecturers is explicitly demanded, or when it is part of the e-learning strategy to equip each student with a PDA).

As a result, the following recommendations should always be adapted to the specific requirements of the implementing institution. They can present various options which may then be discussed, but they should not be understood to be an optimal solution for every project.

Infrastructure

The choice for an *infrastructure* must consider the cost situation, the range required, and the targeted end user devices. Basically, the lower the costs, the shorter the achievable range. Under certain conditions, costs could be considered under long-term aspects. For example, when the mobile education system is implemented within the scope of a project and funding would include infrastructure costs. Then it is important how the cost structure will evolve after funding has expired, and users are required to bear their own costs. In the long run, a decrease in data transmission costs might be expected due to market demand—already some mobile network providers offer data flat rates for WAP pages today.

When initiating a mobile education system, several paths can be taken. If no infrastructure exists and funding is limited, transferring files via synchronization or personal area networks (e.g., Bluetooth or IR transmission) reduces initial costs and does not require extensive knowledge of mobile technologies, but it does provide first experiences and feedback from users. Such systems can later be upgraded by including wireless networks (WLAN or cellular) for extended coverage.

Another option in this situation is offering a cellular network-based service, for example a WAP service. Public wireless networks provide network connectivity, and the costs are carried by the users. As a result, acceptance will most likely be low. Cellular networks could extend file transfers; they should not be the only way to access the system.

With the success of i-mode in Japan, additional business models for mobile services have been discussed and established. Educational institutions function as content and service providers in these mobile value chains, theoretically allowing them to charge users for their learning content and services. I-mode network providers (Japans NTT DoCoMo, or a licensee) handle the billing, so content providers are not required to charge each user individually (Barnes & Huff, 2003; Sharma & Nakamura, 2003). Therefore, it would be possible to create revenue with e-learning content and services, assuming that a demand for these exists. Creating content for i-mode requires only basic

knowledge of Internet technologies, as pages are implemented using cHTML (compressed HTML), which is a subset of HTML. Some browsers in i-mode phones are also able to render XHTML.

If sufficient funding exists, an educational institution could opt for installing a campus-wide wireless LAN. Such a WLAN is more attractive for universities than for high schools, as the number of privately owned mobile devices (especially laptops) is higher, campuses are usually larger, and students work at less determined times and places. Similarly, if a WLAN already exists at the campus, a mobile education system should include it as a distribution channel.

Should an infrastructure and additional funding be available, another strategy could be to subsidize network connectivity for users. This might include wireless network traffic via cellular networks or hardware (e.g., PDAs, WLAN cards, or mobile phones). Such incentives can lead to an increased acceptance of the mobile education system.

End User Devices

The cost factor also affects the type of usable *mobile devices.* Short range, cost efficient infrastructures primarily target PDAs as end user devices. Most students and staff do not have such a device, and additional costs would be incurred for purchasing or lending a device. Again, this factor could be absorbed within a project or by a hardware sponsor. Mobile phones, on the other hand, are widespread and therefore are an attractive target base. However, they have technical limitations which must be taken into account.

When end user devices are issued by the educational institution, organizational tasks increase effort for and, possibly, resistance by teachers. Purchasing, keeping track of, and supporting mobile devices are additional challenges for teachers and lecturers. If damaged, devices have to be replaced, which may incur costs for either the student or the organization. Small devices are also prone to loss or theft. Curtis et al. (2003) argue that students, after having proven able to utilize their devices carefully and adequately, should be allowed to keep their devices provided by the school even outside the classroom. When mobile services target devices which are already owned by the students, this issue is less of a problem.

Device support is determined by the availability of the device, the infrastructure, and the implemented services. Providing access to mobile phones allows a higher number of users to employ the mobile education system, increasing acceptance. Push services, like SMS information, can be adopted intuitively without extensive training. Complex services, for example WAP sites or video streaming, create usability challenges ranging from device configuration to GUI design.

These complex services are easier to implement for PDAs and smartphones, as display capabilities (higher resolution, touch sensitive screen, etc.) allow intuitive usability design. Furthermore, multimedia content can be shown in a higher quality. PDAs can also be extended with external devices, such as probes, graphic cards for beamer connection, or GPS. They are more versatile than traditional mobile phones, but this added versatility has to be exploited by offering appropriate services.

Processes

Availability of mobile education systems is the key factor for anytime, anywhere learning. Students are given the ability to engage in *educational processes* regardless of time and place. Weiss (2002) notes that mobile usage patterns focus on retrieving a specific answer within a short time frame ("hunting"), instead of browsing through data provided by the system. As a result, m-education should reflect this behavioural pattern by offering short learning segments, and the ability to query personalized information.

All educational processes benefit from the achievable individuality mobile devices can pro-

vide. Since each mobile device is—as a personal item—usually associated with a single user, processes can be automatically adapted to individual preferences. From a usability viewpoint, this benefit is also a requirement. Display size and input limitations prohibit extensive user interaction with the system. Therefore, users should be presented the functions they are most likely to choose in a manner that reduces navigational efforts and speeds up transactions.

Extending the idea of adaptation to a technical level leads to the concept of "transcoding" (Sharma & Nakamura, 2003). Multimedia content can be created once, and will then be transcoded from this single source to the capabilities and requirement of a specific device and network. Thus, the need for producing identical information in varying formats, sizes, and for different bandwidths is eliminated and replaced by a (semi) automatic conversion process.

For mobile education, this means that multimedia educational content—audio and video lectures, slides, or pictures—only have to be created once by the lecturer. This content is then available, without further modifications, for PC-based and PDA- or mobile phone-based e-learning systems. For textual content, where automatic adaptation from a source text is not feasible, a multi channel delivery approach is possible. Here, the length and depth of a text is tailored to the end user device. A short summary is available for mobile phones, a longer text for PDAs, and the full text for PCs. Apparently, with such a solution mobile devices can only be an addition to existing e-learning efforts, and not a replacement.

Usually, at educational institutions the number of students is considerably higher than the number of lecturers and teaching staff. Therefore, learning-oriented services are relevant to the largest user group and could include access to learning material and self tests. Teaching-oriented services (e.g., Myers, 2001) primarily focus on support during lectures, as the preparation of content—including the production of multimedia data—requires the capabilities of stationary PCs. Furthermore, teaching staff is less mobile than students, reducing the need for mobile solutions. Administrative-oriented services depend on a connection to existing systems. Since these systems contain sensitive material, security is an essential factor for such services. Several security approaches are available (Dornan, 2002; Hansmann, Merk, Nicklous, & Stober, 2003; Sharma & Nakamura, 2003), but not all might be available for a specific wireless technology mix.

Summary

Currently, there is not a single combination of mobile technologies which would fit the needs of all mobile education projects. Both the choice for an infrastructure and end user device depend on the specific situation. Also, some infrastructures (e.g., WLAN) can be used more flexibly than others, as they offer access to a larger variety of device types. As a result, many projects start with applications that are easy to implement and cost-effective to distribute. J2ME programs with educational purpose provide a good starting point to get familiar with the characteristics of mobile services development. Other approaches may include adapting an existing Web-based e-learning system for mobile access, either through file transfer (using Plucker or AvantGo), i-mode, or WML.

With regard to processes, it is probably best to start with learning-oriented processes, from which a large number of users can benefit. Mobile classroom applications can also be implemented easily, especially when they are not dependent on or connected with the mobile education system. This way, dependencies can be reduced, and introduction of new technologies is alleviated. Mobile learning, in its simplest form, can be achieved by offering short audio clips of lectures for download as MP3 files. Students can then transfer these audio files to their portable MP3 player or burn them on a CD as a type of audio book. Other means

include converting lecture notes to e-books, which can be read on a PDA.

Table 7 summarizes possible software choices for each infrastructure—end user device combination. Since this sector continually develops, the items in the table are not meant to be exhaustive. Furthermore, not all devices in a category support all software solutions: not all mobile phones are yet capable of playing audio- or video-files, and only certain handsets can use i-mode. Personal area network connectivity, such as Bluetooth, is mainly used to transfer data directly between devices. Thus, they can be included in the "file transfer" column.

"Multimedia" comprises, as a generic term, educational content in various formats: for example, audio, video, animations, pictures, e-books. "Plucker" and "AvantGo" are applications which store HTML pages on the mobile device and let the user browse them off-line. AvantGo is available for some mobile phone models and can be synchronized over a wireless network.

Table 8 lists examples for m-education software solutions. Not included are multimedia applications (e.g., video or audio players). A discussion of these can be found in Lehner, Nösekabel, & Bremen (2004) and in Nösekabel (2005). The list is by no means exhaustive, but shows the wide range of available software solutions. Most m-learning approaches make use of stand-alone application that are installed on a mobile device, and they either already contain learning content or retrieve them via a wireless network connection. With ImagiProbe, a hardware/software combination is also included in the list. ImagiWorks offers various probes (e.g., for temperature measurement) that can be connected to a mobile device. The software records and analyzes data collected with the hardware probe, and results can be shared with other users.

If possible, a new mobile education system should be organizationally integrated into an existing e-learning system. Some e-learning systems already offer mobile access to their services and content, such as Stud.IP and Blackboard. This access is mostly, but not exclusively, based on WML. Such additional modules are easy to integrate into the existing system. However, they require a specific combination of infrastructure and end user devices. If this combination is feasible for a specific project, then these models could be the best way to implement mobile services for e-learning. Furthermore, they could be the only solution, if the existing system is not accessible and does not provide interfaces for data exchange.

FUTURE TRENDS

Technological progress will increase both device and network capabilities, and demand for data services may lead to an attractive pricing struc-

Table 7. Mobile technology matrix

	File transfer	WLAN	Cellular network
Mobile phone	Applications (J2ME) Multimedia download		SMS WAP i-mode Multimedia streaming AvantGo
PDA	Applications Multimedia download Plucker/AvantGo	HTML WAP Multimedia streaming Plucker/AvantGo	HTML WAP Multimedia streaming Plucker/AvantGo
Other	Multimedia download	Multimedia streaming	Multimedia streaming

ture. Already a few network operators offer flat rates for WAP data transmission, although these are restricted to WML pages and do not include HTTP traffic or other downloads. Should this change, mobile services in general, and mobile education specifically, will be more attractive than they are today.

Additionally, an increased adoption of learning objects (Dodero, Aedo, & Diaz, 2002; Rosenberg, 2001; Wagner, 2005; Wiley, 2002) would solve some content related issues. Learning objects are small units of information or instruction, and they can be assembled into structures of a higher order, for example, courses. Learning objects are scalable and each of these objects can be created for mobile deployment, taking device capabilities and limitations into account. Learners can then decide which learning objects they require and transfer them to their mobile devices. If a learning path or another connection between learning objects is defined, succeeding objects can be suggested to the student. As a result, data transmission and memory requirements are reduced because only needed learning objects are retrieved.

Another area which has received limited research so far is mobile edutainment. Edutainment combines entertainment applications (e.g., games) with an educational background. Software for stationary PCs is already well established, especially with learning content appropriate for children. For mobile devices, however, few concepts exist (Bellotti, Berta, De Gloria, & Margarone, 2003; Feix, Göbel, & Zumack, 2004; Ströhlein, 2004). This could prove to be a potential market gap, as mobile phones and mobile entertainment products (ring tones, themes, and games) are popular among younger users.

CONCLUSION

Both the survey and the SWOT analysis show that costs are a deciding factor for m-education projects. Costs are primarily determined by the infrastructure, which is also limiting the supported devices (e.g., only very few mobile phones can be used in WLANs). Furthermore, both the providing institution and the users act on a limited budget. As a result, file transfers, which incur low costs to both the provider and the user, are a popular choice.

Regarding end user devices, a dependency on industrial partners can be observed. Since mobile phones are the devices mostly available to students,

Table 8. Exemplary software solutions for mobile education

Product	URL	Infrastr.			Devices			Processes		
		F	W	C	P	M	O	L	T	A
Mobile Learning Engine	www.elibera.com			x		x		x		
AvantGo	www.avantgo.com	x	x	x	x	x		x		
Plucker	www.plkr.org	x	x	x	x			x		
Stud.IP	www.studip.de		x		x	x		x		x
Blackboard	www.blackboard.com		x		x	x				x
Pebbles	www.pebbles.hcii.cmu.edu		x		x				x	
Hi-CE	www.hi-ce.org	x			x			x	x	
Hot Lava	www.hotlavasoftware.com	x	x	x	x	x		x		
Four.OStudent	www.fourostudent.net	x			x					x
ImagiProbe	www.imagiworks.com	x			x			x	x	

supporting them would ensure a large, although technically somewhat limited, user base. Additionally, users are already competent in operating their devices. PDAs offer more capabilities, but have to be provided through partnership programs because they are not as widespread as mobile phones or laptops among students. A problem is that mobile phones are less suitable for a file transfer infrastructure than PDAs.

Most projects initially focused on supporting learning processes. This can be considered to be feasible, as immediate benefits are likely to be gained here. Lecturers and administrative staff are also working at stationary PCs and would be unable to realise the full potential of a mobile solution. However, employing mobile services in the teaching process has proven to be viable. Offering mobile access to administrative data requires additional efforts concerning security.

One important factor is the combination of technical and pedagogical aspects. Systems providing only a technical solution without sensible pedagogical integration do not address the needs of learners and thus reduce acceptance. Mobile technologies are therefore only a part of the solution. *"Technology in and of itself may not guarantee better learning. But when effectively deployed, technology can help focus attention while attracting and maintaining a learner's interest"* (Wagner, 2005, p. 48). All in all, there is no immediate pressure to implement mobile education services. Still, if an e-learning infrastructure has already been successfully established, a next step could consist of planning and extending this infrastructure with mobile services. In this case, it is important to select an appropriate wireless infrastructure in order to create a significant additional value for the users. If the costs outweigh the perceived benefit, usage of the system will remain low.

An important aspect of any e-learning or m-education system is thus the alignment with pedagogical theories and practices. Technology-driven projects are useful for discovering new potentials and possibilities. Long-term solutions, however, require an integrated and interdisciplinary approach.

REFERENCES

Back, A., Bendel, O., & Stoller-Shai, D. (2001). *E-learning im unternehmen. Grundlagen – Strategien – Methoden – Technologien.* Zurich, Switzerland: Orell Füssli.

Balamuralikrishna, R., & Dugger, J.C. (1995). SWOT analysis: A management tool for initiating new programs in vocational schools. *Journal of Vocational and Technical Education, 12*(1), 1-5.

Barnes, S.J., & Huff, S.L. (2003). Rising sun: iMode and the wireless Internet. *Communication of the ACM, 46*(11), 79-84.

Bellotti, F., Berta, R., De Gloria, A., & Margarone, M. (2003). MADE: Developing edutainment applications on mobile computers. *Computers & Graphics, 27*(4), 617-634.

Clarke, I., & Flaherty, T.B. (2002). mLearning: Using wireless technology to enhance marketing education. *Marketing Education Review, 12*(3), 67-76.

Curtis, M., Williams, B., Norris, C., O'Leary, D., & Soloway, E. (2003). *Palm handheld computers – A complete resource for classroom teachers.* Eugene: International Society for Technology in Education.

Dodero, J.M., Aedo, I., Diaz, P. (2002). Participative knowledge production of learning objects for e-books. *The Electronic Library, 20*(4), 296-305.

Dornan, A. (2002). *The essential guide to wireless communications applications: From cellular to WiFi* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Dourish, P. (2001). *Where the action is: The foundations of embodied interaction*. MA: MIT Press.

Feix, A., Göbel, S., Zumack, R. (2004, June 24-26). DinoHunter: Platform for mobile edutainment applications in museums. In *Proceedings of the Second International Conference on Technologies for Interactive Storytelling and Entertainment* (pp. 264-269). Berlin, Germany: Springer.

Gorski, S.E. (1991). The SWOT team: Focusing on minorities. *Community, Technical, and Junion College Journal, 63*(3), 30-33.

Hansmann, U., Merk, L., Nicklous, M.S., & Stober, T. (2003). *Pervasive computing* (2nd ed.). Berlin, Germany: Springer.

Hummel, K.A., & Hlavacs, H. (2003, January 6-12). Anytime, anywhere learning behavior using a Web-based platform for a university lecture. In *Proceedings of the SSGRR 2003 Winter Conference*, L'Aquila (pp. 1-6).

Larsson, M. (2000). Wireless telephony application: Telephony in WAP. In M. van der Heijden & M. Taylor (Eds.), *Understanding WAP* (pp. 65-96). Boston: Artech House Publishers.

Le Bodic, G. (2003). *Mobile messaging technologies and services: SMS, EMS, and MMS*. Chichester: John Wiley & Sons.

Lehner, F., Nösekabel, H., & Bremen, G. (2004). *M-learning und M-education – Mobile und drahtlose anwendungen im unterricht*. Passau: Research Report Business Computing W-08-04.

Lehner, F., Nösekabel, H., & Lehmann, H. (2004). Wireless e-learning and communication environment – WELCOME at the University of Regensburg. *E-service Journal*, *2*(3), 23-41.

Lyytinen, K., & Yoo, Y. (2002). Issues and challenges in ubiquitous computing. *Communications of the ACM, 45*(12), 63-65.

Myers, B.A. (2001). Using handhelds and PCs together. *Communications of the ACM*, *44*(11), 34-41.

Nösekabel, H. (2005). *Mobile education*. Berlin, Germany: GITO.

Novicevic, M.M., Harvey, M., Autry, C.W., & Bond, E.U. (2004). Dual-perspective SWOT: A synthesis of marketing intelligence and planning. *Marketing Intelligence & Planning, 22*(1), 84-94.

Nyiri, K. (2002, August 29-30). Towards a philosophy of m-learning.In *Proceedings of the IEEE International Workshop on Wireless and Mobile Technologies in Education* (pp. 121-124), Växjö.

Rosenberg, M.J. (2001). *E-learning. Strategies for delivering knowledge in the digital age*. New York: McGraw-Hill.

Schmatz, K.-D. (2004). *Java 2 micro edition*. Heidelberg, Germany: dpunkt.verlag.

Sharma, S.K., & Kitchens, F.L. (2004). Web service architecture for m-learning. *Electronic Journal on e-Learning, 2*(1), 203-216.

Sharma, C., & Nakamura, Y. (2003). *Wireless data services – Technologies, business models and global markets*. Cambridge University Press.

Sikora, A. (2001). Wireless *LAN - Protokolle und anwendungen*. Munich, Germany: Addison-Wesley.

Ströhlein, G. (2004). HistoBrick: Mobile edutainment into descriptive statistics. *I-com*, *3*(2), 53-56.

Valentin, E.K. (2001). SWOT analysis from a resource-based view. *Journal of Marketing Theory and Practice*, *9*(2), 54-69.

Wagner, E.D. (2005). Enabling mobile learning. *EDUCAUSE Review, 40*(3), 40-53.

Weiss, S. (2002). *Handheld usability.* Chichester: John Wiley & Sons.

Wiley, D.A. (2002). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In D.A. Wiley (Ed.), *The instructional use of learning objects* (pp. 3-23). Bloomington: Association for Educational Communications and Technology.

APPENDIX I: INTERNET SESSION: SUMMARIZE MOBILE AND WIRELESS LEARNING APPROACHES

Interaction

The Web site lists several resources regarding learning with mobile technologies. Compile and compare typical definitions of mobile and wireless learning (*http://www.E-Learningcentre.co.uk/eclipse/Resources/mlearning.htm*). What are key differences and similarities? Use the strategic discussion in this chapter to apply it to one project of your choice from the Web site. Would you have suggested another kind of implementation? If so, which, and why?

APPENDIX II: CASE STUDY

Enhancing an Existing E-Learning Solution with Mobile Services

A university with approximately 16,000 students on a single but spacious campus possesses an established Web-based e-learning portal. Although usage of the system is not compulsory, over 8,000 students have registered with the system, which provides learning material in the form of audio and video recordings, and lecture slides. Multimedia files can be streamed or downloaded, and slides are available as PDFs.

Plans of the university are to enhance the current system by offering mobile services. The existing portal is open sourced, and the underlying data model is well documented. Wireless access points have been installed at hot spots on the campus (e.g., cafeterias, larger classrooms). There is no further funding for additional infrastructure measures—including mobile devices—and currently no knowledge about writing mobile applications exists, although there is know-how about Internet technologies like HTML and PHP.

The university is well equipped with PCs and laser printers which are grouped in several pools all over the campus. A survey has shown that most students own a mobile phone, but no PDA. In the past, the university has cooperated loosely with a company offering mobile marketing services via SMS. Even though the financial funds of the company are very limited, it has signalled interest in a joined project should the opportunity arise.

Questions

1. What strategy (infrastructure, end user devices, and processes) of implementing mobile education services would you recommend and why?
2. Which key factors could result in success or failure of a mobile education project at the university?
3. Which other current (mobile and non-mobile) technologies could be integrated into the system to provide learning services?

APPENDIX III: POSSIBLE PAPER TITLES/ESSAYS

- Mobile learning and education: Benefits and limitations
- Embedding mobile devices in learning processes
- Mobile education and e-learning strategies
- Location based services for mobile education
- Mobile edutainment: Concepts and implementations

This work was previously published in Ubiquitous and Pervasive Knowledge and Learning Management: Semantics, Social Networking and New Media to Their Full Potential, edited by M. Lytras and A. Naeve, pp. 67-93, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 1.13
Anywhere, Anytime Learning Using Highly Mobile Devices

Mark van 't Hooft
Kent State University, USA

Graham Brown-Martin
Handheld Learning, London, UK

Karen Swan
Kent State University, USA

INTRODUCTION

In a world that is increasingly mobile and connected, the nature of information resources is changing. The new information is networked, unlimited, fluid, multimodal, and overwhelming in quantity. Digital technologies, such as mobile phones, wireless handheld devices, and the Internet, provide access to a wide range of resources and tools, anywhere and anytime. This type of access and connectivity has also had an impact on how we collaborate on projects and share media and therefore, greatly increases opportunities to learn inside *and* outside institutionalized school systems. Learners now have the tools to take learning beyond classrooms and the school day.

The development of handheld devices can be traced back to Alan Kay's vision of the Dynabook. As early as the 1970s, Kay envisioned a mobile, kid-friendly, notebook-sized computer with artificial-intelligence capabilities that would support children's learning inside and outside of school. Similar ideas soon followed in the form of devices such as the Psion I (1984), the GRiDPaD (1988), Amstrad's PenPad, and Tandy's Zoomer (1993), the Apple Newton (1993-1995), and the eMate (1997-1998). During the 1990s and early 2000s, Palm developed a series of handheld devices that defined the handheld market in North America, while Microsoft developed several versions of its Windows Mobile software that could be found

on mobile devices made by such companies as HP, Dell, and more recently, Fujitsu Siemens (Bayus, Jain, & Rao, 1997; HPC Factor, 2004; Williams, 2004).

There are also many devices whose primary function is entertainment or communication, including media players such as Apple iPods, portable gaming devices like the Sony PSP and the Nintendo DS, and, of course, mobile phones. These types of devices are becoming increasingly popular and multifunctional, with iPods being able to store and play music, pictures, and video; portable gaming devices sporting wireless capabilities for interaction between devices (and in the case of the PSP, Internet access); and mobile phones being used to shoot pictures and video, upload content to the Web or e-mail it elsewhere, do text messaging, and make phone calls. Whatever the device, convergence seems to be increasingly important, and growing numbers of young people are using these mobile, digital, and connected tools daily, whenever and wherever they need them, and this includes schools.

BACKGROUND

Mobile computing enthusiasts have advocated the use of highly mobile devices for teaching and learning to get closer to a ubiquitous computing environment, defined in 1991 by Mark Weiser as a setting in which "a new way of thinking about computers in the world … allows the computers themselves to vanish into the background" and become indistinguishable from everyday life (p. 94). Weiser emphasized that ubiquitous computing does not just mean portability, mobility, and instant connectivity, but also the existence of an environment in which people use many computing devices of varying sizes that interact with each other, combined with a change in human psychology, to the point where users have learned to use the technology well enough that they are no longer consciously aware of its presence and do not have to be. This version of ubiquitous computing has recently been revisited by scholars such as Yvonne Rogers (2006), who proposes a modified version in which

UbiComp technologies are designed not to do things for people but to engage them more actively in what they currently do (p. 418);

and Bell and Dourish (2007), who argue that ubiquitous computing is characterized by power-geometries (the ways in which spatial arrangements, access, and mobility reflect hierarchies of power and control); heterogeneity (as opposed to standardization and consistency in technology, use, and regulation); and management of ubiquitous computing that is messy.

Weiser's somewhat revised vision of ubiquitous computing fits well with current visions of technology integration in education and its potential impact on teaching and learning. Academic research has shown that computer use and student learning gains are "closely associated with having computers accessible to all students in teachers' own classrooms" (Becker, Ravitz, & Wong, 1999; see also Shin, Norris, & Soloway, 2007). Highly mobile devices provide a solution because of their small size and comparatively low cost in acquisition and ownership (Norris & Soloway, 2004; Sharples, 2000a), and they supplement the existing technology infrastructure. Some scholars have defined the resulting learning environment as "handheld-centric," "providing all students with access to valuable resources on a shared but timely basis," where each tool has been earmarked for its intended use (Norris & Soloway, 2004; Tatar, Roschelle, Vahey, & Penuel, 2003). Another group of scholars is looking at learning with highly mobile devices from a broader perspective. They have coined the term m-learning, "the processes of coming to know through conversations across multiple contexts amongst people and personal interactive technologies" (Sharples, Taylor, & Vavoula, 2007).

Highly mobile devices are also altering the nature of technology integration in teaching and learning, and can act as catalysts for radical changes in pedagogical practices (Fung, Hennessy, & O'Shea, 1998). Their fundamental difference from more traditional desktop computing environments lies in the fact that users "interacting with a mobile system interact with other users [and] interact with more than one computer or device at the same time" (Roth, 2002, p. 282; see also Cole & Stanton, 2003). Consequently, highly mobile devices lend themselves well for both individual and collaborative learning, if used appropriately. Roschelle and Pea (2002), for example, highlight three ways mobile devices have been used to enhance collaborative learning–classroom response systems, participatory simulations, and collaborative data gathering–and suggest there are many more uses (see also Roschelle, 2003).

Moreover, because of their small size, portability, and connectivity, highly mobile devices do not constrain users like desktops and laptops do. As such, they encourage learners to use technology across the curriculum and in everyday activities, and embrace it as a lifelong-learning tool to be used anywhere and anytime (Inkpen, 2001; Sharples, 2000b), eventually leading to the type of ubiquitous computing that Weiser envisioned and Rogers, and Bell and Dourish advocate.

TEACHING AND LEARNING WITH MOBILE DEVICES

Highly mobile devices possess certain characteristics that allow for frequent and immediate access to a variety of tools and information sources for teachers and students, and their use in classrooms and other learning environments is bringing about many changes. However, it is important to understand that simply putting more digital tools in schools is not the solution to making technology use for teaching and learning meaningful and effective. Rather, teaching, learning, and technology need to be reconceptualized before the full educational possibilities inherent in small, versatile, and mobile digital technologies can be realized.

In *The Educators Manifesto* (1999), McClintock proposes that digital technologies change what is pedagogically possible. To take advantage of these possibilities, teaching must be continuously redefined within the changing context that new tools such as handheld computers create. Teaching should be reconceptualized as "conducting learning," thereby putting more responsibility for learning on the learner. Second, teaching must no longer be thought of as restricted by the spatial and temporal boundaries that current educational systems impose. Third, the content and focus of teaching must be redefined to meet the needs of the 21st century world (Swan, Kratcoski, & van 't Hooft, 2007).

If teaching is to be reconceptualized to take full advantage of mobile tools, so should learning. As digital tools are becoming increasingly mobile, connected, and personal, they have the potential to make learning student-centered, and can support both individual and social construction of knowledge. In particular, students need to be given more responsibility for their own learning. Four areas in which learning should be redefined as more student-centered are engagement and motivation, individualization and choice, collaboration and peer learning, and learning for all students (Swan et al., 2006).

Mobile technology has the potential to have a substantial and positive impact on teaching and learning. Merely introducing the tools in the classroom will not suffice; it is even more important that educators think about how teaching and learning need to change in order to take full advantage of the good things that digital technology has to offer for students and teachers alike.

The first step in rethinking teaching and learning within a context that includes the latest digital tools is simple, yet radical. Educators need to embrace the technology and learn about the ways

in which younger generations are using it. Current students live in a world that is connected 24/7 and high tech, with an overwhelming amount of communication devices and information channels. Within this context, digital tools are increasingly personal, mobile, networked, social, accessible, flexible, multimodal, and contextual (see e.g. Roush 2005, Thornburg, 2006; van 't Hooft & Vahey, 2007).

Second, we need to rethink the role of technology in schools and the fundamental impact this changing role is going to have on teaching and learning. Too often, we look at technology as being integrated in the existing curriculum, which entails doing the same things we were doing, and using technology as an add-on. Indeed, we probably need to stop thinking about technology *integration* altogether, but instead see technology as an agent of *transformation* that will enable us to do new things in new ways. As stated above, for example, mobile technology has the potential to break through the temporal barriers of the school day and the brick and mortar of school walls, making learning an authentic and relevant aspect of everyday life, and not just schooling (Alexander, 2004; Breck, 2006).

Third, fundamental changes in teaching and learning as brought about by pervasive digital tools require that teachers carefully reexamine how they view and use technology, and how this impacts their teaching philosophy, curriculum, and practices. This type of examination is not going to take place overnight. It takes time and effort. It takes motivation and engagement, individualization and choice, collaboration, and a group effort by all. In the end, it may, and probably will, require fundamental changes in the ways in which we teach our children.

Fourth, there are always the technical and logistic issues to be overcome. These include more traditional issues related to networking, compatibility, security, maintenance, and training, as well as new problems created by new technologies, such as copyright infringement, violation of privacy, and cyberbullying.

Fifth, while highly mobile devices provide affordances that many other technologies cannot, there are always limitations on their use. Therefore, it is essential that teachers (and) students consider when it is appropriate to use a mobile tool for purposes of learning and when it is not. Whatever the choice of tool, it should not get in the way of learning. For example, it would be unwise to try to do extended video editing or high-end graphics design on a mobile device.

Finally, we cannot overlook the most important partner in all of this, the students. Current generations of students prefer quick and easy access; communication and networking; digital, hyperlinked, and multimedia content; and just-in-time learning that is relevant and useful. In addition, in a digital and connected world, *learners* are mobile; active, communicative, and resourceful; and construct context through interaction (Alexander, 2004; Roush, 2005; Sharples, 2005). How will they be affected by fundamental changes in teaching and technology use for formal *and* informal learning?

FUTURE TRENDS

Various pilot and research projects have attempted and are attempting to bring about changes in teaching and learning by introducing highly mobile devices. In *classroom settings*, a large-scale implementation of handheld computers has investigated what happens to teaching and learning when many devices are introduced in formal educational settings (Vahey & Crawford, 2002). Other examples include RoomQuake, which used handheld devices and a variety of artifacts to simulate earthquakes; and the application of handheld computers in combination with scientific probes. In *informal environments*, we have seen mixed reality games that combine the real world

with virtual environments (often through the use of digital overlays or location-based resources) to enable users to experience and reflect on both. Examples of such projects include Environmental Detectives (explore imaginary scientific problems or environmental disasters using Pocket PCs and GPS in a real setting), Frequency 1550 (using cell phones and GPS to learn collaboratively about the history of Amsterdam), and MobiMissions (a game in which players create missions for others and can choose which missions to take on). Most of the initiatives listed here are described in greater detail by Rogers and Price (2007).

However, as admirable as these projects are, the real work needs to be done in bridging the gap between learning in formal and informal settings. This could consist of the use of highly mobile devices to augment field trip experience and, at the same time, provide students with resources for learning, upon return to the classroom, in the form of digital data collected during the field trip. An early example of such a project is Ambient Wood, an attempt to digitally augment a woodland habitat. A more recent example is MyArtSpace, in which students choose which data to collect and store during a museum visit, for later use in the classroom (Vavoula, Sharples, Lonsdale, Rudman, & Meek, 2007), but these types of examples are still few and far between.

Research can be helpful here. Future inquiries in the area of wireless mobile learning devices should be focused on how this technology is changing interactions between learners, digital content, and technology, and how education will need to adapt to a world that is increasingly mobile and connected (van 't Hooft & Swan, 2007). Other questions of interest include: How can we create the best possible tools for learning without the technology getting in the way? How can mobile technologies best accommodate and support active and collaborative learning? How does context affect learning, especially when it constantly changes?

Finally, the current dearth of large-scale implementations of highly mobile devices can be blamed on a variety of reasons. For one, educational institutions usually do not have the resources to provide every student with a digital tool. Second, they haven't figured out yet how to take advantage of the mobile devices that many students already own or have access to. In fact, in many instances, it is not only inability, but also unwillingness on the part of the "traditional" education sectors to perceive these same devices that are an integral part of the everyday life of a young learner as a viable platform. As a result, schools are banning devices such as mobile phones, when they could be used as mixed-media creators and communicators, and are instead trying to hold on to a computing model based on desktops and laptops that is slowly coming to its demise. Ultimately, the plethora of mobile devices and the manner in which they are embedded in the lives of young learners will raise the question, "Who supplies education?"

CONCLUSION

Younger generations are not fazed by constant change. They are growing up in societies that are in constant flux, where access to information is overwhelming, and technology is mobile, connected, and constant; they do not know a world without it. They know how to use the hardware and software, and are not afraid to learn to use new tools. However, they need guidance in learning how to use digital tools in ways that are meaningful, productive, responsible, and safe. In order for this to happen, teaching and learning in educational institutions will have to change to accommodate the use of highly mobile devices anytime and anywhere. Only then will students gain the knowledge, skills, and attitudes that are needed to be successful in the 21st century world.

REFERENCES

Alexander, B. (2004). Going nomadic: Mobile learning in higher education. *EDUCAUSE Review, 39*(5), 29-35.

Bayus, B. L., Jain, S., & Rao, A. G. (1997). Too little, too late: Introduction timing and new product performance in the personal digital assistant industry. *Journal of Marketing Research, 34*(1), 50-63.

Becker, H., Ravitz, J. L., & Wong, Y. (1999). *Teacher and teacher-directed student use of computers and software*. Report #3, Teaching, learning, and computing: 1998 national survey. Irvine, CA: Center for Research on Information Technology and Organizations, University of California, Irvine.

Bell, G., & Dourish, P. (2007). Yesterday's tomorrows: Notes on ubiquitous computing's dominant vision, *Personal and Ubiquitous Computing, 11*(2), 133-143.

Breck, J. (2006). Why is education not in the ubiquitous web world picture? *Educational Technology, 46*(4), 43-46.

Cole, H., & Stanton, D. (2003). Designing mobile technologies to support co-present collaboration. *Personal and Ubiquitous Computing, 7*, 365-371.

Fung, P., Hennessy, S., & O'Shea, T. (1998). Pocketbook computing: A paradigm shift? *Computers in the Schools, 14*, 109-118.

HPC Factor. (2004). *A brief history of Windows CE: The beginning is always a very good place to start.* Retrieved October 12, 2004, from http://www.hpcfactor.com/support/windowsce/

Inkpen, K. (2001). Designing handheld technologies for kids. *Personal Technologies Journal, 3*, 81-89. *Proceedings of CHI, Conference on Human Factors in Computing Systems.* Seattle: WA.

McClintock, R. (1999). *The Educator's Manifesto: Renewing the progressive bond with posterity through the social construction of digital learning communities*. New York: Institute for Learning Technologies, Teachers College, Columbia University. Retrieved March 21, 2007, from http://www.ilt.columbia.edu/publications/manifesto/contents.html

Norris, C., & Soloway, E. (2004). Envisioning the handheld-centric classroom. *Journal of Educational Computing Research, 30*(4), 281-294.

Rogers, Y. (2006). Moving on from Weiser's vision of calm computing: Engaging ubicomp experiences. In P. Dourish & A. Friday (Eds.), *Ubicomp, LNCS 4206* (pp. 404-421). Berlin: Springer-Verlag.

Rogers, Y., & Price, S. (2007). Using ubiquitous computing to extend and enhance learning experiences. In M. van 't Hooft & K. Swan (Eds), *Ubiquitous computing in education: Invisible technology, visible impact* (pp. 460-488). Mahwah, NJ: Lawrence Erlbaum Associates.

Roschelle, J. (2003). Unlocking the value of wireless mobile devices. *Journal of Computer Assisted Learning, 19*, 260-272.

Roschelle, J. & Pea, R. (2002). A walk on the WILD side: How wireless handhelds may change computer-supported collaborative learning. *International Journal of Cognition and Technology, 1*(1), 145-272.

Roth, J. (2002). Patterns of mobile interaction. *Personal and Ubiquitous Computing, 6*, 282-289.

Roush, W. (2005). Social machines. *Technology Review, 108*(8), 45-53.

Sharples, M. (2000a). *Disruptive devices: Personal technologies and education*. (Educational Technology Research Paper Series 11). Birmingham, United Kingdom: University of Birmingham.

Sharples, M. (2000b). The design of personal mo-

bile technologies for lifelong learning. *Computers and Education, 34*, 177-193.

Sharples, M. (2005, October 5). *Re-thinking learning for the mobile age*. Retrieved September 26, 2006, from http://www.noe-kaleidoscope.org/pub/lastnews/last-0-read159-display

Sharples, M., Taylor, J., & Vavoula, G. (2007). A theory of learning for the mobile age. In R. Andrews & C. Haythornthwaite (Eds.), *The Sage handbook of elearning research* (pp. 221-47). London: Sage.

Shin, N., Norris, C., & Soloway, E. (2007). Findings from early research on one-to-one handheld use in K-12 education. In M. van 't Hooft & K. Swan (Eds.), *Ubiquitous computing in education: Invisible technology, visible impact* (pp. 19-39). Mahwah, NJ: Lawrence Erlbaum Associates.

Swan, K., Cook, D., Kratcoski, A., Lin, Y., Schenker, J., & van 't Hooft, M. A. H. (2006). Ubiquitous computing: Rethinking teaching, learning and technology integration. In S. Tettegah & R. Hunter (Eds.). *Education and technology: Issues in applications, policy, and administration* (pp. 231-252). New York: Elsevier.

Swan, K., Kratcoski, A., & van 't Hooft, M. (2007). Highly mobile devices, pedagogical possibilities, and how teaching needs to be reconceptualized to realize them. *Educational Technology, 47*(3), 10-12.

Tatar, D., Roschelle, J., Vahey, P., & Penuel, W. R. (2003). Handhelds go to school: Lessons learned. *IEEE Computer, 36*(9), 30-37.

Thornburg, D. D. (2006). Emerging trends in educational computing. *Educational Technology, 46*(2), 62-63.

Vahey, P., & Crawford, V. (2002). *Palm education pioneers: Final report*. Menlo Park, CA: SRI International.

van 't Hooft, M., & Swan, K. (Eds.). (2007). *Ubiquitous computing in education: Invisible technology, visible impact*. Mahwah, NJ: Lawrence Erlbaum Associates.

van 't Hooft, M., & Vahey, P. (2007). Introduction to the special issue. *Educational Technology Magazine, 47*(3), 3-5.

Vavoula, G., Sharples, M., Lonsdale, P., Rudman, P., & Meek, J. (2007). Learning bridges: A role for mobile technology in education. *Educational Technology Magazine, 47*(3), 33-37.

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265* (3), 94-95, 98-102.

Williams, B. (2004). *We're getting wired, we're going mobile, what's next?* Eugene, Oregon: ISTE Publications.

KEY TERMS

Bluetooth: An industrial specification for wireless personal area networks (PANs). Bluetooth allows devices to connect and exchange information over a secure, globally unlicensed short-range radio frequency.

GPS: Global positioning system. It consists of a receiver that uses three or more GPS satellites to calculate its location.

Highly Mobile Devices: Digital devices that have high mobility, a small footprint, computational and display capabilities to view, collect, or otherwise use representations and/or large amounts of data; and the ability to support collaboration and/or data sharing. Devices include PDAs, mobile phones, some tablet computers, networked graphing calculators, UMPCs, the new generation of handheld gaming systems, iPods, motes, and data loggers.

Informal Learning: Learning in which both goals and processes of learning are defined by the learner, and where the learning is situated rather than preestablished.

M-Learning: "The processes of coming to know through conversations across multiple contexts amongst people and personal interactive technologies" (Sharples, Taylor, & Vavoula, 2007).

Mobile Phone: A portable electronic device for personal telecommunications over long distances, often supplemented by features such as instant messaging, Internet and e-mail access, global positioning (GPS), and a digital camera. Most mobile phones connect to a cellular network.

PDA: Personal digital assistant. A handheld computing device that is characterized by a touch screen, a memory card slot and Infrared, Wi-Fi, and/or Bluetooth for connectivity. Data can be synchronized between PDAs and desktop or laptop computers.

UMPC: Ultra mobile personal computer. A small form-factor tablet PC (larger than a PDA but smaller than a tablet PC) that features a touch screen no larger than 7 inches, flexible navigation and input options, and WiFi connectivity.

WiFi: Short for "wireless fidelity" and a popular term for a high-frequency wireless local area network (WLAN), using the 802.11 protocol.

This work was previously published in Encyclopedia of Information Technology Curriculum Integration, edited by L. Tomei, pp. 37-42, copyright 2008 by Information Science Publishing (an imprint of IGI Global).

Chapter 1.14
Current Status of Mobile Wireless Technology and Digital Multimedia Broadcasting*

J. P. Shim
Mississippi State University, USA

Kyungmo Ahn
Kyunghee University, Korea

Julie M. Shim
Soldier Design LLC, USA

ABSTRACT

The purpose of this chapter is to present an overview of wireless mobile technology, its applications, with a focus on digital multimedia broadcasting (DMB) technology. The chapter also explores the research methodology regarding users' perception on DMB cellular phones and presents empirical findings. Implications for future research are presented. The report attempts to provide stimulating answers by investigating the following questions: (1) Do users perceive easy access to DMB applications as a satisfactory service offered by DMB service providers? (2) Do users perceive high-quality DMB program content as a satisfactory service offered by the DMB service providers? (3) Are there differences between different age groups in terms of their perception of DMB phone prices, phone usage time, program content, and services?

INTRODUCTION

Wireless mobile technology and handheld devices are dramatically changing the degrees of interaction throughout the world, further creating a ubiquitous network society. The emergence of these wireless devices has increased accuracy, ease-of-use, and access rate, all of which is increasingly essential as the volume of information handled by users expands at an accelerated pace. Mobile TV broadcasting technology, as a nascent industry, has been paving a new way to create an intersection of telecommunication and media industries, all of which offers new opportunities

to device makers, content producers, and mobile network operators.

There are currently various wireless connectivity standards (e.g., Wi-Fi, Bluetooth, Radio Frequency Identification [RFID], etc.), which have been expanding across all vertical industries, in an era of mobile and ubiquitous computing, which provides access to anything, anytime, and anywhere. Mobile TV technologies have been creating a buzz, as it adds a new dimension to the "on the go" mobility factor—simultaneous audio and video services are broadcasted in real-time to mobile devices in motion, such as mobile TV-enabled phones, PDAs, and car receivers.

There are currently three major competing standards: digital video broadcasting for handhelds (DVB-H), which is going through trial phases in Europe; digital multimedia broadcasting (DMB), which has been adopted in South Korea and Japan; and MediaFLO (QUALCOMM Inc., 2005), which is currently in trial phase in the United States with plans to launch by late 2007. The competition scheme is further intensified given the challenge of how quickly terrestrial and satellite DMB can be deployed and commercialized throughout countries such as Korea, Japan, and Europe. Additionally, there is pressure to recoup the costs with creating the network and catapult the technology to the ranks of industry standard.

The purpose of this chapter is to present an overview of wireless mobile technology, its applications, with a focus on DMB technology. The chapter also explores the research methodology regarding users' perception on DMB cellular phones and presents empirical findings from Study Phases I and II, along with actual DMB subscriber usage results. Implications for future research are presented.

Given that the research topic of DMB has not yet been covered extensively, the use of qualitative methods is considered advantageous when exploring the topic to develop theoretical variables, which may then be employed in quantitative research. Thus, with the difference found between the DMB cellular phone usage experience and traditional cellular phone usage, qualitative methodology was applied to Study Phase I. The project was then triangulated by the use of quantitative methodology in Study Phase II to develop an additional understanding of the DMB cellular phone users' experiences as identified in Study Phase I.

The report attempts to provide stimulating answers by investigating the following questions: (1) Do users perceive easy access to DMB applications as a satisfactory service offered by DMB service providers? (2) Do users perceive high-quality DMB program contents as a satisfactory service offered by the DMB service providers? (3) Are there differences between different age groups in terms of their perception of DMB phone prices, phone usage time, program contents, and services?

WIRELESS MOBILE TECHNOLOGIES: CURRENT STATUS AND CONCEPTS

Over the last decade, wireless technologies have attracted unprecedented attention from wireless service providers, developers, vendors, and users. These wireless technologies provide many connection points to the Internet between mobile phones and other portable handheld devices to earpieces and handsets. These technologies include Wi-Fi hotspots, Bluetooth, WiMAX, wireless broadband Internet (WiBRO), RFID, and others. Wi-Fi hotspots, with a distance and penetration of approximately 50 feet, are physical addresses where people can connect to a public wireless network, such as a cafe, hotel, or airport. WiMAX is a metropolitan-scale wireless technology with speeds over 1Mbps and a longer range than Wi-Fi. WiBRO, the Korean version of WiMAX, allows users to be connected to the Internet while in motion, even in cars traveling up to 100 kilometers

per hour. It is anticipated that users may one day seamlessly switch between networks multiple times per day, depending on the service offered by a specific network service provider.

Many industries have seen the benefits of these wireless technology applications, of which some will be described here. For local, federal, and state agencies, wireless connections provide for GPS functionality, along with real-time vehicle tracking, navigation, and fleet management. For automated logistics and retail industries, RFID tags will give information on just-in-time inventory or shipment location, security status, and even environmental conditions inside the freight. In the health care industry, the wireless applications include patient and equipment monitoring, and telemedicine through the monitoring of an outpatient's heart via continuous electrocardiograms (ECG). Other applications already on the radar: handsets that function as a blood pressure monitor, a blood glucose meter, and wireless pacemaker. One of the hurdles that wireless solution carriers have to overcome is the cost of the devices, and whether insurance companies are willing to cover or share the costs. The wireless technology allows government officials and emergency response teams to stay informed of critical information in the event of an emergency or a disaster that affects wire line services, much like Katrina; these include advanced warnings and public alerts, emergency telecommunications services, global monitoring for environment, and assistance with search and rescue (SAR).

PC World, an online technology magazine, recently reported that the number of Wi-Fi hotspots reached the 100,000 mark globally.[1] Businesses are realizing the value-added service by offering free or paid wireless services to attract customers. Analysts believe that locations such as school campuses and citywide deployment of WiMax technology will benefit users.

a. **United States:** Wi-Fi integration into retail, hospitality, restaurant, and tourism industries has been instrumental for marketing plans, particularly for franchise venues, including Starbucks and McDonald's.
b. **Asia/Pacific:** An article in *The Australian* (2003, March 4) described that 200 restaurants in Australia have migrated away from taking orders via pen and paper to using wireless handhelds to relay orders to the kitchen/bar staff. In addition to offering this type of service, Japan's NTT DoCoMo introduced its iMode Felica handset, enabling users to scan their handsets as their mobile wallets (m-wallets), eliminating the need to carry a credit card, identification, and keys. The feature allows for conducting financial transactions, purchasing services/products, or opening electronic locks.[2] The issue at hand is the different business models of the wireless carrier and that of the credit card companies.

DIGITAL MULTIMEDIA BROADCASTING: CURRENT STATUS AND CONCEPTS

Digital multimedia broadcasting (DMB) is a process of broadcasting multimedia over the Internet or satellite that can be tuned in by multimedia players, capable of playing back the multimedia program.[3] DMB is an extension of digital audio broadcasting (DAB), which is based on the European Eureka 147 DAB Radio standard. DMB technology has two sub-standards: satellite-DMB [S-DMB] and terrestrial-DMB [T-DMB]. While both S-DMB and T-DMB broadcasts television to handheld devices in motion, the difference lies in the transmission method: via satellite versus land-based towers. These real-time transmissions allow users to view live TV programs, including news, reality shows, or sports games on their DMB cellular phones in the subway.

With mobile growth two or three times that of Europe and North America (Budde, 2002), Japan

and Korea have been known for their cutting edge technological innovations and tech-savvy consumers. Korea is one of the world's most broadband-connected countries, with a high penetration rate (Lee, 2003; Shim et al., 2006a; Shim et al., 2006b). The government initiatives have been instrumental in this arena, as the government's hands-on style has created the IT infrastructure necessary to power the latest technological tools. The mobile markets in Japan and Korea have become optimal testing grounds for mobile operators and manufacturers before rolling out products in the rest of the world, given the consumers' insatiable appetite of acquiring the latest technologies, early acceptance behavior, and education fever.

In Asia-Pacific and Europe, considered to be the power houses of the mobile gaming industry, wireless gaming and instant messages have exceeded expectations. In North America, music downloads and e-mails have become essential. As the market for mobile applications, (including short message service [SMS], ring-tones, games, music, videos) is becoming more saturated, more wireless applications have become integrated into most consumer electronics devices, from digital cameras to video game consoles. With over 85% cellular phone penetration rate, Korea introduced the world's first DMB mobile-enabled phone, or "TV-on-the-go" in 2005.[4] While Japan currently provides S-DMB services designed for car receivers, Korea has been the only country to provide full-blown S-DMB and T-DMB services on cellular phones while in motion (including car receivers) by late 2006. With T-DMB and S-DMB services already launched in Korea, several countries in Europe, and the U.S. are planning to launch DVB-H services by the end of 2007. Informa, a consultancy, says there will be 125 million mobile TV users by 2010.[5]

The history of DMB began with the development of DAB services during the mid-1990s in the U.S. and Europe (Korean Society for Journalism and Communication Studies, 2003; Nyberg, 2004). The current status of Mobile TV services in the U.S., Europe, Japan, and Korea is shown in Table 1 (Shim, 2005b).

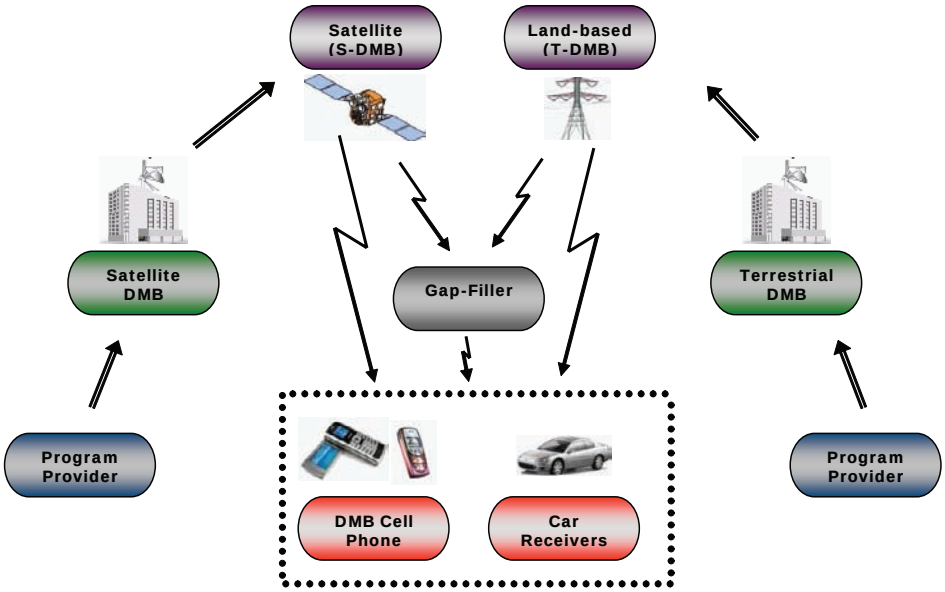
As shown in Figure 1, DMB program producers provide a variety of programs and content to the DMB center, which broadcasts through either

Table 1. Current status of mobile TV services in various countries

Country	USA		Europe		Japan	Korea	
Mobile TV technology	MediaFlo	DVB-H	DVB-H	T-DMB	S-DMB	S-DMB	T-DMB
Receiving device	Car receiver	Car receiver	Mobile TV- phone, Car receiver	Car receiver	Car receiver	Mobile TV- phone, Car receiver	Car receivers
Service launch date	2006	2006	2006	2006	2004	May 2005	Dec 2005

Sources: The Korea Times, (2005, January 18) "Korea's Free Mobile Broadcasting Faces Snag".
KORA Research 2003-10,. (2004, May). "A Market Policy Study on DMB".
M. H. Eom, "T-DMB Overview in Korea," (2006, April). Proceedings of 2006 Wireless Telecommunications Symposium, Pomona, CA,

Figure 1. An overview of the mobile wireless framework



satellites or towers. Thus, the DMB cellular phone users receive content and programs through satellites, towers, or "gap-fillers" (small base stations) to ensure there are no reception problems, even in underground subways (Shim, 2005a).

Consumers are increasingly gravitating towards customized devices and features, as a miniaturized interactive entertainment center is packaged into the cellular phone, complete with an MP3 player, multi-megapixel camera, digital video recorder, CD-quality audio, and a selection of satellite broadcast television and audio channels (Olla & Atkinson, 2004) as they can choose from television and audio on-demand and simultaneously make phone calls. The mobile TV-enabled phone, equipped with these features, has become more than integrated into one's lifestyle, as it becomes an extension of the consumer's identity. The handset carriers are in the process of yet again trying to capitalize on producing fashion-forward phones and portable gaming consoles.

DMB data service is a framework of the following groups: data provider, audio/video content producer, DMB producer, advertiser, and customer. A schematic view of DMB data service and the components, shown in Figure 2, provides a basic understanding of the general structure of the DMB business model. The figure also shows interaction of the DMB producer with other groups of DMB data services.

For example, the DMB producer provides various content and programs to customers for a service fee. The DMB producer charges an advertising fee to the advertiser, from whom customers can purchase directly for advertised services via the DMB device. The audio/video content producer and data provider each provide various contents to the DMB producer for a fee. The perceived richness of the medium should have an impact on the use of the communication medium (Daft & Lengel, 1986; Smagt, 2000). The rich media is more appropriate in ambiguous

communications situations, which emphasizes Daft and Lengel's valuable contribution of placing equivocality high in the business and information systems field.

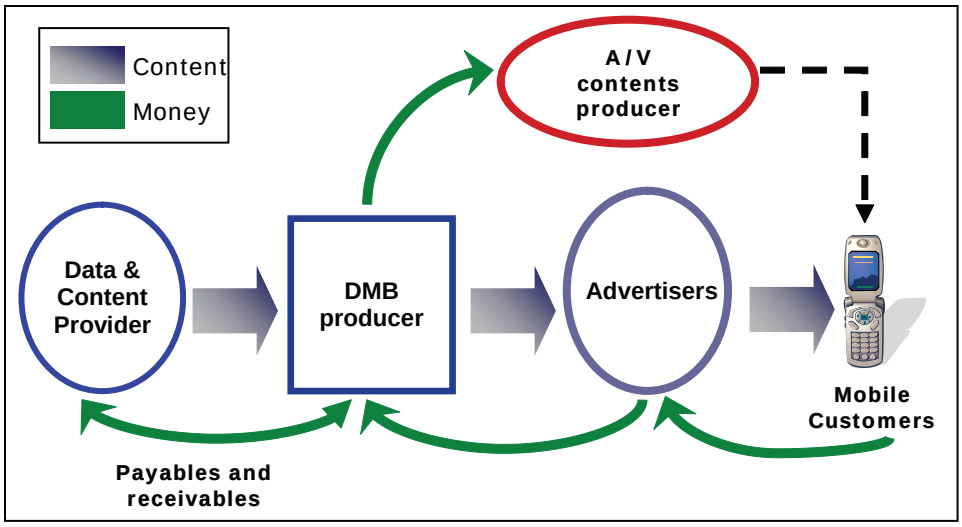
There exists a rich body of knowledge of technology adoption and diffusion, including the digital multimedia broadcasting technology. For example, several theoretical backgrounds, such as institutional theory, technology acceptance model (TAM) (Venkatesh & Davis, 2000), and diffusion of innovation theory (Gharavi, Love, & Cheng, 2004; Rogers, 1983) explain the DMB technology adoption at an individual, organizational, and industry level (Lee, 2003; Shim, 2005a, 2005b). Among the theories, Lee and Shim both describe the major factors behind Korea's information and communication technology diffusion such as: external factors (global economy, government policies), innovation factors (usefulness, ease of use, self-efficacy), and imitation factors (subjective norm of belongingness, word of mouth). The authors believe that either the diffusion theory (such as external, internal, and mixed influence models), or TAM (such as perceived usefulness and perceived ease of use), or the combination of both can be applied behind DMB cellular phone adoption and diffusion.

RESEARCH METHODOLOGY

A recent study demonstrates a higher number of DMB viewers than regular TV viewers during the daytime (Figure 3). Since the DMB cellular phone captures the content-on-demand aspect, the DMB phone service (S-DMB) are optimal for the on-the-go daytime enthusiasts.

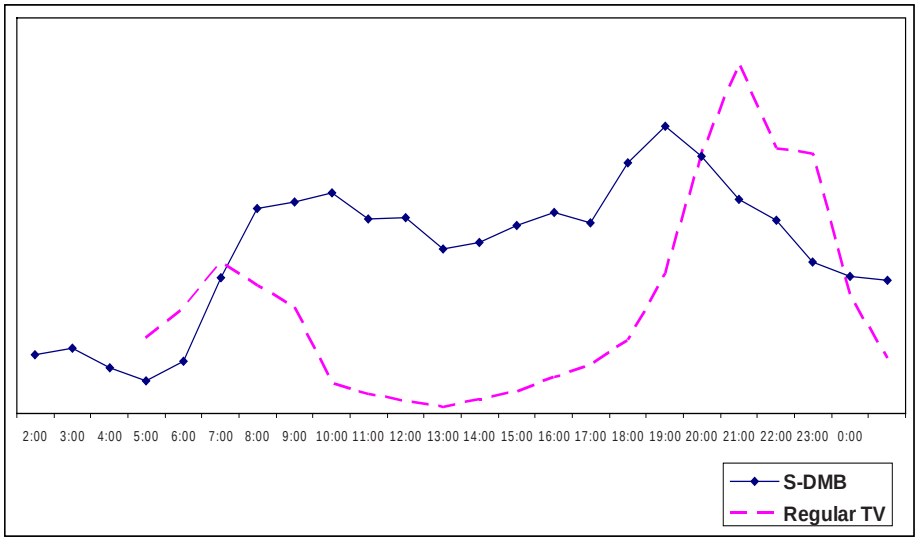
To determine how integral DMB phones have been and will be in consumers' daily lives, the authors conducted qualitative and quantitative

Figure 2. A schematic view of DMB data service business model



Source: Modified from KORA Research 2003-10. (2004, May). "A Market Policy Study on DMB," Research Report of Korea Radio Station Management Agency.

Figure 3. The percentage of viewing on S-DMB vs. regular TV



Source: Suh, Y. (2005, November). "Current Overview of S-DMB," TU Media.

analyses. Study Phase I describes the use of the qualitative research method, specifically the existential phenomenological method. Study Phase II describes the quantitative research methods including the survey questionnaire (Shim, Shin, & Nottingham, 2002).

STUDY PHASE I: QUALITATIVE ANALYSIS

Although quantitative instruments serve as valid methods to study the perceived use of DMB phones, qualitative research methods, such as interviewing, can reveal the function of variables perhaps overlooked by survey designers. The current project was designed to employ the qualitative technique of existential phenomenology. Thus, it develops an in-depth understanding of the new concept of DMB usage by investigating respondents' reports of their DMB phone usage experiences. With this data collection technique, the respondent is encouraged to describe in-depth the personally experienced phenomenon (Thompson, Locander, & Pollio, 1989).

Existential phenomenology was selected among various qualitative methods, such as case studies and ethnography, because of its attention to a respondent's individualistic, subjective expression of an actual live experience of the situation of interest. Such reflection on a single experience encourages the perceiver to focus on nuances that would likely escape the broader brush of a researcher's selection of choices among a pre-set list of quantitative dimensions or escape even the surface comparison of reports of respondents' experiences. Existential phenomenology encourages the respondent to consider specific and live events. The goal is to discover patterns of experiences (Thompson et al., 1989).

Since the purpose of existential phenomenology is to describe the experience as it is lived, the interview has been found to be a powerful tool for attaining in-depth understanding of another person's experience (Kvale, 1983). Research analysis of interview-derived information is considered

valid because the respondents' own words are used to understand their experiences (Feagin, Orum, & Sjoberg, 1991). Accordingly, respondents in this research were presented with a set of open-ended questions designed to encourage them to discuss and describe their experiences with DMB phone usage. To determine a specific set of key factors that would be of critical concern to DMB users, 19 respondents in Korea were enlisted in Study Phase I.

A purposive sample is deemed appropriate for exploratory research designed to query respondents who have experienced a phenomenon of interest. Thus, the networking technique was utilized to obtain a purposive sample of individuals who had interacted with the DMB cellular phone services. These respondents were then asked to name additional individuals who had experienced DMB services. Thus, aside from the requirement of the respondents' familiarity with the DMB services, demographic characteristics of the sample resulted by random chance.

The majority of respondents were well-educated young professionals with a zealous tech-gadget nature, affluent, computer proficient, and somewhat knowledgeable about DMB services. Although this sample clearly is not representative of the population at large, the sample profile corresponds with what the authors presumed to be identified as a typical DMB service user. Thus, the experiences relayed by these respondents are considered to be a reasonable representation of a random sampling of regular DMB cellular phone users. After respondents were assured of confidentiality and protection of their privacy, each tape-recorded interview lasted 20-30 minutes. Each interview began with open-ended questions posed in a conversational format to encourage the respondent to develop a dialogue with the interviewer, providing the context from which the respondent's descriptions of his or her own DMB service experience could flow freely and in detail. Participants were encouraged to discuss not only their DMB services experiences, but also their attitudes and perceptions regarding negative and positive aspects of DMB services.

Such in-depth descriptions have been found to be beneficial in revealing emotional and behavioral underpinnings of overt user behavior. In reality, the act of a respondent's description of a specific experience in-depth, frequently results in further personal insights that arise through the revival of the experience. The respondents were asked to describe their main reasons for purchasing DMB cellular phones, which varied: "to gain information access," "to spot the latest trends," "for education or entertainment," "to watch TV while commuting," and "for movies, dramas, and shopping." Their personal positive experiences were: "mobility—a deviation from a fixed location point," "high quality reception," "convenience," "accessible anytime/anywhere," "lifestyle change," "great for commuting," and "good for managing time." On the other hand, the negative aspects they experienced included: "expensive device," "reception problem," "low battery hours with limited usage time (e.g., 2-3 hours)." Most respondents reported that the following areas would have great potential for future DMB applications and content: information access, education/learning, e-trading, retail, tourism, and entertainment. In Study Phase II, these themes were reconstructed to set up independent variables for the quantitative analysis.

STUDY PHASE II: QUANTITATIVE ANALYSIS

To determine the extent to which DMB phones are being used as the latest multimedia product, the authors developed a questionnaire. DMB has a wide array of advantages: personalized, live media (television, radio, or data broadcasts) that can be viewed on-demand anytime; the mobility of the phone which receives satellite and terrestrial television broadcast signals even at high speeds or underground; and an interactive handset

into which one can speak via the handset while watching TV programs. The research instrument underwent two pretests. The first pretest involved administering the questionnaire to 25 graduate and undergraduate students at a large university in Seoul, Korea. The questions, which concerned price, usage time, program content, and services were modified to reduce the effects of proximity bias on the responses, with several questions reworded for clarity. The second pretest was conducted at a DMB phone service provider company to ensure the content validity. A five-point Likert scale was used for recording the responses.

DMB will not be successful if content and service providers fail to provide high quality service, a variety of content, and reasonable prices for services and handsets (Teng, 2005). Several research studies demonstrated that there are differences among age groups on factors such as technology adoption and usage (Larsen & Sorebo, 2005; Ventatesh, 2000). It is believed that older generations are more anxious about the use of technologies than the younger generations. A number of research studies have supported this belief (Gilbert, Lee-Kelley, & Barton, 2003). Based on the theories and research questions along with Study Phases I and II, the authors developed the following six hypotheses:

H_1: The user's easy access to DMB service is perceived as a satisfactory service offered by the DMB service provider.

H_2: Premium (excellent) content of DMB programs corresponds with a good quality DMB service provider.

H_3: There is a difference between different age groups and their perceived value of DMB handset price.

H_4: There is a difference between different age groups and their perceived value of DMB phone usage time.

H_5: There is a difference between different age groups and their perceived value of DMB program content.

H_6: There is a difference between different age groups and their perceived value of DMB services.

The authors and their research assistants distributed the questionnaire to 300 randomly selected individuals inside the Korea Convention Exhibition Center (COEX) and Korea World Trade Center during January and February 2005. Of the 300 randomly selected individuals' responses, 264 were valid. The two-page questionnaire was divided into three sections with a total of 32 questions. In Section 1, the authors asked the randomly selected participants about DMB services, such as information sources about DMB services, user satisfaction ratings, influential factors when choosing DMB services, DMB applications, and others. The questions in Section 2 covered the participants' perceived values of DMB application services. Section 3 inquired of participants' demographics.

DATA ANALYSIS AND FINDINGS

The 264 usable research instruments collected from the respondents were well represented in terms of gender, age, and occupation. Statistical Package for the Social Sciences (SPSS) was used to calculate descriptive statistics and perform a confirmatory factor analysis. The respondents' primary occupations included: students (51.9%), IT staff (15.2%), government employees (13.3%), professionals (7.6%), self-employees (4.1%), housewives (3%), and others (4%). Approximately 73.8% of the sample respondents indicated that they had either undergraduate (64%) or graduate school (9.8%) education.

The respondents were well represented in terms of gender and age. About 30% of the sample

respondents had not heard about DMB. Of the 70% of respondents who had heard about DMB, the main sources included: TV (26%), newspaper (20%), Internet (15%), friends (6%), and others (33%). About one-fifth (20.1%) of the respondents were utilizing DMB services. Of those respondents, 62.2% were satisfied with their current DMB service whereas 30.3% were only satisfied on a mediocre level. In other words, only 7.5% of the current DMB users were not satisfied with their DMB services. The current users accessed their DMB phones for news and information; leisure and tourism; public relations (marketing); shopping; games; and education. The users believed that the DMB services would impact service industries such as tourism and retail.

The results also indicated that among the sample respondents, the non-users felt that the following major factors would be taken into consideration when choosing DMB services for the future: (1) pricing of DMB cellular handset, (2) video quality, (3) program content, (4) quality [of DMB cellular handset], (5) ease of use, and (6) others [e.g., customer service by the DMB cell phone manufacturer or service provider; brand image and perception]. The aforementioned results from the sample respondents were very astonishingly similar to the 19 interviewees' perceived values.

The independent variables that determine DMB services are: price, usage time, and program content. The dependent variable is DMB service. Table 2 provides a definition of each of these variables. The reliability measure (construct validity) for these constructs was Cronbach's coefficient (alpha). Even though the general rule of thumb for reliability is a value of 0.8 (alpha), values of 0.6 or 0.7 may be considered adequate in some cases (Hair, 1998). Overall, the model provides a valid representation of the data and the constructs are reliable. The reliability test generated Cronbach's coefficient alpha of .7343 for the 12 items. From the analysis, it was concluded that the measure of 12 items was reliable. Coefficient alphas for the four constructs are shown in Table 3.

Table 2. Model construct

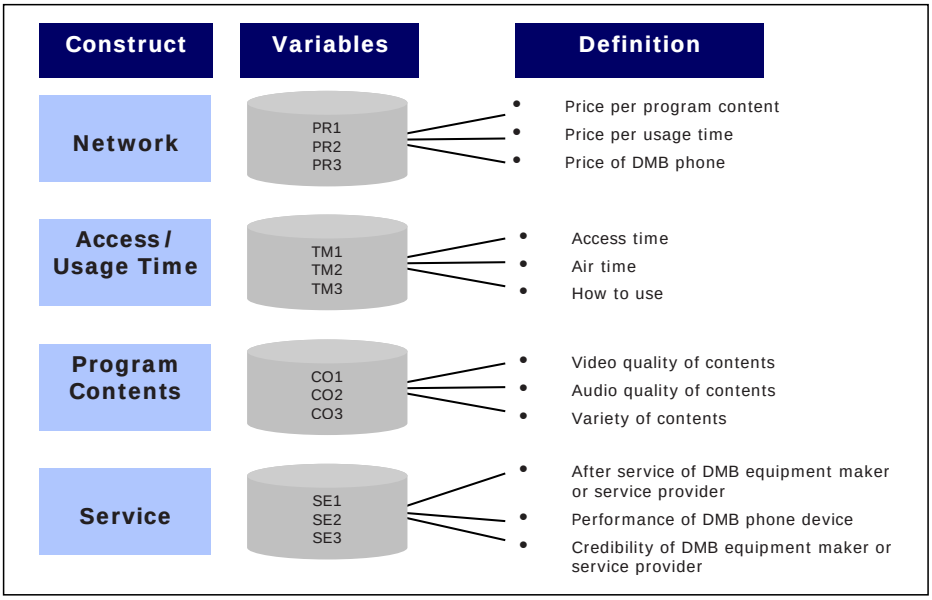


Table 3. Coefficient alpha for construct

Construct	Variables	Cronbach's alpha
Price	PR1, PR2, PR3	.7970
Access/Usage time	TM1, TM2, TM3	.6218
Program content	CO1, CO2, CO3	.8104
Service	SE1, SE2, SE3	.7081

Table 4. Construct: Factor loadings

Constructs	Loading	Eigenvalue	Communality (%)
PR1	.875		
PR2	.863	2.137	71.223
PR3	.791		
TM1	.792		
TM2	.744	1.710	56.986
TM3	.727		
CO1	.908		
CO2	.907	2.189	72.970
CO3	.737		
SE1	.838		
SE2	.782	1.897	63.219
SE3	.764		

Table 5. Correlation matrix for the constructs

	Price	Usage time	Program content	Service
Price	1.000			
Access/Usage time	.296**	1.000		
Program content	.454**	.497**	1.000	
Service	.266**	.481**	.511**	1.000

** *P<0.01*

Table 6. Analysis of service performance

Independent Variable	Dependent Variable: Service	
	Beta	t-value
Price	0.008	0.141 (sig = .888)
Access/Usage time	0.300	5.104 (sig = .000)
Program contents	0.358	5.689 (sig = .000)
R^2	0.330	
F	42.602	
Sig.	.000	

A series of principal components factor analyses using a VARIMAX rotation were used to assess the unidimensionality in this study. Eigenvalues of at least 1.0 were used to assess the number of factors to extract. The dimensionality of each factor was assessed by examining factor loadings. Factor loadings on construct are shown in Table 4.

Assessing dimensionality involves examining the inter-correlations among the major constructs. A correlation matrix for the constructs is shown in Table 5. The inter-construct correlation coefficients were all positive and significant at less than 0.01 (see Table 5).

The *t* test was used in the quantitative analysis. The price factor of the DMB phone usage is not an issue if the user perceives the DMB program content to be valuable. Table 6 also showed that DMB service was affected by program content (beta=0.358, t-value=5.689). The users associate easy access/connection time to the DMB services with reliability provided by DMB equipment makers or service providers (beta=0.300, t-value=5.104). H_1 and H_2 were supported. ANOVA and Duncan test were used to evaluate hypotheses H_3, H_4, H_5, and H_6.

DMB PHONE PRICE AND RELATED FEES

The users were asked to rate the importance of price issues of the DMB handset and related service fees when selecting a DMB cell phone. These issues include price per program content, price

Table 7. ANOVA and Duncan Test of DMB phone price and related fees

	Sum of squares	df	Mean Square	F	Sig.
Between groups	11.161	3	3.720	12.583	.000
Within groups	76.879	260	.296		
Total	88.040	263			

7a. ANOVA

Age	N	Subset for alpha = .05		
		1	2	3
Teens	45			4.6889
20s	116		4.2328	
30s	52		4.3269	
40s and older	51	4.0261		
Sig.		1.000	.354	1.000

7b. Duncan Test

per usage time, and price of the DMB handset. The mean response among the teens was 4.46 (on a scale of 1=unimportant and 5=very important); 4.23 for 20s, and 4.33 for 30s. The mean response among the older generations (40s and older) was 4.0. Table 7b showed that there were significant differences among teens and the other age groups (20s, 30s, 40s, and older). And the 20s and 30s age group perceived the DMB phone price and related fees differently, when compared with teens, and the 40s and older age group.

In an effort to explain this unexpected finding, the authors used analysis of variance (ANOVA) to see if there were any significant differences between the DMB handset price/related fees and age group. As shown in Table 7a, the difference is statistically significant (F=12.583, df=3, 260, p=0.000), which demonstrates that the younger generation is willing to pay the current market price for the DMB handset and related services, given that they perceive the content to be useful and worthwhile. This supports H_3, as it validates that there is a difference between the age groups and their perceptions of DMB phone price.

ACCESS/USAGE TIME

The users were asked to rate the importance of access/usage time issues of DMB services and handset when selecting a DMB cellular phone. These issues include access time, air time, and the time it takes to get familiarized with the DMB handset and services. The mean response among the teens was 4.02 (on a scale of 1=unimportant and 5=very important); 3.82 for 30s, 3.86 for 40s and older. Table 8b demonstrates a slight discrepancy between teens and those in their 20s, but no significant divergences among other age groups (30s, 40s and older). The ANOVA test showed that there was not a significant difference between age groups and their judgments of the importance placed on the DMB access/usage time (see Table 8a). This does not support H_4, as it validates that there is no

Table 8. ANOVA and Duncan Test of access/usage time

	Sum of Squares	df	Mean Square	F	Sig.
Between groups	2.943	3	.981	2.502	.060
Within groups	101.935	260	.392		
Total	104.878	263			

8a. ANOVA

Age	N	Subset for alpha = .05	
		1	2
Teens	45		4.0296
20s	116	3.7328	
30s	52	3.8269	3.8269
40s and older	51	3.8627	3.8627
Sig.		.298	.102

8b. Duncan Test

Table 9. ANOVA and Duncan of program content

	Sum of Squares	df	Mean Square	F	Sig.
Between groups	5.870	3	1.957	6.304	.000
Within groups	80.689	260	.310		
Total	86.559	263			

9a. ANOVA

Age	N	Subset for alpha = .05	
		1	2
Teens	45		4.6963
20s	116	4.2902	
30s	52	4.3013	
40s and older	51	4.3268	
Sig.		.743	1.000

9b. Duncan Test

Table 10. ANOVA and Duncan Test of Service

	Sum of Squares	df	Mean Square	F	Sig.
Between groups	3.428	3	1.143	3.355	.019
Within groups	88.557	260	.341		
Total	91.985	263			

10a. ANOVA

Age	N	Subset for alpha = .05	
		1	2
Teens	45		Teens
20s	116	4.2241	20s
30s	52	4.3782	30s
40s and older	51	4.3137	40s and older
Sig.		.185	Sig.

10b. Duncan Test

difference between age groups and their perceptions of the DMB phone access/usage time. When focusing on strategic moves, the major players in the DMB market do not have to place as much emphasis on the end-users' access/usage time.

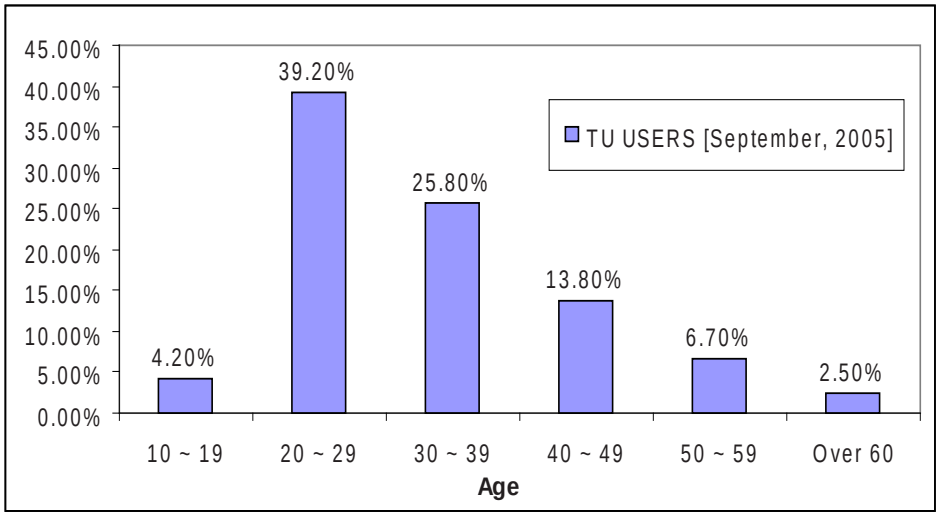
PROGRAM CONTENT

The users were asked to rate the importance of program content of DMB handsets when selecting a DMB cellular phone. These issues include video quality of content, audio quality of content, and a selection of content. As shown in Table 9a, the ANOVA test showed that there was a difference between various program contents and age groups ($F = 6.30$, $df = 3, 260$, $p=0.000$). The mean response among the teens was 4.69 (on a scale of 1=unimportant and 5=very important); 4.29 for the 20s age group, 4.30 for the 30s age group, and 4.32 for the 40s age group and older. Table 9b showed significant deviation between teens and the other age groups (20s, 30s, and 40s and older). The perception on program content for each age group (20s, 30s, and 40s and older) differed from the teens' perceptions. This supports H_5.

SERVICE

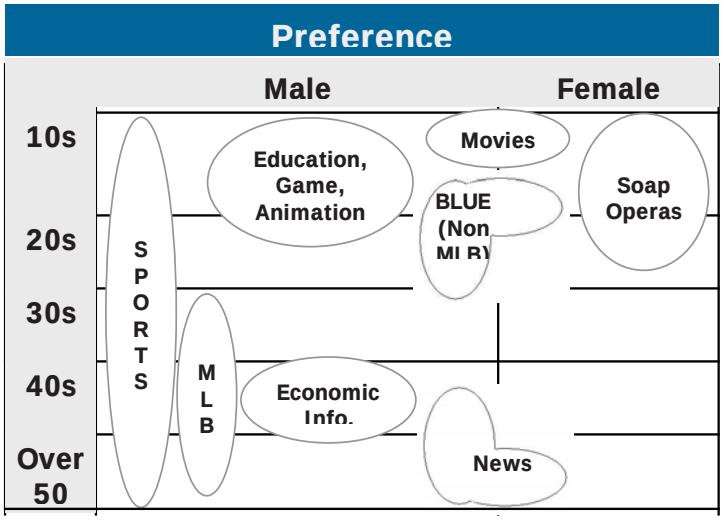
The users were asked to rate the importance of DMB services when selecting a DMB cellular phone. These services include after-service of the DMB equipment maker or service provider, performance of the DMB phone device, and the credibility of the DMB equipment maker or service provider. As shown in Table 10a, the ANOVA test showed that there was a difference between the age groups and their approach to the importance of DMB services ($F = 3.355$, $df = 3, 260$, $p <0.019$). The mean response among the teens was 4.54. Table 10b showed that there are significant differences among teens and other age groups (20s, and 40s and older). And each age group's (20s, and 40s and older) perception on services deviated significantly from the teens' perceptions. This supports H_6.

Figure 4. Actual DMB statistical usage data by age



Source: TU Media. (2005, September)

Figure 5. Preferences for various age groups and genders



Source: TU Media (2006)

As shown previously, all hypotheses (except H_4) were supported.

ACTUAL USAGE AND IMPLICATIONS

Although the exploratory study's results showed teens and the 20s age group as heavy users of DMB services, the actual DMB statistical usage data (see Figure 4) differed. The actual usage data was recently released from TU-Media (S-DMB service provider), which revealed that those in their late 20s, 30s, 40s, and 50s represent a large percentage of users of the following DMB services: soap operas, sports, and music program content (Suh, 2005). The results in Figure 5 show a correlation between preferences across age and gender. While the sports channel was the only preferred program among males to spread across all age groups, soap opera programs were preferred by female teens and those in their 20s. The authors believe that there are several reasons as to why the younger consumers may not be currently subscribed to S-DMB services: (1) S-DMB handsets (which retails for $600-$800) are too expensive; (2) The teens lack the extra out-of-pocket money to pay for the S-DMB $13 monthly service fee (and $20 activation fee); and (3) The parents do not feel justified in purchasing a DMB handset for their children's TV and gaming purposes. Additionally, most of the school-age children have little time to watch DMB program content due to the academic load. Furthermore, the actual usage results among various demographic groups for T-DMB services, once released, are expected to differ from that of S-DMB services given that T-DMB services are free and advertiser-supported.

CONCLUSION

The mobile TV standards (e.g., DMB, DVB-H, and MediaFLO) and wireless technologies will add a new dimension to the connectivity between enterprises and consumers as well as their access to information and entertainment. Given the demand for ubiquitous computing in an impatient,

technology-hungry, instant gratification-seeking population, the desire for mobile TV will continue to grow and soon mobile TV will be synonymous to today's radio in the long run (Kim, 2004). Similar to the interactive TV (iTV) (Tsaih, Chang, & Huang, 2005), the DMB has implications, which include: (1) service and content providers use the DMB as a vehicle for business-to-consumer (B2C) commerce via programs, content, and services; (2) consumers have real-time access to DMB services and programs on mobile phones, PDAs, and other mobile devices anytime and anywhere.

As mentioned earlier, the key issues for the DMB market include: (1) optimal capital investment levels to achieve adequate service coverage for T- DMB and S- DMB services; and (2) appropriate business models, with respect to advertising-supported vs. subscription services. The wireless mobile service industry has very complex issues, which span across technical, logistical, social, and cultural issues (Trappey, Trappey, Hou, & Chen, 2004). Thus, this requires cooperation among the cellular and network service providers, service developers, and equipment makers to collaborate with the government and users to create growth in the cellular telecommunications industry.

Although this research is based on exploratory methods, it still has its limitations. For example, the sample size was collected during the experimental/trial stages of S-DMB services in Korea. The authors reinforce the continuation of this research to solidify findings with an increased sample size of respondents collected during the actual stage of S-DMB and T-DMB services. In addition to this belief, the authors endorse the notion of longitudinal studies conducted to obtain more results. Furthermore, the authors strongly believe that the findings from this exploratory research will be valuable for the DMB service and content providers to gain insight into various age groups and their perceptions.

One of the implications of this paper of wireless mobile technologies and mobile TV is to demonstrate how important the government initiative can benefit less-developed and developing countries. For example, South Korea's Ministry of Information & Communication (MIC) established the "IT839" Strategy—"8" services; "3" infrastructures; "9" growth engines—as a roadmap for Korea's future IT development plan (MIC, 2005). The authors hope that this discussion will be beneficial for the mobile wireless industry and the academia for insight and understanding of the trends of the ubiquitous computing era.

NOTE

* A portion of this chapter is based on an earlier work: Shim, Ahn, & Shim. (2006). Empirical Findings on Perceived Use of Digital Multimedia Broadcasting Mobile Phone Services. *Industrial Management & Data Systems, 106* (2).

REFERENCES

The Australian. (2003, March 4th).

Budde, P. (2002). Asia and Australia telecommunications industry overview. *Annual Review of Communications, 55,* 243-250.

Daft, R., & Lengel, R. (1986). Organizational formation requirements, media richness and structural design. *Management Science, 32*(5), 555-571.

Eom, M. (2006). T-DMB overview in Korea. In *Proceedings of 2006 Wireless Telecommunications Symposium*, Pomona, CA.

Feagin, J. R., Orum, A. M., & Sjoberg, G. (1991). *A case for the case study.* Chapel Hill, NC: The University of North Carolina Press.

Gharavi, H., Love, P., & Cheng, E. (2004). Information and communication technology in the

stockbroking industry: An evolutionary approach to the diffusion of industry. *Industrial Management & Data Systems, 104*(9), 756-765.

Gilbert, D., Lee-Kelley, L., & Barton, M. (2003). Technophobia, gender influence and consumer decision-making for technology-related products. *European Journal of Innovation Management, 6*(4), 253-263.

Hair, J. F. (1998). *Multivariate data analysis.* Prentice Hall.

Kim, J. (2004). Terrestrial DMB's effects on the electronics industry. In *Proceedings of 2004 Terrestrial DMB International Forum* (pp. 131-142).

KORA Research. (2004, May). *A market policy study on DMB* (Rep. No. 2003-10).

Korea Radio Station Management Agency. (2004). *A market policy study on DMB.*

Kim Tae-gyu, K. (2005, January 18). Korea's free mobile broadcasting faces snag. *The Korea Times.*

Korean Society for Journalism and Communication Studies. (2003). *A study on satellite DMB.*

Kvale, S. (1983). The qualitative research interview: A phenomenological and a hermeneutical mode of understanding. *Journal of Phenomenological Psychology, 14*(2), 171-196.

Larsen, T., & Sorebo, O. (2005). Impact of personal innovativeness on the use of the Internet among employees at Work. *Journal of Organizational and End User Computing, 17(*2), 43-63.

Lee, S. M. (2003). South Korea: From the land of morning calm to ICT hotbed. *Academy of Management Executive, 17(*2), 7-18.

Ministry of Information and Communication. (2005). *IT 839 Strategy.* Republic of Korea.

Nyberg, A. (2004). Positioning DAB in an increasingly competitive world. In *Proceedings of 2004 Terrestrial DMB International Forum* (pp. 131-142).

Olla, P., & Atkinson, C. (2004). Developing a wireless reference model for interpreting complexity in wireless projects. *Industrial Management & Data Systems, 104*(3), 262-272.

QUALCOMM Incorporated. (2005). *MediaFLO: FLO technology brief.* Retrieved from www.qualcomm.com/mediaflo

Rogers, E. M. (1983). *Diffusion of innovation* (3rd ed.). New York: The Free Press.

Shim, J. P. (2005a). Korea's lead in mobile cellular and DMB phone services. *Communications of the Association for Information Systems, 15,* 555-566.

Shim, J. P. (2005b). Why Japan and Korea are leading in the mobile business industry. *Decision Line, 36*(3), 8-12.

Shim, J. P., Ahn, K. M., & Shim, J. (2006). Empirical findings on the perceived use of digital multimedia broadcasting mobile phone service. *Industrial Management & Data Systems, 106*(2), 155-171.

Shim, J. P., Shin, Y. B., & Nottingham, L. (2002). Retailer Web site influence on customer shopping: An exploratory study on key factors of customer satisfaction. *Journal of the Association for Information Systems, 3,* 53-75.

Shim, J. P., Varshney, U., & Dekleva, S. (2006a). Wireless evolution 2006: Cellular TV, wearable computing, and RFID. *Communications of the Association for Information Systems, 18,* 497-518.

Shim, J. P., Varshney, U., Dekleva, S., & Knoerzer, G. (2006b). Mobile and wireless networks: Services, evolution & issues. *International Journal of Mobile Communications, 4*(4), 405-417.

Smagt, T. (2000). Enhancing virtual teams: Social relations v. communication technology. *Industrial Management & Data Systems, 100*(4), 148-156.

Suh, Y. (2005). *Current overview of S-DMB.* TU Media.

Teng, R. (2005, January). Digital multimedia broadcasting in Korea. *In-Stat Report No. IN-052469WHT.* Retrieved from http://www.cctv.org/InStatPaper.pdf

Thompson, C. J., Locander, W. B., & Pollio, H. R. (1989). Putting consumer experience back into consumer research: The philosophy and method of existential-phenomenology. *Journal of Consumer Research, 16,* 133-146.

Trappey, A., Trappey, C., Hou, J., & Chen, B. (2004). Mobile agent technology and application for online global logistic services. *Industrial Management & Data Systems, 104(*2), 169-183.

Tsaih, R., Chang, H., & Huang, C. (2005). The business concept of utilizing the interactive TV. *Industrial Management & Data Systems, 105*(5), 613-622.

Ventatesh, W. (2000). Age differences in technology adoption decisions: Implications for a changing work force. *Personnel Psychology, 53,* 375-403.

ENDNOTES

1. http://www.pcworld.com/news/article/0,aid,124478,00.asp (2006, January 24)
2. Wi-Fi Hotstats. *Wireless Review, 22*(8) (August, 2005)
3. www.scala.com/vignettes/digital-multimedia-broadcasting.html
4. www.chiefexecutive.net/depts/technology/197a.htm
5. http://www.economist.com/business/displaystory.cfm?story_id=5356658&no_jw_tran=1&no_na_tran=1 (2006 Jan 5)

Chapter 1.15
Mobile Portals

Ofir Turel
California State University, USA

Alexander Serenko
Lakehead University, Canada

INTRODUCTION

The diffusion of mobile services is one of important technological phenomena of the twenty-first century (Dholakia & Dholakia, 2003). According to the International Telecommunication Union,[1] the number of mobile service users had exceeded 1.5 billion individual subscribers by early 2005. This represents around one-quarter of the world's population. The introduction of .mobi, a new top-level domain,[2] is expected to further facilitate the usage of mobile services. Because of their high penetration rates, mobile services have received cross-disciplinary academic attention (e.g., Ruhi & Turel, 2005; Serenko & Bontis, 2004; Turel, Serenko & Bontis, 2007; Turel, 2006; Turel & Serenko, 2006; Turel & Yuan, 2006; Turel et al., 2006). While the body of knowledge on mobile services in general is growing (Krogstie, Lyytinen, Opdahl, Pernici, Siau, & Smolander, 2004), there seems to be a gap in our understanding of a basic, yet important service that mobile service providers offer, namely mobile portals (m-portals).

M-portals are wireless Web pages that help wireless users in their interactions with mobile content and services (based on the definition by Clarke & Flaherty, 2003). These are a worthy topic for investigation since, in many cases, they represent the main gate to the mobile Internet and to wireless value-added services (Serenko & Bontis, 2004). Particularly, users of premium wireless services typically employ m-portals to discover and navigate to wireless content such as news briefs, stock quotes, mobile games, and so forth. Given this, m-portals have a strong value proposition (i.e., a unique value-added that an entity offers stakeholders through its operations)

for both users and service providers. These value dimensions, which drive the implementation and the use of m-portals, are explored in the subsequent sections.

Despite that a number of publications solely devoted to the topic of m-portals already exist, there are very few works that not only present the concept of mobile portals, but also portray their characteristics and discuss some of the issues associated with their deployment by service providers and employment by individual users. The value proposition of mobile portals was rarely explored in depth, and some motivational factors for developing and using mobile portals still remain unclear. To fill this gap, this article explores value proposition of mobile portals from both a wireless service provider and an individual user perspective. Based on this discussion, two conceptual frameworks are suggested.

The rest of this article is structured as follows. First, the key value drivers of m-portals from a wireless service provider's viewpoint are portrayed. Second, a framework that depicts the unique attributes of mobile portals and their impact on the value users derive from these services is offered. This framework is then utilized for discussing some of the challenges mobile portal developers and service providers currently face. These obstacles need to be overcome in order for service providers and users to realize the true value of mobile portals.

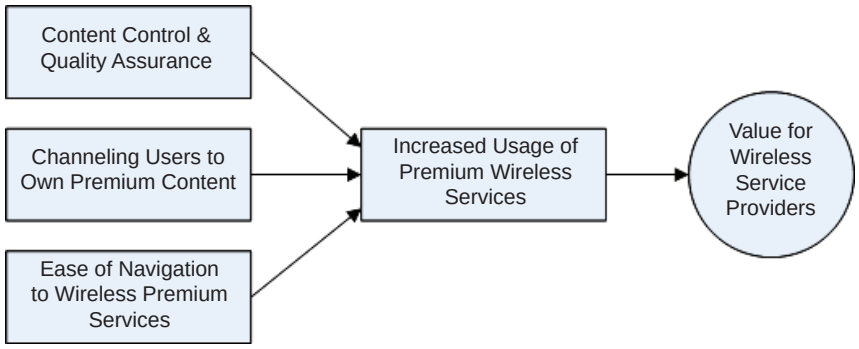
WHAT ARE MOBILE PORTALS?

As defined earlier, m-portals are wireless Web pages especially designed to ease the navigation and interaction of users with mobile content and services. They are either based on existing Internet resources adjusted to the format of mobile networks or developed from scratch for wireless networks exclusively. Occasionally, m-portals are formed by aggregating several applications together, for example, e-mail, calendars, instant messaging, and content from different information providers in order to combine as much functionality as possible. Usually, mobile portals offer basic information on news, shopping, entertainment, sports, yellow pages, and maps. M-portals can provide access to specific niche content such as health care publications information (Fontelo, Nahin, Liu, Kim, & Ackerman, 2005), public services (Philarou & Lai, 2005), travel services (Koivumäki, 2002), and so forth, or offer general access to the mobile Internet (Jonason & Eliasson, 2001).

Although the field of research pertaining to mobile portals is relatively new, a number of studies have recently investigated the concept of mobile portals from both the technical and system adoption perspectives. From the technical standpoint, scholars have investigated various aspects required for service delivery including the development of the infrastructure required for m-portal services, hypertext languages for wireless content, personalization principles, and device optimization. For example, a context-aware mobile portal was developed (Mandato, Kovacs, Hohl, & Amir-Alikhani, 2002). It automatically adapts to user needs based on explicit preferences and implicit information derived from the content viewed by individuals and is achieved through the incorporation of leading-edge technologies and principles. This allows users to receive customized portal services in real-time at no cost. The usage of mobile agents was also offered as a solution to develop a personalization mechanism that considers both user and device profiles (Samaras & Panayiotou, 2002). From the technology adoption perspective, most scholars are concerned with the acceptance of wireless portals by individuals and organizations. For instance, a conceptual model of m-portal adoption was offered (Serenko & Bontis, 2004) and the role of marketing in the promotion of wireless portals was studied (Blechar, Constantiou, & Damsgaard, 2005).

Despite the differences in research directions, all academics agree that having mobile portals

Figure 1. A conceptual framework of the value drivers of m-portals from the wireless service provider perspective



available is not sufficient to ensure the commercial success of this novel technology. As such, m-portals should present strong value proposition for both end users and service providers. The following section discuses the value proposition of mobile portals in detail.

THE VALUE PROPOSITION OF MOBILE PORTALS

M-portals offer various value propositions for both wireless service providers and users. These value dimensions are essential for driving the development, deployment, acceptance and usage of mobile portals by various stakeholders. Value perceptions are a key driver of consumer behavior in terms of services and products in general (Zeithaml, 1988), and with regards to mobile value-added services in particular (Turel & Serenko, 2006; Turel, Serenko, & Bontis, 2007). Service providers are also motivated by value when implementing and offering services (Afuah & Tucci, 2001; Porter, 1980, 1985). To better understand the value of these services for the two key stakeholders, namely, wireless service providers and users, the following two subsections outline some of the key value drivers of m-portals.

Value for Wireless Service Providers

From the wireless service provider perspective, m-portals are important since they enable providers to create a "walled garden" of services,[3] direct users to their controlled premium content, and maximize their revenues. The voice communications market has become extremely competitive in most developed countries (Paltridge, 2000). This results in price wars and a steady decline in the average voice-communications based revenue per user (ARPU) (Hatton, 2003; Swain et al., 2003). To stay competitive, wireless service providers have begun offering value-added services (VAS), such as mobile gaming, music downloads, and so forth (Barabee, 2003). Typically, these premium wireless services are facilitated through branded m-portals of the service providers. This makes it easy to access these premium services since they are readily accessible from the first screen of a portable device. In contrast, it is relatively difficult to access external Web sites (i.e., outside of the "walled garden") since it requires more tedious navigation, especially when a 10-button keypad is used for data entry.

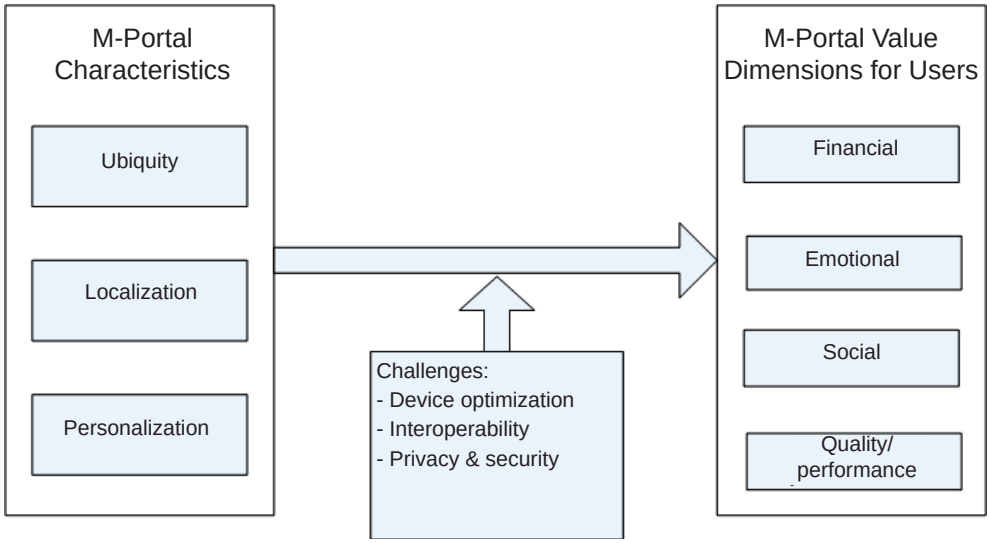
M-portals enable service providers to increase their revenues from value-added services due to three unique service characteristics. *First*, m-portals make it easier to navigate to the desired wireless content because the portal groups its

content in a meaningful way (e.g., games, news, finance, etc.). That is, users do not have to search for specific content using the QUERTY keypad. Instead, they can use hierarchical tree menus to navigate through the content by using only the OK button. For example, to reach a specific stock quote, users may choose finance, then select latest stock quotes, browse through the list of stocks and finally click on the preferred one. It should be noted that although usability is considered one of the growth drivers for wireless devices adoption (Guy, 2003), mobile services are still relatively difficult to use and fail to fit various important tasks (Buchanan, Farrant, Marsden, & Pazzani, 2001; Perry & Ballou, 1997). Thus, to help people partially overcome the usability and accessibility barriers of the wireless Internet, service providers offer m-portals.

Second, m-portals enable service providers to direct users to the premium content for which the service providers have revenue sharing. Mobile service providers may not only charge users for pure connectivity services or traffic (per minute in circuit switched second generation networks such as GSM or CDMA, or per kilobyte in packet switched networks such as GPRS or UMTS), but also profit from the actual content. For instance, people may access the premium content of a wireless service provider, such as ringtones and icons, and pay a premium fee. This fee is typically shared between the content aggregator or provider, and the wireless carrier. Therefore, the carrier may gain revenue from two sources: connectivity/traffic fees and premium content charges. The wireless carriers' share of the content revenue is flexible and may range from 9% to 80% (ARC Group, 2001; MacDonald, 2003).

Third, m-portals enable content quality control. That is, wireless service providers can ensure that the content presented on their portal is appropriate (e.g., no offensive content) and meets their service standards and portfolio of handsets. This is important since unlike the regular Internet, which is mostly free of charge, users of mobile services may pay connectivity, transmission, and premium content fees. In addition, interoperability issues may affect service quality. For example, a polyphonic ringtone that is converted for the use with a handheld device that supports only simple ringtones may cause incompatibility, lose

Figure 2. A conceptual framework of the value drivers of m-portals from the user perspective



its value, and lead to customer complaints. Therefore, service providers want to ensure the quality of their offerings. This is especially true since it was empirically shown that value-added services are perceived as the most important dimension of wireless service quality, and that they have a strong positive effect on subscribers' satisfaction (Kim, Park, & Jeong, 2004). Such a quality control approach was proven successful in the case of i-Mode in Japan (Barnes & Huff, 2003; Jonason & Eliasson, 2001; MacDonald, 2003).

Overall, wireless carriers provide m-portals for quality assurance of premium content, traffic channeling for maximizing their premium revenues, and access control. In addition, m-portals are utilized for easing the wireless Web content search experience for both novice and expert users. This is expected to increase the usage of premium wireless services that, in turn, may affect service providers' revenues. Figure 1 presents a framework of the value drivers of m-portals from the wireless service provider perspective.

Value for Users

Mobile portals allow subscribers to realize value beyond that delivered by the regular Internet or traditional commerce. Users' value perceptions are defined as an "overall assessment of the utility of a product (or service) based on perceptions of what is received and what is given" (Zeithaml, 1988, p. 14). Value perceptions are important since they determine customer satisfaction (Anderson & Fornell, 2000; Fornell, Johnson, Anderson, Cha, & Bryant, 1996; Turel & Serenko, 2004), influence brand loyalty (Yang & Peterson, 2004), and affect user acceptance of wireless value-added services (Turel et al., 2007). Particularly, it has been demonstrated that a user's assessment of the value of wireless value-added services has four dimensions: financial value (i.e., value-for-money), social value (i.e., the enhancement of the social self-concept yielded by the service), emotional value (i.e., the value derived from the affective states generated by the service), and quality/performance value (i.e., the utility derived from quality perceptions and performance expectations) (Turel et al., 2007). Based on strong empirical evidence, the value assessment of m-portals should encapsulate the abovementioned four value dimensions.

It is believed that the ubiquity, localization and personalization of mobile portals differentiate them from other Web portals. As such, these attributes are expected to be key value drivers for mobile users. Ubiquity is the ability of mobile subscribers to access information or services from anywhere at any time, and also, to be reachable at anyplace at any time (Watson, Pitt, Berthon, & Inkhan, 2002). Mobile portals are not limited to a permanent location or time zone, and therefore can support "any time" services. The notion of "any time" in the wireless services context goes beyond simple time issues because it encapsulates simultaneity (Jaureguiberry, 2000). While the wired Internet offers a limited capacity to perform simultaneous tasks (e.g., searching the Internet for a stock quote while walking), mobile portals can facilitate full simultaneity and support the broader "any time" concept. Given the increased ease of use provided by mobile portals through the presentation of efficient hierarchical tree menus, it is also expected that relevant information can be sent or received in a timely manner.

Localization is the presentation of relevant, timely location-specific information. Wireless networks are capable of determining the location of users (Karagiozidis, Markoulidakis, Velentzas, & Kauranne, 2003) and provide location-relevant services based on this information (Barnes, 2003). Services that utilize callers' location information may include emergency caller location, asset tracking, navigation, location-sensitive wireless promotions, and so forth. Mobile portals can add location-based values to the overall service experience by tailoring service menus to a current user's location. For example, airport-relevant hyperlinks (e.g., arrivals and departures, check in, transporta-

tion from the airport, etc.) may appear on the front page of the portal when the system identifies that the user is located near an airport.

Personalization is the utilization of personal profiles, needs and preferences for providing user-specific information or services over the wireless network. The need for personalization of mobile services is driven by various contextual dispositions; it can lead to cognitive, social and emotional effects (Blom & Monk, 2003). In the context of mobile portals, personalization is relatively easy to implement since most wireless devices are carried and used by a single person. The input for personalizing m-portal services can come from various sources. First, users can build a static profile. For this, they can enter their general preferences through a call center, a registration Web site, or a wireless device. These preferences may include the look and feel of the service and a general interest profile. This list of interests can be translated into the structure of the menu so that top menu items match the user's interests. Second, the service provider can produce a dynamic profile, based on past user behavior, location data and other contextual inputs. For example, a stock quote that has been frequently viewed by a user can appear on the first page of the portal. Other contextual dimensions, such as time and location, can be added to the user profile. That is, the m-portal may provide a personalized menu only in certain times or locations. For instance, a menu for the retrieval of sports news can be provided only on weekday mornings when a person commutes. Note that this personalized menu approach may substantially improve the ease of use of mobile services because navigating to the desired wireless content by using a handheld device may be much more tedious than similar navigations by using a PC.

It should be noted that it is not easy for wireless service providers to deliver this value proposition to mobile subscribers. While the telecommunication infrastructure is mostly in place, various issues, such as device optimization, interoperability, privacy and security, still need to be overcome before users and service providers are able to fully realize the value proposition of m-portals. Device optimization refers to tailoring the same wireless content to multiple handhelds in an optimal manner. Due to a variety of handheld devices, service providers need to find a way to ensure usability across them. For example, one screen may contain up to 10 lines of content and another up to four lines only. In this case, the service provider needs to decide if a 10-line content item (e.g., news brief) should be summarized or presented with a scroll bar. Interoperability refers to the exchange of content from different networks and devices. For instance, service providers need to ensure that a CHTML[4] Web site can be accessed from a GSM handset that supports WAP only. Privacy and security refer to the protection of user personal information and ensuring individuals have full control over their static and dynamic personal usage profiles. This is especially important in the wireless context since service providers have sensitive information such as user location. To summarize these value drivers and potential barriers, Figure 2 depicts the value dimensions of m-portals from a user perspective, taking into account the issues that service providers need to consider.

SUMMARY

The purpose of this article was to introduce the concept of mobile portals and discuss several current issues associated with the employment of m-portals by individuals. For this, two conceptual frameworks were constructed. The first one refers to the value drivers of m-portals from the service provider perspective. Three drivers that increase service usage and improve profitability are suggested: (1) content control and quality assurance; (2) channeling users to their own premium content; and (3) ease of navigation to wireless premium services. The second framework relates to

the value drivers from the end-user perspective. It is argued that mobile portal characteristics, such as ubiquity, localization and personalization, represent value for individuals. The m-portal value is described by financial, emotional, social and quality/performance dimensions. The relationship between m-portal characteristics and user value is moderated by several challenges such as device optimization, interoperability, and privacy/security.

Mobile portals are a novel technology that has become very popular among mobile device users. In order to deliver high-quality m-portal services and to meet customer expectations, providers should pay attention to the academic works emerging in this area. It is hoped that this article may potentially contribute in our understanding of this important phenomenon.

REFERENCES

Afuah, A., & Tucci, C. L. (2001). *Internet business models and strategies. Text and cases*. New York: McGraw-Hill.

Anderson, E. W., & Fornell, C. (2000). Foundations of the American Customer Satisfaction Index. *Total Quality Management & Business Excellence, 11*(7), 869-882.

ARC Group. (2001). *Content and applications.* London: Author.

Barabee, L. (2003). *Carriers make a play in wireless entertainment.* Boston: The Yankee Group.

Barnes, S. J. (2003). Developments in the m-commerce value chain: Adding value with location-based services. *Geography, 88*, 277-288.

Barnes, S. J., & Huff, S. L. (2003). Rising sun: iMode wireless Internet. *Communications of the ACM, 46*(11), 76.

Blechar, J., Constantiou, I., & Damsgaard, J. (2005). *The role of marketing in the adoption of new mobile services: Is it worth the investment?* Paper presented at the International Conference on Mobile Business, Sydney, Australia.

Blom, J. O., & Monk, A. F. (2003). Theory of personalization of appearance: Why users personalize their PCs and mobile phones. *Human-Computer Interaction, 18*(3), 193-228.

Buchanan, G., Farrant, S., Marsden, G., & Pazzani, M. (2001, May). *Improving mobile Internet usability.* Paper presented at the WWW10, Hong Kong, China.

Clarke, I., III, & Flaherty, T. B. (2003). Mobile portals: The development of m-commerce gateways. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce: Technology, theory and applications* (pp. 185-210). Hershey, PA: Idea Group Publishing.

Dholakia, R. R., & Dholakia, N. (2003). Mobility and markets: Emerging outlines of m-commerce. *Journal of Business Research, 57*(12), 1391-1396.

Fontelo, P., Nahin, A., Liu, F., Kim, G., & Ackerman, M. (2005). *Accessing MEDLINE/PubMed with handheld devices: Developments and new search portals.* Paper presented at the 38th Hawaii International Conference on System Sciences, Hawaii.

Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American Customer Satisfaction Index: Nature, purpose, and findings. *Journal of Marketing, 60*(7), 7-18.

Guy, A. (2003). *Industry players stress standards and usability as growth drivers at Yankee Group Mobile Messaging Forum* (Yankee Group Research Note in Wireless Mobile Services). Boston: The Yankee Group.

Hatton, M. (2003). *Pricing becomes the keystone of mobile operators' consumer strategy.* Boston: The Yankee Group.

Jaureguiberry, F. (2000). Mobile telecommunications and the management of time. *Social Science Information (Sur Les Sciences Sociales), 39*(2), 255-268.

Jonason, A., & Eliasson, G. (2001). Mobile Internet revenues: An empirical study of the I-Mode portal. *Internet Research: Electronic Networking Applications and Policy, 11*(4), 341-348.

Karagiozidis, M., Markoulidakis, Y., Velentzas, S., & Kauranne, T. (2003). Commercial use of mobile, personalised location-based services. *Journal of the Communications Network, 2* (3), 15-20.

Kim, M. K., Park, M. C., & Jeong, D. H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunication Policy, 28*(2), 145-159.

Koivumäki, T. (2002). Consumer attitudes and mobile travel portal. *Electronic Markets, 12*(1), 47-57.

Krogstie, J., Lyytinen, K., Opdahl, A. L., Pernici, B., Siau, K., & Smolander, K. (2004). Research areas and challenges for mobile information systems. *International Journal of Mobile Communications, 2*(3), 220-234.

MacDonald, D. J. (2003). NTT DoCoMO's i-Mode: Developing win-win relationships for mobile commerce. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce: Technology, theory and applications* (pp. 1-25). Hershey, PA: Idea Group Publishing.

Mandato, D., Kovacs, E., Hohl, F., & Amir-Alikhani, H. (2002). CAMP: A context-aware mobile portal. *IEEE Communications Magazine, 40*(1), 90-97.

Paltridge, S. (2000). Current statistics; Mobile communications update. *Telecommunication Policy, 24*(5), 453-456.

Perry, E. L., & Ballou, D. J. (1997). The role of work, play, and fun in microcomputer software training. *Data Base for Advances in Information Systems, 28*(2), 93-112.

Philarou, R., & Lai, F. L. (2005). *Behind e-governments of less advantageous nations: A report of the meanings of e-government portals to lower and medium DAI nations.* Paper presented at the Hong Kong Mobility Roundtable, Hong Kong, China.

Porter, M. E. (1980). *Competitive strategy: Techniques for analyzing industries and competitors.* New York: Free Press.

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance*. New York: The Free Press.

Ruhi, U., & Turel, O. (2005). Driving visibility, velocity and versatility: The role of mobile technologies in supply chain management. *Journal of Internet Commerce, 4*(3), 97-119.

Samaras, G., & Panayiotou, C. (2002). *Personalized portals for the wireless user based on mobile agents.* Paper presented at the 2nd International Workshop on Mobile Commerce, Atlanta, Georgia.

Serenko, A., & Bontis, N. (2004). A model of user adoption of mobile portals. *Quarterly Journal of Electronic Commerce, 4*(1), 69-98.

Swain, W., Entner, R., Guy, A., Barrabee, L., Yunus, F., Hatton, M., et al. (2003). *Data ARPU saves the day for wireless operators.* Boston: The Yankee Group.

Turel, O. (2006). Contextual effects on the usability dimensions of mobile value added services: A conceptual framework. *The International Journal of Mobile Communications, 4*(3), 309-332.

Turel, O., & Serenko, A. (2004, July 12-13). *User satisfaction with mobile services in Canada.* Paper presented at the Third International Conference

on Mobile Business, M-Business 2004, New York City, New York.

Turel, O., & Serenko, A. (2006). Satisfaction with mobile services in Canada: An empirical investigation. *Telecommunication Policy, 30*(5-6), 314-331.

Turel, O., Serenko, A., & Bontis, N. (2007). User acceptance of wireless short messaging services: Deconstructing perceived value. *Information & Management, 44*(1), 63-73.

Turel, O., Serenko, A., Detlor, B., Collan, M., Nam, I., & Puhakainen, J. (2006). Investigating the determinants of satisfaction and usage of Mobile IT services in four countries. Journal of *Global Information Technology Management, 9*(4), 6-27.

Turel, O., & Yuan, Y. (2006). Investigating the dynamics of the m-commerce value system: A comparative viewpoint. *International Journal of Mobile Communications, 4*(5), 532-557.

Watson, R. T., Pitt, L. F., Berthon, P. Z., & Inkhan, G. M. (2002). U-commerce: Extending the universe of marketing. *Journal of the Academy of Marketing Science, 30*(4), 329-343.

Yang, Z., & Peterson, R. T. (2004). Customer perceived value, satisfaction, and loyalty: The role of switching costs. *Psychology & Marketing, 21*(10), 799-822.

Zeithaml, V. A. (1988). Consumer perceptions of price, quality and value: A means-end model and synthesis of evidence. *Journal of Marketing, 52*(3), 2-22.

KEY TERMS

Compact Hyper-Text Markup Language (CHTML): A subset of HTML for small portable devices. (http://www.Webopedia.com/TERM/C/cHTML.html)

General Packet Radio Service (GPRS): A standard for wireless communications which runs at speeds up to 115 kilobits per second, compared with current GSM's (Global System for Mobile Communications) 9.6 kilobits. GPRS, which supports a wide range of bandwidths, is an efficient use of limited bandwidth and is particularly suited for sending and receiving small bursts of data, such as e-mail and Web browsing, as well as large volumes of data. (http://www.Webopedia.com/TERM/G/GPRS.html)

Global System for Mobile Communications (GSM): One of the leading digital cellular systems. GSM uses narrowband TDMA, which allows eight simultaneous calls on the same radio frequency. GSM was first introduced in 1991. (http://www.Webopedia.com/TERM/G/GSM.html)

Mobile Portals (M-Portals): Wireless Web pages especially designed to assist wireless users in their interactions with wireless content and services (based on the definition by Clarke & Flaherty, 2003).

"Walled Garden": Refers to the content that wireless device users are able to see. The availability and selection of this content is limited by a service provider. (http://www.Webopedia.com/TERM/G/GSM.html)

Wireless Application Protocol (WAP): A secure specification that allows users to access information instantly via handheld wireless devices such as mobile phones, pagers, two-way radios, smart-phones and communicators. (http://www.Webopedia.com/TERM/W/WAP.html)

Universal Mobile Telecommunications System (UMTS): A 3G mobile technology that will deliver broadband information at speeds up to 2 Mbit/sec. Besides voice and data, UMTS will deliver audio and video to wireless devices anywhere in the world through fixed, wireless

and satellite systems. (http://www.Webopedia.com/TERM/U/UMTS.html)

Value Proposition: The primary benefit of a product or service. (http://www.pcmag.com/encyclopedia_term/0,2542,t=value+proposition&i=53664,00.asp)

ENDNOTES

1. http://www.itu.int
2. For more information, refer to the Domain Name Web site at http://www.domainbank.net/mobi/index.cfm
3. The term "walled garden" refers to the content that wireless device users are able to see. The availability and selection of this content is limited by a service provider. More information is available on the Webopedia Web site at http://www.Webopedia.com/TERM/W/walled_garden.html.

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 587-593, copyright 2007 by Information Science Publishing (an imprint of IGI Global).

Chapter 1.16
Mobile Portals as Innovations

Alexander Serenko
Lakehead University, Canada

Ofir Turel
California State University, Fullerton, USA

INTRODUCTION

The purpose of this chapter is to analyze mobile portals (m-portals) as an innovation. M-portals are wireless Web pages that help portable device users interact with mobile content and services (based on the definition by Clarke & Flaherty, 2003). Previous works in the area of mobile portals mostly concentrated on their technical aspects, implementation issues, classifications, and user acceptance (e.g., Gohring, 1999; GSA, 2002; Koivumäki, 2002). At the same time, these studies did not view mobile portals as innovations themselves, nor discussed the innovative potential of this novel technology. Analyzing technological artifacts as innovations is important for two reasons. First, such analysis can help m-portal developers and providers pinpoint the salient m-portal characteristics that drive service diffusion. Second, it can assist potential m-portal developers and providers understand the risks associated with entering this segment of wireless services.

This study attempts to contribute to the knowledge base by discussing various dimensions of the innovativeness of mobile portals and predicting the commercial success as well as potential risks of designing m-portals. Specifically, this investigation utilizes two innovation-based models as a lens of analysis. The first is the Moore and Benbasat's (1991) list of perceived characteristics of innovating (PCI), which is adapted to assess the innovation features of mobile portals. The second is the Kleinschmidt and Cooper's (1991) market and technological newness map. By applying these frameworks, the study attempts to develop a better understanding of individual innovation characteristics and the innovation typology of mobile portals that is important for both theory and practice.

Mobile portals are a fruitful area of growth and interest. Even though the technology has been in use for only several years, both researchers and practitioners have devoted substantial efforts to design m-portals that would meet end-user requirements. To ensure the success of this technology, it is important to further understand its innovative potential. However, little work has been done in this area. A discussion grounded on

the existing innovation schools of thought would help to bridge that gap.

M-PORTALS AS INNOVATIONS

There are several works that have already discussed the importance of mobile data innovations. This line of research was inspired by the continuous breakthroughs in the mobile telecom sector (Berkhout & van der Duin, 2004). Several factors facilitate constant innovation in the telecommunications industry. *Bandwidth* is the first one. For the past years, the bandwidth of both wired and wireless networks has been continuously increasing by mostly following the Gilder's Law. It states that bandwidth grows three times as fast as the CPU speed. This trend facilitates the development of various innovative technologies, including wireless Internet access and mobile portals. *Industry structure* is the second factor inspiring innovation. Currently, the North American and European industries are, to some extent, de-regulated, restructured, and consist of numerous independent service providers (Turel & Serenko, 2006). There are certain advantages of this industry structure. It increases competition among individual players that have to constantly innovate to stay competitive. At the same time, there are innovations created by partnerships with organizations in the same or different sectors. In the case of mobile portals, this is transparent in alliances between infrastructure, technology, media, and content providers who combine their efforts to deliver a single innovative product on the market (Turel & Yuan, 2006). There are various new business models that may be implemented with the employment of mobile portals. For example, revenues from services accessed through a mobile portal are usually shared between a wireless carrier and service provider (ARC Group, 2001; MacDonald, 2003). *Agent-based technologies* are the third factor fostering innovations in the mobile services industry (Al agha & Labiod, 1999; Kotz et al., 2002). Especially, agent-based computing is an important tool to enhance the functionality of mobile portals and enable new business models (Chen, Joshi, & Finin, 2001; Panayiotou & Samaras, 2004). An agent is a software entity that is autonomous, continuous, reactive, collaborative; it constantly works in the background of a computer system, such as a mobile application, analyzes all user actions, develops user profiles, communicates with other agents or systems, and acts on behalf of the user by making recommendations (Detlor, 2004; Serenko, 2006). Agent technologies are considered an important innovation that may contribute substantially in the development of new computer technologies, business models or human-computer interaction approaches (Serenko & Detlor, 2004; Serenko, Ruhi, & Cocosila, 2007). For example, an agent that learns a user's profile over time may design personalizable mobile portals tailored to the needs of each particular individual; as user behavior changes, the agent adjusts the content of a portal.

In order to better understand the innovating characteristics of mobile portals, Moore et al.'s (1991) list of perceived characteristics of innovating is employed. Their approach originates from diffusion of innovations theory introduced by Rogers (1983) and Rogers and Shoemaker (1971), and concentrates on technology innovation adoption research (Plouffe, Hulland, & Vandenbosch, 2001). A list of perceived characteristics of innovating applied to mobile portals is presented next:

- *Relative advantage* is the degree to which an innovation is superior to the ideas, practices, or objects it supersedes. In terms of mobile portals, a relative advantage of using this technology is evident in ubiquity, localization, and personalization. Ubiquity allows users to access mobile portals from anywhere at anytime given that a wireless connection is established. Localization is

the generation of a portal targeted to the current location of a mobile device user, and personalization is the employment of user profiles to deliver portals tailored to the needs of each person individually (Clarke et al., 2003; Serenko & Bontis, 2004; Watson, Pitt, Berthon, & Inkhan, 2002). As such, this is a vital feature of m-portals.

- *Compatibility* is the degree to which an innovation is consistent with the existent values, previous experiences, and current needs of adopters. In the case of mobile portals, compatibility has two key dimensions: technical compatibility and needs compatibility. First, the m-portal technology should be compatible with various mobile devices, such as wireless PDAs or cell phones. At the same time, most existing WWW portals cannot be directly displayed on mobile devices. The concept of m-portals is not entirely new; it is assumed that the majority of mobile device users are familiar with WWW portals. Thus, m-portals are partially compatible with mobile devices. Second, m-portals should be compatible with life-styles and needs of many individuals in countries in which wireless phones have highly penetrated (e.g., Italy, Singapore, etc.). Users in these countries are accustomed to wireless applications, and learned to appreciate the ubiquity offered by wireless content and services (Turel, 2006).
- *Ease of use* is the degree to which an innovation is perceived as being relatively difficult to understand and use. There are two aspects of m-portal technologies relating to this characteristic. On the one hand, mobile portals are more difficult to navigate by using a mobile device than a regular WWW portal. On the other, m-portals improve the ease of use of the mobile Internet by organizing important content and making it easier to access.
- *Results demonstrability* is the degree to which the benefits and utilities of an innovation are readily apparent to the potential adopter. M-portals save time and money (airtime fees) by easing and accelerating the navigation to the desired mobile application or content. As such, m-portal users may quickly observe the benefits by locating information and services more effectively, economically, and efficiently.
- *Image* is the degree to which innovation usage is perceived to enhance adopters' image, prestige, or status in their social system. With respect to m-portals, this is not a major benefit of the technology. In developed countries, mobile device users are not currently perceived as highly innovative individuals by the other members of their social group. Recently, Turel, Serenko, and Bontis (2007) conducted an empirical study of short messaging services (SMS) adoption in Canada and concluded that social value of SMS, which was defined as the enhancement of one's social self-concept provided by the usage of SMS, does not have an impact of SMS usage intentions given that SMS is not perceived as a highly innovative technology. It is suggested that the same holds true in the case of mobile portals, and image is not the key reason for m-portal employment.
- *Visibility* is the degree to which the results of an innovation are visible to others. Given the low image enhancement associated with m-portals (see the previous paragraph), m-portal users are not likely to brag about the use of this service. Thus, the outcomes of the employment of this technology will be hardly visible to other wireless WWW users, colleagues, or friends. Indeed, it is up to m-portal users to communicate the visibility of portal usage to the others.
- *Trialability* is the degree to which a potential adopter believes that an innovation may be experimented with on a limited basis before

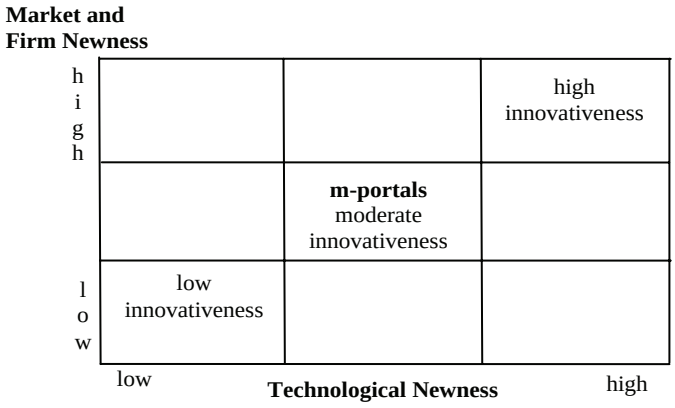
an adoption decision needs to be made. Currently, there are both free and fee-based mobile portals. In the case of free portals, there is a limited financial risk associated with the service because users may try it out, pay a marginal airtime fee, and discontinue without consequences of any kind. At the same time, some users may not feel comfortable signing up for the usage of commercial mobile portals before having some exposure to the actual m-portal services. The latter type of portals presents a higher financial risk.

- *Voluntariness* is the degree to which innovation use is perceived as being voluntary, or of free will. In terms of m-portals, the individual-level usage is voluntary; it is a person's decision whether to access a portal. At the same time, the organizational-level use may be both voluntary—when the access of an organizational wireless portal is optional, and mandatory—when employees must access specific m-portals for their work.

Overall, these characteristics of m-portals, as perceived by both end users and other members of a social system, affect the rate of m-portal adoption. It is believed that the higher the levels of these innovative attributes, the faster mobile portals are accepted. Based on the previous discussion, researchers and practitioners may potentially facilitate fast adoption of m-portals. However, this approach does not allow them to accurately predict the commercial success and potential risks associated with the development of mobile portals by the wireless industry players. For this, the categorization schema developed by Kleinschmidt et al. (1991) is applied. Figure 1 presents Kleinschmidt et al.'s market and technological newness map.

According to this typology, there are three categories of innovativeness: low, moderate, and high that are positioned along two axes of technological and market/manufacturer newness. Highly innovative products and services are comprised of new to the customers, markets, and manufactures products and services. Moderately innovative offerings consist of less innovative

Figure 1. Kleinschmidt et al.'s (1991) market and technological newness map applied to mobile portals



products and services that are not already new to both businesses and consumers. Low innovative items represent modifications, revisions, and improvements of existing offerings. The major advantage of using this model is that it allows approximating the amount of uncertainty and risk involved in the commercialization of an innovation. Kleinschmidt et al. (1991) argue that moderately innovative items are less likely to succeed and are accompanied by a greater risk than low and high innovative ones. With respect to mobile portals, it is hypothesized that they represent a moderately innovative offering. First, most of the technologies to deliver m-portals have been developed earlier, and they were only adjusted to support mobile portal deployment. Second, from the mobile device user perspective, the concept of portals has been well known; the novelty is the delivery of portals over a hand-held device. Location and personalization services are relatively newer; overall, this reflects a moderate degree of innovativeness. This demonstrates that mobile portal providers face the highest extent of risk as suggested by the model.

CONCLUSION AND IMPLICATIONS

The utilization of the PCI and Newness Map frameworks to analyze mobile portals has some managerial and research implications. First, information systems researchers may employ the concepts proposed in the PCI framework, as applied to m-portals, to identify the antecedents of user intention to adopt this innovation. A model explicating the relationships between these factors and user behavior with m-portals may be proposed and tested. The finding of such analyses can advance the technology adoption research stream and offer some insights for m-portal service developers and providers as well as for wireless carriers.

Second, strategy and marketing researchers may use the Newness Map applied to m-portals to investigate the market dynamics of the m-portals sector. Such analyses may lead to better business models, and a well-thought-of risk taking approach employed by industry participants.

The previous conceptualization has several limitations that may be addressed in future research. First, driving factors in adopting information technology innovations change over time (Waarts, van Everdingen, & van Hillegersberg, 2002) that will dramatically affect the predicted diffusion of mobile portals in future. As mobile technologies advance, the importance of perceived characteristics of innovating will change and new factors will emerge. Second, there are other alternative innovation theories that may also enhance our understanding of the field (Abernathy & Clark, 1985; Chandy & Tellis, 2000; Utterback, 1994). For example, Garcia and Calantone (2002) report that in innovation research there are at least 15 constructs and 51 distinct scale items that have been applied in 21 empirical investigations. Despite these limitations, it is believed that this chapter sheds some light on an important area, suggests implications for managers, and inspires academics to conduct further research.

REFERENCES

Abernathy, W. J., & Clark, K. B. (1985). Innovation: Mapping the winds of creative destruction. *Research Policy, 14*(1), 3-22.

Al agha, K., & Labiod, H. (1999). MA-WATM: A new approach towards an adaptive wireless ATM network. *Mobile Networks and Applications, 4*(2), 101-109.

ARC Group. (2001). *Content and Applications.* London, UK: ARC Group.

Berkhout, G., & van der Duin, P. (2004, March 2004). *Mobile data innovation: Lucio and the cyclic innovation model.* Paper presented In Proceedings of the 6th International Conference

on Electronic Commerce, Cape Town, South Africa.

Chandy, R. K., & Tellis, G. J. (2000). The incumbent's curse? Incumbency, size, and radical product innovation. *Journal of Marketing, 64*(3), 1-17.

Chen, H., Joshi, A., & Finin, T. (2001). Dynamic service discovery for mobile computing: Intelligent agents meet Jini in the Aether. *Cluster Computing, 4*(4), 343-354.

Clarke III, I., & Flaherty, T. B. (2003). Mobile portals: The development of m-commerce gateways. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce: Technology, theory, and applications* (pp. 185-210). Hershey, PA: Idea Group Publishing.

Detlor, B. (2004). *Towards knowledge portals: From human issues to intelligent agents.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garcia, R., & Calantone, R. (2002). A critical look at technological innovation typology and innovativeness terminology: A literature review. *Journal of Product Innovation Management, 19*(2), 110-132.

Gohring, N. (1999). Mobile portals on the rise. *Telephony, 237*(1), 26.

GSA. (2002). *Survey of mobile portal services* (Vol. 8). Sawbridgeworth, UK: Global Mobile Suppliers Association.

Kleinschmidt, E. J., & Cooper, R. G. (1991). The impact of product innovativeness on performance. *Journal of Product Innovation Management, 8*(4), 240-251.

Koivumäki, T. (2002). Consumer attitudes and mobile travel portal. *Electronic Markets, 12*(1), 47-57.

Kotz, D., Cybenko, G., Gray, R. S., Jiang, G., Peterson, R. A., Hofmann, M. O., et al. (2002). Performance analysis of mobile agents for filtering data streams on wireless networks. *Mobile Networks and Applications, 7*(2), 163-174.

MacDonald, D. J. (2003). NTT DoCoMO's i-Mode: Developing win-win relationships for mobile commerce. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce: Technology, theory, and applications.* Hershey, PA: IRM Press.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

Panayiotou, C., & Samaras, G. (2004). mPERSONA: Personalized portals for the wireless user: An agent approach. *Mobile Networks and Applications, 9*(5), 663-677.

Plouffe, C. R., Hulland, J. S., & Vandenbosch, M. (2001). Research report: Richness versus parsimony in modeling technology adoption decisions—Understanding merchant adoption of a smartcard-based payment system. *Information Systems Research, 12*(2), 208-222.

Rogers, E. M. (1983). *Diffusion of innovations* (3rd ed.). New York: Free Press.

Rogers, E. M., & Shoemaker, F. F. (1971). *Communication of innovations* (2nd ed.). New York: Free Press.

Serenko, A. (2006). The importance of interface agent characteristics from the end-user perspective. *International Journal of Intelligent Information Technologies, 2*(2), 48-59.

Serenko, A., & Bontis, N. (2004). A model of user adoption of mobile portals. *Special Issue of the Quarterly Journal of Electronic Commerce, 4*(1), 69-98.

Serenko, A., & Detlor, B. (2004). Intelligent agents as innovations. *AI & Society, 18*(4), 364-381.

Serenko, A., Ruhi, U., & Cocosila, M. (2007). Unplanned effects of intelligent agents on Internet

use: Social informatics approach. *AI & Society, 21*(1-2), 141-166.

Turel, O. (2006). Contextual effects on the usability dimensions of mobile Value Added Services: A conceptual framework. *The International Journal of Mobile Communications, 4*(3), 309-332.

Turel, O., & Yuan, Y. (2006). Investigating the dynamics of the M-Commerce value system: A comparative viewpoint. *International Journal of Mobile Communications, 4*(5), pp. 532-557.

Turel, O., & Serenko, A. (2006). Satisfaction with mobile services in Canada: An empirical investigation. *Telecommunications Policy, 30*(5-6), 314-331.

Turel, O., Serenko, A., & Bontis, N. (2007). User acceptance of wireless short messaging services: Deconstructing perceived value. *Information & Management, 44*(1), 63-73.

Utterback, J. M. (1994). *Mastering the dynamics of innovation: How companies can seize opportunities in the face of technological change.* Boston: Harvard Business School Press.

Waarts, E., van Everdingen, Y. M., & van Hillegersberg, J. (2002). The dynamics of factors affecting the adoption of innovations. *Journal of Product Innovation Management, 19*(6), 412-423.

Watson, R. T., Pitt, L. F., Berthon, P. Z., & Inkhan, G. M. (2002). U-commerce: Extending the universe of marketing. *Journal of the Academy of Marketing Science, 30*(4), 329-343.

KEY TERMS

Kleinschmidt and Cooper's (1991) Market and Technological Newness Map: A categorization schema that defines three categories of innovativeness: low, moderate, and high, positioned along two axes of technological and market/manufacturer newness. The major advantage of using this model is that it allows approximating the amount of uncertainty and risk involved in the commercialization of an innovation.

Mobile Portals (M-Portals): Wireless Web pages especially designed to assist wireless users in their interactions with wireless content and services (based on the definition by Clarke et al., 2003).

Moore and Benbasat's (1991) List of Perceived Characteristics of Innovating (PCI): A list of important characteristics of an innovation that affect its diffusion rate. The factors include relative advantage, compatibility, ease of use, results demonstrability, image, visibility, trialability, and voluntariness.

Short Messaging Services (SMS): Short messaging service (SMS), also known as text messaging, is one of the most frequently utilized mobile services. SMS enables sending and receiving text messages of up to 160 characters to and from mobile devices. The text is entered by using a phone keypad or a PC keyboard, and it may consist of words, numbers, or alphanumeric combinations. SMS was created as part of the GSM Phase 1 standard. It uses the network-signalling channel for data transmitting and receiving.

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 594-598,

Chapter 1.17
Mobile Portals for Knowledge Management

Hans Lehmann
Victoria University of Wellington, New Zealand

Ulrich Remus
University of Erlangen-Nuremberg, Germany

Stefan Berger
Detecon International GmbH, Germany

INTRODUCTION

More and more people leave their fixed working environment in order to perform their knowledge-intensive tasks at changing locations or while they are on the move. Mobile knowledge workers are often separated from their colleagues, and they have no access to up-to-date knowledge they would have in their offices. Instead, they rely on faxes and messenger services to receive materials from their home bases (Schulte, 1999). In case of time-critical data, this way of communication with their home office is insufficient.

Mobile knowledge management (KM) has been introduced to overcome some of the problems knowledge workers are faced when handling knowledge in a mobile work environment (e.g., Berger, 2004; Grimm, Tazari, & Balfanz, 2002,). The main goal of mKM is to provide mobile access to knowledge management systems (KMS) and other information resources, to generate awareness between mobile and stationary workers by linking them to each other, and to realize mobile KM services that support knowledge workers in dealing with their tasks (see chapter, "A Mobile Portal for Academe: The Example of a German University" in the same book).

So far, most of the off-the-shelf KMS are intended for the use on stationary desktop PCs or laptops with stable network access, and provide just simple access from mobile devices. As KMS are generally handling a huge amount of information (e.g., documents in various formats, multimedia content, etc.) the limitations of (mobile) information and communication technologies (ICTs), like mobile devices such as PDAs and

mobile phones, becomes even more crucial (Hansmann, Merk, Niklous, & Stober, 2001). Mobile devices are usually not equipped with the amount of memory and computational power found in desktop computers; they often provide small displays and limited input capabilities, in comparison to wired networks, wireless networks generally have a lower bandwidth restricting the transfer of large data volumes and due to fading, lost radio coverage, or deficient capacity, wireless networks are often inaccessible for periods of time.

Today, many KMS are implemented as knowledge portals, providing a single point of access to many different information and knowledge sources on the desktop together with a bundle of KM services. In order to realize mobile access to knowledge portals, portal components have to be implemented as mobile portlets. That means that they have to be adapted according to technical restrictions of mobile devices and the user's context.

This contribution identifies requirements for mobile knowledge portals. In particular, it reviews the main characteristics of mobile knowledge portals, which are considered to be the main ICT to support mobile KM. In addition, it outlines an important future issue in mobile knowledge portals: The consideration of location-based information in mobile knowledge portals.

MOBILE KNOWLEDGE PORTALS

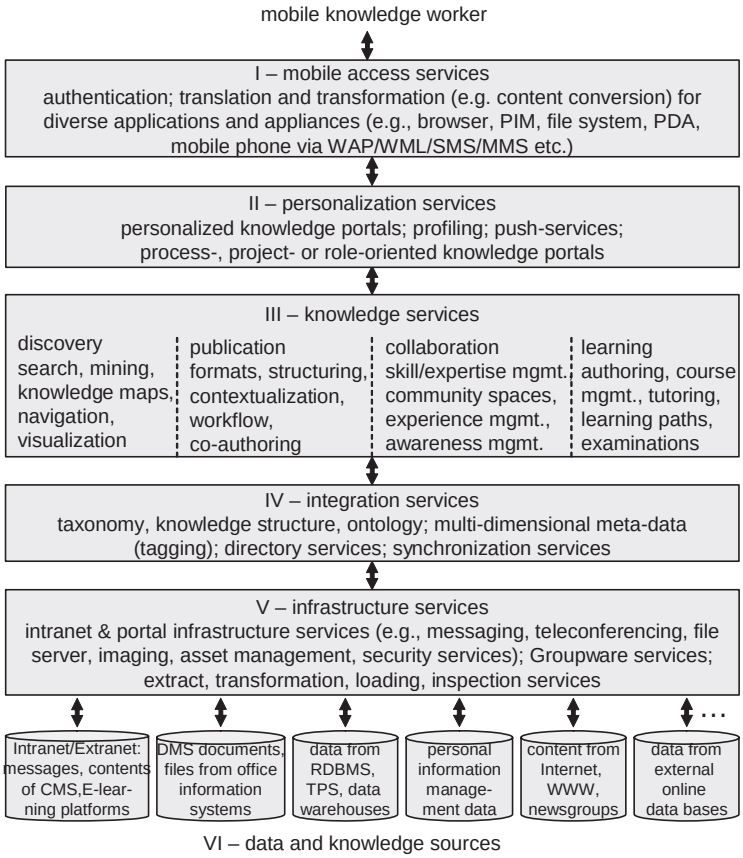
Most knowledge management systems (KMS) are implemented as centralized client/server solutions (Maier, 2004) using the portal metaphor. Such knowledge portals provide a single point of access to many different information and knowledge sources on the desktop, together with a bundle of KM services (cf. Collins, 2003; Detlor, 2004), for example, contextualization, semantic search, collaboration, visualization and so forth. The added value of these portals compared to other KM tools is the integration of technologies for storage of, and access to, information and knowledge, with the ones for support of the interaction and collaboration activities in a unique entity (Loutchko & Birnkraut, 2005). Typically, the architecture of knowledge portals can be described with the help of KMS-layers (Figure 1, Maier, 2004).

The first layer includes data and knowledge sources of organizational-internal and external sources. Examples are database systems, data warehouses, enterprise resource planning systems, content and document management systems. The next layer provides intranet and portal infrastructure services as well as groupware services, together with services to extract, transform, and load content from different sources. On the next layer, integration services are necessary to organize and structure knowledge elements according to a taxonomy or ontology.

The core of the architecture consists of a set of knowledge services in order to support discovery, publication, collaboration, and learning. Personalization services are important to provide a more effective access to the large amounts of content, that is, to filter knowledge according to the knowledge needs in a specific situation, and offer this content by a single point of entry (portal). In particular, personalization services, together with mobile access services, become crucial for the use of KMS in mobile environments.

Portals can be either developed individually or by using off-the-shelf portal packages, such as BEA WebLogic, IBM Portal Server, Plumtree Corporate Portal, Hyperwave Information Portal, or SAP Enterprise Portal. Most of these commercial packages can be flexibly customized in order to build up more domain-specific portals by integrating specific portal components (so called "portlets") into a portal platform. Portlets are more or less standardized software components that provide access to various applications and (KM) services, for example, portlets to access enterprise resource planning systems, document management sys-

Figure 1. Layer architecture of knowledge portals (Adapted from Maier, 2004)



tems, personal information management, and such like. In order to realize mobile access to knowledge portals, portlets have to be implemented as mobile portlets. That means that they have to be adapted according to technical restrictions of mobile devices and the user's context.

REQUIREMENTS FOR MOBILE KNOWLEDGE PORTALS AND PLATFORMS

Typical requirements for mobile knowledge portals and platforms can be derived from our definition of mobile KM. Note that these requirements are not restricted to a mobile environment, but cater to the special needs of a mobile work environment, for example, speech technology is a crucial service in order to overcome typical input limitations. A mobile knowledge portal should provide specific services (cf. Berger, 2004):

- to support the *social networking* of knowledge worker and to *create awareness* (mobile access to employee yellow pages, skill directories, directories of communities, knowledge about business partners focus-

Table 1. Selected portal packages

	social networking, create awareness	*mobile access*	*location-orientation*	*proactive information delivery*	*Hetero-genous technolo-gies*	*adaptive information delivery*	*speech technology*
Autonomy Portal-in-a-Box	■	■		■		■	
Livelink Portal / Wireless	■	■		■	■	■	
Hyperwave Information Portal		■					
Hummingbird Enterprise Portal		■		■	■	■	
Plumtree Portal /Wireless	■	■		■	■	■	
IBM Websphere Every Place Access / Voice		■		■		■	■

ing on asynchronous (e-mail, short message service) and synchronous communication (chat), collaboration, cooperation, and community support);

- to enable *mobile access* on various knowledge sources via different devices (e.g., knowledge about organization, processes, products, internal studies, patents, online journals, ideas, proposals, lessons learned, best practices, community home spaces (mobile virtual team spaces), evaluations, comments, feedback to knowledge elements) focusing on services for presentation (e.g., summarization functions, navigation models);
- to support *location-oriented information delivery* (adaptation of documented knowledge according to the user's current location, locating people according to the user's location, for example, locating colleagues, knowledge experts, personalization, profiling according to the user's location and situation, providing proactive mobile KM services);
- to *support heterogeneous technologies* and standards, for example, different devices, protocols, and networks;
- to provide *proactive information delivery* (using mobile devices focusing on push services);
- to provide *adaptive information delivery* (using mobile devices focusing on profiling, personalization, contextualization); and
- to use *speech technology* in order to enable mobile access of knowledge portals. The portal should provide advanced services, for example, to read out e-mails and information subscriptions, use speech-to-text technologies.

EXAMPLES OF MOBILE KNOWLEDGE PORTALS AND PLATFORMS

More and more application server platforms, for example, IBM Websphere, Oracle Application Server, and SAP Mobile Business Platform, are enhanced by mobile business components and mobile interfaces to other back-end systems, enabling the development of comprehensive mobile knowledge portal solutions. The IBM Websphere Everyplace Access Platform, for example, provides prepacked mobile portlet applications (e.g., LDAP-Search Portlet, Lotus Notes, and MS Exchange Portlet), synchronization services to synchronize dates and addresses, content adaptation services, offline Web content browsing and common services for user authentification.

In order to get an idea about features and functions offered by existing mobile knowledge portals, we briefly describe selected commercial portal solutions and classify these solutions according to their main focus with regard to mKM requirements. However, none of the commercial available portal solutions is meeting all of these mobile KM requirements:

- **Hyperwave Information Portal:** Hyperwave's WAP (wireless application protocol) framework, for example, enables mobile users to browse the hyperwave information server with WAP-enabled devices. Special WAP-tracks are provided in order to access the portal. Currently, only a limited number of out-of-the-box tracks are offered, for example, find-people portlet, news-changer (Hyperwave, 2002).
- **Livelink Wireless:** At present, the arguably most comprehensive support for mobile KM seems to be provided by the Livelink portal from Open Text Corporation. With the help of the wireless server, users can access discussion boards, task lists, user directories (MS Exchange, LDAP, Livelink User Directory), e-mails, calendar, and documents (Figure 2). In addition, it provides some KM services specially developed for mobile devices, for example, automatic summarization of text. Hence, even longer texts can be displayed on smaller screens (Figure 3).
- **Autonomy Portal-in-a-Box:** This portal provides typical KM functions, for example, automated content aggregation and management, intelligent navigation and presentation, personalization, role-based access, and so forth. The IDOL mobile is an extension of the portal solution and enables the access to specific portlets via WAP browser. The retrieval portlet is able to search the knowledge base of portal-in-a-box using common search options, for example, keywords, metadata, full text, and summarizes the query results. In order to support the networking between knowledge workers, autonomy provides a special community portlet (Autonomy, 2005).
- **Hummingbird Enterprise Portal:** The mobility solution enables users to securely browse access enterprise content no matter which device they use (Palm, Pocket PC,

Figure 2. Tasklist, calender, and discussion board of Open Text's Livelink Wireless (Open Text, 2003, p. 12)

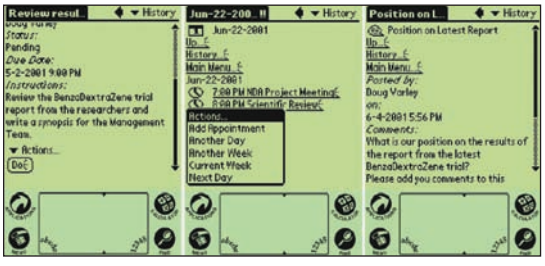
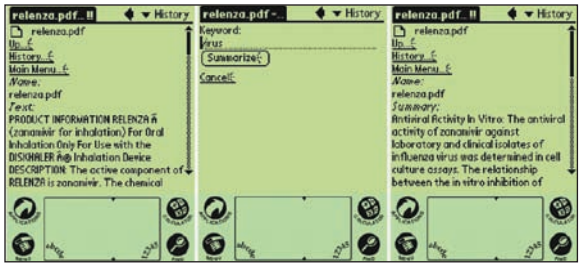


Figure 3. Automatic text summarization (Open Text, 2003, p. 11)



Smart Phones) on any network-connected drive. Search results can also be viewed with a summary. It performs common actions, such as check-in, check-out, e-mail, publish, uses multiple view options, native format, as HTML or PDF, preview, metadata, history, versions. The system provides functions to manage workflows, document reviews, and escalations, as well as instant messaging and intelligent notifications. The delivery of content to the device can be controlled with rules based on priority, size, and sender (Hummingbird, 2005).

- **Plumtree Wireless Device Server:** The main focus lies on social networking, mobile, proactive access on information sources, and the support of heterogeneous technologies and standards. Customers can retrieve portal resources from virtually anywhere by using the wireless device server, an add-on component to the Plumtree corporate portal that allows users to access supported gadget Web services from mobile devices, such as WAP-enabled mobile phones, wireless-enabled Palm handheld computers, and BlackBerry wireless handheld (Plumtree, 2005).

CONCLUSION AND OUTLOOK

At the moment, commercial portal packages cannot sufficiently fulfil the needs of mobile KM. Most of the systems are enhanced by mobile components, which are rather providing mobile access to stationary KM services instead of implementing specific mobile KM services. Hence, a full mobile KM solution should make use of some specific characteristics of mobile technology like permanent connectivity, anytime accessibility, or exploit location-related context of the users to provide, them with some additional value, like delivering location-related information or providing anytime connectivity to domain experts.

In particular, today's knowledge portals are ill-suited to support aspects of KM derived from a location-oriented perspective (Berger, 2004). One reason is that the context, which is defined by the corresponding situation (tasks, goals, time, and identity of the user), is still not extended by location-oriented context information (Abecker, van Elst, & Maus, 2002). The field of location-oriented KM draws attention from research in mobile knowledge management, ubiquitous computing, location-based computing, and context-aware computing (Lueg & Lichtenstein, 2003).

Some research projects are already addressing the issue of location-oriented information delivery. The vision of the EU-funded project MUMMY, for example, is to enable mobile, personalised knowledge management based on the usage of rich multimedia to improve the efficiency of mobile business processes. The portal prototype will enable, for instance, a facility manager to have situation-aware mobile access to up-to-date project data, such as a construction plan, multimodal annotations, and deficiency lists, or to collaborate on acquired material and plans with remote experts (Grimm et al., 2002).

However, the explicit consideration of the user's location could make business process more efficient, as times for searching can be reduced

due to the fact that information about the location might restrict the space of searching (e.g., an engineer might get information about a system that he/she is currently operating). Possibly, redundant ways between mobile and stationary work place are omitted when the information is already provided on the move. Another advantage is seen in the portals personalisation services: When considering the user's location, information can be delivered to the user in a much more customized and targeted way (Rao & Minakakis, 2003). Finally, the integration of common knowledge services, together with location-oriented mobile services, may also extend the scope for new applications in KM, for example, the use of contextual information for the continuous evolution of mobile services for mobile service providers (Amberg, Remus, & Wehrmann, 2003). One can also think of providing a more "intelligent" environment, where information about the user's location, combined with sophisticated knowledge services, adds value to general information services (e.g., in museums, where customized information to exhibits can be provided according to the user's location).

To build up mobile knowledge portals that can support the scenario described, mobile portlets are needed that can realize location-oriented KM services. In case of being implemented as proactive services (in the way that a system is going to be active by itself), these portlets might be implemented as push services. In addition, portlets have to be responsible for the import of location-oriented information, the integration with other contextual information (contextualization), and the management and exploitation of the location-oriented information. Of course the underlying knowledge base should be refined in order to manage location-oriented information.

With respect to mobile devices, one has to deal with the problem of locating the user and sending this information back to the knowledge portal. Mobile devices might be enhanced with systems that can automatically identify the user's location. Dependent on the current net infrastructure (personal, local, or wide-area networks), there are many possibilities to locate the user, for example, WiFi, GPS, or radio frequency tags (Rao & Minakakis, 2003).

Loutchko and Birnkraut (2005) identified another important issue, the opportunity to change access devices and protocols on-the-fly, depending on users' current location and environment. This, however, requires that mobile knowledge portals provide tools and services for device and session management. Moreover, the mobile technology could even add more value to the functionalities of the knowledge portal by providing him/her with location- and context-related knowledge both through the push- and pull-based mechanisms.

All in all, even though research in the field of mKM is increasing (e.g., FieldWise (Fagrell, Forsberg, & Sanneblad, 2000), MUMMY, (Grimm et al., 2002), Shark (Schwotzer & Geihs, 2003), K_Mobile (Gronau, Laskowski, & Martens, 2003)) there is still a long way to go until the potentials of mobile technologies are fully realized in mobile knowledge portals. More applied research work is needed in the future to address the adaptation of mobile services, the consideration of the user and work context for KM, and the design of highly context-aware knowledge portals.

REFERENCES

Abecker, A., van Elst, L., & Maus, H. (2002). *Exploiting user and process context for knowledge management systems.* Paper presented at the 8th International Conference on User Modeling, Sonthofen, Germany.

Amberg, M., Remus, U., & Wehrmann, J. (2003). *Nutzung von Kontextinformationen zur evolutionären Weiterentwicklung mobiler Dienste*. Paper presented at the 33rd Annual Conference "Informatics 2003", Workshop Mobile User—Mobile

Knowledge—Mobile Internet, Frankfurt a.M., Germany.

Autonomy. (2005). *Autonomy portlets. Technical brief.* Retrieved October 18, 2005, from http://www.autonomy.com/downloads/

Belotti, V., & Bly, S. (1996). *Walking away from the desktop computer: Distributed collaboration and mobility in a product design team* (pp. 209-218). Paper presented at the CSCW'96. Boston: ACM Press.

Berger, S. (2004). *Mobiles Wissensmanagement. Wissensmanagement unter Berücksichtigung des Aspekts Mobilität.* Berlin: dissertation.de.

Collins, H. (2003). *Enterprise knowledge portals* (1st ed.). New York: American Management Association.

Detlor, B.(2004). Towards knowledge portals: From human issues to intelligent agents. In *Information science and knowledge management* (Vol. 5). Berlin: Springer.

Fagrell, H., Forsberg, K., & Sanneblad, J. (2000). FieldWise: A mobile knowledge management architecture. In *Proceedings of the ACM conference on Computer supported cooperative work* (pp. 211-220). Philadelphia.

Grimm, M., Tazari, M.-R., & Balfanz, D.(2002). Towards a framework for mobile knowledge management. Paper presented at the *Fourth International Conference on Practical Aspects of Knowledge Management 2002 (PAKM 2002)* (pp. 326-338). Vienna, Austria.

Gronau, N., Laskowski, F., & Martens, S. (2003). K_Mobile: Betriebliche Informationsinfrastruktur und mobiler Wissenszugang. *Industrie Management, 19*(6), 21-24.

Hansmann, U., Merk, L., Niklous, M. S., & Stober, T. (2001). *Pervasive computing handbook*. Berlin: Springer.

Hummingbird. (2005). *Hummingbird enterprise mobility data sheet* Retrieved October 18, 2005, from http://mimage.hummingbird.com/alt_content/binary/pdf/collateral/ds/he04/mobility_datasheet.pdf

Hyperwave. (2002). *HIP hyperwave information portal: User guide.* Munich.

Loutchko, I., & Birnkraut, F. (2005). Mobile knowledge portals: Description schema and development trends. In K. Tochtermann & H. Maurer, H. (Eds.), Paper presented at the I-KNOW'05 (5th International Conference on Knowledge Management), Graz, Austria. *Journal of Universal Computer Science (J.UCS),* 187-196.

Lueg, C., & Lichtenstein, S. (2003, November 26-28). *Location-oriented knowledge management.* Paper presented at the 14th Australasian Conference on Information Systems (ACIS 2003), Perth, WA, Australia.

Maier, R. (2004). *Knowledge management systems, information and communication technologies for knowledge management.* Berlin: Springer.

Open Text Corporation. (2003). *Livelink wireless: Ubiquirous access to Livelink information and services.* White Paper. Waterloo, ON Canada.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: Understanding access anytime, anywhere. *ACM Transactions on Human-Computer Interaction, 8*(4), 323-347.

Plumtree. (2005). Plumtree software wireless device server, IT research library @forbes.com. Retrieved October 18, 2005, from http://itresearch.forbes.com/detail/PROD/1025302061_754.html

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services, *Communications of the ACM, 46*(12), 61-65.

Schulte, B. A. (1999). *Organisation mobiler Arbeit. Der Einfluss von IuK-Technologien*. Wiesbaden: DUV.

Sherry, J., & Salvador, T. (2001). Running and grimacing: The struggle for balance in mobile work. *Wireless World: Social and Interactional Aspects of the Mobile Age* (pp. 108-120). New York: Springer.

Schwotzer, T., & Geihs, K. (2003). Mobiles verteiltes Wissen: Modellierung, Speicherung und Austausch. *Datenbank Spektrum, 3*(5), 30-39.

KEY TERMS

Enterprise Portal: An application system that provides secure, customizable, personalized, integrated access to a variety of different and dynamic content, applications, and services. They provide basic functionality with regard to the management, structuring, and visualization of content, collaboration, and administration.

Knowledge Management System (KMS): Knowledge management systems (KMS) provide a single point of access to many different information and knowledge sources on the desktop, together with a bundle of KM services, in order to support the main KM activities, that is, capture, organise, store, package, search, retrieval, transfer, (re-) use, revision, and feedback.

Location-Orientation: Location-orientation explicitly considers the location of the mobile worker and adapts mobile services accordingly.

Mobile KM Service: The core of the KMS architecture consists of a set of knowledge services in order to support discovery, publication, collaboration, and learning. Personalization services are important to provide a more effective access to the large amounts of content, that is, to filter knowledge according to the knowledge needs in a specific situation, and offer this content by a single point of entry (portal). In particular, personalization services, together with mobile access services, become crucial for the use of KMS in mobile environments.

Mobile Knowledge Management: Mobile knowledge management is a KM approach focusing on the usage of mobile ICT in order to provide mobile access to knowledge management systems and other information resources, generate awareness between mobile and stationary workers by linking them to each other, and realize mobile KM services that support knowledge workers in dealing with their tasks.

Mobile Portal: A mobile portal is an enterprise portal focusing on the mobile access of applications, content, and services, as well as the consideration of the location while on the move. Mobile access is about accessing stationary KMS, whereas location-orientation explicitly considers the location of the mobile worker.

Mobile Portlet: Mobile portlets are portlets enabling the mobile access of mobile workers. Special portlets can be implemented to support location-orientation in mobile portals.

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 599-605,

Chapter 1.18
Mobile Knowledge Management

Volker Derballa
University of Augsburg, Germany

Key Pousttchi
University of Augsburg, Germany

INTRODUCTION

Whereas knowledge management (KM) has gained much attention in the field of management science and practice as the eminent source of competitive advantage (e.g., Davenport & Prusak, 1998; Drucker, 1993; Nonaka & Takeuchi, 1995; Probst, Raub, & Romhardt, 2003), one issue has been largely neglected: The aspect of mobility.

Conventional solutions for knowledge management systems (KMSs) have in common that they are designed for stationary workplaces and consequently require the corresponding infrastructure—that is, personal computers and fixed-line network access. Thus, they do not cater for business processes in which workers move around in or outside the premises. The result is that knowledge support for mobile workers is often rather restricted, once a task has to be performed outside of the office. Organizations in which parts of the workforce belong to one of the following classifications are concerned in that context:

- Specialists, mobile on the premises (e.g., in-house technicians)
- Specialists, mobile outside the premises (e.g., members of the sales force)
- Specialists and executives in companies with mobile operations (e.g., organizations like contracting business, police, or armed forces)
- Decision makers (e.g., CEOs who are required to make timely and well-funded decisions disregarding their current position)

The need for mobile KM stems from one of the most prominent challenges in KM: ensuring the availability of knowledge in the moment of knowledge demand. Insufficient knowledge at "point-of-action" is the wording Wiig (1995) uses to describe that problem. There exist situations in the course of daily work that require particular knowledge that is not owned by the individual actor. As long as organization members are located

at the same place, knowledge repositories can be easily accessed. In some cases it might for example be sufficient to walk down the office floor and ask colleagues for help in order to establish a basic form of knowledge exchange. Another example is the access of best practices databases using a stationary computer.

Analyzing business processes with mobile elements, it is obvious that the insufficient integration of many mobile workplaces leads to suboptimal processes. It is usually required to interrupt the actual task in order to feed knowledge into or retrieve it from repositories. A mobile worker can access his company's knowledge infrastructure not at all or only indirectly. This leads to a time-consuming process in which workers spend valuable working hours searching for knowledge instead of pursuing their actual job. That is exactly what has to be avoided, considering the imperative of making access to knowledge as simple as possible. Figure 1 illustrates the break existing in the process chain due to the insufficient integration of mobile workplaces: as the mobile worker is not integrated into the process chain, the information and knowledge flows in mobile business processes are equally disrupted.

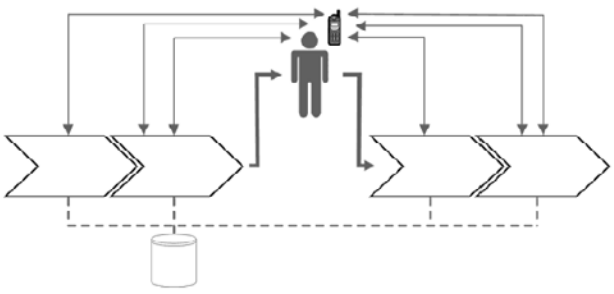
As the aspect of mobility is underrepresented in KM literature, we aim at providing an evaluation framework for managing knowledge in mobile settings (i.e., mobile KM). In order to do that, we will resort to the insights gained in the discussion of mobile techniques. As both concepts have not been sufficiently put together, we think that substantial benefits can be derived by merging the ideas behind mobile techniques and KM.

BACKGROUND

As a survey of KM literature shows, mobile KM has been largely neglected. The following section presents an overview of exemplary articles dealing with mobile KM.

The works of Fagrell (2000) can be regarded as some of the first valuable approaches to address the area of mobile KM. With NewsMate, Fagrell is presenting a KMS application that aims at supporting mobile knowledge workers. In this system KM is integrated with the relevant task that needs to be supported. A proof of concept is given by presenting a working prototype. This prototype allows journalists to access internal and external information. Further, NewsMate acts as an expert finder by automatically identifying colleagues who have worked on the same topic.

Figure 1. Non-integration of mobile workplaces into the process chain



Grimm, Tazari, and Balfanz (2002) are discussing limitations of mobiles devices for the purposes of mobile KM and present a framework for the implementation of mobile KMSs. In the course of that, they address technological as well as human limitations and thus touch the area of human computer interaction (HCI). The authors aim at using the potential of mobile technology to deliver context-specific information by using the user's location to determine relevant context. Using a "virtually centralized" context manager to handle profiles of relevant objects, a situation recognition engine enables the context-specific provision of knowledge.

Martens and Gronau (2003) introduce the potentials of mobile KM by primarily referring to the dimension of ubiquity. They base their analysis on specific characteristics of mobile technology as discussed in literature dealing with the mobile economy. As a result, the reductions of spatial and temporal limitations, as well as persistent connectivity, are isolated as relevant factors. Further, the authors present short examples of how those potentials can be used in the area of KM.

Two areas can be pointed out as deficits in existing mobile KM literature. There is a lack of a holistic concept, as the focus is put either on specific technological or on KM-related issues. Additionally, it can be observed that the term "mobile KM" is used for a wide variety of cases (e.g., Lehner & Berger, 2001). That said, some of those cases do not deserve the term mobile KM, as they represent simple data integration processes. In order to structure the different applications termed mobile KM, the following three categories are developed:

1. **Mobile information exchange** includes the transfer of data and information using e-mail as well as the access to operational systems used in an organization to retrieve sales figures or market data.
2. **Mobile business intelligence** refers to the access of processed enterprise data using mobile devices. It involves the technologies introduced earlier (e.g., data mining and data warehouses).
3. **Mobile KM** describes that management process in the course of which mobile communication techniques in conjunction with mobile devices are employed for the creation, validation, presentation, distribution, or application of knowledge.

The basis for those categories is the knowledge versus information view (Holsapple, 2003). One of the authors representing this point of view is North (1999). He argues that information—derived from signs and data—is the basis for knowledge as soon as it is associated with context or other information (North, 2001). Starting with signs (e.g., "0123") and structuring them with a certain syntax will result in data (e.g., "10,23"). Data plus semantic becomes information (e.g., "10.23 refers to the percent improvement in sales figures"). This information is relatively useless, because up to that point it cannot be assessed whether—in this example—the increase can be judged as being sufficient enough. Only in context with other information and experiences is one able to determine that an increase of 10.23% is positive indeed for a company that operates in a shrinking or stable market, whereas it would considered below average for a company in booming business. Included in the knowledge stair is the idea that knowledge is a direct precursor of competence, because it enables competent action. Information on the other hand is relatively useless (Sveiby, 1997), as long as it is not processed and linked with other information, judgments, or personal opinions.

Talking about mobile KM in the narrower sense, we think that only those processes dealing with representations that have been to some degree mentally processed by human actors can be considered mobile KM. Additionally, mobile KM must be integrated into a holistic KM concept. Mobile data access as well as mobile business

intelligence serve as supporting techniques in the context of mobile KM as they provide the input for human knowledge creation and thus can be referred to as mobile KM in the broader sense.

MAIN FOCUS OF THE ARTICLE

We state that for the success of mobile KM, it is not sufficient to merely make a conventional KM application available with new media. Instead, the use of mobile communication technology is only remunerative if it results in obtaining distinct supplementary added value. In order to verify the contribution mobile technology can make to KM, we are referring to the theory of informational added values (IAVs) which has been augmented with electronic added values (EAVs) and mobile added values (MAVs) (Bazijanec, Pousttchi, & Turowski, 2004).

In his theory of informational added values, Kuhlen (1996) discusses the impacts of information work in information markets. In this context we will introduce the categories of the supplementary gains of utility. Kuhlen terms resulting gains as informational added values and classifies them into eight main types: organizational, strategic, innovative, macroeconomic, efficiency, effectiveness, aesthetic-emotional, and flexible added values.

Efficiency added values cover the increase of operating efficiency and cost effectiveness. Effectiveness added values cover an augmentation in output quality. Aesthetic-emotional added values cover increase of subjective factors as wellbeing, job satisfaction, or acceptance of performance. Flexible added values cover a shift to a higher level of flexibility. Organizational added values cover the opportunity to build new forms of organization through the use of information and communication systems. Innovative added values cover the creation of an entirely new product or service (or combination of both) through the usage of new means of communication. Strategic added values qualify advantages that go beyond the operational and tactical level by creation of a significant competitive edge. Macroeconomic added values qualify advantages that go beyond the level of single companies and result in impacts on occupational images, economy, or society as a whole. IAV can be described as the resulting benefit of electronic or mobile solutions.

EAV refers to typical characteristics of electronic solutions leading to supplementary IAV. EAV results from the advantages of the utilization of fixed-line Internet access. Four EAVs can be differentiated: reduction of spatio-temporal restrictions, multimediality and interaction, equality of access, and reduction of technical restrictions. As the focus of this article is on mobile KM, we are not going to discuss EAV further. Interested readers can find further information in Turowski and Pousttchi (2003).

MAV refers to properties of mobile technology and its utilization leading to supplementary IAV like gains in efficiency or effectiveness in comparison to the use of fixed networks. MAV however only represents a potential, and a mobile solution does not have to take advantage of any MAV. But in order to gain supplementary IAV, at least one MAV has to be employed. Otherwise, the use of mobile technology is not remunerative. In the following we introduce the MAV ubiquity, context-sensitivity, identifying functions, and command and control functions.

Ubiquity

Ubiquity is the possibility to send and receive data anytime and anywhere, and thus eliminate any spatiotemporal restriction. It is originated not only in the technical possibility, but also in the typical usage of mobile devices, which accompany their user nearly anytime and anywhere. It permits the reception of time-critical and private information. Ubiquity effects in accessibility of mobile services anytime, anywhere for the user which affects reaction time and convenience aspects of

services. But it affects also in reachability of users. This means primarily to reach a single user anytime, anywhere.

Context-Sensitivity

Another typical attribute is context-sensitivity, which describes the delivery of customized products or services fitting the particular needs of the user in his current situation. This is particularly enabled by three features.

Personalization allows creating specific services through preference profiles. These may be generated by information the user provides about him, but also by applications tracking his attitude. As on one hand a mobile device is typically used only by a single user and on the other hand one user typically uses only one mobile device, resulting data is of high quality. *Interactivity* enables specific services through direct information exchange. Both sides can react without any delay on actions or requests of the other. *Location determination* allows specific products and services for the user to be created, in the context of his current location or by referencing on the location of other users. In particular, combinations of these concepts allow determining a user's context. Typical applications based on the MAV of context-sensitivity are location-based services.

Identifying Functions

The ability to authenticate the user as well as the device is already immanent to a mobile network. Together with the aforementioned typical 1:1-attribution of a mobile device to its user (which is perhaps not true for any other technical device except a wristwatch), this provides a capability to authenticate the actual user with a feasibility already sufficient for most applications. In case it may be necessary, it is also easily possible to apply further means of authentication on the device, from a personal identification number to biometric identification or mobile signatures. This allows much easier than conventional Internet techniques to use mobile devices for critical processes.

Command and Control Functions

The last properties to present are command and control functions of mobile devices. Mobile devices can be used as remote control for almost any application or device. For this purpose they use networking capabilities of any range, from the personal or local area network up to the wide area network. If the target is an application, it has just to be connected to the Internet. If the target is a device (which can be almost any electrical device), control may be realized, for example using networking capabilities via ubiquitous computing technology or embedded mobile devices.

MAV-BASED ANALYSIS

For analytical purposes it was necessary to choose a KM process model that serves as a framework for the evaluation of the potential mobile techniques can contribute towards KM. In the next section the process model according to Bhatt (2001)—consisting of the elements knowledge creation, validation, presentation, distribution, and application—is introduced, before the results of the MAV-based analysis are discussed (examples for mobile KM use cases demonstrating the effect on KM processes can be found in Derballa & Pousttchi, 2004).

Knowledge creation refers to a process in which new knowledge is created by combining and integrating different modes of knowledge. Knowledge validation describes controlling activities like testing new and removing old knowledge. Knowledge presentation refers to the display of knowledge—that is, different formats, data standards, and so forth. Knowledge distribution deals with sharing and distributing knowledge between organization members. Knowledge application is the term for the use of knowledge in a particular

context. Taking into account two approaches—the technical as well as the social strategy—we have examined each KM sub-process regarding the mobile added values that can be generated through the use of mobile technology. Table 1 presents an overview of the results.

Knowledge creation is supported through the mobile added value of ubiquity, as this aspect allows the creation of knowledge regardless of spatial and temporal restrictions. This refers to the enabler function; mobile technology is inherent when it comes to virtual teamwork and the mobile access to knowledge repositories. Context-sensitivity and identifying functions act as supporting factors in that context. They facilitate the documentation of the knowledge creation process. Using those values, it becomes possible to gather information in which that knowledge was created as well as on the participating users.

Knowledge validation benefits from the aspect of ubiquity as the verification of knowledge becomes possible immediately in that moment an event has occurred that leads to a new judgment of existing knowledge. Furthermore, the MAV identification function enables an accurate documentation of the user responsible for the validation.

Knowledge presentation is only supported to a very low degree regarding all four MAVs.

Knowledge distribution is improved by the ability of mobile technology to deliver knowledge everywhere (MAV ubiquity), adjusted/aligned to the relevant context (MAV context-sensitivity), and appropriate for the individual user (MAV identification functions). Taking that into account, it becomes possible to employ push approach and deliver the knowledge to the user, instead of requiring the user actively to retrieve knowledge. Thus the overall KM process is considerably improved as it is no longer necessary for knowledge seekers to be actively involved in the process of determining what knowledge is relevant for them. Instead, the relevancy of knowledge for a particular actor can to some degree be determined by the context. Thus the knowledge seeker is relieved from that burden. Further, to retrieve knowledge, the knowledge seeker has to have a certain understanding of what he is looking for. Without that, it is almost impossible to find that knowledge, which is relevant in a particular context. By switching from pull to push, this problem can be attenuated. In addition the MAV control and command functions enable the control of KMS using mobile devices.

Knowledge application is enhanced indirectly by the fact that mobile technologies make it possible to have relevant knowledge delivered to the individual user regardless of spatial and temporal

Table 1. Results of MAV-based analysis

	Mobile Added Values			
KM Process	Ubiquity	Context-Sensitivity	Identifying Functions	C&C Functions
Creation	X	X	X	
Validation	X		X	
Presentation				
Distribution	X	X	X	X
Application	X	X	X	

restrictions and thus ensure that "insufficient knowledge in time-of-action" is avoided.

The results of the MAV-based analysis demonstrates the substantial impact of mobile techniques on the process of knowledge distribution. Considering the different roles of individual MAVs in the context of mobile KM, an order of relevancy can be identified. The primary MAV is ubiquity as it extends the reach of KM. Due to that MAV, KM solutions become available in situations that otherwise could not have been included in the KM process. The other MAVs act as supporting factors, with context-sensitivity and identifying functions coming second and control and command functions ranking third. Those MAVs improve the quality as well as the effectiveness of KM solutions.

FUTURE TRENDS

Current research projects (e.g., the EU-funded project MUMMY) demonstrate the relevancy of mobile KM. As mobile KM is no isolated application, further research is necessary to determine how KM can be fully integrated into a holistic KM concept. In the technological domain this includes questions of data and application integration. Regarding the process perspective the integration of knowledge flows into mobile business processes needs to be analyzed.

Mobile techniques offer a great potential to KM. The impact on organizations and its actors however is not clear. The Adaptive Structuration Theory as presented in DeSanctis and Poole (1994) can be used to assess possible effects. Further, empirical studies need to be conducted to investigate the usability of mobile technology in the conjunction with different KM techniques.

Two major trends can be identified as factors that influence future developments in the area of mobile KM. In the field of KM, a stream of research with focus on process orientation of KM is evolving (e.g., Becker, Hinkelmann, & Maus, 2002). On the other hand, enterprises are increasingly considering mobile business processes and aim at integrating the necessary applications into their business information systems (Gruhn & Book, 2003). If those trends can successfully be integrated, mobile KM is on the way to establish itself as an integral part of KM.

CONCLUSION

As more and more business processes are conducted by organization members who are locally and temporally dispersed, it is obvious that KM restricted to stationary workplaces alone can not cater for knowledge support requirements of mobile workers. The gaps existing due to the insufficient integration of mobile actors can be filled using MAV. With the MAV ubiquity, context-sensitivity, identifying functions, and command and control functions, serious KM deficits can be reduced, and thus the overall KM scope and effectiveness can be improved.

This article contributes to the area of mobile KM by introducing the relevant values a mobile KM solution has to provide in order to be remunerative. By doing so, the basis for an evaluation of mobile KM is laid. Further, the ideas presented in this article can be used to support the development of a holistic mobile KM framework.

REFERENCES

Bazijanec, B., Pousttchi, K., & Turowski, K. (2004). *An approach for assessment of electronic offers. Proceedings of FORTE 2004,* Toledo, OH.

Becker, A., Hinkelmann, K., & Maus, H. (2002). Integrationspotenziale für Geschäftsprozesse und Wissensmanagement. In A. Becker, K. Hinkelmann, & H. Maus (Eds.), *Geschäftsprozessorientiertes Wissensmanagement: Effektive wis-*

sensnutzung bei der Planung und Umsetzung von Geschäftsprozessen. Berlin.

Bhatt, G.D. (2001). Knowledge management in organizations: Examining the interaction between technologies, techniques, and people. *Journal of Knowledge Management, 5*(1), 68-75.

Davenport, T.H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* Boston.

Derballa, V., & Pousttchi, K. (2004). Extending knowledge management to mobile workplaces. *Proceedings of the 6th International Conference of Electronic Commerce,* Delft, The Netherlands.

DeSanctis, G., & Poole, M.S. (1994). Capturing the complexity in advanced technology use: Adaptive Structuration Theory. *Organization Science, 5*(2), 121-145.

Drucker, P.F. (1993). *Post-capitalist society.* New York.

Fagrell, H. (2000). *Mobile knowledge (no. 18): Gothenburg studies in informatics.*

Grimm, M., Tazari, M.-R., & Balfanz, D. (2002). Towards a framework for mobile knowledge management. *Proceedings of the PAKM,* Vienna.

Gruhn, V., & Book, M. (2003). Mobile business processes. *Proceedings of the Innovative Internet Community Systems, 3rd International Workshop* (IICS 2003), Leipzig.

Holsapple, C.W. (2003). Knowledge and its attributes. In C.W. Holsapple (Ed.), *Handbook on knowledge management* (Vol. 1, pp. 165-188). Berlin.

Kuhlen, R. (1996). *Informationsmarkt: Chancen und Risiken der Kommerzialisierung von Wissen*. Konstanz.

Lehner, F., & Berger, S. (2001). *Mobile knowledge management.* Regensburg: Universität Regensburg, Lehrstuhl für Wirtschaftsinformatik III.

Martens, S., & Gronau, N. (2003). Erschließung neuer Potentiale im Wissensmanagement über den mobilen Kanal. *Proceedings der GI,* Bonn.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating-company: How Japanese companies create the dynamics of innovation.* New York.

North, K. (1999). *Wissensorientierte Unternehmensführung: Wertschöpfung durch Wissen*. Wiesbaden.

Probst, G., Raub, S., & Romhardt, K. (2003). *Wissen managen: Wie Unternehmen ihre wertvollste Ressource optimal nutzen.* Wiesbaden.

Sveiby, K.E. (1997). *The new organizational wealth.* San Francisco.

Turowski, K., & Pousttchi, K. (2003). *Mobile commerce*. Heidelberg.

Wiig, K.M. (1995). *Knowledge management methods: Practical approaches to managing knowledge*. Arlington.

KEY TERMS

Electronic (EC) and Mobile Commerce (MC): EC is defined as any kind of business transaction, in the course of which transaction partners employ electronic means of communication, may it be for initiation, arrangement, or realization of performance. MC is a subset of these, on the condition that at least one side uses mobile communication techniques (in conjunction with mobile devices).

Mobile Added Values (MAVs): Those properties (ubiquity, context-sensitivity, identifying functions, and command and control functions) of mobile technology and its utilization which are responsible for gaining supplementary IAV in comparison to EC solutions.

Mobile Business Intelligence: Refers to the access of processed enterprise data using mobile devices. Involves different technologies (e.g., data mining and data warehouses).

Mobile Information Exchange: Includes the transfer of data and information using e-mail as well as the access to operational systems used in an organization to retrieve data or information (e.g., sales figures or market data).

Mobile Knowledge Management: Describes that management process in the course of which mobile communication techniques in conjunction with mobile devices are employed for the creation, validation, presentation, distribution, or application of knowledge. An important issue is the integration of knowledge flows and mobile business processes to ensure knowledge support for mobile workers.

Theory of Informational Added Values (IAVs): Concept discussing the impacts of information work in information markets comprising the following eight types: organizational, strategic, innovative, macroeconomic, efficiency, effectiveness, aesthetic-emotional, and flexible added values. IAVs may represent the resulting benefit of an EC solution as well as of an MC solution.

This work was previously published in Encyclopedia of Knowledge Management, edited by D. Schwartz, pp. 645-650, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.19
Assessing Human Mobile Computing Performance by Fitts' Law

Thomas Alexander
FGAN - Research Institute for Communication, Information Processing, and Ergonomics, Germany

Christopher Schlick
RWTH Aachen University, Germany

Alexander Sievert
German Sport University Cologne, Germany

Dieter Leyk
German Sport University Cologne, Germany
Central Institute of the Federal Armed Forces Medical Services Koblenz, Germany

ABSTRACT

This chapter describes the interdependence between locomotion while walking and human input performance in mobile Human-Computer-Interaction (HCI). For the analysis of the interdependence, appropriate performance measures, for example, subjective workload ratings or error rate, have to be applied. The way in which Fitts' law can enhance the analysis is explained. In an experiment with n=18 participants, the general indices of performance (bits per second) were measured while standing and walking with constant speed (2, 3.5, 5 km/h). Results show a significant increase of the error rate and a significant decrease of the index of performance for increased walking speed. Subsequent regression analyses allow quantitative estimation of these effects. The results show a division of the interdependence in two parts, based on the difficulty of the input task; they define threshold values for accuracy of user input. These values can be applied for the implementation and design of future Graphical User Interfaces (GUI) for mobile devices.

INTRODUCTION

Flexibility, variability, and mobility are topics of growing importance for today's society. This trend affects work with modern IT-systems (Goth, 1999): There is a growing availability and market for portable and mobile devices. They facilitate ubiquitous information access throughout customers' visits, while traveling, wandering through a production plant, or for working at home offices. It is expected that the market share of telecommuting and according devices for information access will increase. IT-developers and providers share this optimistic estimation of the growth potential (Business Week, 2006). They assume that today's mobile computers already have a market share of 40% (Microsoft, 2006). This requires special information infrastructures and personal mobile devices.

Common portable and mobile devices are notebooks, tablet-PCs, personal digital assistants (PDA), and so-called smartphones, which are cellular phones with enhanced functionality. For applications while standing or on the move (walking), when no tables or horizontal racks are available, weight and size issues are most relevant. They reduce the available devices to small, lightweight PDAs and smartphones.

PDAs and especially smartphones often rely on direct keypad input. Keypads allow a fast selection of a limited number of special functions. However, this is hardly sufficient for a more complex interaction. In that case, a point-and-click procedure is applied, which requires special touch-sensitive screens and pens for HCI and a WIMP-metaphor (windows, icons, menus, and pointer) is implemented (well known from most desktop systems) for the graphical user interface (GUI). Required training is reduced and most users can instantly use the device. Text input is facilitated by a miniaturized (virtual) keyboard or handwriting recognition. The keyboard solution displays a miniaturized QWERTY-keyboard and keys are selected by pointing and clicking. Handwriting recognition requires a stable position of the base for a precise text input. In both cases, pointing and pointing accuracy are essential.

Most of today's GUIs of mobile devices are simply adapted from stationary desktop systems. Characteristics of mobile use and their effects have not been considered (Berteksen & Nielsen, 2000; Crowley et al., 2000; Danesh et al., 2001; Dunlop & Brewster, 2002; Lumsden & Brewster, 2003; York & Pendharkar, 2004). However, effects of mobility on input performance are likely because of various reasons. First, walking itself causes distracting movements and forces on the arm and hand system. This leads to reduced input performance. Second, pointing and moving are two concurrent tasks, both of which require attention and processing resources. As a matter of fact, performance in either of the two tasks is reduced. This can be observed when users either stop when working with a mobile device or quit working with the device.

These observations show an overall need for the inclusion of general ergonomic findings, results, and models to optimize HCI and the according GUI on the move. As a first step, valid measures have to be analyzed in order to quantify input performance under different mobile conditions. They must be sensitive enough to detect even small effects. Based on these measures, subsequent analyses will give more detailed recommendations for the design of the GUI. This way a real mobile use can be achieved.

ASSESSING INPUT PERFORMANCE OF MOBILE DEVICES

Using computers on the move is a combination of walking and HCI. There are several reasons to assume interdependences between both. The extent of these interdependences varies with the degree of mobility. For quantifying it is necessary to identify and measure HCI performance correctly. There are different methods for doing

so. This section addresses these issues and proposes the inclusion of general characteristics of the human operator.

Mobility and HCI: Two Concurrent Tasks

Mobile computing consists of the parallel processing of two tasks: walking or being "on the move" on the one hand, and HCI on the other. Such parallel processing results in interference between the two tasks (Hinckley et al., 2000; Navon & Gopher, 1979; Wickens, 1984). The extent of the interference varies with the similarity of the task.

Walking is a complex task. It requires multimodal (visual, auditory) encoding of environmental stimuli, complex central processing for reacting to the encoded stimuli and navigational orientation, and finally, controlled and coordinated motor output (i.e., walking itself). Most people are very well trained in performing this complex task so that several sub-tasks are delegated to lower processing levels. Posture control is one of those. When walking becomes more difficult due to higher speed or obstacles, a higher level of attention is required. This leaves few or no resources for additional processing (Baber et al., 1999; Kristoffersen & Ljungberg, 1999; Oulasvirta, 2005).

Interacting with the computer also requires visual encoding, cognitive processing, and motor responding. The extent of resource exploitation varies with task difficulty. While simple tasks can be handled with little attention, difficult or more complex tasks require a lot of attention and processing resources.

Consequently, both tasks compete for similar attentional and processing resources. This hypothesis is supported by the following observation: If walking speed increases or walking becomes more difficult in an environment with a lot of obstacles, people quit working with the PDA. This is because walking demands the major part of the available resources. These resources are drawn from HCI, which is the secondary task. However, if HCI becomes the main task because of prioritization, people often stop walking. In this case, available resources are moved to HCI. Too few resources are left for the walking task.

In general, the shifting of resources towards one task or the other is a deliberate decision. It is influenced by various factors (motivation, deadlines, instruction, etc.), which are beyond conscious control. However, due to safety reasons and the importance of avoiding accidents, walking is usually the primary task and HCI the secondary. This is especially true for moving in traffic situations. The highly dynamic, stimuli-rich environment causes a reduction of the available resources. There are situational induced impairments and disabilities (SIID) which hinder mobile computing (Sears et al., 2003). During movements, spatial orientation and self-movements are clearly main tasks, and only limited attention and processing resources are available for interacting with the computer (Kristoffersen & Ljungberg, 1999; Lumsden & Brewster, 2003; Pascoe et al., 2000; Perry et al., 2001). There are many attentional shifts between environment and task processing on the mobile device (Oulasvirta, 2005).

In addition to resource sharing, there is a biomechanical coupling between walking and HCI. The walking movement itself causes additional interfering forces on the hand-arm-system. They are permanently present during walking. The result is a task that includes hitting a target on the moving display with a moving input device. This is much more difficult than stationary HCI and requires additional processing.

There are many further interrelationships between walking and mobile computing. Ebersbach (1995) found changes of gait kinematics and reduced task performance while walking. Targets on the display are more frequently missed while walking (Beck, 2002). In experiments, Brewster (2002) found lower workload and higher input performance for standing rather than walking. This

increases with complexity of the task. Although there are no differences between stationary and mobile processing of single tasks, task performance of more complex tasks varies (Barnard et al., 2005).

Mobile Input Performance

Consequently, input performance for evaluating mobile computing should be examined in a realistic setting, ideally during walking. However, an extensive research of Kjeldskov & Stage (2004) reveals that this was only done in few publications, referring to an evaluation of a mobile system and an acoustic navigation system for blind persons. Our similar research identified a few more publications which found limited usability, reduced task performance, and higher workload for mobile computing (Barnard et al., 2005; Beck et al., 2002; Brewster, 2002; Lin et al., 2005; Oulasvirta, 2005).

In general, there are two different methods for analyzing mobile input performance: laboratory studies and field experiments. In a quick comparison, the latter obtain higher face validity because measures are taken under real-life conditions (Thomas et al., 2002). Field experiments allow real walking, where kinematics differ from those when walking on treadmills (Nigg et al., 1995; Schache et al., 2001; Vogt et al., 2002). With treadmills, step length is reduced and different hip, ankle, and spine motions occur (Arlton et al., 1998; Murray, 1985).

In contrast to this, laboratory tests offer a wider control of environmental conditions (Alton et al., 1998; Johnson, 1998; Kjeldskov et al., 2004; Pirhonen et al., 2002). Key situations can be analyzed relatively simply and some experimental methods such as observation or thinking aloud can be applied (Rantanen et al., 2002; Sawhney & Schmandt, 2000). The participant is in a safe state throughout the experiment. This cannot always be guaranteed in all environments in field experiments (Kjeldskov & Stage, 2004). To gain more general insight into mobile interaction performance, laboratory studies are selected rather than field experiments (Alton et al., 1998; Kjeldskov & Graham, 2003; Petrie et al., 1998).

There are many ways for measuring human input performance. A common method for assessing input performance is subjective rating. In this case, participants are simply interviewed after the experiment about their own performance (Barnard et al., 2005; Mustonen et al., 2004). The participants' comments are recorded and transferred into a qualitative or quantitative measure afterward. Questionnaires are another method for assessing subjective input performance. Here, questions about special qualities are presented to the participants. The participants estimate numerical values for the extent of this quality. There are standardized questionnaires available for measuring general usability, workload, and so forth (Barnard et al., 2005; Beck et al., 2002; Kjeldskov & Stage, 2004). A further method is observation by experimenters (Jacobsen et al., 2002; Kjeldskov & Stage, 2004; Oulasvirta, 2005). The participants are observed while performing their specific HCI task. Observers report characteristics, categorize, and rate characteristics. This method seems to be more objective than the ratings described before. However, observations remain subjective because they are strongly dependent on the observer.

For an objective assessment of input performance, various performance variables can be applied. Most of them are fully integrated into the main experimental task and refer to accuracy, time, and task-specific measures.

A common measure for accuracy is error rate (r_E), which is defined as the ratio of missed targets to total targets. Obviously, higher error rates mean reduced performance. Barnard et al. (2005) and Beck et al. (2002) report higher error rates for mobile than for stationary use. Other error measures are more task-dependent and refer

to wrong command selection or wrong ways of task processing. It must be noted that error rate is always dependent on the difficulty of the task. If a task is simple, the error rate is low, while a difficult task results into a higher error rate. There is a close dependence between accuracy and time. Obviously, tasks are performed more accurately when more time is taken and vice versa. Consequently, accuracy measurements should involve an explicit prioritization in the experimental task and task times should be recorded as well.

In terms of time, the time to complete a special task or sub-task is frequently used (e.g., Barnard et al., 2005; Cerney et al., 2004; Kjeldskov & Stage, 2004). Time measurement starts with the task and ends as soon as the task is completed. Temporal performance measures are typically task-dependent and serve as intermediate results for more general findings. When recording temporal measures, it is important to consider and determine the overall temporal accuracy of the system. This is especially important for mobile devices. Although today's devices run with high-speed CPUs, the system update is not identical to the maximum resolution of temporal measures. The actual measurement rate is smaller because the systems are not real-time systems and there are many basic I/O system processes running in parallel. Moreover, at least the double measurement frequency is required for an exact representation of the measured signal (Nyquist, 1928).

The last category includes task-specific measures. Such measures are more complex and refer to complex problem-solving approaches (Dahm et al., 2004; Oppermann, 2003; Ziefle, 2002). An example for a task-specific measure is linguistic processing and semantic understanding. In this case, a special scenario is described on the mobile display and the participant answers questions about it. Evidently, time to complete the special task and error rate are dependent on the application. Results of these task-specific measures are often difficult to generalize.

Applying Fitts' Law for Assessing HCI Performance

For a more comprehensive quantitative analysis of mobile HCI performance, general laws of human information processing have to be referenced. A basic law that was first discovered by Fitts (1954) describes the relation between movement time and index of difficulty for goal-directed movements. Thus, it facilitates an estimation of human sensorimotor performance. Fitts' law has been frequently applied for researching HCI issues of many kinds. Applying it for mobile GUIs is a way of quantifying human performance under mobile conditions.

Common user input on a PDA includes pointing and clicking on targets. The necessary movements are goal-directed and visually controlled arm-hand movements. For estimating the performance of these movements, Fitts' law can be applied (Fitts, 1954; Fitts & Peterson, 1964). According to this, the movement time (MT) is linearly dependent on the index of difficulty (ID) of a movement. It is:

$$MT = a + b\ ID \qquad (1)$$

Fitts' coefficients a and b are determined by regression. The first coefficient is a theoretical intercept for ID=0. It can be interpreted as reaction time. The second coefficient b is a measure of input performance and it characterizes the slope of the curve. Its reciprocal value (1/b) describes the index of performance in bits per second (bps) (Card et al., 1978). The ID of a movement is the logarithm of the quotient of target width (W) and amplitude of the movement (A). It is:

$$ID = \log_2 (2\ A/W) \qquad (2)$$

Goal-directed movements are divided into an initial ballistic and a final visually controlled

phase with different characteristics. The ballistic phase is fast and inexact. Its purpose is to move the finger towards the target region as fast as possible. The visually controlled phase takes longer but is more precise. It is used to hit small targets. The ID determines the more dominant phase of the movement. Simple movements with ID<3.58 are primarily ballistic, while more difficult movements with higher ID are visually-controlled (Wallace & Newell, 1983). Gan & Hoffmann (1988) found a threshold between ID=3 and 4 for the transition between ballistic and visually controlled movements.

Fitts' law has been used for many analyses of goal-directed movements, several of which have a background in HCI (compare MacKenzie, 1995), and allows a quantification of input performance for precise manual inputs.

However, despite the frequent applications of Fitts' law for stationary HCI, it has only rarely been used for investigating mobile HCI or different GUI layouts. In their publication, Lin et al. (2005) verify it for general pointing tasks on a PDA. There is a linear interrelationship between movement time and index of difficulty. It was determined that performance was slightly reduced with mobile conditions. The according indices of performance decrease from 7.8 bps (sitting), to 6.5 bps (slow walking), and, finally, to 6.6 bps (fast walking). In contrast to this, Cerney et al. (2004), with their study about different mobile text entry methods, found that Fitts' law was not applicable for PDA interaction. Measured movement times were independent from the index of difficulty. A reason for this was that the analyzed IDs were below 3. As stated before, this characterizes ballistic movements where the applicability of Fitts' law is limited. As of yet, there is no general consensus about the effect of the user's movement on Fitts' law and derived measurements of HCI performance.

Experiment on Mobile Input Performance on the Move

This work focuses on a common application of mobile devices: orientation and navigation in an unknown area. An electronic map of the surroundings is shown on the PDA-display. The user can select targets and the system calculates waypoints toward this target accordingly. A high level of precision is required for target input. For this application, user input performance should be analyzed while standing and during walking at different speeds. This ensures that the mobile characteristics are considered accordingly. The results, however, are not valid for this specific task alone. The resulting values are benchmarks for each point-and-click procedure for GUIs. In this case, they are used to determine size of icons and menu items for an optimized GUI.

Hypothesis

Walking and HCI share human information processing resources to a special extent. Higher processing demands on one task result in decreased performance in the other. If, for instance, walking becomes more demanding because of a higher walking speed, HCI performance will be reduced. It is postulated:

Walking affects input performance of a targeting task on a PDA. The effect varies with the extent of walking difficulty (i.e., walking speed).

Input performance can be characterized in general by two variables. It is accuracy/error rate r_E, on the one hand, and movement time on the other. Fast input leads to a reduced precision with a higher error rate and vice versa. For quantifying input performance it is necessary to consider both variables. In order to avoid misunderstandings between the different types of movements

(walking and hand movement), the term input time (IT) will be used in the future, instead of movement time.

As a result of different IDs, there is a separation and transition between simple and difficult movements. In this study, this separation had to be considered for the analysis of r_E and IT. The separation might result in changes of the dependency between ID and input performance. If the results were to show such differences, the further analyses would be performed separately for each movement type.

For specifying the effect on error rate r_E, a general linear model is selected. The model with intercept c_E, and the coefficients c_v, c_{ID} and c_{vID} includes linear terms for walking speed v, for index of difficulty ID, and for the interaction of both terms v x ID. It is:

$$r_E(v, ID) = c_E + c_v\, v + c_{ID}\, ID + c_{vID}\, v\, ID \qquad (3)$$

Error rate and walking speed can be measured instantly.

For assessing human input performance, the according index of performance (1/b) is determined in [bps]. This requires additional pre-analyses. At first, separate regression analyses are performed in order to estimate the parameters of Fitts' law separately for each walking speed. A subsequent inclusive regression analysis considers additional terms for the influence of walking speed. Thus, both coefficients of Fitts' law, a and b, are split into a velocity-independent (index ID) and a velocity-dependent fraction (index v). It is:

$$IT(v, ID) = (a_{ID} + a_v\, v) + (b_{ID} + b_v\, v)\, ID \qquad (4)$$

Notice that the effect of ID is considered in the function already so that there is no need to include an additional term for the interaction of walking speed and ID. By comparing both linear coefficients b_{ID} and b_v, it is possible to put the effect of the velocity-dependent term and the velocity-independent term in perspective.

Participants

Eighteen members of the participating institutions took part in the experiments. Mean age was 29 +/- 10.6 yrs (AM +/- 1 sd). Fifteen participants were male, 3 female. Each participant was familiar with computer work, computer interaction, and interaction with a PDA. None was involved in designing this study or in preparing the specific experiment. Each participant completed the total experimental trials in a randomized, balanced order.

Apparatus

A standard Dell Axim X50v-Pocket PC was utilized for the experiment (Processor Intel Xscale PXA270, 624 MHz, dimensions: 73 x 119 x 18 mm, weight: 165 g, display diagonal: 9.4 cm, resolution: 480 x 640 pixel). Participants used the standard input pen (10 cm length, 3 mm diameter).

Figure 1. Participant and HP Cosmos Pulsar™ treadmill

A H/P/Cosmos Pulsar™ treadmill, as shown in Figure 1, was used for providing and controlling a constant walking speed. The size of the treading surface was 190 cm x 65 cm. For safety reasons, it was equipped with a safety belt, fall stop, and breast clamp. The device is certified for medical treatment.

Procedure

The experiment includes walking on the treadmill and a simultaneous HCI task. Each experiment starts with a HCI trial during standing. This trial serves as a baseline. Subsequent trials on the treadmill follow in a balanced randomized order. Walking speed is varied between 0 (i.e., standing), 2, 3.5, and 5 km/h (first experimental factor). Higher walking speeds are not considered because of safety reasons.

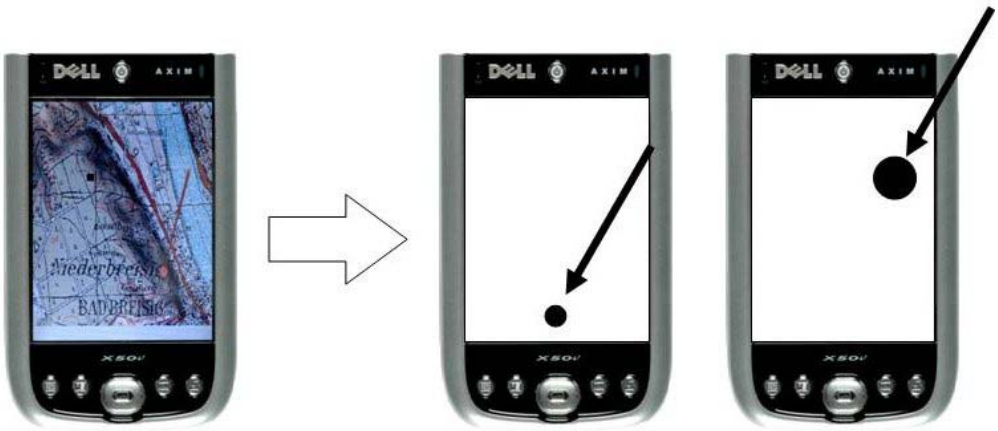
The information input task on the PDA is a target assignment task, as it is shown in Figure 2. It simulates interaction with navigation software on a PDA. During the experimental task, a starting point and a single target point appear sequentially. Participants are instructed to click the appropriate points. The size and position of the starting point remain constant. The size and position of the target vary.

Based on the results from pre-tests, circular target sizes from d=10 pixel (2.4 cm) to d=50 pixel (12 mm) are used. The distance between the starting and the target point is varied, so that according IDs vary from 2.0 to 5.6. The step size between each ID condition is 0.2 ID. This results in 19 steps for the ID (second experimental factor).

Participants are instructed to hit the target point as accurately as possible (prioritizing of precision against input speed). Each ID-step is repeated three times. In total, 57 single start-target movements are carried out per trial. The order of movements is randomized. After each trial the actual error rate [percent] is fed back to the participants. Three trials are carried out directly following each other in one session.

During the trials, relative error rate (missing target point) and time between appearance of the

Figure 2. Dell Axim™ PDA showing map (left) and experimental pointing task (middle: start position; right: target position)



target point and hitting the touch-screen with the pen in [ms] are measured. The accuracy of the time measurement is set to 55 ms.

A single session takes about 15 minutes. At the end of each session a break of at least 5 minutes follows. During the break, the participants rate the subjective workload or task difficulty on a two-level judgment scale (Käppler, 1993). The scale has verbal and numerical descriptors, which allows a differentiation between "simple" (value 0) and "difficult" (value 100) tasks.

Statistical Analysis

The statistical analysis was performed with the statistical software package SYSTAT™ Version 11.0 (Systat, 2004).

The error rate was analyzed by a two-way analysis of variance (ANOVA) with the two independent variables walking speed and ID. A subsequent pairwise comparison with Sidak adjustment was used to identify significant differences between the two different factor levels. When a separation into subgroup was revealed, the subsequent analyses were performed separately for each subgroup. The postulated relationship between relative error rate, ID, and walking speed was specified by a final multiple regression analysis for each of the subgroups.

The second part of the analysis referred to Fitts' law. At first, separated regression analyses tested the applicability and validity of Fitts' law for each walking speed. Errors were omitted from the analyses, as proposed by Card et al. (1978). Afterwards, the individual function for input time and index of performance (1/b) were calculated for each participant. Differences were tested for significance by a one-way ANOVA (factor: walking speed) and subsequent pairwise comparison (including Sidak adjustment). In case a grouping between factor levels appeared, the subsequent analyses were performed separately for each group. For the specification of the relationship between input time, ID, and walking speed, a final multiple regression analysis was carried out.

The subjective ratings were tested for significance by a one-way ANOVA (factor: walking speed) and subsequent pairwise comparisons (including Sidak adjustment). The results were used to compare the sensitivity of subjective ratings to the sensitivity of the objective ratings mentioned.

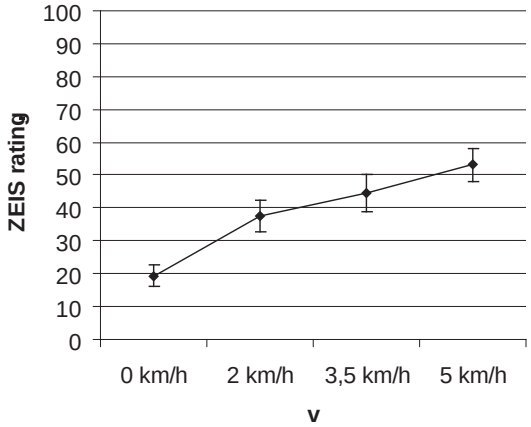
The chosen level of significance for each analysis was p=0.05.

RESULTS

Subjective Workload

The ratings for the subjective workload and task difficulty increase from 19 (standing) to 53 (5 km/h). This corresponds to an increase from "easy" to "predominantly difficult." Workload differs between standing and walking. There is only a small linear increase with increased walking

Figure 3. Mean and standard error of subjective workload (ZEIS rating) at four walking speeds (v)



speeds. In this case, the ratings are similar. The according distributions are shown in Figure 3.

The ANOVA reveals a significant effect of walking speed on the subjective rating ($F_{3,51}$=13.5; p<0.01). The factor explains o^2=41% of the observed variance.

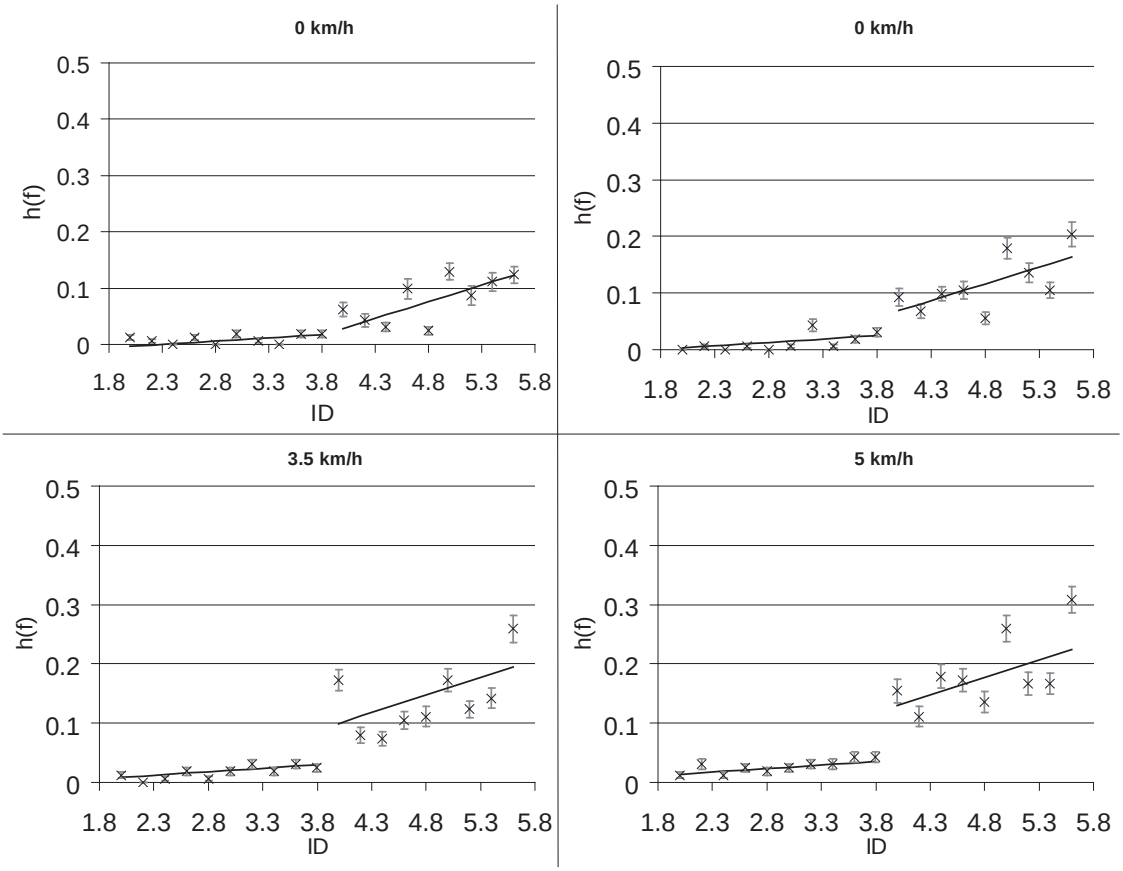
A final pairwise comparison put this result into perspective. It shows only significant differences between standing and walking (p<0.01), but not between different walking speeds. These results show that the amount of subjective workload is similar during walking. It does not increase for walking speeds between 2 km/h and 5 km/h.

Error Rate

Error rate was measured for each walking speed and each of the different IDs. Figure 4 shows the distribution of the relative error rates for each walking speed level.

For simple interaction tasks, no errors occurred. Therefore, two conditions were merged into a single group for the subsequent ANOVA. Both factors, ID ($F_{9,153}$=36.0; p<0.01) and walking speed ($F_{3,51}$=15.7; p<0.01) have a significant effect on error rate. It increases from 4.2% (standing) to 10.1% (5 km/h) on the average. The interaction of

Figure 4. Mean, standard error, and regression function (separately for ID<4 and ID≥4) of error rate r_E for four walking speeds



both factors affects error rate as well ($F_{27,459}$=2.0; p<0.01). The factor ID explains o^2=80.6% of the variance, while the factor walking speed only explains o^2=9.3%. The interaction explains o^2=4% of the variance.

However, the subsequent pairwise comparison reveals significant differences only between two different ID-sections. There are no significant differences of error rate within each section. One section includes simple movements with ID<4 and the other, more difficult movements with ID≥4. This result makes a separate analysis for each section necessary.

For movements with ID<4 the relative error rate is nearly constant. It varies between 0% and 4.3% for all IDs and walking speeds. As shown in figure 3, there is a linear effect of both factors. Both factors, ID ($F_{4,68}$=3.7; p<0.01) and walking speed ($F_{3,51}$; p<0.01) affect error rate. There is no significant interaction ($F_{12,204}$=0.5; p=0.9). Therefore, it will be omitted for the further analysis. The ID explains o^2=31.4%, and the walking speed o^2=34.8% of the variance.

The regression analysis estimates the relative error rate r_E based on ID and walking speed v. For movements with ID<4 it is r_E:

$$r_E(v, ID) = - 0.027 + 0.012\ ID + 0.003\ v \qquad (5)$$

According to this regression function, error rates between 0% and 3.6% are expected for ID<4 and v<5 km/h. This is nearly a constant relationship.

The constant behavior of the error rate changes for movements with ID≥4. As shown in Figure 4, error rates of these movements increase much stronger with growing ID and walking speed. Between ID=4 and ID=5.6, the error rate doubles or increases even more. The ANOVA reveals a significant effect of ID ($F_{4,64}$=14.5; p<0.01) and walking speed (F3,51=15.5; p<0.01) on error rate. The according values of the explained variance are o^2=46.6% for ID and o^2=39.6% for walking speed. Since the interaction is not significant ($F_{12,204}$=1.2; p=0.30) it is omitted from the further analysis.

The regression analysis for estimating the relative error rate r_E for movements with ID≥4 determines the parameters of the model function as:

$$r_E(v, ID) = -0.208 + 0.059\ ID + 0.020\ v \qquad (6)$$

This estimates absolute values for the analyzed inputs between 2.8% (ID=4; standing) and 22.2% (ID=5.6; 5 km/h). In comparison to movements with smaller IDs, the effect of walking speed is much stronger. This is because the coefficient of the walking speed increases with a factor of (0.02: 0.003 = 6.7) compared to an increase of (0.059: 0.012 = 4.9) of the coefficient of the ID.

Input Time and Fitts' Law

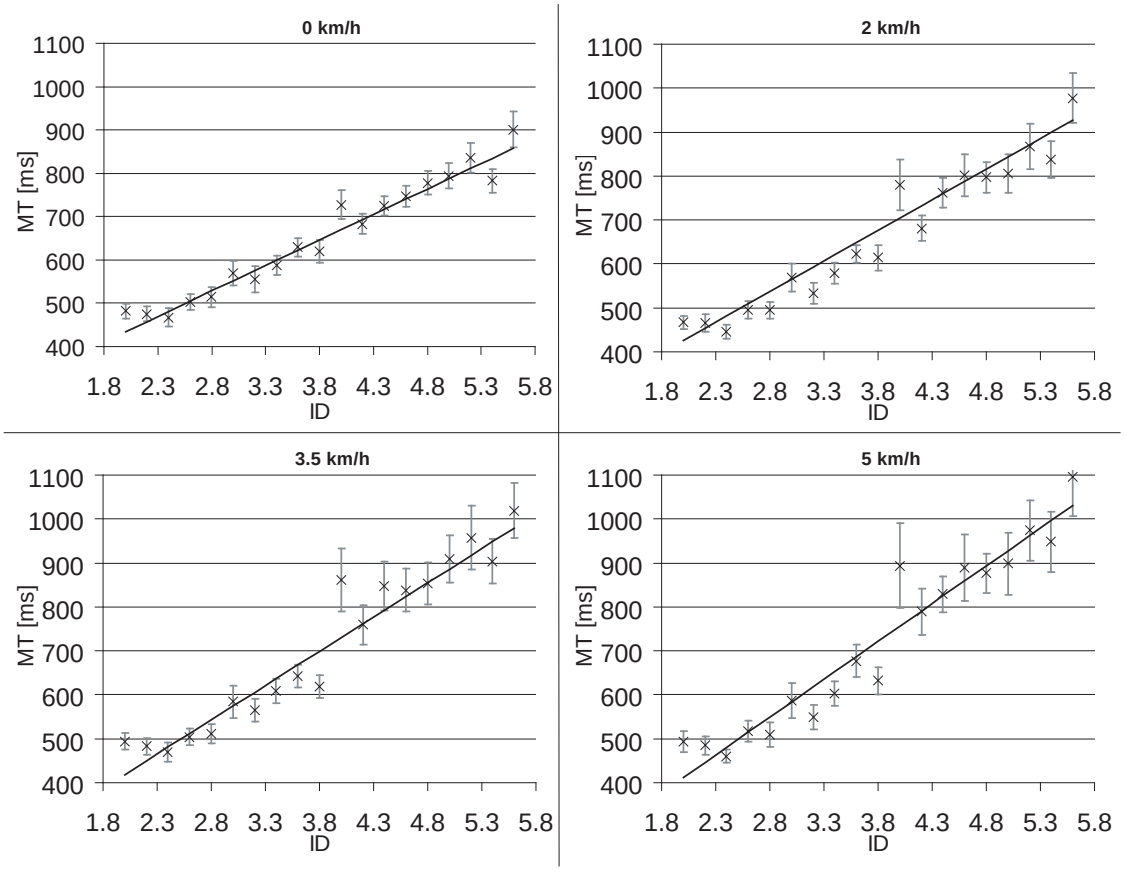
The initial correlation analysis for each walking speed condition reveals a high probability of a linear relationship between input time IT and ID (r>0.955; $r_{0.1\%}$=0,693). This confirms applicability of Fitts' law for the subsequent analysis.

Figure 5 shows the empirical values for each walking speed. Notice that there are constant input times for interaction tasks with ID < 2.6. In contrast to the general linear relationship, no or just a small effect of ID or walking speed on movement time can be observed. For higher IDs the linear relationship is valid again. Apart from this observation, no evidence for a separation into two or more groups is found, so the following analysis refers to the total continuous data.

Based on the estimated Fitts' linear coefficient b, the index of performance (1/b) can be easily determined. It varies between 8.8 bps (standing) and 6.0 bps (5 km/h). Since the goal of this study was a specification of Fitts' law, the further analysis refers to the linear coefficient b again.

The ANOVA shows a significant effect of walking speed on the linear coefficient b ($F_{3,51}$=6.7; p<0.01). It explains o^2=23.9% of the observed

Figure 5. Mean, standard error, and regression function of input time [ms] according to Fitts' Law for four walking speeds



variance. The following pairwise comparison reveals significant differences only for differences of walking speeds above 3 km/h. This can be observed throughout the total ID range. There are no evidences for differences that would require separate analyses.

The final regression analysis extends Fitts' law by additional terms, which are velocity-dependent. They specify the effect of the walking speed on movement time. Accordingly, input time IT in [ms] can be estimated by walking speed v and ID as:

$$IT(v, ID) = (198.4 - 26.2\,v) + (117.8 + 10.9\,v)\,ID \quad (7)$$

The according index of performance (1/b) for information input while standing is 8.5 bps. It decreases to 5.8 bps at a walking speed of 5 km/h.

The model function allows a comparison of the absolute effect between walking speed and ID. The effect of ID on input time is slightly stronger than the effect of walking speed. For a higher walking speed (v = 5 km/h) the according linear coefficient

(b_2 = 10.9 x 5 = 54.5) reaches half the value of the velocity-independent coefficient (117.8). Although ID remains the dominant term, walking speed still has a large effect on input time. For high walking speed, input time increases to 1.5 times the initial input time when standing. This must be considered when designing HCI with more complex and difficult user input.

DISCUSSION

The analysis supports the use of objective variables to analyze the effect of motion status and walking speed on input performance. It is possible to quantify the effect of walking speed on mobile HCI performance by relative error rate and by applying Fitts' law for visually controlled movements.

In general, error rate increases with increased walking speed, as already observed by Barnard et al. (2005) and Beck et al. (2002). The experiment described in the previous section confirms the relationship and revealed a division of this relationship in two parts. For simple interaction tasks (ID<4), error rate remains practically constant and is affected little by ID and walking speed. The results show that both factors have a similar effect. However, error rate is lower than 4% throughout the range of both experimental factors.

The behavior for interaction tasks with ID≥4 is different. For these tasks, both factors have a stronger effect. This results in an increased error rate with up to 22%. The results show that walking speed has an effect similar to that of ID. This finding is important for practical GUI design since participants were explicitly instructed to focus on accuracy. In real applications, this is not likely. Consequently, even higher error rates can be expected.

A possible explanation for this separation into two parts is based on the division of movements into a ballistic and a visually controlled part. According to this, the main characteristic of a movement changes between ID=3 and ID=4 (Wallace & Newell, 1983; Gan & Hoffmann, 1988). Lower IDs are characteristically ballistic movements; higher IDs are visually controlled. A visually controlled movement requires more perceptual resources than a ballistic movement. Therefore, the parallel processing of a walking task leads to reduced PDA-input performance and higher error rates.

With regard to Fitts' law, there is no statistically significant division between the two parts, although there are some hints from the observation. Instead, a single, continuous function can be applied. This applicability of Fitts' law for describing interaction performance is similar to the results of Lin et al. (2005) with a simulated PDA, though it is in contrast to the findings of Cerney et al. (2004), who have not found a linear relationship between movement time and ID. An explanation for this might be Cerney's simple targeting tasks with ID<3. As shown in the previous section, there was also a constant error rate observed for ID<2.6. The constant behavior did not continue for more difficult movements with higher ID, though.

For the total experimental range of IDs between 2.0 and 5.6, Fitts' law was found to be applicable. Indices of performance could be determined based on the specification and the calculation of the constant and linear coefficient. The value of 8.5 bps for the standing condition is comparable to the value of 10.4 bps for computer mouse interaction (Card et al., 1978). It is higher than the values given in other studies for various interaction devices (compare MacKenzie, 1995). A comparison of the values with the empirical values of Lin et al. (2005) for a simulated PDA shows close similarity as well. In this study, indices of performance between 8.8 bps (standing) and 6.0 bps (5 km/h) were determined. Lin et al. (2005) observed indices of 7.8 bps (sitting), 6.5 bps (slow walking), and 6.6 bps (fast walking).

Referring to the subjective ratings, the participants were able to differentiate between "easy" (standing) and "predominantly difficult" (walking). Barnard et al. (2005) and Kjeldskov & Stage (2004) made a similar observation. Mustonen et al. (2004) found subjective ratings to be more sensitive than objective performance measures. In this study it was found that subjective ratings only allow a differentiation between motion statuses, but not between different walking speeds. There are two possible explanations for this. First, it could be that subjective ratings are not sensitive enough to detect changes. Second, subjective workload does not change between different walking speeds. However, both explanations postulate that the changes of workload caused by walking speed are relatively small.

The results support the initial hypothesis of an influence of walking speed on input performance. It is also shown that this effect is separated into two, possibly three sections with different characteristics. In the first section (ID<2.6), error rate and movement speed are nearly constant. They are independent from walking speed and ID. With increased ID (2.6≤ID<4), Fitt's index of performance decreases with higher walking speeds. Error rate remains nearly constant. With increasing visually controlled movements (ID≥4), error rate increases strongly.

CONCLUSION

Using mobile IT-devices on the move, so-called mobile computing, is a challenging topic which requires novel approaches for assessing HCI performance and optimizing GUI design. Simple performance measures such as subjective ratings, error rate, or task-dependent time measures can be used for rough qualitative estimations, but they are often not sufficient for quantitative analyses of greater detail.

It was found that Fitts' law can be successfully applied for this. It is possible to estimate input performance and input time for easy and difficult movements and to differentiate even between small effects. Fitts' law facilitates modeling input performance (e.g., input time) based on target size and movement distance. In contrast to this, simple performance measures such as error rate, are less sensitive. They just change for more difficult input movements with ID>4. Subjective ratings for workload allow only a rough distinction between standing and moving. These results argue in favor of applying Fitts' law and derived objective measures as sensitive variables for mobile GUI design.

A simple way to optimize HCI performance is to stop and stand. Attention and information processing resources shift accordingly and enhance performance. This can frequently be observed in reality when people stop walking as soon as a task gets more complicated and they have to shift information processing resources towards HCI. However, this behavior severely constrains working on the move.

From a designer's perspective, the GUI of a device has to fulfill two important requirements for mobile computing. First, the target areas (icons, menus) must be large enough so that the targeting movement is easy. It is recommended that the index of difficulty of the required movements should remain below ID=4. As a result, high error rates can be avoided and user movements affect general input performance only marginally. Second, an error-tolerant behavior of the software is required because of the increased error rates with increasing walking speed. In the present experiment, error rates reached up to 22%, which hampers practical usability. It must be considered that input accuracy is usually prioritized in laboratory experiments. In reality, environmental stimuli are likely to significantly disturb the user.

The threshold value for ID defines the minimum size of target areas as icons or menus items. For pointing and selecting tasks it can be used to define "snap" areas. In contrast to current GUIs for PDAs which require a precise hitting of the

target, the closest target within the snap area of the pen position is assigned.

This shows that a simple adaptation of desktop paradigms as GUIs must be considered insufficient for mobile computing. Moreover, it is necessary to develop and specify new paradigms which take the mobile environment and its special characteristics into consideration. Otherwise, high input performance requires stopping and standing. This inhibits real mobility and strongly contradicts the need for mobility in our society. General relationships from information processing can be applied as quantitative measures for defining prospective ergonomic design criteria and for evaluating new GUIs. Thereby, effective and efficient mobile computing on the move is facilitated.

FUTURE TRENDS

As described in the introduction of this chapter, mobility and the support of mobile working will gain importance in future. It will affect mobile computers and put high demands on mobile HCI. GUIs are required that take the characteristics of both mobility and the user's self-movement, into account. This results in the consideration of many aspects, ranging from the devices' weights and the dynamics of self-movement (walking) to sharing information processing resources. A great deal of research is still required in order to optimize GUIs for maximum human performance. There is a need for such studies since mobile devices such as PDAs or smartphones are broadly available and already used.

Furthermore, valid and accurate measures of human performance are required for prospective ergonomic GUI design and retrospective evaluation of implementations. This is not limited to specific measures, but refers to the general experimentation method, whether it is in laboratory studies or field experiments. The experiment described in this chapter was performed in a laboratory setting with a treadmill. Walking speed was controllable and additional environmental stimuli were reduced to a minimum. In contrast to such pre-defined settings, individual walking speeds will have to be analyzed in field experiments in the future. Since the method of applying measures such as subjective workload, error rate, and input time has been found to be successful and sensitive, a similar method may be applicable for outdoor and field experiments. By comparing the respective performance measures it will be possible to estimate and quantify the effect of the experimentation method (treadmill vs. field experiment).

More research is needed to specify similar requirements for HCI and GUI design. The experiment described in this chapter can be considered as an example of such a study. It gives detailed recommendations for the size of target areas that affect icon and menu item size. Another relevant aspect is visualization, because GUIs serve to visualize information as well. Daily experience shows that visual perception and cognitive processing is not independent from other environmental stimuli. Movement is likely to affect visual perception performance as well, and further empirical analyses are needed to specify the interdependence. Thereby, it will be possible, for instance, to define minimum icon and font sizes for mobile computing, independently from mobility status and walking speed. This is a more elegant way than just maximizing sizes or having the users adjust them manually (which is practically limited to changing font size).

In general, mobile HCI-analyses have to take characteristics of portability, that is, computing at different locations, and mobility, that is, computing while on the move, into account. By applying objective measures of performance, such as error rate or Fitts' law, it is possible to define ergonomic requirements for GUI design. Moreover, it is possible to evaluate different designs and optimize them.

REFERENCES

Alton, F., Baldey, L., Caplan, S., & Morrisey, M. C. (1998). A kinematic comparison of overground and treadmill walking. *Clinical Biomechanics, 13*(6), 434-440.

Baber, C., Knight, J., Haniff, D., & Cooper, L. (1999). Ergonomics of wearable computers. *Mobile Networks and Applications, 4*(1), 15–21.

Barnard, L., Yi, J. S., Jacko, J. A., & Sears, A. (2005). An empirical comparison of use-in-motion evaluation scenarios for mobile computing devices. *International Journal of Human-Computer Studies, 62*, 487-520.

Beck, E. T., Christiansen, M. K., & Kolbe N. (2002). *Metoder til Brugbarhedstest af Mobile Apparaten*. Aalborg University: Department of Computer Science (in Danish).

Bertelsen, O. W., & Nielsen, V. (2000). Augmented reality as a design tool for mobile interfaces. In *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (pp. 185–192).

Brewster, S. A. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing, 6*(3), 188–205.

Business Week. (2006, March 2). *Why Google's going mobile*.

Card, S. K., English, W. K., & Burr, B. J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. *Ergonomics, 21*, 601-613.

Cerney, M. M., Mila, B. D., & Hill, L. C. (2004). Comparison of mobile text entry methods. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting.* Santa Ana: HFES.

Crowley, J. L., Coutaz, J., & Blerard, F. (2000). Perceptual user interfaces: things that see. *Communications of the ACM, 43*(3), 54–64.

Dahm, M., Felken, C., Klein-Bösing, M., Rompel, G., & Stroick, R. (2004). Handyergo: Breite Untersuchung über die Gebrauchstauglichkeit von Handys. In R. Keil-Slawik, H. Selke, & G. Szwillus (Eds.), *Mensch & Computer 2004: Allgegenwärtige Interaktion* (pp. 75–84). München: Oldenbourg Verlag (in German).

Danesh, A, Inkpen, K., Lau, F., Shu, K., & Booth, K. (2001). Geney: Designing a collaborative activity for the palm handheld computer. In *Proceedings of CHI 2001* (pp. 388-395). New York: ACM.

Dunlop, M., & Brewster, S. (2002). The challenge of mobile devices for human-computer interaction. *Personal and Ubiquitous Computing, 6*, 235-236.

Ebersbach, G., Dimitrijevic, M. R., & Poewe, W. (1995). Influence of concurrent tasks on gait: A dual task approach. *Perceptual and Motor Skills, 81*, 107-113.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47*, 381-391.

Fitts, P. M., & Peterson, J. R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology, 67*, 103-112.

Gan, K. C., & Hoffman, E. R. (1988). Geometrical conditions for ballistic and visually controlled movements. *Ergonomics, 31*, 829-839.

Goth, G. (1999). Mobile devices present integration challenges. *IT Pro*, 11–15.

Hinckley, K., Pierce, J., Sinclair, M., & Horvitz, E. (2000). Sensing techniques for mobile interaction.

In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (Vol. *2*(2), 91–100.

Initiative D21. (2006). *Press release of the D21-innovation workshop on mobile society*. Retrieved January 26, 2006, from http://www.initiatived21.de

Jacobsen, C. S., Langvad, P., Sorensen, J. J., & Thomsen H. B. (2002). *Udvikling og Vurdering af Brugbarhedstest til Mobile Apparater*. Aalborg University: Department of Computer Science (in Danish).

Johnson, P. (1998). Usability and mobility: interactions on the move. In *Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices.*

Käppler, W. D. (1993). *Beitrag zur Vorhersage von Einschätzungen des Fahrerverhaltens.* Fortschritt Berichte VDI-Reihe 12, Nr. 198. Düsseldorf: VDI-Verlag (in German).

Kjeldskov, J., & Graham, C. (2003). A review of mobile HCI research methods. In *Proceedings of the 5th International Mobile HCI 2003 conference*, Udine, Italy. Lecture notes in computer science. Berlin: Springer.

Kjeldskov, J., Skov, M. B., Als, B. S., & Hoegh, R. T. (2004). Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In *Proceedings of the 6th International Mobile HCI Conference*, Glasgow.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal Human-Computer Studies*, 599-620.

Kristoffersen, S., & Ljungberg, F. (1999). Making place to make IT Work: empirical explorations of HCI for mobile CSCW. In *Proceedings of the International Conference on Supporting Group Work (GROUP '99)* (pp. 276–285).

Lin, M., Price, K. J., Goldman, R., Sears, A., & Jacko, J. (2005). Tapping on the move—Fitts' law under mobile conditions. In *Proceedings of Information Resources Management Association International Conference* (pp. 132-135). San Diego, CA.

Lumsden, J., & Brewster, S. (2003). A paradigm shift: alternative interaction techniques for use with mobile & wearable devices. In *Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative research (CASCON'03)* (pp. 197-210). Toronto, Canada.

MacKenzie, I. S. (1995). Movement time prediction in human-computer interfaces. In R. M. Baecker, W. A. S. Buxton, J. Grudin, & S. Greenberg (Eds.), *Readings in human-computer interaction* (2nd ed.) (pp. 483-493). Los Altos, CA: Kaufmann.

Microsoft. (2005). Press release: *Go mobile*. Retrieved March 05, 2006, from http://msdn.microsoft.com/windowsvista/mobile/

Murray, M. P., Spurr, G. B., Sepic, S. B., Gardner, G. M., & Mollinger, L. A. (1985). Treadmill vs. floor walking: Kinematics, electromyogram, and heart rate. *Journal of Applied Physiology, 59*, 87-91.

Mustonen, T., Olkkonen, M., & Hakkinen, J. (2004). Examining mobile phone text legibility while walking. In *Proceedings of CHI 2004* (pp. 1243-1246). Vienna: ACM Press.

Navon, D., & Gopher, D. (1979). On the exonomy of the human processing system. *Psychological Review, 86*, 214-255.

Nigg, N. M., DeBoer, R., & Fisher, V. (1995). A kinematic comparison of overground and treadmill running. *Medicine and Science in Sports and Exercise, 27*, 98-105.

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Trans. AIEE, 47,* 617-644.

O'Donnell & Eggemeier. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance: Cognitive processes and performance*, vol. 2. New York: Wiley.

Oppermann, R. (2003). Ein Nomadischer Museumsführer aus Sicht der Benutzer. In G. Szwillus, J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 31-42). Stuttgart: B. G. Teubner (in German).

Oulasvirta, A. (2005). The fragmentation of attention in mobile interaction, and what to do with it. *Interaction, 12*(6), 16-18.

Pascoe, J., Ryan, N., & Morse, D. (2000). Using while moving: HCI issues in fieldwork environments. *ACM Transactions on Computer–Human Interaction (TOCHI), 7*(3), 417–437.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: understanding access anytime, anywhere. *ACM Transactions on Computing–Human Interaction (TOCHI), 8*(4), 323–347.

Petrie, H., Johnson, V., Furner, S., & Strothotte, T. (1998). Design lifecycles and wearable computers for users with disabilities. In *Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices*, GIST Technical Report G98-1, Glasgow.

Pirhonen, A., Brewster, S. A., & Holguin, C. (2002). Gestural and audio metaphors as a means of control for mobile devices. In *Proceedings of CHI' 2002*. New York*:* ACM.

Rantanen, J., Impio, J., Karinsalo, T., Reho, A., Tasanen, M., & Vanhala, J. (2002). Smart clothing prototype for the Artic environment. *Personal and Ubiquitous Computing, 6*, 3-16.

Sawhney, N., & Schmandt, C. (2000). Nomadic radio: speech and audio interaction for contextual messaging in normadic environments. *Transactions on Computer-Human Interaction, 7*(3), 353- 383.

Schache, A. G., Blanch, P. D., Rath, D. A., Wrigley, T. V., Starr, R., & Bennell, K. L. (2001). A comparison of overground and treadmill running for measuring the three-dimensional kinematics of the lumbo-pelvic-hip complex. *Clinical Biomechanics, 16*(8), 667-680.

Sears, A., Lin, M., Jacko, J., & Xias, Y. (2003). When computers fade … pervasive computing and situationally induced impairments and disabilities. In C. Stephanidis & J. Jacko (Eds.), *Human-computer interaction: Theory and practice* (Part II) (pp. 1298-1302). London: Lawrence Erlbaum Associates.

SYSTAT. (2004). *SYSTAT Version 11.* SYSTAT Software Inc. Richmond, CA.

Thomas, B., Grimmer, K., Zucco, J., & Milanese, S. (2002). Where does the mouse go? An investigation into the placement of a body-attached TouchPad mouse for wearable computers. *Personal and Ubiquitous Computing, 6*, 97-112.

Vogt, L., Pfeifer, K., & Banzer, W. (2002). Comparison of angular lumbar spine and pelvis kinematics during treadmill and overground locomotion. *Clinical Biomechanics, 17*(2), 162-165.

Wallace, S. A., & Newell, K. M. (1983). Visual control of discrete aiming movements. *Quarterly Journal of Experimental Psychology, 35a*, 311-321.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of Attention* (pp. 483-493). New York: Academic Press.

York, J., & Pendharkar, P. C. (2004). Human-computer interaction issues for mobile computing in

a variable work context. *International Journal of Human-Computer Studies, 60*, 771-797.

Ziefle, M. (2002). The influence of user expertise and phone complexity on perfomance, ease of use and learnability. *Behaviour & Information Technology, 21*(5), 303–311.

KEY TERMS

Fitts' Law: Based on Shannon's theorem for information processing, *Fitts' law* allows the estimation of required times for rapid movements between a starting point and a target area. Fitts introduced the index of difficulty (ID) as a characteristic measure for such movements. The ID is defined as:

$ID = \log_2 (2 A / W)$.

The term A describes the amplitude of the movement, or the distance between starting point, and target area, and the term W, the width of the target area. The movement time MT is linearly dependent on the ID. It is:

$MT = a + b\ ID$.

The coefficients *a* and *b* are both regression coefficients. The coefficient [a] defines the intercept for ID=0 and the coefficient b the steepness of the relationship.

Index of Performance: The *index of performance* is a measure for characterizing the speed of a visually controlled movement. Thus, it is also a measure of movement performance. It is calculated by the reciprocal value of the linear coefficient b of Fitts' law. Another definition is the ratio of $ID_{average}$ to $MT_{average}$. Both definitions allow a comparison of input performance under different movement conditions.

Mobile Computing: In contrast to stationary computing, which comprises stationary working with a computer at the same location, and portable computing, which refers to stationary working with a computer at different locations, the term *mobile computing* describes working with a computer while moving. This leads to weight and size requirements and special demands for human-computer interaction. The user interface has to consider the characteristics of mobility, for example, disturbing, random external forces because of the movement, parallel processing of orientation tasks, and so forth, for an optimal user input performance.

Subjective Workload: There are many definitions of the term *Workload*. One concise definition was given by O'Donnell & Eggemeier (1986): "… Workload refers to that portion of the operator's limited capacity actually required to perform a particular task." Subjective workload describes the effort invested by human operators into task performance. It can be assessed by subjective ratings.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 830-846, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.20
Evaluating Mobile Human–Computer Interaction

Chris Baber
The University of Birmingham, UK

ABSTRACT

In this chapter the evaluation of human computer interaction (HCI) with mobile technologies is considered. The ISO 9241 notion of 'context of use' helps to define evaluation in terms of the 'fitness-for-purpose' of a given device to perform given tasks by given users in given environments. It is suggested that conventional notions of usability can be useful for considering some aspects of the design of displays and interaction devices, but that additional approaches are needed to fully understand the use of mobile technologies. These additional approaches involve dual-task studies in which the device is used whilst performing some other activity, and subjective evaluation on the impact of the technology on the person.

INTRODUCTION

This chapter assumes that 'usability' is not a feature of a product, that is, it does not make sense to call a product itself 'usable'. Rather, usability is the consequence of a given user employing a given product to perform a given activity in a given environment. Holcomb and Tharp (1991) proposed a 'model' of interface usability, which is illustrated by Table 1. The definitions presented in Table 1 arose from consideration of the user interface of desk-based computers. However, it ought to be apparent that the majority of the components are defined in terms of an individual's perceptions of features of the user interface.

The International Standards Organization has a number of standards relevant to human-computer interaction (Bevan, 2001). Current standards for mobile devices tend to focus on product attributes, for example, *ISO 18021: Information Technology —User Interface for Mobiles* (2001) provides interface specifications for Personal Digital Assistants. Other Standards have recognized the multifaceted nature of usability and have sought to encourage an approach that is similar to Quality Assessment (Earthey et al., 2001). Demonstrating compliance with the standards requires analysts to document their evaluation, demonstrating how it meets the objectives of the standard. The

Table 1. Holcomb and Tharp's (1991) "model" of interface usability

Component	Term
Functional	Able to accomplish tasks for which software is intended Perform tasks reliably and without errors
Consistent	Consistent key definitions Show similar information at same place on screens Uniform command syntax
Natural and Intuitive	Learnable through natural conceptual model Familiar terms and natural language
Minimal memorization	Provide status information Don't require information entered once to be re-entered Provide lists of choices and allow picking from the lists Provide default values for input fields
Feedback	Prompt before destructive operations like DELETE Show icons and other visual indicators Immediate problem and error notification Messages that provide specific instructions for action
User help	Online help system available Informative, written documentation
User control	Ability to undo results of prior commands Ability to re-order or cancel tasks Allow operating system actions to be performed within the interface

definition of usability offered by the International Standards Organization, that is, in ISO9241, part 11, is, *"... the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."* (ISO9241-11, 1998). The implications are that, first, usability is the consequence of a given user employing a given product to perform a given activity in a given environment (as stated) and, second, that it is possible to measure aspects of this relationship in terms of effectiveness, efficiency, and user satisfaction. It is important to note that these three aspects are inter-connected and that any evaluation activity ought to try to measure some aspect of each (Frøkjær et al., 2000).

Defining Evaluation Targets

If one is able to speak of measures, then it makes sense to be able to determine some criteria that indicate good or poor performance on these measures. Good et al. (1986) proposed that it is important to define both evaluation targets and metrics that relate to these targets. For example, in a study of conferencing systems, Whiteside et al. (1988) identified 10 attributes that they felt reflected the use of the conferencing system, for example, ranging from a fear of feeling foolish to a number of errors made during task performance. For each attribute, Whiteside et al. (1988) defined a method for collecting data about that attribute, for example, questionnaires, observation, and so

Table 2. Defining evaluation targets

Factors	Method	Metrics	Worst	Target	Best	Current
Performance						
Task	CPA	Time	-15%	0	+5%	-2%
Practice	User trials	Time	-15%	0	+5%	-10%
	1st vs. 3rd trial	% change	1st > 3rd	3rd > 1st	0	3rd>1st
Subjective	SUS[1]	Scale: 0-100	50	60	70	65
evaluation	SUMI[2]	Scale: 0-100	50	60	70	60
	Heuristics	Scale: 0-10	<6	6	>6	8

forth, and then set performance limits relating to best, worst, and planned levels. A study of a wearable computer for paramedics (Baber et al., 1999) used this concept to produce Table 2. In Table 2, three measures of performance were undertaken, that is, predictive modeling (using critical path analysis), user trials, and performance improvement arising from practice. In addition, three subjective evaluation methods were used. Table 2 shows how the system met (or exceeded) some of the target criteria but fell below the target for time (although it is not within the 'worst' case range). One benefit of such a technique is to allow the design team to decide whether there is a need for more effort to refine the device, or whether, having met (some or all of) the requirements, the design process can be closed.

The ISO9241 notion of usability requires the concept of evaluation targets, for example, one could begin with a target of "66% of the specified users would be able to use the 10 main functions of product X after a 30 minute introduction." Once this target has been met, the design team might want to increase one of the variables, for example, 85% of the specified users, or 20 main functions, or 15 minute introduction, or might want to sign-off that target.

Why Conduct Evaluation?

The concept of usability that is used in this chapter (and in ISO9241) implies that changing any one of the variables {user, activity, device, environment} can have an impact on usability. This implication points to the well-known assertion that an activity that a designer of the product might find easy to perform could prove problematic for a user who has had little or no previous experience of the product. It also points to potential issues relating to the usability of mobile technology, particularly through consideration of the environment. If we think about sending a text-message from a handheld device, such a mobile telephone or a Blackberry™, the activities involved could be somewhat different while sitting on a train versus walking down a busy street. This change in environmental setting will have a marked effect on usability of the device. This does not necessarily result from the design of the device itself but rather from the interactions between design, use, and environment. As Johnson (1998) pointed out, *"HCI methods, models and techniques will need to be reconsidered if they are to address the concerns of interaction on the move."* (Johnson, 1998). The question for this chapter, therefore,

is how best to address the relationship between user, activity, product, and environment in order to evaluate the usability of mobile technology. Related to this question is how evaluation might capture and measure this relationship, and then what can designers do to improve usability. This latter point is particularly problematic if one assumes that design is about creating a product rather than about creating an interaction.

Before considering these questions, it is worth rehearsing why one might wish to conduct evaluation. Baber (2005) notes that the primary reason for conducting evaluation, in HCI, is to influence design (ideally, to *improve* the product). This implies that evaluation ought never to be a one-off activity to be conducted at the end of the design lifecycle in order to allow a design to be signed-off (Gould & Lewis, 1985; Johnson, 1992). Rather, it means the following:

1. Evaluation is a recursive activity that cuts across the entire design lifecycle, for example, software engineers will run versions of the code to debug and check; product designers will continually critique and refine their concepts. What is not always apparent is the manner in which these processes could (or indeed ought) be made formal and to result in something that can be communicated to other members of the design team.
2. Evaluation should be incorporated into as many stages of design as possible—this points to (i) but also raises that questions of recording and communicating the results of evaluation in a many that can be beneficial to the design process.
3. Evaluation should be designed to maximize the impact of the evaluation of the design stage in which it is used—the suggestion is that, rather engaging in evaluation as a mandated exercise to allow sign-off between stages, it ought to be an activity that positively advances the design process.
4. Evaluation should guide and inform design activity—the results of any evaluation should be reported in a manner that can lead to change in the design and can be reported in a manner that is transparent and reliable.

A final point to note is that evaluation is a process of comparing the product against something else, for example, other products, design targets, requirements, standards. Thus, evaluation requires a referent model (Baber, 2005). It is naïve to believe that one can "evaluate" something in a vacuum, that is, to think that one can take a single product and "evaluate" it only in terms of itself. In many ways this is akin the concept of a control condition in experimental design; one might be able to measure performance, but without knowing what would constitute a baseline for the measure, it is not possible to determine whether it is good or bad.

Defining Referent Models

While it might be fairly clear as to why comparison requires a referent model, there is a problem for novel technologies. After all, the point of these technologies is to move beyond the conventional desk-bound personal computers and this will ultimately create new forms of interaction. However, the move to different technologies makes it hard to establish a sensible basis for evaluation. What is the referent model for mobile HCI?

A common form of mobile technology is the digital tour-guide that, knowing where the user is (using Global Positioning Satellite (GPS) to determine location) and what the user is doing, can provide up-to-the-minute information to help the user. There are few, if any, products that are like these concepts, so what constitutes a referent? At one level, this is simply because future HCI is attempting to develop approaches to interaction with technology for which there are no existing models. The answer to this question, the author

suggests, comes from the assertion at the start of this chapter: usability is not a characteristic of the product, but the result of the interactions between user, product, activity, and environment. If we assume that tourists have a variety of strategies and artifacts that they currently use to find out where they are or to find out interesting information about a particular location, for example, maps, books, leaflets, other people. One could ground an initial evaluation of using the product to perform a given set of activities in comparison with existing practices. Conducting evaluation against other products in terms of a set of activities offers the analyst the following benefits:

1. The evaluation will cover a range of functions on the products. It is important to ensure that the comparison provides a fair and accurate view of the product. After all, it is not really the point of evaluation to just demonstrate the product X is better than product Y—partly because there are bound to be occasions when products X and Y are similar, or where product Y is better than product X, and partly because simply knowing that X > Y tells us very little about how to improve X (or Y) or why X is superior.
2. The focus of the evaluation is less on product functioning than on user activity. This might appear, at first glance, to be tautological—surely product evaluation is about evaluating the product? This is, of course, true in a technical sense. However, HCI is about human-computer *interaction*, and the defining feature of this relationship is the interaction (rather than either human or computer). If one is concerned with technical evaluation then, perhaps some of the features to be included in a comparison table (like the one shown in Table 2) would be some of the technical features, for example, processor speed, RAM, memory, and so forth.
3. As the evaluation is concerned with user activity (as opposed to product functioning), the type of metrics that could be applied may well change. When comparing user activity on two or more products, it is important to decide what information is really being sought. Do we want to know only that X>Y? Or do we want to know that using product X or Y have differing effects on user activity?

In the field of mobile and wearable computers, much of the evaluation research has focused on comparing performance on a wearable computer with performance using other media. Thus, studies might compare performance using a wearable computer, say to perform a task that involves following instructions, and find that sometimes performance is superior in the paper condition (Siegel & Bauer, 1997; Baber et al., 1999) and sometimes it is superior in the wearable computer condition (Bass et al., 1995, 1997; Baber et al., 1998). This highlights the potential problem of comparing disparate technologies in an evaluation; it is not clear that any differences in performance are due to the experiment favoring one technology over another or whether there are other factors at play here. For example, a common observation is that people using the wearable computer tend to follow the instructions laid out on the display, whereas people using paper tend to adopt a more flexible approach (Siegel & Bauer, 1997; Baber et al., 1999). The notion that technology influences the ways in which people work is often taken as 'commonsense' by Human Factors engineers. However, the question of how and why such changes arise ought to have a far deeper impact on evaluation than is currently the case. As mentioned earlier, one way to deal with this problem is to focus on activities that people are performing using a variety of products. However, this will only cope with part of the problem. For instance, the electronic tour-guide given could be evaluated in comparison with other ways of performing activities, but this does not tell us whether any differences between the electronic tour-guide and the other products

are due to the concept or to the realization of the concept or to changes in the activity arising from the use of the device. In other words, if we find that the electronic tour-guide performs less well than speaking to someone, is this because the tour-guide lacks information, or because it lacks clear presentation of information, or because it lacks speedy access to the information, or because it lacks flexibility of response, or because of some other reason (the evaluation could point to all of these, not to specific reasons).

At one level, the evaluation of mobile HCI calls for the application of current evaluation techniques. However, there are other aspects of future HCI that call for rethinking of evaluation. In other words, it might not be entirely appropriate to take methods that have proven useful for evaluating desktop HCI and apply these to future HCI. As Wilson and Nicholls (2002) point out in discussing the evaluation of virtual environments:

There are only a limited number of ways in which we can assess people's performance...We can measure the outcome of what they have done, we can observe them doing it, we can measure the effects on them of doing it or we can ask them about either the behavior or its consequences. (Wilson & Nicholls, 2002)

The underlying assumption here is that human behavior is measurable in a finite number of ways. Combining this assertion with the need to study the relationship between user, activity, device, and environment, it becomes apparent that evaluation of user interaction with mobile activity can be reduced to a small number of requirements. Furthermore, the ISO 9241 notions of efficiency, effectiveness, and satisfaction point to the approaches outlined by Wilson and Nicholls (2002). For example, efficiency could be considered in terms of the amount of resource expended in order to achieve a goal (perhaps in terms of time to complete a task), and effectiveness could relate to the quality of this performance (perhaps in terms of the amount of activity completed or the quality of the outcome), and satisfaction would relate to the user perception of the activity (perhaps in terms of a judgment relating to their own performance or effort, perhaps relating to some aspect of using the device). What is required is not so much a battery of new measures, so much as an adaptation of existing approaches that pay particular attention to the relatively novel aspects of the environment and activity that pertain to mobile devices.

MAKING SENSE OF HUMAN ACTIVITY WITH MOBILE TECHNOLOGY

The argument so far is that what needs to be evaluated is not simply the product, but the interaction between user, activity, device, and environment. This raises the question of what can be defined as appropriate forms of activity. The first issue for mobile technology is the assumption that it is to be used on the move, which raises two possibilities: (1) 'on the move' means physically moving, for example, walking, driving a car, traveling as a passenger; (2) 'on the move' means being in different places away from 'normal' office environments. One problem relating to both of these possibilities is the difficulty of collecting data in the field—there are problems arising from recording the data, managing the collection of data, and controlling experimental conditions that are far from trivial. However, if evaluation studies involve managing the interactions between user, activity, device, and environment, then it might not be possible to concentrate efforts on specific aspects of the interactions, for example, comparing the use of the device under different mobility conditions.

Interacting with Mobile Technology while Walking

Consideration of interaction while moving immediately suggests that asking people to evaluate a product whilst sitting down in a laboratory might lead to different results than when using the product 'on the move'. This is just what Kjeldskov and Stage (2004) demonstrated. Indeed, they found that having participants report usability problems while sitting down in the laboratory led to more usability problems being reported than when the participants performed the evaluation while walking. They suggested that this result might have arisen from different demands on attention—in the seated condition there was little distraction from the product and so participants were able to devote most of their attention to it, but in the walking conditions, attention needed to be divided between the device and the task of walking. This effect can be compounded by variation in other contextual factors, such as lighting levels and complexity of the path that one is following (Barnard et al., 2007).

It has been demonstrated that walking can impact cognitive tasks (Ebersbach et al., 1995), and so the use of a mobile device could be thought of in terms of a 'dual-task'. A common methodological approach in Ergonomics/Human Factors involves asking participants to perform one task while attending to another, for example, tracking a line on a screen while also performing mental arithmetic. There are several reasons why this approach is useful, both in terms of developing theory of human performance and in terms of considering how combinations of tasks can be modified. In broad terms, the assumption is that the human ability to process information from several sources can be compromised under conditions of increasing complexity. Complexity might arise from the difficulty of one or both of the tasks, from the quality of the signals being attended to, from the amount of interference between the two tasks, and so forth. By measuring performance on the tasks under different levels of complexity, it is possible to judge the person's ability to perform and the amount of interference that could occur.

Taking the dual-task paradigm as a starting point, one can consider many forms of mobile technology to be used not only in different places but also while the person is physically moving, for example, walking down a busy street or following a predefined route or walking on a treadmill. Thus, one approach to studying mobile technology from a dual-task perspective would involve measuring some aspect of walking and some aspect of using the technology. Barnard et al. (2005) compared reading tasks on a Personal Digital Assistant (PDA) while walking on a treadmill and walking along a defined path. They found a reduction in walking speed (by around 33%) compared to walking without performing the tasks on the device. This indicates that using the device leads to measurable changes in walking activity. They found no difference in comprehension between conditions (although it is often difficult to find measurable differences in comprehension in experiments that involve reading from screens, see Dillon, 1992), but they did find that word search took significantly longer when walking along the path than on the treadmill. This result suggests that participants walking the path had more need to divide their attention between the world and the device, and indeed, path following correlated with the use of scroll bars on the device, suggesting that more attention on the world led to more need to scroll the text to find one's place while reading. What is interesting about this particular study is that it reports objective results on both primary (using the device) and secondary (walking under different conditions) tasks, and shows some interactions between the two. This shows that modifying the environment has a bearing on activity which, in turn, affects user performance (even with the same device requiring the same activity).

Using Mobile Technology while On-the-Move

Prototypical mobile technologies often address scenarios related to tourists because this emphasizes the need to move around an unfamiliar environment and the desire for information relating to the world around us, for example, routes to places, interesting information about landmarks, advice on traveling, and so forth. Considering the scenario from the perspective of usability, evaluation could allow the design team to agree on 'benchmark' levels of performance using existing practices, and to then consider what benefits might accrue from modifying those practices through the introduction of technology.

While the activity of walking can interact with the use of the mobile device, there are other aspects of use on-the-move that can also play a role. Duh et al. (2006) compared the evaluation of a mobile telephone, used to perform a set of activities, in the laboratory and on a Mass Rapid Transit (MRT) train in Singapore. The study showed that participants encountered significantly more problems in the train condition than in the laboratory. The authors relate the problem to five primary areas: ambient noise levels, movement of the train, issues relating to privacy, increase in effort needed to perform the activity, additional stress, and nervousness. Of these factors, the main ones relate to aspects of the environment, viz. noise and movement of the train, and these, in turn, have a bearing of the ability of participants to complete the activity. In addition to the affect of movement on the performance of the users, the impact on the performance of the technology is equally important, for example, what happens when wireless networks do not cover the whole of the area and the user encounters 'shadows', or what happens when positioning systems have drift or inaccuracies. One approach might be to attempt to guarantee optimal delivery of service at all times by modifying the infrastructure rather than the device, for example, with boosters located in the environment. Another approach would be to provide ways of informing the user about accuracy of the data on which the system is working (Bell et al., 2006).

Devices can also be used while driving automobiles and there is evidence that interference between the activity of driving and the activity of using the device are more than simply physical (Boase et al., 1988; Brookhuis et al., 1991; Svenson & Patten, 2005; Wikman et al., 1998). This means that using a 'hands-free' kit will not eliminate all forms of interference. For example, Nunes and Recarte (2002) show that the more cognitively demanding a telephone conversation, the greater the reduction in the user's ability to attend to the environment while driving. This research further highlights the problem of isolating the usability of the device itself from the interactions between user, activity, and environment.

SUBJECTIVE EVALUATION OF TECHNOLOGY

"[U]ltimately it is the users of a software system [or any product] who decide how easy its user interface is to manipulate…" (Holcomb & Tharp, 1991). Thus, one might feel that asking people about the product would be the obvious and most useful approach to take. However, there are several problems with this approach, for example, people might not always be able to articulate how they feel about the product (so the reports might be incomplete or inconsistent), people might use a variety of previous experiences as their referent models (so it might be difficult to generalize results across respondents), people might not be able to respond critically to the product (so there might be a 'halo-effect' with the participant responding to the novelty of the device rather than considering issues of usability). For these and other reasons, it is common practice to provide some structure to subjective evaluation, usually through some form of procedure or checklist. Furthermore, it would be suggested that subjective evaluation should be used as a secondary measure as far as

practicable, with the primary focus on data collected from user trials.

Subjective Response to the Device

Participants could be asked to walk-through the performance of a given activity using the device, by explaining what they are doing and why. Monk et al. (1986) presented a detailed set of guidelines on how to use walk-through approaches to the evaluation in their Cooperative Evaluation method. The main aim of the approach is to capture problems that users experience when using a product.

In terms of checklists, a great deal of research effort from the late 1980s to the mid 1990s led to the development of a number of usability surveys. Some, like CUSI-Computer User Satisfaction Inventory (Kirakowski & Corbett, 1988) and QUIS-Questionnaire for User Interface Satisfaction (Chin et al., 1988), are designed to capture user response to an interface, particularly in terms of affective components (such as satisfaction). Others, like the checklist of Ravden and Johnson (1989) or SUS (Brooke, 1996), have been designed to cover both aspects of the interface and characteristics of usability. While these surveys are based on sound HCI principles, interpretation is left to the analyst who could lead to potential bias or misinterpretation. The SUMI checklist (Kirakowski, 1996) was developed using a rigorous approach to defining appropriate components of usability and presents results in terms of a comparison with a database of previous evaluations.

Subjective Responses to Using the Device To Perform an Activity

In addition to eliciting opinions from users regarding the device, researchers are also keen to obtain reactions of some of the consequences of using the device. By way of analogy, if we consider the virtual reality research community, we can see efforts to elicit reaction to either the physical effects of using virtual reality, for example, Cobb et al.'s (1999) Virtual Reality Induced Symptoms and Effects (VRISE) or the measurement of 'presence'(Slater et al., 1994; Witmer & Singer, 1998). In the domain of wearable computers, physical effects have been evaluated using self-report on a comfort rating scale (Knight et al., 2002).

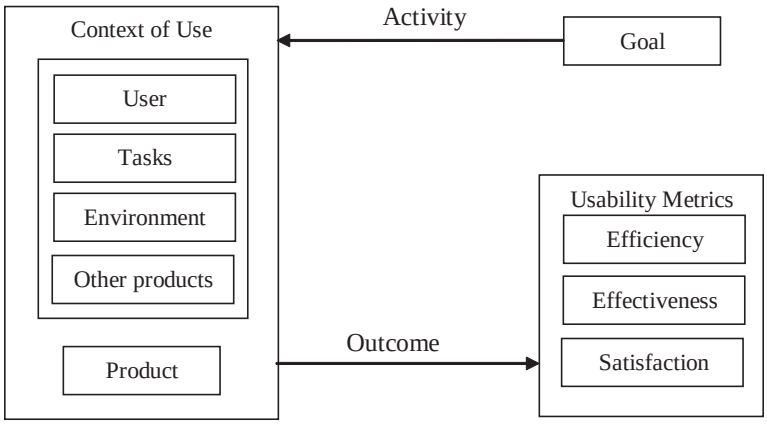
In terms of performing an activity, researchers often make use of the NASA-TLX (Hart & Staveland, 1988) which measure subjective response to workload. The basic notion is that activities make different demands on people in terms of time pressure or mental effort, and can lead to different responses such as frustration or perceived level of performance. The NASA-TLX captures these responses and can be used to compare perceptions of users with combinations of different devices or activities.

DESIGNING AN EVALUATION PROTOCOL

Throughout this chapter, emphasis has been placed on the notion of 'context of use' and the concept of usability defined by ISO 9241, pt. 11. The relationship between these concepts is illustrated by Figure 1. In order to evaluate any item of technology, one needs to plan an appropriate campaign of evaluation—this means consideration of the evaluation from the start of the design process and performance of evaluation as often as practicable during the course of development. Assume that, whatever design process is being followed, there will be four primary phases: initial concept development, prototyping, specification and build. At each phase, the form and type of evaluation will change (depending on access to functionality on the product as much as anything else), but the basic considerations remain constant, that is, adequately defining context of use and applying appropriate usability metrics.

Before elaborating on Figure 1 as a process, it is worth re-emphasizing the point made earlier that usability evaluation always involves comparison with the product being considered against some

Figure 1. ISO9241 usability evaluation process



referent model. The referent model could be other products, but is equally likely to be a set of design targets (see Table 2). In terms of comparison, a set of usability metrics can be applied. The 'efficiency' metric relates to the manner in which resources are applied during the activity in order to achieve the outcome; the 'effectiveness' metric relates to the completion of the outcome; the 'satisfaction' metric relates to the user's response to performing the activity. Needless to say, all metrics apply to a given user performing a given activity in order to achieve a given goal in a given context of use with a given product. In terms of what to measure, each metric has several options. For the sake of brevity, in this chapter, one quantitative and one qualitative measure for each metric will be considered (the reader is encouraged to review ISO 9241, pt. 11 as a starting point for considering alternatives). For 'efficiency', a quantitative metric could be the number of mistakes a person made when using the product, and a qualitative metric could be a subjective workload (using the NASA-TLX mentioned); for 'effectiveness', a quantitative metric could be time to achieve the goal and a qualitative metric could be a subjective rating of performance; for 'satisfaction', a quantitative metric could be time spent using the device (over the course of several days) and a qualitative metric could be a self-report of how pleasant the product was to use. It should be apparent that the distinction between efficiency, effectiveness, and satisfaction is somewhat arbitrary, which is why is important to make sure that all three metrics are applied during evaluation.

The idea that evaluation requires a 'protocol' is meant to imply that one ought to approach it in much the same way that one approaches the design of an experiment, that is, by defining independent variables, which are the goal, activity, and context of use, and by defining dependent variables, which are the usability metrics. The notion of the referent model also makes sense in terms of experimental design because the 'hypothesis' under test is that the outcome will be equal to or better than the referent model.

Initial Concept Development

During the 'initial concept development' phase, it is possible that one of the components of 'context of use' will dominate the others. For example, a designer might have an idea about how the product will function or how to perform a particular task or how to help a particular user. In order to explore this concept, designers make use of scenarios in various forms, for example, storyboarding, sketching, rich pictures, illustrative stories, and so forth. From Figure 1, it can be argued that a good scenario would include (as a minimum) some consideration of the type of person who would be likely to use the product, the tasks that the person would perform in order to achieve specific goals (as well as any other tasks that might need to be performed concurrently), the environment in which they might be performing these tasks, and the presence or use of other products to support this activity. In the domain of 'traditional' computer systems, the 'environment' can be assumed to be more or less constant, that is, the computer would be used on a desk in an office. In mobile computing, the 'environment' will have a significant impact on how the product will be used, as will the range of tasks that the person will be performing. It is this impact of the environment and the increasing range of concurrent tasks that makes evaluating mobile technology different from other computer applications. One way in which these aspects can be considered is to develop a scenario in which a person achieves the defined goal using no technology, another in which they use 'contemporary' technology and another in which they use the concept product. By storyboarding these different scenarios, the design team gets a feeling for the main benefits to be gained from using the product (and an appreciation as to whether or not to pursue its development). During this stage, the usability metrics can be defined in terms of what measures can sensibly differentiate the product from any alternative ways of performing the task.

Prototyping

During 'prototyping' different versions of the product are developed and tested. The prototype need not be a fully-functioning product. Indeed, Nilsson et al. (2000) shows how very simple models can be used to elicit user responses and behaviors. Their study involved the development of a handheld device (the 'pucketizer' for use in water treatment plants and the initial studies had operators walking around the plant with a non-functioning object to simulate the device. From this experience, the design team went on to implement a functioning prototype based on an 8-bit microcontroller, wireless communications, and a host computer running a JAVA application). This work is interesting because it illustrates how embedding the evaluation process in the environment and incorporating representative end-users lead to insights for the design team. Taking this idea further, it is feasible for very early prototyping to be based on paper versions. For example, one might take the form factor of the intended device (say a piece of wood measuring 5" x 3" x ½"—which is approximately the size of Personal Digital Assistant) and then placing 3" x 2" paper 'overlays' to represent different screen states—change the 'screens' is then a matter of the user interacting with buttons on the 'product' and the evaluator making appropriate responses. Of course, this could be done just as easily using an application in WinCE (or through the use of a slideshow on the device), but the point is that initial concepts can be explored well before any code is written or any hardware built.

Specification and Build

'Specification and build' is the phase that one might traditionally associate with evaluation. Evaluation activity at this phase of the design process would be 'summative' (i.e., occur at the summation of the process), and would usually be used to confirm that the design was accept-

able prior to committing to manufacture. At this stage, the main concerns regarding hardware and software would have been dealt with and so any usability evaluation that would call for significant change to hardware or software is likely to be ignored (unless the product is scrapped and the process started again, or unless these recommendations are filed for the next version of the product). However, usability evaluation can play an important role in this phase because it will form part of the acceptance testing of end-users and could, if positive, play a role in defining marketing activity or informing training requirements.

CONCLUSION

While the concept of usability as multi-faceted might seem straightforward, it raises difficult problems for the design team. The design team focuses its attention on the device, but the concept of usability used in this chapter implies that the device is only part of the equation and that other factors relating to the user and environment can play significant roles. The problem with this, of course, is that these factors lie outside the remit of the design team. One irony of this is that a well-designed device can 'fail' as the result of unanticipated activity, user characteristics, and environmental features.

The issue raised in this chapter is that evaluating mobile technology involves a clear appreciation of the concept of usability, in line with ISO standard definitions. The ISO9241 concept of usability emphasizes the need to clearly articulate the 'context of use' of the device, through consideration of user, activity, device, and environment. This means that evaluation has to take account of the interactions between user, activity, device, and environment. What is essential is that evaluation is conducted in a way that ensures a good fit between the 'context of use' in the real-world and that simulated in the laboratory. This does not mean that one needs to include all aspects of the real-world in the laboratory but that one is able to reflect key variables and that the evaluation is designed to ensure a balanced comparison. It would be easy to 'prove' that a given device was superior to any other device simply by ensuring that the test favored the device in question. It is equally easy to 'prove' that evaluation in the 'laboratory' do not reflect performance in the 'real-world'. However, such studies often reflect a limited grasp of adequate experimental design and, ultimately, a poor understanding of science. One is not 'proving' that a product is well designed through evaluation. Rather one is demonstrating 'fitness-for-purpose' under a well-defined context of use.

REFERENCES

Baber, C. (2005). Evaluation of human-computer interaction. In J. R. Wilson & E. N. Corlett (Eds.), *Evaluation of human work* (pp. 357-388). London: Taylor and Francis.

Baber, C., Arvanitis, T. N., Haniff, D. J., & Buckley, R. (1999). A wearable computer for paramedics: Studies in model-based, user-centered and industrial design. In M. A. Sasse & C. Johnson (Eds.), *Interact'99* (pp. 126-132). Amsterdam: IOS Press.

Baber, C., Haniff, D. J., Knight, J., Cooper, L., & Mellor, B. A. (1998). Preliminary investigations into the use of wearable computers. In R. Winder (Ed.), *People and computers XIII* (pp. 313-326). Berlin: Springer-Verlag.

Baber, C., Haniff, D. J., & Woolley, S. I. (1999). Contrasting paradigms for the development of wearable computers. *IBM Systems Journal, 38*(4), 551-565.

Barnard, L., Yi, J. S., Jacko, J. A., & Sears, A. (2005). An empirical comparison of use in- motion evaluation scenarios for mobile computing devices. *International Journal of Human-Computer Studies, 62,* 487-520.

Barnard, L., Yi, J. S., Jacko, J. A., & Sears, A. (2007). Capturing the effects of context on human performance in mobile computing systems. *Personal and Ubiquitous Computing, 11*(2), 81-96.

Bass, L., Kasabach, C., Martin, R., Siewiorek, D., Smailagic, A., & Stivoric, J. (1997). The design of a wearable computer. In *Proceedings of the CHI '97* (pp. 139-146). New York: ACM.

Bass, L., Siewiorek, D., Smailagic, A., & Stivoric, J. (1995). On site wearable computer system. In *Proceedings of the CHI '95* (pp. 83-88). New York: ACM.

Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D., Benford, S., Hampshire, A., & Capra, M. (2006). Interweaving mobile games with everyday life. In *Proceedings of the CHI 2006* (pp. 417-426). New York: ACM.

Bevan, N. (2001). International standards for HCI and usability. *International Journal of Human Computer Interaction, 55,* 533-552.

Boase, M., Hannigan, S., & Porter, J. M. (1988). Sorry, can't talk now…just overtaking a lorry: The definition and experimental investigation of the problem of driving and hands-free car phone use. In E. D. Megaw (Ed.), *Contemporary ergonomics.* London: Taylor and Francis.

Brooke, J., (1996). SUS: a quick and dirty usability scale. In P. W. Jordan, B. Weerdmeester, B. A. Thomas, & I. L. McLelland (Eds.), *Usability evaluation in industry* (pp. 189-194). London: Taylor and Francis.

Brookhuis et al, (1991). The effects of mobile telephoning on driving performance. *Accident Analysis and Prevention, 23*

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface In *Proceedings of the CHI '88* (pp. 213-218). New York: ACM.

Cobb, S. V. G., Nichols, S. C., Ramsey, A. D., & Wilson, J. R. (1999). Virtual reality induced symptoms and effects (VRISE). *Presence: Teleoperators and Virtual Environments, 8,* 169-186.

Dillon, A. (1992). *Designing usable electronic text.* London: Taylor and Francis.

Duh, H. B-L., Tan, G. C. B., & Chen, V. H-H. (2006). Usability evaluation for mobile devices: a comparison of laboratory and field tests. In *Proceedings of Mobile HCI '06* (pp. 181-186). Berlin: Springer-Verlag.

Earthey, J., Sherwood Jones, B., & Bevan, N. (2001). The improvement of human-centered processes—Facing the challenges and reaping the benefit of ISO 13407. *International Journal of Human Computer Studies, 55,* 553-585.

Ebersbach, G., Dimitrijrvic, M. R., & Poewe, W. (1995). Influence of concurrent tasks on gait: A dual task approach. *Perceptual and Motor Skills, 81,* 107-113.

Frøkjaer, E., Hertzum, M., & Hornbaek, K. (2000). Measuring usability: are effectiveness, efficiency and satisfaction really correlated? In *Proceedings of the CHI '2000* (pp. 345-352). New York: ACM.

Good, M., Spine, T. M., Whiteside, J., & George, P. (1986). User-derived impact analysis as a tool for usability engineering. In *Proceedings of the CHI '86* (pp. 241-246). New York: ACM.

Gould, J. D., & Lewis, C. H. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM, 28*(3), 300-311.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*(3), 203-261.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A.

Hancock & N. Meshkati (Eds.), *Handbook of mental workload*. Amsterdam: North Holland.

Holcomb, R., & Tharp, A. L. (1991). What users say about software usability. *International Journal of Human-Computer Interaction, 3*(1), 49-78.

ISO 18021. (2001). Information technology—User interface for mobiles. ISO9241-11 1998.

ISO 9241. (1998). Ergonomics of office work with VDTs—Guidance on usability. Geneva: International Standards Office.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques *(CHI'91)* (pp. 119-124). New York: ACM.

Johnson, P. (1992). *Human-computer interaction: psychology, task analysis and software engineering.* London: McGraw-Hall.

Johnson, P. (1998). *Usability and mobility.* Paper presented at the 1st Workshop on Human-Computer Interaction with Mobile Devices, Glasgow.

Kirakowski, J., & Corbett, M. (1988). Measuring user satisfaction. In D. M. Jones & R. Winder (Eds.), *People and computers IV* (pp. 329-430). Cambridge: Cambridge University Press.

Kirakowski, J., & Corbett, M. (1993). SUMI: the software usability measurement inventory. *British Journal of Educational Technology, 24,* 210-214.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human-Computer Studies, 60,* 599-620.

Knight, J. F., & Baber, C. (2005). A tool to assess the comfort of wearable computers. *Human Factors, 47,* 77-91.

Knight, J. F., Baber, C., Schwirtz, A., & Bristow, H. (2002). The comfort assessment of wearable computers. *Digest of Papers of the 6th International Symposium on Wearable Computing* (pp. 65-74). Los Alamitos, CA: IEEE Computer Society.

Monk, A. F., Wright, P. C., Haber, J., & Davenport, L. (1986). *Improving your human-computer interface: A practical technique.* London: Prentice-Hall.

Nielsen, J. (1993). *Usability engineering.* London: Academic Press.

Nilsson, J., Sokoler, T., Binder, T., & Wetcke, N. (2000). Beyond the control room: Mobile devices for spatially distributed interaction on industrial process plants. In P. Thomas & H. W. Gellerson (Eds.), *Handheld and ubiquitous computing: Second international symposium* (pp. 30-45). Berlin: Springer.

Nunes, L., & Recarte, M. A. (2002). Cognitive demands of hands-free phone conversation while driving. *Transportation Research Part F: Traffic Psychology and Behavior, 5,* 133-144.

Ravden, S. J., & Johnson, G. I. (1989). *Evaluating usability of human-computer interfaces.* Chichester: Ellis Horwood.

Slater, M., Usoh, M., & Steed, A. (1994). Depth of presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 3,* 130-144.

Svenson, O., & Patten, C. J. D. (2005). Mobile phones and driving: A review of contemporary research. *Cognition, Technology and Work, 7,* 182-197.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability engineering: Our experience and evolution. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 791-817). Amsterdam: Elsevier.

Wichansky, A. M. (2000). Usability testing in 2000 and beyond. *Ergonomics, 43*(7), 998-1006.

Wikman, A., Nieminen, T., & Summala, H. (1998). Driving experience and time sharing during in-car tasks on roads of different widths. *Ergonomics 4,* 358-372.

Wilson, J. R., & Nichols, S. C. (2002). Measurement in virtual environments: Another dimension to the objectivity/subjecitivity debate. *Ergonomics, 45,* 1031-1036.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: a presence questionnaire. *Presence: Teleoperators and Virtual Environments, 7,* 225-240.

KEY TERMS

Context of Use: The combination of user, task, product, and environment during the achievement of a desired goal

Dual-Task: The performance of two (or more) tasks at the same time. This could involve simultaneous performance or could involve some form of time-sharing between the tasks.

Effectiveness: The ability of a given user to employ a given product to achieve a desired goal in a given context of use

Efficiency: The optimal expenditure of resources by a given user in using a given product to achieve a desired goal in a given context of use

Referent Model: A product (or set of metrics) against which a given product can be compared

Satisfaction: The subjective response of a user to interacting with a product

Usability: *"... the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."* (ISO9241-11, 1998)

ENDNOTES

1. SUS: Software Usability Scale (Brooke, 1996)
2. SUMI: Software Usability Metrics Inventory, Kirakowski and Corbett (1993)

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 731-744, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.21
Mobile Public Relations Strategies

Chris Galloway
Monash University, Australia

INTRODUCTION

Public relations is about the "ethical and strategic management of communication and relationships" (Johnston & Zawawi, 2004, p. 6) with individuals and groups ("publics") important to an organization. At one time such publics could safely be thought of in relatively static terms such as geographic location. This is, of course, still possible—but such fixed categories are of diminishing importance when it comes to building relationships with modern publics and communicating organizational messages to them. Even the motor vehicles that facilitate physical movement are becoming "smarter" and converging with technologies such as mobile telephony, personal entertainment systems and handheld computing (Sherry & Urry, 2000, as cited in Sheller, 2002).

This article aims to explore the idea that mobile technologies mean PR practitioners must rethink both the notion of publics and also how to relate to them. A "mobile PR" will undermine taken-for-granted views about the nature of media, messages, and the kinds of relationships public relations people can expect to create on behalf of their clients. Many practitioners are still getting to grips with the online public relations they have known—through activities such as arranging the building of corporate Web sites, monitoring online discussions relevant to client interests and both disseminating company information online and responding to inquiries about it. The idea of an even more flexible communications environment enabled by mobile technologies may seem very daunting. No-one has so far worked out how to "do" PR in this new communications climate—there are no prescriptions or generally accepted approaches. Yet if practitioners do not confront the dilemma of how to reach mobile audiences they risk becoming irrelevant to many clients who must communicate in the mobile space or face unacceptable decay in their business.

BACKGROUND: WHY MOBILE IS DIFFERENT

The Internet and mobile marketplaces have important differences. As Lindgren, Jedbratt, and Svensson note, "The mobile marketplace has a much wider reach than anything before it. It becomes synonymous with everywhere" (2002, p. 5). Siau, Lim, and Shen agree that, "the emerging mobile commerce operates in an environment very different from e-commerce conducted over the wired Internet" (2003, p. 2). It is important that PR practitioners understand the differences and adjust their thinking and practice accordingly.

Siau et al. list mobile market features they consider are *not* characteristic of "traditional" e-commerce:

- **Ubiquity:** Users can get any information they want, whenever they want it, wherever they are (including, now, RSS feeds delivered via mobile internet services).
- **Reachability:** Businesses can reach customers anywhere, anytime—equally, a user can be in touch with and available for other people anywhere, anytime.
- **Localization:** Knowledge of a user's physical location allows locality-specific services to be provided.
- **Personalization:** Mobile commerce applications can be personalized to represent information or provide services in ways appropriate to a specific user.
- **Dissemination:** some wireless infrastructures support simultaneous delivery of data to all mobile users within a specific geographical region. (Siau et al., 2003, pp. 2-3)

As wireless technologies evolve, customer relations will experience dramatic change (Siau et al., 2003, p. 16). To what extent can these relations be called *public* relations and be the domain of PR practitioners rather than marketers? Lines of demarcation can be very blurry. But PR is foregrounded when the prime purpose is to use mobile technologies to create connections–relationships—where the end game is the nature of the relationship rather than an immediate output such as a purchase decision. Examples include:

- Viewer voting via SMS or mobile calls associated with television pop contests, where the goal is to create a broad community of fans who feel they are helping shape the outcome. Their sense of involvement and connection to the contestants—and to other contest followers—is something the promoters hope will morph into long-term enthusiasm for, and purchasing of, products associated with the winner.
- Free delivery of information to mobile devices where consumers have opted-in to receive it—such as notifications of airline or public transport schedule changes. Individuals who are thereby saved a fruitless trip to the train station will value such a convenience—and the organization which provides it, enhancing the organization's reputation for being accessible (if not always for making the trains run on time).
- Personalized, user-specified content delivery. This can range from downloading a daily prayer to getting top news points from the daily paper sent to a mobile device. The organization providing the service is extending its reach to consumers who may not otherwise stop to access their content, or in fact may not be able to access it by other means. This extended reach is valuable to the organizations concerned as a larger pool of customers creates more opportunities to piggyback paid services on free ones.
- Research conducted by SMS, where members of a group which may be dispersed through several countries both receive and respond to researchers' questions through their mobile devices. The group is connected

> to the research company solely through their mobile equipment but their relationship with the company is essentially no different in nature from that of a focus group meeting in the company's home office.

These examples draw on some of the characteristics of mobile commerce noted by Siau et al. (2003) such as ubiquity, reachability, and personalization to highlight the fact that mobile technologies are enabling new forms of connectivity between organizations and publics that differ from the previous concept of cyberspace as something entered through fixed, location-specific devices. In their difference, the examples also highlight the need for PR practitioners to rethink their approaches to electronic public relations, recognising they now need to do more than designing conventional public relations campaigns that may (or may not) incorporate some online activity. "Mobile PR" may constitute only one part of a public relations initiative—but it should be seen as one that cannot be ignored.

Public relations has developed a range of strategies and tactics to influence publics, most focused on using mass media. Audiences are assumed to be susceptible to media-based persuasion expressed in fact and logic-based statements that frame a particular advocacy position. When the Internet became widely available, public relations practitioners began using it as just a new tool for doing what they had long done, such as publishing media statements and other corporate information and disseminating advocacy material. E-mail meant that some interactivity could be introduced. A new field known variously as cyber-PR, online public relations, electronic PR or "E-PR" developed. Online press conferences were held and some companies began monitoring online chat groups where discussions could highlight an emerging issue that might affect their business. E-PR was used alongside traditional public relations approaches in campaign implementations. Discourse about it focused on translating questions of communication efficiency to the online environment, such as how organizations can integrate the Internet into their existing investor relations activities (Kuperman, 2000) and how they may identify issues that need to be managed (de Bussy, Watson, Pitt, & Ewing, 2000). Interest has focused on exploiting the technology to deliver communication efficiencies for the organization rather than on delivering experiences consumers may want, such as a sense of "connectedness", which Dholakia et al. (2000) describe as "the feeling of being linked to a world outside the specific site" (Gustafson & Tilley, 2003).

RECONCEPTUALIZING PR FOR A MOBILE WORLD

Existing E-PR tactics are not longer sufficient for mobile-driven markets. E-PR needs to encompass "M-PR" (mobile public relations). The fluid nature of mobile communications means some core, generally accepted notions of public relations planning need redefinition. PR campaign planning models vary (Bobbitt & Sullivan, 2005, p. 32). There are, however, common elements used in communication processes that aim to build a mutually beneficial relationship with a public – a best-practice goal of contemporary public relations. They include:

- The need to identify priority audiences ("target publics")
- Selecting appropriate media
- Designing effective messages

Mobile communication also reworks the idea of relationship.

Table 1 attempts to point to both similar and dissimilar aspects of the mobile and traditional electronic public relations environments, necessarily making generalizations as it does so.

Table 1.

Mobile PR	**'Traditional' Electronic PR**
Consumers access cyberspace from mobile devices	Consumers access cyberspace from fixed devices
Target publics form fluid communities linked by mobile communication	Target publics self-select by "pulling" information to themselves
Messages abbreviated, "burst-y", often in TXT and symbols – "emoticons"	Messages use symbols and plain text; facility for delivery of extensive content
Devices allow multimedia experience	Multimedia delivered via fixed devices
Demand to satisfy emotional needs such as sense of involvement is a strong driver for relationship formation	Greater opportunity to influence relationships with rationales for advocacy positions

Target Publics

The world of mobile commerce calls for a wider way of thinking about relating to publics, one that recognizes that mobile communicators—one description of them is "global knowledge-nomads" (Lindgren, Jedbratt, & Svensson, 2002, p. 10)—are consuming content and managing relationships, including with commercial organizations, every bit as much as those consumers who are working with fixed technologies. Mobile communicators cannot be unambiguously defined by familiar psychographic, demographic or even geographic characteristics. "Fleetingness" is the key characteristic, with audiences seeming to be always just beyond reach (Proctor & Kitchen, 2002).

This means the idea of "target publics" must be reconceptualized, taking into account that "in a global context mobile telephony is used by a far broader spectrum of the population than PCs and the Internet" (Lacohee, Wakeford, & Pearson, 2003, p. 206). To PR practitioners raised on linear, structured campaign planning models, elusiveness of audiences is as frustrating as it is demanding. Yet, as Proctor and Kitchen suggest, inconstancy of targets need not justify inactivity: a key may be to "adopt an open, untargeted, ill-defined approach which leaves scope for imaginative consumer participation" (2002, p. 154). Such an approach is ideal for the interactivity of mobile communication, acknowledging that "mobility does not only relate to our physical bodies. We are also mentally mobile and are adept at migrating between various mental states, various identities" (Lindgren et al., 2002, p. 24). Computer and mobile phone-supported social networks are enabling communities of shared interest to form in both physical and online spaces (Wellman, 2001, as cited in Sheller, 2002) and such communities may form the basis of new publics (Sheller, 2002, p. 46).

Media

Commonly used PR planning approaches based on a stage-by-stage implementation concept include a point where media are chosen to carry the messages intended to influence a target audience. Complementing traditional print and broadcast media, the internet has been used as a medium to reach online audiences. Mobile devices capable of supplying voice, video, and text (including RSS feeds) to a user constitute a new, rich and challenging medium for the E-PR practitioner. It is demanding because familiar tactics such as issuing press statements cannot be deployed as before: they require reconstitution for convergent mobile media use.

Messages

The "intensity, brevity, and the absence of narrative continuity" (Lash, 2002, p. 206) that characterize mobile-enabled communication are incompatible with delivering extended logic-based PR advocacy material. Post-modern audiences are likely to be influenced as much by emotion, experience and a desire to find meaningful connection with others as they are by logic (compare Caru & Cova, 2003; Ito & Daisuke, forthcoming, p. 3). Mobile communication's interactivity can offer this kind of intense connectivity.

On-the-move consumers not only seek to be accessible themselves most of the time—Ito and Daisuke's "persistent connectivity" (forthcoming, p. 19)—increasingly they also want to draw down needed information to their mobile devices. Many may remain consumers of traditional media such as newspapers, television and radio while relying largely on mobile devices to function in cyberspace. The condensed nature of much mobile communication means mobile communicators may be less exposed to the extended information and advocacy material available on organizational Web sites, thus obsoleting the press release in its traditional form. That in itself is a big challenge for E-PR practitioners, who must also take into account the mobile field's rapid technological change, which may lead to unexpected applications (Lacohee, Wakeford, & Pearson, 2003, p. 206).

PR people now face the task of designing and delivering content to mobile publics that meets their less tangible need for connection with others: according to Fox (2001) mobile telephones provide a "'social lifeline' in a fragmenting and isolating world" and even the gossip that comprises much mobile-based chat "restores our sense of connection and community" (2001, p. 1). They may need to consider an approach that has been dubbed "dynamic touch" (Galloway, 2005) because it relies on creating an *experience* of connectivity rather than on getting prose published or images broadcast. One example is the sense of connection with contestants that fans develop during the *American Idol* television contests through voting via SMS, e-mailing messages of support and discussing their favorite's progress on their telephones.

Public relations people may need to design more such virtual experiences as part of building and maintaining relationships with mutable, mobile publics, recognizing that "the mobile marketplace merges the virtual and physical marketplaces" (Lindgren et al., 2002, p. 6). This is not the stuff of PR "as we know it" but rather, largely uncharted territory for professional communicators who will need to learn how to imagine and deliver experiences that connect with people in ways that have a planned, overall consistency to them.

Relationships

Increasingly, today's markets are driven by many organizations' perceived need for a "continuous personalized dialog with customers" (Lindgren et al., 2002, p. 17). In the light of this pressure, any concept of E-PR as merely managing familiar tasks in an electronically-enabled environment

falls far short. It does so for a number of reasons. One is simply the strong and sustained growth in mobile telephony and in wireless data usage. More than 20% of the world's population uses a mobile telephone and usage is increasing at more than 10% a year (Rerisi, 2003, as cited in Grant & Meadows, 2004). Such surging growth and the changes it is bringing in the way people connect with organizations and individuals that matter to them is simply too significant safely to ignore.

As Levy points out, "major technological inventions not only enable us to do 'the same things' more quickly, better or on a greater scale, but also allow us to do, feel or organize ourselves differently" (2001, p. 199). People now use mobile technologies to negotiate coordination of their activities, permitting "direct contact that in many ways is more interactive and flexible than time-based co-ordination" (Ling, 2004, p. 58), enabling users to "have a foot in both the here and now as well as the there and now" (2004, p. 190). According to Lacohee, Wakeford, and Pearson (2003), the linking of communication to mobility is central to contemporary social networks. As they point out, such communication need not be in a close social network: the BBC has found with text messages that "New technologies are giving us a level of interaction with out audiences that we have never seen before" (Chapman, quoted in Lacohee, Wakeford, & Pearson, p. 206).

Internet and mobile-linked consumers are primarily interested in building a sense of *connectedness* with others (individuals, groups, and organizations) rather than in consuming particular types of media content. As Proctor and Kitchen note, "Postmodern consumers seek to feel good in separate, different moments by acquiring self-images that make them marketable, likeable, and/or desirable in each situation or moment" (2002, p. 148). People choose to group themselves in shifting "neo-tribes" which share emotions and a "more spiritual sense of community" (compare Patterson, 1998, p. 70). While text messages are often low in informational value they are considered high in social grooming (Lacohee, Wakeford, & Pearson, 2003, p. 206).

FUTURE TRENDS

In the United States, the number of mobile device users exceeds the number of people who use personal computers (Lim & Siau, 2003). As more computing features become not only available on mobile devices but also easier and cheaper to use, this trend can be expected to continue. Wireless capabilities are being incorporated into more technologies and extended for others—and there is the prospect of "unheard of transmission rates" (Meadows, 2004, p. 356, as cited in Grant & Meadows). As more and more functions are offered on wireless devices, it is likely that those with one predominant use will give way to those with multiple capabilities (compare Banks & Fidoten, 2004, as cited in Grant & Meadows, 2004).

For public relations practitioners this means coming to terms with more than a need to understand these converging capabilities and how to deliver content in a way that is both technically and "culturally" compatible—culturally in the sense of being viewed by users as appropriate to the environment of the hybrid communication and computing devices. It also means learning to surf the fluid waves of the mobile market as "members of the communities in which they wish to communicate" (Nicovich & Cornwell, as cited in in Rettie, 2002, p. 261).

CONCLUSION

Public relations practitioners must learn to use mobile devices as new media for campaigns, not just as tools for managing their schedules or arranging the next client function. Doing so will demand a fundamental reworking of familiar concepts to adapt to the new mutable world of mobiles.

REFERENCES

Bobbitt, R., & Sullivan, R. (2005). *Developing the public relations campaign: A team-based approach*. Boston: Pearson.

Caru, A., & Cova, B. (2003). A critical approach to experiental consumption: Fighting against the disappearance of the contemplative time. Paper presented to the Critical Management Studies Conference, 2003. *Electronic Journal of Radical Organization Theory, 8*(1) 2004. Retrieved June 1, 2005, from http://www.mngt.waikato.ac.nz/research/ejrot/cmsconference/2003/proceedings/criticalmarketing/Caru.pdf

de Bussy, N. M., Watson, R. T., Pitt, L. F., & Ewing, M. T. (2000). Stakeholder communication management on the Internet: An integrated matrix for the identification of opportunities. *Journal of Communication Management, 5*(2), 138-146.

Fox, K. (2001). Evolution, alienation, and gossip. *Social Issues Research Centre.* Retrieved June 2, 2005, from http://www.sirc.org/publik/gossip.shtml

Galloway, C. J. (2005, November). Cyber-PR and dynamic touch. *Public Relations Review, 31*(4), 572-577.

Grant, A. E., & Meadows, J. H. (2004). *Communication technology update* (9th ed). Burlington; Oxford, MA: Focal Press.

Gustavsen, P. A., & Tilley, E. (2003). *Public relations communication through corporate Web sites: Towards an understanding of the role of interactivity.* Retrieved February 23, 2005, from http://www.praxis.bond.dedu.au/prism/papers/refereed/paper5.pdf

Haig, M. (2000). *e-PR: The essential guide to public relations on the Internet.* London; Milford, CT: Kogan Page.

Ito, M., & Daisuke, O. (forthcoming). Mobile phones, Japanese youth, and the re-placement of social contact. Forthcoming in R. Ling & P. Pedersen (Eds.), *Mobile communications: Renegotiation of the social sphere.* Retrieved June 2, 2005, from http://www.itofisher.com/mito/archives/mobile youth.pdf

Johnston, J., & Zawawi, C. (2004). *Public relations theory and practice* (2nd ed.). Crows Nest: Allen & Unwin.

Kuperman, J. C. (2000). The impact of the Internet on the investor relations of firms. *Journal of Communication Management, 5*(2), 147-159.

Lacohee, H., Wakeford, N., & Pearson, I. (2003, July). A social history of the mobile telephone with a view of it future. *BT Technology Journal, 21*(3), 203-211.

Lash, S. (2002). *Critique of information.* London: Sage.

Levine, M. (2002). *Guerilla P.R. Wired.* New York: McGraw-Hill.

Levy, P. (2001). *Cyberculture*. (Boonoono, R,Trans.). Minneapolis; London: University of Minnesota Press.

Lim, E.-P., & Siau, K. (Eds.). (2003). *Advances in mobile commerce technologies.* Hershey, PA: Idea Group Publishing.

Lindgren, M., Jedbratt, J., & Svensson, E. (2002). *Beyond mobile: People, communications, and marketing in a mobilized world.* Basingstoke, Hampshire: Palgrave.

Ling, R. (2004). *The mobile connection: The cell phone's impact on society.* San Francisco: Elsevier.

Paterson, R. (2005). *Robert Paterson's Weblog.* Retrieved May 31, 2005, from http://smartpei.typepad.com/robert_patersons_weblog/2004/07/mag

Proctor, T., & Kitchen, A. (2002). Communication in postmodern integrated marketing. *Corporate*

Communications: An International Journal, 7(2), 144-154.

Rettie, R. (2002). Net generation culture. *Journal of Electronic Commerce Research, 3*(4), 254-264.

Sheller, M. (2002). Mobile publics: Beyond the network perspective. *Environment and planning D: Society and space, 22*, 39-52.

Taylor, E. (2004) Love e, Love e not. *Social Issues Research Centre*. Retrieved June 2, 2005, from http://www. sirc.org/articles/love_e_love_e_not. shtml

KEY TERMS

Cyberspace: The boundaryless virtual world accessed through computer networks, whether one's access device is fixed or mobile.

Dynamic Touch: An experience in cyberspace designed by a professional communicator to stimulate emotional responses that will help advance the interests of the communicator's client organization.

Media: In a public relations context, these are channels for delivery of organizational messages to target publics.

Message: Any content (audio/written/visual) an organization wishes to deliver to a public in order either to inform or to motivate them to a desired response.

Mobile-PR: The application of public relations strategies and tactics in cyberspace accessed through mobile devices.

Public: Any group of people an organization wishes to reach because their interest or influence is relevant to the organization in some way. People may belong to more than one public and shift between them quickly, depending on the issue in question. For example, an Internet user who is also an environmental activist who uses SMS to coordinate times and places of protests with other activists.

Public Relations: The planned and sustained effort to build mutually beneficial relationships between an organization and those individuals or groups whose interest or influence makes them relevant to the organization.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 805-810, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.22
Introducing Mobile Government

M. Halid Kuscu
Mobile Government Consortium International, UK

Ibrahim Kushchu
Mobile Government Consortium International, UK

Betty Yu
The Chinese University of Hong Kong, Hong Kong

ABSTRACT

This chapter introduces the mobile government concept and creates a context for discussing various applications, services, and the relevant technologies. The context presented also includes the introduction of ideas on the characteristics of mobile government and some major issues that justify the necessity of the mobile government and identify the potential that it possesses. This chapter should serve as a foundation for the discussions on the further chapters.

INTRODUCTION

E-government efforts aim to benefit from the use of most innovative forms of information communication technologies, particularly Web-based Internet applications, in improving governments' fundamental functions. These functions are now using mobile and wireless technologies and creating a direction for e-government: mobile government. Mobile government (m-government) may be defined as a *strategy and its implementation involving the utilization of all kinds of wireless and mobile technology, services, applications, and devices for improving benefits to the parties involved in e-government including citizens, businesses, and all government units* (Kushchu & Kuscu, 2003).

Despite its early stage, m-government seems to have a substantial influence on the generation of set of complex strategies and tools for e-government efforts and on their roles and functions. M-government is inevitable. The number of people having access to mobile phones and mobile Internet connection is increasing rapidly. The mobile access—anywhere any time—is becoming a natural part of daily life, and the governments

will have to transform their activities according to this demand of convenience and efficiency of interactions for all parties.

The coming age of m-government raises several interesting questions. Will m-government replace the e-government activities? Despite its significance, m-government cannot be seen as replacing e-government, and in many cases it will be complementary to e-government efforts. The conventional e-government efforts provide services through wired network with interactive and relatively intelligent Web applications. The value of m-government comes from the capabilities of applications supporting mobility of the citizens, businesses, and internal operations of the governments. For example, supporting law enforcement agents who are on patrol is a distinctive advantage of mobile government services over conventional e-government implementations. Wireless applications may enable greater mobilization of the government officials with the ability to handle real-time information concerning, for example, crimes, accidents, safety, and other public issues.

This chapter provides an introduction to m-government. In the following section, a number of m-government applications from various countries are briefly presented in order familiarize the readers with what is actually involved in m-government. Next, the distinctive characteristics that differentiate m-government from e-government are stated. Finally, a discussion of current issues in m-government is presented.

SOME MOBILE GOVERNMENT APPLICATIONS AND SERVICES

The technology and the services landscape is slowly taking its place in various m-government implementations. Some of the early adopters of m-government services include law enforcements, fire fighting (Easton, 2002), emergency medical services, education, health, and transportation (Yu & Kushchu, 2004). The following tables provide examples from each group of applications on where the application is developed, and a brief description of each.

Instant Information Release

Mobile devices are often carried by users and are always turned on. This characteristic enables mobile devices to serve as a warning or reminder to users with quick and specific information release (Table 1).

Quick Information Collection

The possibility of retrieving information while on the move is one of the major characteristics of mobile government applications. This allows civil servants to collect necessary information to provide more efficient and effective service to the public. Examples are shown in Table 2.

Mobile Transactions

Transactions are essential parts of doing business. Currently, many companies are exploring the possibility of mobile commerce and some governmental organizations have already taken the initiative to utilize this opportunity (Table 3).

Faster Information Exchange

In case the speed of information exchange is important but not critical, applications are defined as enhancing the pleasure value for users (Table 4).

Fighting Against Crime

To fight against crime, law enforces need citizens' corporation to provide information. The reports from citizens can increase the chance for police

Table 1. Instant information release

Where	Description	Relevance
SMS for people with hearing disabilities		
Great Britain	• citizens with hearing problems can be contacted by the police with SMS.	• may not be replaced by the wired Internet
Amsterdam	• SMS message are sent to citizens with hearing problem in emergencies • includes instruction such as "go home" or "leave the place", so they understand how to react.	• people with hearing problems cannot listen to warning bells and the only way to warn them of danger is by the use of mobile devices which can vibrate to notify the users • satisfies the need to warn citizens with hearing problems about potential danger
Special notification cases		
California, USA	• SMS are sent to citizens in case of energy black-outs.	• governments try every communication channel to notify citizens during emergencies
London	• Police may send notifications to citizens about potential terrorist threats or attacks.	• yet working people are too concentrated on work and may not receive the warnings • notice danger when friends or families call • more effective to spread the message directly
SMS floods warning systems		
Malaysia	• automatic measuring devices are installed to monitor water level. • when flooding rises to certain level, the control centre sends a message to all the affecting citizens.	• citizens may aware of danger in day time even without warnings from government • in night time, mobile device which are often with the user serve as a very good way to warn the users of potential danger.
United Kingdom	• In case of emerging floods, information are sent via SMS, emails, fax and television.	

to arrest suspects, find missing people and better investigate the cases. Table 5 shows some examples.

THE CHARACTERISTICS OF M-GOVERNMENT

M-government involves a strategy and implementation of governmental services through a mobile platform to provide its users, both citizens and civil servants, the benefit of getting services and information from anywhere at anytime (Kushchu & Kuscu, 2003). The use of mobile technologies and applications differentiates m-government from any other developments in the public sector using new technologies, including e-government. Based on various studies on mobile government applications (Yu & Kushchu, 2004), and their use in practice (Cilingir & Kushchu, 2004), a num-

Table 2. Quick information collection

Place	Description	Relevance
Fire fighting		
USA	• firemen receive critical information on their way to the site using mobile devices • get information such as structure of building, presence of toxic materials, surrounding environment and number of people trapped. • can connect mobile devices to camera in the buildings on fire and observe the interior environment	• in the short time between receiving fire alarm and arriving at the site, fire fighters may have only a few minutes to form a strategy based on limited information. • fire fighters can receive more information and forms better strategy. • impossible to use wired internet
Search for missing children / citizens and criminals		
Germany	• when police are searching for missing person or criminals, SMS message will be sent to registered bus and taxi drivers. • includes relevant information such as description of the person and possible location to be aware.	• increase the chance of finding missing person by extending the search from police to drivers • minimize the searching time for missing people • minimize the potential danger of criminals posed on the public

Table 3. Mobile transaction

Place	Description	Relevance
Mobile automobile parking		
Sweden	• registered driver can log in and log out a parking space using a mobile phone. • fee is automatically charged to the driver's account • receipt is sent via SMS	• drivers can skip the painful process of carrying loose changes and searching around for the nearest parking machines • the convenience and time saving create a better experience
Tax declaration		
Norway	• pre-filled tax declaration form is mailed to the citizen in advance • if the person has nothing to change in the form, he can send a SMS message with specific code and complete the entire tax declaration procedure.	• simplify the tax declaration • is feasible in e-government context, but mobile technology improves users' experience because they can complete the whole process even during his way to office.

Table 4. Faster information exchange

Place	Description	Relevance
Mobile hospital staff		
Sweden	• doctors and nurses can catch, deliver and receive care data at the point of care • equipped with pocket PCs which are connected through wireless LAN to the central database	• mobile technology enables hospital staff to have faster information flow • decrease in time for transferring data results in better decision in shorter time
Mobile elderly care workers		
Sweden	• field workers are equipped with mobile devices • provide updated information on elderly, ailing or handicapped people in need of home care	• ability to access data from service site allows care workers to spend more time on their job rather than travelling around for information
SMS for higher rate of employment		
Australia	• target: citizens, mainly teenagers • on potential offers or updates	• allows job seekers to reach information in a timely manner
Sweden	• SMS were sent to a pool of registered workers who are willing to work as temporary	• improves users' experience in accessing the information.
More efficient garbage collection via SMS		
Quezon, Philippines	• reports need for cleaning services in given areas.	• allows citizens to communicate with dustmen service, so the environment can be improved.

Table 5. Fighting against crime

Place	Description	Relevance
Reporting crime		
Manila, Philippines	• can report suspicious activities via SMS • receive SMS messages on the increase in crime rate in particular region	• encourage people to report criminal activities by simplified procedures, easier channel and faster response from the
Italy	• a couple of thieves are caught after photos of criminal act were taken by others and sent as MMS to the police.	• improved the participation in crime prevention from citizens

ber of differentiating factors can be identified in terms of better precision and personalization in targeting users and in delivering content, more convenient accessibility and availability, and a larger and wider user base.

- More convenient accessibility and availability (power of pull):
  - M-government enhances the adoption of online governmental services by citizens through the improved convenience it offers. Citizens can use the online governmental services not only "anytime" but also "anywhere".
  - Mobile devices are always on. This is different from personal computers where most mobile devices are always switched on. Usually, these devices stay at an inactive state, but applications can "wake up" the device. This is very different from e-government applications.
  - Mobile devices are designed to be carried around. As mobile devices are always carried around by the user, applications can be designed to provide instant information to the users. An example is to send out warnings during emergencies.
- Better precision and personalization in targeting users and delivering content (power of push):
  - A computer can be shared among different users, but mobile devices are designed to be used by a single user. This means that personalized information can reach the same user at any time through that one specific device.
  - M-government increases the acceptance, adoption, and the usage of online governmental services by reaching the citizens through a more personal, familiar, and friendly device.
- Larger and wider user base (power of reach):
  - M-government reaches a larger number of people through mobile devices, which often far exceeds the wired Internet user community.
  - M-government reaches a variety of audiences, including people who have no training or experience with computers and the Internet, but are active users of mobile communication.

MAJOR ISSUES IN M-GOVERNMENT

The Drivers of M-Government

There are various technological and non-technological driving forces for m-government. These forces will place severe influence on the new and existing e-government efforts to move towards adoption of mobile applications and services. Some of these forces include:

- increasing mobile infrastructure and mobile device penetration in Europe and in the world;
- evolution of mobile Internet technologies, standards, and protocols toward faster and more sophisticated applications; and
- adoption of mobile Internet applications and services by individuals and businesses.

The Transition from E-Government to M-Government

M-government is building upon e-government efforts, and there are basically two important issues related to the transition from and the relationship between e-government to m-government:

- **M-government is inevitable.** The major forces influencing m-government adoption include: (a) current technological advances in the areas of wireless World Wide Web and the Internet; (b) benefits to be gained from value added business models stemming from these developments; and (c) the citizen's rising expectations for a better and convenient government services.
- **M-government will be complimentary to e-government.** Some of m-government services are replications of e-government services on the mobile platforms. However, the real value of m-government efforts surfaces with those services and applications which are only possible using wireless and mobile infrastructure.

  The synergy between e-government and m-government may be of concern especially for those countries that have already gone ahead in making substantial investments in e-government implementations. Now that m-government is inevitable, extending activities to wireless devices and networks will enable these countries to be more proactive in their operations and services by providing real-time and up-to-date information to officials on the move and by offering citizens a broader selection of choices of interaction. For these countries, m-government implementations are emerging as one of the additional value-added features for the integrated and flexible data communication and exchange mechanism among government units. They may use more advanced wireless applications such as location-based information exchanges. These emerging applications are expected to stimulate m-government by enhancing location-based services such as fire fighting and medical emergencies. If requested, these technologies may be used to transfer location-specific information to mobile device users (i.e., information about traffic conditions or the weather). How about the implications for those countries that have not yet started or are at the early stages of e-government strategy and implementation processes? These countries may have more advantages depending on the type of the issues faced by the governments. In developing countries, mobile government applications may become a key method for reaching citizens and promoting exchange of communications especially when used in remote areas. In such countries with insufficient conventional telecom infrastructures and greater acceptance of mobile phones, ability of reaching rural areas may be considered an important feature of m-government.

Implementation Issues

Implementing m-government will also bring a series of challenges. Some of the typical challenges for e-government are naturally shared by m-government efforts. Lanwin (2002) states some of these challenges. Among them, we will visit those which are most relevant to m-government including infrastructure development, privacy and security, legal issues, mobile penetration rate, and accessibility.

- **Developing wireless and mobile networks and related infrastructure:** For m-government to flourish, the information technology infrastructure must be present. This infrastructure is both physical and "soft". The physical infrastructure refers to the technology, equipment, and network required to implement m-government. Equally important are soft infrastructures such as institutional arrangements, and software that make m-government transactions possible. Even though m-government is in its initial stage, various software are available

for m-government services. PacketWriter, Pocket Blue, and Pocket Rescue are a few examples of m-government software developed by Aether systems (for more information, please visit http://www.aethersystems.com/webfiles/industries/government/#roi).

- **Promoting mobile penetration and increasing accessibility:** The success of mobile government will depend largely on the number of its users: the citizens. But socio-economic factors such as income, education level, gender, age, handicap, language differences, and regional discrepancies will affect the citizens' attitude toward mobile government. In order to increase citizen participation and provide citizen-oriented services, governments need to offer easy access to m-government information in alternative forms, possibly, using video and voice communications.
- **Protecting privacy and providing security for the data and interactions:** Privacy and security are the most significant concerns citizens have about m-government. The general fear is that their mobile phone numbers will be traced, when they send their opinions and inquiries to the government. The government and related parties must overcome the mistrust, and assure mobile users that people's privacy is protected and that the information will not be sold to third parties.
- **Regulating and developing legal aspects of mobile applications and use of the services:** Many countries around the world have not yet adopted legislations for data and information practices, which spell out the rights of subjects (citizens) and the responsibilities of the data holders (government). In some cases, the law does not recognize mobile documents and transactions. There is yet no clear legal status for government's online publications, insufficient regulations and laws for online form filings, online signatures, and on online taxable transactions.

CONCLUSIONS

The recent developments in business models, services, and technologies of the WWW and Internet created new dimensions on the interactivity, mobility, and intelligence of Web-based solutions. As e-business evolves towards m-business (Sadeh, 2002), e-government seems to follow the trend with a few but significant mobile government (m-government) applications. Millions of mobile phone users, equipped with Internet connections, will put severe pressure on the government to extend appropriate e-government services into the mobile platform. It is now inevitable for e-government professionals, practitioners, and researchers to acquire necessary skills to face the new move toward m-government.

There already exist various m-government applications and business models in the areas of law enforcement, education, transport, health, and firefighting. M-government business models will typically follow an enhanced version of e-government models (Abramson & Means, 2001) where appropriate. We will see applications enabling governments to perform better:

1. in serving the citizens using mobile information and communication models;
2. in doing business with the citizens and other government and business organizations using mobile transactions models;
3. in integrating various government units and officials through mobile portals; and
4. in promoting active participation in the government affairs establishing m-democracy models.

The existing technological foundations, applications, and services support the idea that m-government will be a significant part of e-government

efforts. The policy makers and IT professionals need to get ready to embrace these developments and participate in the ways to enhance e-government activities through m-government.

As this brief overview suggests, m-government is in its early stage of development. The developments in e-business and m-business areas are influencing mobile technology adoption by governments. In parallel, the existing research in the m-government field often focuses a few applications (Easton, 2002) and mobile business issues as they relate to e-government (Holmes, 2001). There is now a growing need to examine m-government-related issues from the perspectives of their own and build a framework for the study of m-government efforts. Recently established Mobile Government Consortium International (MGCI—www.mgovernment.org) and the research resources site www.mGovlab.org are two leading groups aiming to build and guide the developments in mobile government research and practice. The interest to the m-government field is growing significantly from mainly three groups: IT and Telco's, government organizations and academia, and series of conferences (i.e., International Conferences on Mobile Government) are providing a forum for exchange of ideas among these groups.

REFERENCES

Abramson, M., & Means, E. G. (Eds.) (2001). *E-government 2001.* New York: Rowman and Littlefield Publishers, Inc.

Cilingir, D., & Kushchu, I. (2004). E-government and m-government: Concurrent leaps by Turkey. In D. Remenyi (Ed.), *Proceedings of European Conference on E-Government (ECEG 2004),* Trinity College, Dublin, June 17-18 (pp. 813-821). Department of the Taoiseach, Dublin, Ireland; Reading, UK: Academic Conferences International.

Easton, J. (2002). *Going wireless: Transform your business with wireless mobile technology.* USA: HarperCollins.

Kushchu, I., & Kuscu, H. (2003). From e-government to m-government: Facing the inevitable. In the *Proceeding of European Conference on E-Government (ECEG 2003),* Trinity College, Dublin, July 3-4 (pp. 253-260). Reading, UK: Academic Conferences International.

Holmes, D. (2001). *eGov: eBusiness strategies for government.* London: Nicolas Brealey.

Lanwin, B. (2002). *A project of info dev and The Center for Democracy & Technology: The e-government handbook for developing countries.* Retrieved February 15, 2004, from http://www.cdt.org/egov/handbook/2002-11-14egovhandbook.pdf

Sadeh, N. (2002). *M-commerce: Technologies, services and business models.* Canada and USA: John Wiley and Sons, Inc.

Yu, B., & Kushchu, I. (2004). The value of mobility for e-government. In the *Proceedings of European Conference on E-Government (ECEG 2004),* Trinity College, Dublin, June 17-18 (pp. 887-899). Department of the Taoiseach, Dublin, Ireland; Reading, UK: Academic Conferences International.

This work was previously published in Mobile Government: An Emerging Direction in E-Government, edited by I. Kushchu, pp. 1-11, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 1.23
Key Issues in Mobile Marketing:
Permission and Acceptance

Stuart J. Barnes
University of East Anglia, UK

Eusebio Scornavacca
Victoria University of Wellington, New Zealand

ABSTRACT

The growth and convergence of wireless telecommunications and ubiquitous networks has created a tremendous potential platform for providing business services. In consumer markets, mobile marketing is likely to be a key growth area. The immediacy, interactivity, and mobility of wireless devices provide a novel platform for marketing. The personal and ubiquitous nature of devices means that interactivity can, ideally, be provided anytime and anywhere. However, as experience has shown, it is important to keep the consumer in mind. Mobile marketing permission and acceptance are core issues that marketers have yet to fully explain or resolve. This chapter provides direction in this area. After briefly discussing some background on mobile marketing, the chapter conceptualises key characteristics for mobile marketing permission and acceptance. The chapter concludes with predictions on the future of mobile marketing and some core areas of further research.

INTRODUCTION

The proliferation of mobile Internet devices is creating an extraordinary opportunity for e-commerce to leverage the benefits of mobility (Chen, 2000; Clarke, 2001; de Haan, 2000; Durlacher Research, 2002; Evans & Wurster, 1997; Kalakota & Robinson, 2002; Siau & Shen, 2003; Yuan & Zhang, 2003). Mobile e-commerce, commonly known as m-commerce, is allowing e-commerce businesses to expand beyond the traditional limitations of the fixed-line personal computer (Barnes, 2002a; Bayne, 2002; Clarke, 2001; Lau, 2003; Siau & Shen, 2003; Sigurdson & Ericsson, 2003). According to a study by Telecom Trends

International (2003), global revenues from m-commerce could grow from $6.8 billion in 2003 to over $554 billion in 2008.

Mobile commerce has a unique value proposition of providing easily personalized, local goods and services, ideally, at anytime and anywhere (Durlacher Research, 2002; Newell & Lemon, 2001). Due to current technological limitations, some problems, such as uniform standards, ease of operation, security for transactions, minimum screen size, display type, and the relatively impoverished web sites, are yet to be overcome (Barnes, 2002b; Clarke, 2001).

As each mobile device is typically used by a sole individual, it provides a suitable platform for delivering individual-based target marketing. This potential can improve the development of a range of customer relationship management (CRM) tools and techniques (Seita, Yamamoto, & Ohta, 2002). It is believed that in the near future marketing through the mobile phone will be as common a medium as the newspaper or TV. However, mobile marketing is unlikely to flourish if the industry attempts to apply only basic online marketing paradigms to its use; the medium has some special characteristics that provide quite a different environment for ad delivery, including time sensitivity, interactivity, and advanced personalization. Moreover, a key tenet is likely to be that consumers receive only information and promotions about products and services that they want or need; one of the most important aspects to consider is that wireless users demand packets of hyperpersonalized information, not scaled-down versions of generic information (Barnes, 2002c). Sending millions of messages to unknown users (known as spam) or banner ads condensed to fit small screens (Forrester Research, 2001) are doubtless unlikely to prove ideal modes of ad delivery to a captive mobile audience.

This chapter aims to explore the peculiarities of mobile-oriented marketing, focusing on issues of permission and acceptance, and some of the possible business models. The following two sections provide a basic review of the technological platform for mobile marketing and an introduction to marketing on the mobile Internet (focusing on advertising), respectively. The fourth section presents a conceptual definition and model for permission on mobile marketing applications, while section five provides a model for mobile marketing acceptance and examines a number of possible scenarios for mobile marketing, based on the previous analysis. Finally, the chapter rounds off with some conclusions, and further research questions, and provides some predictions on the future of wireless marketing.

THE TECHNOLOGICAL PLATFORM FOR MOBILE MARKETING

Kalakota and Robinson (2002) define mobile marketing as the distribution of any kind of message or promotion delivered via a mobile handset that adds value to the customer while enhancing revenue for the firm. It is a comprehensive process that supports each phase of the customer life cycle: acquisition, relationship enhancement, and retention. A variety of technological platforms are available to support mobile marketing. Here we describe briefly some of the principal components. (For a more detailed discussion, see Barnes [2002b, 2002c].) The m-commerce value chain involves three key aspects of technology infrastructure:

- **Mobile transport.** Current networks have limited speeds for data transmission and are largely based on second-generation (2G) technology. These "circuit-switched" networks require the user to dial up for a data connection. The current wave of network investment will see faster, "packet-switched" networks, such as General Packet Radio Service (GPRS), which deliver data directly to handsets, and are, in essence, always connected. In the near future, third-generation

(3G) networks promise yet higher transmission speeds and high-quality multimedia.

- **Mobile services and delivery support.** For marketing purposes, SMS (a text-messaging service) and WAP (a proprietary format for Web pages on small devices) are considered the key platforms in Europe and the United States, with iMode (based on compact hypertext markup language or cHTML) and iAppli (a more sophisticated version of iMode based on Java) taking precedence in Japan (WindWire, 2000). For PDAs, "Webclipping" is often used to format Web output for Palm or Pocket PC devices.
- **Mobile interface and applications.** At the level of the handset and interface, the brand and model of the phone or PDA are the most important part of the purchase decision, with "image" and "personality" being particularly important to young customers (Hart, 2000).

The next section explores the possibilities and experiences of using wireless marketing on these technology platforms.

MARKETING ON THE WIRELESS MEDIUM

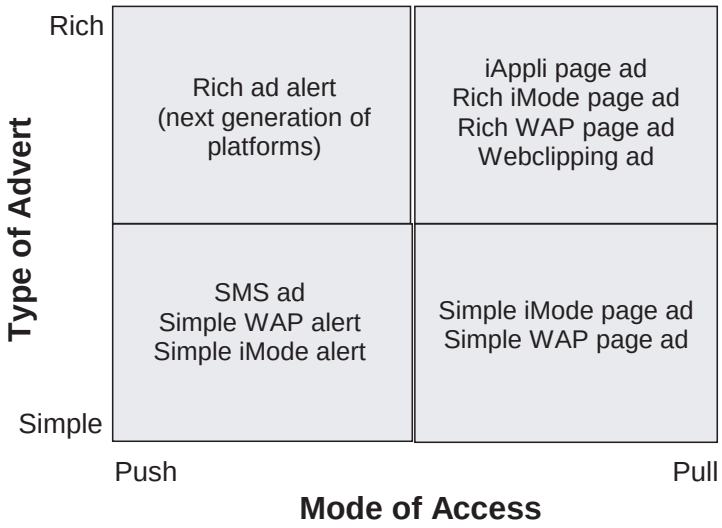
The wireless Internet presents an entirely new marketing medium that must address traditional marketing challenges in an unprecedented way (WindWire, 2000). Key industry players in the value chain providing wireless marketing to the consumer are agencies, advertisers, wireless service providers (WSPs), and wireless publishers. For agencies and advertisers, the wireless medium offers advanced targeting and tailoring of messages for more effective one-to-one marketing. For the WSP, the gateway to the wireless Internet (e.g., British Telecom, AT&T, and Telia-Sonera), wireless marketing presents new revenue streams and the possibility of subsidizing access. Similarly, wireless publishers (e.g., the *Financial Times, New York Times*, and CBS Sportsline), as a natural extension of their wired presence, have the opportunity for additional revenue and subsidizing access to content. At the end of the value chain, there is potential for consumers to experience convenient access and content value, sponsored by advertising (Kalakota & Robinson, 2002; WindWire, 2000).

Like the wired medium, marketing on the wireless medium can be categorized into two basic types: push and pull, which are illustrated in Figure 1. *Push* marketing involves sending or "pushing" advertising messages to consumers, usually via an alert or SMS (short message service) text message. It is currently the biggest market for wireless advertising, driven by the phenomenal usage of SMS—in December 2001, 30 billion SMS messages were sent worldwide (Xu, Teo, & Wang, 2003). An analysis of SMS usage has shown unrivalled access to the 15 to 24 age group—a group that has proved extremely difficult to reach with other media (Puca, 2001).

Pull marketing involves placing advertisements on browsed wireless content, usually promoting free content. Any wireless platform with the capacity for browsing content can be used for pull advertising. WAP and HTML-type platforms are the most widely used. Japan has experienced positive responses to wireless pull marketing, using iMode. Interestingly, wireless marketing in Japan has more consumer appeal than marketing on the conventional Internet. Click-through rates for mobile banner ads during the summer of 2000 averaged 3.6%, whilst those for wireless e-mail on iMode averaged 24.3%. Click-through rates for online banner ads on desktop PCs in Japan often average no more than 0.5 or 0.6% (Nakada, 2001).

Overall, current push services are very much in the lower left-hand quadrant of Figure 1. Until the availability of better hardware, software, and network infrastructure, services will remain basic. With faster, packet-based networks and more

Figure 1. Categorization of wireless marketing—with examples (Barnes, 2002c)



sophisticated devices, protocols and software, richer push-based marketing is likely to emerge, pushing the possibilities into the top left-hand quadrant.

PERMISSION ISSUES FOR MOBILE MARKETING APPLICATIONS

The discussion above has provided some insights about mobile marketing, particularly in terms of the wireless technological platform and basic applications of the medium. However, as yet, we have provided little conceptual discussion. The purpose of this section is to discuss the key variables of mobile marketing and present a conceptual model of permission for applications on this field.

In order for mobile marketing to reach its full potential of personalized information available anytime, anyplace, and on any device, it is necessary to understand the key characteristics of the mobile medium involved. We believe that any mobile marketing application should contemplate the following aspects:

- **Time and Location**. Although two different aspects, we consider them strongly related. An individual's behavior and receptiveness to advertisement is likely to be influenced by their location, time of day, day of week, week of year, and so on. Individuals may have a routine that takes them to certain places at certain times, which may be pertinent for mobile marketing. If so, marketers can pinpoint location and attempt to provide content at the right time and point of need, which may, for example, influence impulse purchases (Kannan, Chang, & Whinston, 2001). Feedback at the point of usage or purchase is also likely to be valuable in

building a picture of time-space consumer behavior.

- **Information**. In particular, data given a context by the user. By itself, data do not contain an intrinsic meaning. It must be manipulated appropriately to become useful. Therefore, information can be defined as the result of data processing, which possesses a meaning for its receiver. Murdick and Munson (1988) point out that quantity of data does not necessarily result in quality of the information. The most important thing is what people and organizations do with the information obtained and its ability of extraction, selection, and presentation of information pertinent to the decision-making process should be considered as a decisive factor.
- **Personalization.** One of the most important aspects to consider is that wireless users demand packets of hyperpersonalized information, not scaled-down versions of generic information (Barnes, 2002c). The nature of the user, in terms of a plethora of personal characteristics such as age, education, socio-economic group, cultural background and so on is likely to be an important influence on how ads are processed. These aspects have already proven to be important influences on Internet use (OECD, 2001), and as indicative evidence has shown above, elements such as user age are proving an important influence on mobile phone usage. The wireless medium has a number of useful means for building customer relationships. Ubiquitous interactivity can give the customer ever more control over what they see, read, and hear. Personalization of content is possible by tracking personal identity and capturing customer data; the ultimate goal is for the user to feel understood and simulating a one-to-one personal relationship. Through relational links of personal preferences, habits, mobile usage, and geographic positioning data the process of tailoring messages to individual consumers can become practical and cost effective.

The combination of the variables mentioned above allows us to understand one of the most important issues in mobile marketing: permission. Godin and Peppers (1999) refer to the traditional way of delivering marketing to customers as "interruption marketing." The authors suggest that instead of interrupting and annoying people with undesired information, companies should develop long-term relationships with customers and create trust through "permission marketing." The concept of permission marketing is based on approaching customers to ask for their permission to receive different types of communication in a personal and intimate way. It is well known among marketers that asking for a customer's permission is better and easier than asking for forgiveness (Bayne, 2002). In the wireless world, there is evidence to suggest that customers do not want to be interrupted—unless they ask to be interrupted (Newell & Lemon, 2001).

A mobile phone is a more personal environment than a mailbox or an e-mail inbox, and an undesired message has a very negative impact on the consumer (Enpocket, 2003; Godin & Peppers, 1999; Newell & Lemon, 2001). As mobile marketing has a more invasive nature than any other media, much attention must be given to permission issues in order to make the mobile marketing experience pleasant to the users. The information received must be of high value to gain the user's permission. It must produce a win–win situation between user and advertiser.

We understand permission as the dynamic boundary produced by the combination of one's personal preferences, that is, personalization, of time, location, and information. The user should be able to indicate when, where, and what information he/she would like to receive. Here are a couple examples of how mobile marketing can help consumers and businesses:

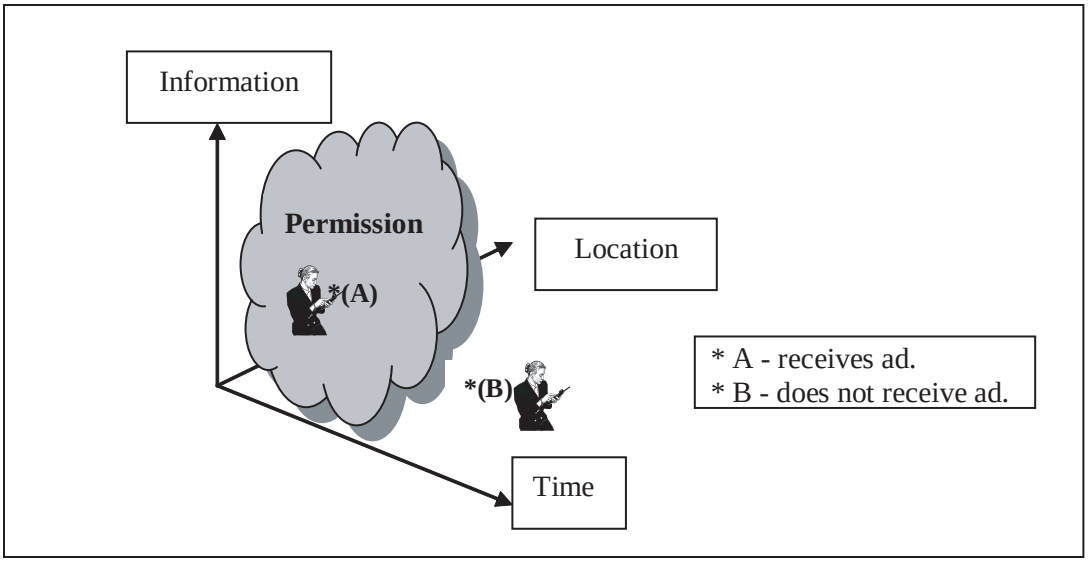
- You are getting ready to go to the airport and you receive a sponsored message saying that your flight is delayed for 4 hours. Because of this information, instead of spending 4 long and boring hours waiting at an airport lounge, you manage to have an enjoyable dinner with your friends.
- You let your wireless service provider know that you would like to receive during weekdays, from 12 p.m. to 1 p.m., information about the menu specials of all Italian restaurants costing less than $20 and within a 1-mile radius of where you are located.

Now, let us consider the situation if this information was not customer relevant, or time and location sensitive. For example, imagine the following scenario. You are on a business trip, it is 3:30 p.m. and you had to forgo lunch due to an important meeting. Next, your cell phone beeps and you receive an offer of a menu special of an unknown restaurant in your hometown. The value to the recipient of this information is zero; moreover, it is more likely to have a negative impact. Figure 2 helps us visualize the concept of permission on mobile marketing.

The idea of a message being sent directly to an individual's phone is not without legislative concerns. Indeed, all over the world, privacy and consumer rights issues lead to the promotion of "opt-in" schemes. In essence, "opt-in" involves the user agreeing to receive marketing before anything is sent, with the opportunity to change preferences or stop messages at any time. Several current initiatives and industry groups, such as the Mobile Data Association, are helping to build standards of best practice for the mobile data industry (MDA, 2003).

As permission for mobile marketing applications should be dynamic, it is important to be able to identify customer responses to events. Stemming from the technological capabilities of

Figure 2. Concept of permission for mobile marketing



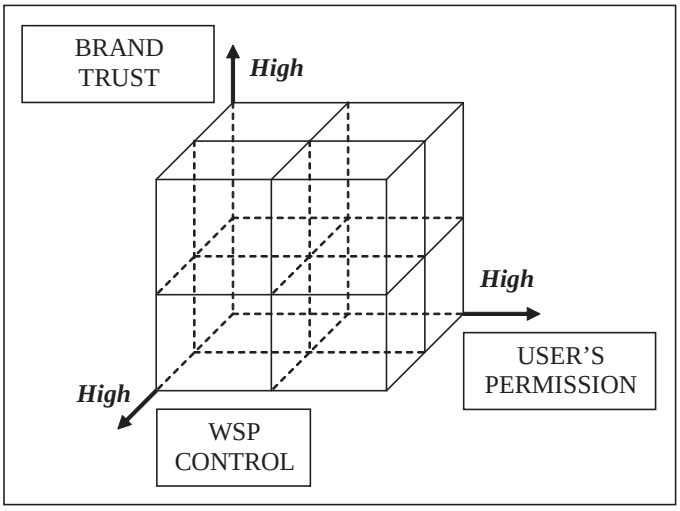
mobile Internet-enabled devices, the measurement of reaction marketing is facilitated. As a consequence, the planning and justification of marketing expenditure becomes more precise. It also will help the identification of which mobile marketing strategies work and which do not. The constant feedback permits marketing strategies to be dynamically adjusted to produce better results for marketers.

ACCEPTANCE OF MOBILE MARKETING

Now we have discussed the technological and conceptual factors surrounding mobile marketing, let us examine the variables that influence customer acceptance. Specifically, this section aims to explore the few studies already accomplished on mobile marketing acceptance and provide a model that summarizes the main variables concerning this issue.

There is no doubt that mobile marketing is still at an embryonic stage. However, several recent studies help us to understand some key factors contributing to the penetration and acceptance of mobile marketing among consumers (Enpocket, 2002a, 2002b; Ericsson, 2000; Godin & Peppers, 1999; Quios, 2000). The study by Ericsson (2000) had a sample of approximately 5,000 users and 100,000 SMS ad impressions in Sweden; the Quios study (2000) examined 35,000 users and 2.5 million SMS ad impressions in the UK; and the Enpocket study (Enpocket, 2002a, 2002b, 2003) researched over 200 SMS campaigns in the UK, surveying over 5,200 consumers—after they had been exposed to some of the SMS campaigns—from October 2001 to January 2003. The results of the three studies tend to converge, each pointing out that more than 60% of users liked receiving wireless marketing. The reasons cited for the favorable attitudes to mobile marketing include content value, immersive content, ad pertinence, surprise factor, and personal context.

Figure 3. Model for mobile marketing acceptance



The Enpocket study (2002a, 2002b, 2003) found that consumers read 94% of marketing messages sent to their mobile phones. It is important to point out that all these customers had given permission to receive third-party marketing. Moreover, the viral marketing capability of mobile marketing was identified by the fact that 23% of the customers surveyed by Enpocket showed or forwarded a marketing message to a friend. Another interesting finding is that the average response rate for SMS campaigns (15%) was almost three times higher than regular e-mail campaigns (6.1%). If delivered by a trusted source such as a wireless service provider (WSP) or major m-portal, acceptance of SMS marketing (63%) was considered comparable to that of TV (68%) or radio (65%). Notwithstanding, SMS marketing delivered by another source was far less acceptable—at just 35% of respondents. Similarly, the rejection level of SMS marketing from a WSP or portal was just 9%, while SMS from other sources was rejected by 31% of those surveyed. Telesales was rejected by 81% of respondents.

The indicative evidence about customer trust was further strengthened by other findings from the surveys. For example, 74% of customers indicated that WSPs were the most trusted organisation to control and deliver SMS marketing to their mobile devices. Major brands such as Coca-Cola and McDonald's were preferred by only by 20% of respondents (Enpocket, 2002a). As a result of the close relationship with the user, SMS marketing typically helps to build stronger brand awareness than other medias (Enpocket, 2002b).

It is important to highlight that the statistics presented above are being materialized in the form of profits mainly by mobile marketing and content sponsorship. Some marketers are using the sponsorship revenue model by conveying brand values through association with mobile content that fits the company's product or corporate image (Kalakota & Robinson, 2002). Features such as mobile barcode coupons are allowing a better measurement and understanding of return on investment (ROI) for mobile marketing (12Snap, 2003).

The indicative evidence and discussion above provide strong hints towards three main variables that influence a consumer's acceptance of mobile marketing: user's permission, WSP control, and brand trust. Figure 3 presents a conceptual model for mobile marketing acceptance based on these factors. Note that user permission is weighted in the model (see below).

The model allows us to forecast eight scenarios for mobile marketing acceptance. Table 1 summarizes the different scenarios. An example for scenario 1 would be if a trusted brand such as Coca-Cola sent a marketing message through the user's WSP (e.g., Vodafone) with his/her permission. In this situation, all the variables have a high level and the message should be highly acceptable to the customer. At the opposite end of the spectrum, in scenario 8, an unknown company (brand) sends a message without WSP control and without the user's permission. Here, the probability of rejection is very high.

Scenarios 4 and 5 point out an element that requires further detailed investigation. We believe that the most important variable in this model is "user permission." For example, if Coca-Cola sends a message via an operator to a user who has not granted permission (scenario 4), it should have a lower acceptance than a brand with low trust that sends a message without WSP control to a customer who granted permission. This assumption is supported by the fact that the great majority of the consumers interviewed by Enpocket (2002a) are fearful that SMS marketing will become comparable to e-mail marketing with high levels of unsolicited messages.

The scenarios presented above are based on literature and on secondary data from the three studies previously approached. It would be interesting in the near future to substantiate this conceptual grid with primary data.

WSP control can directly affect how mobile marketing business models are configured.

Based on the findings from the above analysis, we present two basic business models in which WSP control is the main differentiator (Figure 4). Figure 4a presents a model where the WSP has full control of the marketing delivery. On the other hand, Figure 4b shows a model where marketers can send messages directly to users without the control of the WSP.

The results of the studies presented by Ericsson (2000), Quios (2000), and Enpocket (2002a, 2002b, 2003) allow us to presume that the model presented by Figure 4a should be more successful than the one in Figure 4b. This assumption can also be supported by the fact that a WSP is usually more highly trusted by the consumers and possesses the technological capabilities to limit the delivery of messages. In addition, consumers interviewed by Enpocket (2002a) expressed a strong preference for the WSPs to become the definitive media owners and permission holders—possibly as a consequence of bad experiences with Internet marketing using nontargeted spam mail. Another issue to be taken into consideration is how WSP control can affect the revenue model for mobile marketing. In Figure 4a, the WSP can easily charge marketers for using its services, but in Figure 4b, this becomes a difficult task.

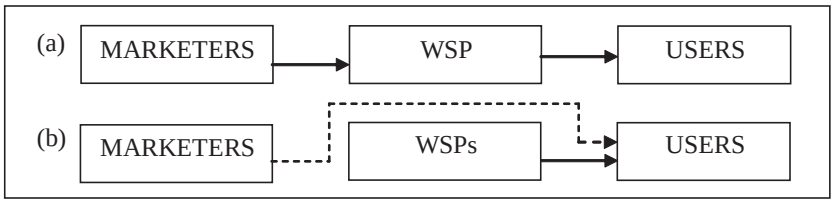
CONCLUSIONS

The immediacy, interactivity, and mobility of wireless devices provide a novel platform for marketing. The personal and ubiquitous nature

Table 1. Scenarios for mobile marketing acceptance

Scenario	Brand Trust	WSP Control	User's Permission	Acceptance
1	High	High	High	High acceptance
2	Low	High	High	Acceptable
3	High	Low	High	Acceptable
4	High	High	Low	Low acceptance
5	Low	Low	High	Acceptable
6	Low	High	Low	Low acceptance
7	High	Low	Low	Low acceptance

Figure 4. Possible business models for mobile marketing



of devices means that interactivity can be provided anytime and anywhere. Marketing efforts are potentially more measurable and traceable. Furthermore, technologies that are aware of the circumstances of the user can provide services in a productive, context-relevant way, deepening customer relationships. The convergence between marketing, CRM, and m-commerce represents a potentially powerful platform for wireless marketing.

Notwithstanding, it is important to keep the consumer in mind; the key to success is the management of and delivery upon user expectations. A key aspect of mobile marketing is likely to be obtaining permission from the users to send information to their mobile devices. Already, the wireless Internet has demonstrated the need for temperance; the wireless Internet is not an emulator of or replacement for the wired Internet, it is merely an additional, complementary channel for services. Further, aside from initial pilot investigations, it is not abundantly clear how consumers will respond to the idea of mobile marketing. Clearly, the issues concerning mobile marketing acceptance need to be further investigated. Alongside, a deeper investigation into business and revenue models is needed; for example, how can companies, marketers, WSPs, and consumers create a win–win environment? In addition, although it is expected that consumers will not tolerate receiving messages without permission, more work is still needed to explain how consumers give permission to receive mobile marketing.

Currently, wireless marketing is embryonic and experimental—the majority of wireless marketing is SMS based (simple push services—lower left-hand quadrant of Figure 1). The next generation of devices and networks will be important in the evolution of wireless marketing; higher bandwidth will allow rich and integrated video, audio and text. In addition, considerable effort is needed in building consumer acceptance, legislation for privacy and data protection, standardizing wireless ads, and creating pricing structures. If these conditions hold, wireless could provide the unprecedented platform for marketing that has been promised. Clearly, it is too early to tell, but future research aimed at examining these fundamental issues will help to further understand the implications of permission-based mobile marketing.

REFERENCES

12Snap. (2003). Mobile barcode coupons—The marketing revolution for marketeers. Retrieved May 18, 2003, from *www.12snap.com/uk/help/couponsshort.pdf*

Barnes, S.J. (2002a). Under the skin: Short-range embedded wireless technology. *International Journal of Information Management*, *22*(3), 165–179.

Barnes, S.J. (2002b). The mobile commerce value chain: Analysis and future developments. *International Journal of Information Management, 22(2),* 91–108.

Barnes, S.J. (2002c). Wireless digital advertising: Nature and implications. *International Journal of Advertising, 21*(3), 399–420.

Bayne, K.M. (2002). *Marketing without wires: Targeting promotions and advertising to mobile device users.* London: John Wiley & Sons.

Chen, P. (2000). Broadvision delivers new frontier for e-commerce. *M-commerce, October,* 25.

Clarke, I. (2001). Emerging value propositions for m-commerce. *Journal of Business Strategies, 18*(2), 133–148.

de Haan, A. (2000). The Internet goes wireless. *EAI Journal, April,* 62–63.

Durlacher Research. (2002). *Mobile commerce report.* Retrieved July 10, 2002, from *www.durlacher.com*

Enpocket. (2002a). Consumer preferences for SMS marketing in the UK. Retrieved March 13, 2003, from *www.enpocket.co.uk*

Enpocket. (2002b). The branding performance in SMS advertising. Retrieved March 13, 2003, from *www.enpocket.co.uk*

Enpocket. (2003). The response performance of SMS advertising. Retrieved March 13, 2003, from *www.enpocket.co.uk*

Ericsson. (2000). *Wireless advertising.* Stockholm: Ericsson Ltd.

Evans, P.B., & Wurster, T.S. (1997). Strategy and the new economics of information. *Harvard Business Review, 75*(5), 70–82.

Forrester Research. (2001). Making marketing measurable. Retrieved February 10, 2002, from *www.forrester.com*

Godin, S., & Peppers, D. (1999). *Permission marketing: Turning strangers into friends, and friends into customers.* New York: Simon & Schuster.

Hart, Peter D. (2000). *The wireless marketplace in 2000.* Washington, DC: Peter D. Hart Research Associates.

Kalakota, R., & Robinson, M. (2002). *M-business: The race to mobility.* New York: McGraw-Hill.

Kannan, P., Chang, A., & Whinston, A. (2001,). Wireless commerce: Marketing issues and possibilities. In *Proceedings of the 34th Hawaii International Conference on System Sciences*, Maui, HI.

Lau, A.S.M. (2003). A study on direction of development of business to customer m-commerce. *International Journal of Mobile Communications, 1*(1/2), 167–179.

Mobile Data Association (MDA). (2003). *Mobile Data Association.* Retrieved May 1, 2003, from www.mda-mobiledata.org/

Murdick, R.G., & Munson, J.C. (1988). *Sistemas de Informacíon Administrativa.* Mexico: Prentice-Hall Hispano Americana.

Nakada, G. (2001). *I-Mode romps.* Retrieved March 5, 2001, from *www2.marketwatch.com/news/*

Newell, F., & Lemon, K.N. (2001). *Wireless rules: New marketing strategies for customer relationship management anytime, anywhere.* New York: McGraw-Hill.

NTT DoCoMo. (2003). Sehin Rain-Apu. Retrieved March 13, from *http://foma.nttdocomo.co.jp/term/index.html* (in Japanese)

Organisation for Economic Co-operation and Development (OEC). (2001). *Understanding the digital divide*. Paris: OECD Publications.

Puca. (2001). Booty call: How marketers can cross into wireless space. Retrieved May 28 2001, from *www.puca.ie/puc_0305.html*

Quios. (2000). *The efficacy of wireless advertising: Industry overview and case study.* London: Quios Inc./Engage Inc.

Sadeh, M.N. (2002). *M commerce: Technologies, services, and business models.* London: John Wiley & Sons.

Seita, Y., Yamamoto, H., & Ohta, T. (2002). Mobairu wo Riyoushitari Aiaru Taimu Maaketingu ni Kansuru Kenkyu. In *Proceedings of the 8th Symposium of Information Systems for Society*, Tokyo, Japan.

Siau, K., & Shen, Z. (2003). Mobile communications and mobile services. *International Journal of Mobile Communications, 1*(1/2), 3–14.

Sigurdson, J., & Ericsson, P. (2003). New services in 3G—new business models for strumming and video. *International Journal of Mobile Communications, 1*(1/2), 15–34.

Telecom Trends International. (2003). M-commerce poised for rapid growth, says Telecom Trends International. Retrieved October 27, 2003, from *www.telecomtrends.net/pages/932188/index.htm*

WindWire. (2000). *First-to-wireless: Capabilities and benefits of wireless marketing and advertising based on the first national mobile marketing trial.* Morrisville, NC; WindWire Inc.

Xu, H., Teo, H.H., & Wang, H. (2003,). Foundations of SMS commerce success: Lessons from SMS messaging and co-opetition. In *Proceedings of the 36th Hawaii International Conference on Systems Sciences*, Big Island, HI.

Yuan, Y., & J.J. Zhang (2003). Towards an appropriate business model for m-commerce. *International Journal of Mobile Communications, 1*(1/2), 35–56.

NOTE

An earlier and shorter version of this paper appeared as Barnes, S. J., & Scornavacca, E. (2004). Mobile marketing: The role of permission and acceptance. *International Journal of Mobile Communications, 2*(2), 128–139.

This work was previously published in Unwired Business: Cases in Mobile Business, edited by S. Barnes and E. Scornavacca, pp. 96-108, copyright 2006 by IRM Press (an imprint of IGI Global).

Chapter 1.24
Dynamic Pricing Based on Net Cost for Mobile Content Services

Nopparat Srikhutkhao
Kasetsart University, Thailand

Sukumal Kitisin
Kasetsart University, Thailand

INTRODUCTION

In the past few years, the mobile phone's performance has increased rapidly. According to IDC's Worldwide Mobile Phone 2004-2008 Forecast and Analysis, sales of 2.5G mobile phones will drive market growth for the next several years, with sales of 3G mobile phones finally surpassing the 100 million annual unit mark in 2007. Future mobile phones can support more than 20,000 colors. With the advancements in functionality and performance of mobile phones, users will use them for all sorts of activities, and that will increase mobile content service requests. Currently, the pricing of mobile content service is up to each provider; typically they implement a fixed price called a market price because the providers do not have a formula to estimate the price according to the actual cost of their services. This article proposes a dynamic pricing model based on net cost for mobile content services.

BACKGROUND

A mobile phone today can support various format data causing mobile content service popularity among all mobile phone users. They can request a music VDO clip, a song, or a mobile phone game program. The price of each mobile content service differs for each different format of data. For example, the price of a true-tone ring tone is 35 baht (Sanook.com, 2005), while a Java game download costs 40 baht (Siam2you, 2005).

Conventionally, an operator set a fixed market price for each mobile content service. The prices can vary from operator to operator. The pricing has not been calculated based on the net cost for

the requested service. Therefore, the set price can be lower or much higher than the actual cost. To come up with a way for a provider to be able to set a mobile content service price based on its actual cost, the provider must be able to quantify its actual cost for service. This article presents mobile content service interaction models and formulas for estimating the actual cost of a mobile content service; a provider can refer to these models and formulas when pricing its services.

Data Formats

The previous section discusses improving the performance of mobile devices and the variety of content available. We can classify mobile content service into four types: audio, image, video, and application (ClearSky Mobile Media, 2005). Users can request an audio clip and use it as their ring tone. They can leave voice messages for each other or download an mp3 song for their entertainment (Nokia, 2005; Sony Ericsson, 2005; Samsung, 2005). The audio content can be of three sub-types: monophonic (Sonic Spot, 2005), polyphonic (Cakewalk, 2005), and true tone. Image format can be either static or dynamic/animation. Users can request a music VDO clip and play it on their mobile phones, and apparently, a few companies have started to provide NetTV on mobile phones as well. Lastly, an example of application content users widely request could be a Java game application.

Parties in Mobile Content Services

Providing mobile content services involves many parties. We consider the following five participants (Bratsberg & Wasenden, 2004; Andreas, 2001). The first party is a user requesting mobile content services. The second party is a mobile operator (MO), which is the owner of a mobile phone service frequency. When a user requests mobile content service, the user will send a request to his or her content provider through the MO's network. The third party is a content provider (CP), which is an organization to serve mobile contents. The MO may or may not have license on the contents. The fourth party is the content owner (CO), which could be a person or an organization that has authorization for legal distribution of the mobile contents. And the last party is a content aggregator (CA), a middleman between a user and a content provider. The CA can help increase the channel to serve mobile content services.

Request and Response Formats

When a user wants to use a mobile content service, he or she makes a request for the desired content from a CA or CP. Then the CA or CP responds to the user with the requested content. Requests and responses can be of the following four types: a Web request through a Web page, a short message service (SMS), an interactive voice responder (IVR), or a WAP request via a mobile internet WAP page. When the CA or CP responds successfully, the response can be sent using one of these four formats: a short message service (SMS), a smart message, a WAP push format, and a multimedia message service (MMS).

Mobile Content Service Models

We categorized all mobile content services into the following interaction models based on involved parties and content providing methods.

Model 1: Parties involved are user, MO, and CP. A user requests content from a CP by using an SMS, IVR, WAP, or Web request. For any request format except a Web request, the request is sent to the CP through the MO's network. For a Web request format, the request is transferred to the CP directly. After the CP processes the request, the CP will reply to the user with the requested content information in the form of a WAP push, a WAP URL, or a bookmark. For a monophonic ring tone request, the CP will send content information in a smart message format.

Other content formats can be replied to with an SMS, a WAP push, or an MMS. The workflow of model 1, as shown in Figure 1, is:

1. User requests content via a Web page.
2. User requests content via SMS, IVR, or WAP page.
3. MO forwards the request to CP.
4. CP sends the content information to user through MO.
5. MO forwards the content information from CP to user.
6. User connects to WAP page or open WAP push for retrieving the content file through MO.
7. MO redirects the file request to CP.
8. CP sends the content file to the user through network of MO.
9. MO transfers content file to the user.

Model 2: Parties involved are user, MO, CP, and CO, with the CP as the content file sender. A user sends a request to a CP in SMS, IVR, WAP or Web format. The CP, after receiving the request,

Figure 1. Mobile content service model 1

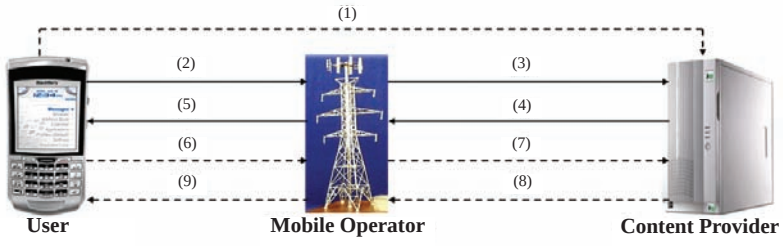
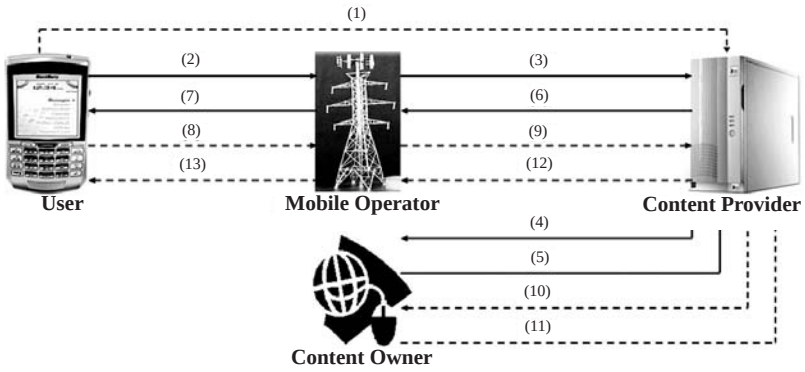


Figure 2. Mobile content service model 2



sends this request to a CO for content information. The CO sends content information to the user via the CP in smart message format, SMS (URL for retrieving content file), or a WAP push. After that, the CP forwards the information to user. The workflow of model 2, as shown in Figure 2, is as follows. Steps 1-3 are the same as in model 1. In step 4, the CP forwards the request to the CO. In step 5, the CO sends the content information to the user via the CP. Steps 6-9 of this model are the same as steps 4-7 of model 1. In step 10, the CP forwards the request to the CO. In step 11, the CO sends the content file to the user via the CP. In step 12, the CP sends the content file to the user through the network of the MO. In step 13, the MO transfers the content file to the user.

Model 3: Parties involved are user, MO, CP, and CO, with the CO as the content file sender. The CO has permission to distribute the content files. The CP is a middleman between the CO and the user. The user requests content from the CP, which then sends the request to the CO. The CO processes the request and sends the content file directly to the user. The workflow for model 3 is: steps 1-10 are the same as steps 1-10 of model 2. In step 11, the CO sends the content file to the user through the network of the MO. In step 12, the MO transfers the content file to the user.

Model 4: Parties are user, MO, CA, CP, and CO, with the CA as the content file sender. For model 4, there is a middleman between the user and the CP. When the user requests a service, the user sends a request to the CA. Then the request is forwarded to the CP and the CO. After the CO processes the request, the content file will be sent to the user via the CP and the CA. The workflow of model 4 is as follows:

1. User requests content via a Web page.
2. User requests content via SMS, IVR, or WAP page.
3. MO forwards request to CA.
4. CA forwards request to CP.
5. CP forwards request to CO.

6-9. CO sends content information to user via CP, CA, and MO.

10-13. User connects to WAP page or open WAP push for retrieving the content file through MO, CA, and CP.

14-17. CO sends the content file to user via CP, CA, and MO.

Figure 3. Mobile content service model 3

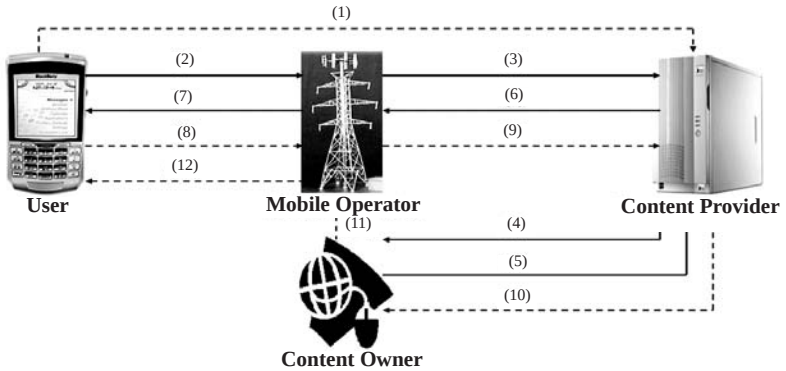


Figure 4. Mobile content service model 4

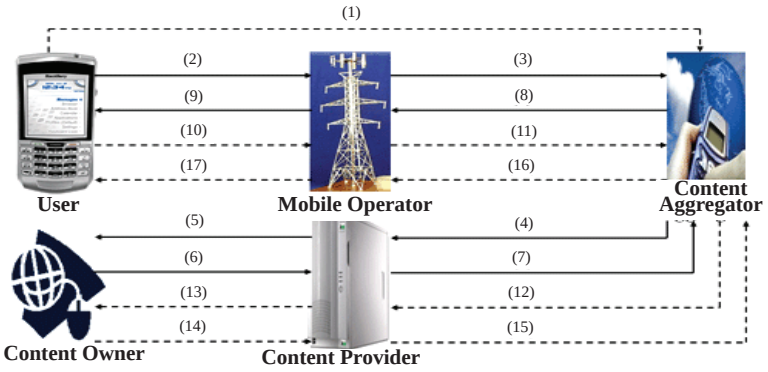
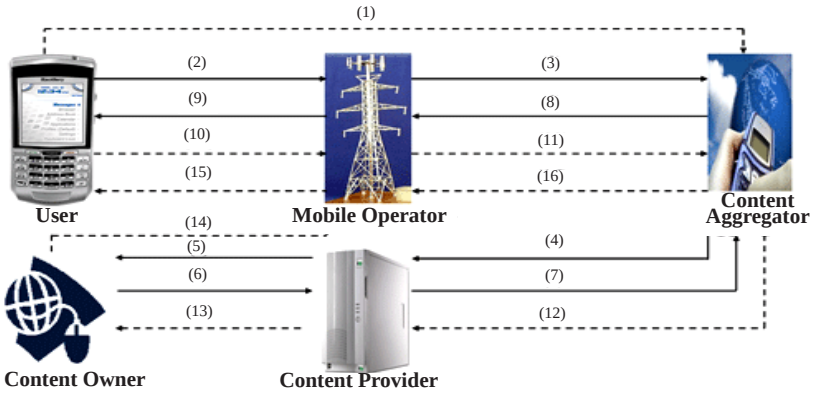


Figure 5. Mobile content service model 5



Model 5: Parties involved are user, MO, CA, CP, and CO, with the CO as the content file sender. Methods for requesting content in this model are the same as those of model 4. They can be via an SMS, IVR, or WAP request. Requests will be sent via the network of the MO. Another option is that a request can be sent to the CA directly. When the CO finishes processing the request, the CP then sends the content file directly to the user. The workflow of model 5 is as follows. Step 1-13 are the same as those in model 4. In step 14, the CO sends the content file to the user through the network of the MO. In step 15, the MO transfers the content file to the user.

Cost of Mobile Content Service

To be able to calculate the actual cost of a service, the providers must know the actual cost of providing the service. Two factors contributing to the

cost of providing a mobile content service are the cost of software or content for the content service and the operational cost. Each party has a different operational cost and pays a different content fee or has a different revenue sharing model for a mobile content service (Smorodinsky, 2002; Kivisaari & Luukkainen, 2003; Stiller, Reichl, & Leinen, 2001). Operating costs for the MO are mainly the bandwidth cost for sending the content to the user. For the CP, the operating costs can be the cost of the bandwidth, the operation cost, the revenue sharing or fee for the MO, and the revenue sharing or fee for the CO. For the CO, its operating costs are from the cost of the bandwidth and the operation cost. And the CA's operating costs come from the cost of the bandwidth, the operation cost, the revenue sharing or fee for the MO, and the revenue sharing or fee for the CP.

Formula for Calculating an Actual Cost

Formulas depend on the mobile content service interaction models, the format of the mobile content, and its transfer venues. Parameters used in the formulas are as follows:

- S refers to the software value or value of a content file.
- I_A refers to the operation cost of CA.
- I_P refers to the operation cost of CP.
- I_O refers to the operation cost of CO.
- I refers to the total operation cost.
- B refers to content file size (Bit).
- D_M refers to bandwidth cost per bit for MO.
- D_A refers to bandwidth cost per bit for CA.
- D_p refers to bandwidth cost per bit of CP.
- D_O refers to bandwidth cost per bit of CO.
- A refers to bandwidth cost for CA.
- P refers to bandwidth cost of CP.
- O refers to bandwidth cost of CO.
- M refers to bandwidth cost of MO.
- C refers to the sum of bandwidth cost for CP and MO.
- R_1 refers to revenue sharing for MO.
- R_{2_1} refers to content fee for CO.
- R_{2_2} refers to revenue sharing for CO.
- R_{3_1} refers to fee for CP.
- R_{3_2} refers to revenue sharing for CP.
- W refers to the sum of dynamic costs before calculation revenue sharing and content fee for CO.
- E refers to the sum of dynamic cost before calculation revenue sharing and content fee for CP.
- T refers to dynamic cost before calculate revenue sharing for MO.
- N refers to the actual cost.

Formula 1: for model 1:

$I_P = I$, $B^* D_p = P$, $B^* D_M = M$, $P+M = C$, $S+I+C = T$, $T /(1-R_1) = N$

Formula 2: for model 2:

$I_P + I_O = I$, $B^* D_O = O$, $B^* D_p = P$, $B^* D_M = M$, $(O+P+M) = C$,

$W + R_{2\_1} = T$ *OR* $W / (1 - R_{2\_2}) = T$, $T/(1- R_1) = N$

Formula 3: for model 3:

$I_P + I_O = I$, $B^* D_O = O$, $B^* D_M = M$, $O+M = C$, $S+I+C = W$,

$W + R_{2\_1} = T$ *OR* $W / (1 - R_{2\_2}) = T$, $T/(1- R_1) = N$

Formula 4: for model 4:

$I_A + I_P + I_O = I$, $B^* D_O = O$, $B^* D_p = P$, $B^* D_A = A$, $B^* D_M = M$, $(O+P+A+M) = C$

$S+I+C = W$, $W + R_{2\_1} = E$ *OR* $W /(1- R_{2\_2}) = E$, $E + R_{3\_1} = T$ *OR* $E / (1- R_{3\_2}) = T$

$T / (1– R_1) = N$

Formula 5: for model 5 :

$I_A + I_P + I_O = I$, $B^* D_O = O$, $B^* D_M = M$, $O+M = C$, $S+I+C = W$,

$W + R_{2\_1} = E$ *OR* $W /(1- R_{2\_2}) = E$, $E + R_{3\_1} = T$ *OR* $E / (1- R_{3\_2}) = T$

$T / (1– R_1) = N$

RESULTS AND ANALYSIS

We analyzed our actual cost formulas presented above by doing experiments based on three different types of transmitting channels: ADSL, leased line, and satellite. Cost for transmitting file content is determined by an average rate from ISPs in Thailand (True Internet, 2005; LOXINFO, 2005; Internet KSC, 2005; Ji-NET, 2005; INET, 2005). Each party uses the same sending channel. For the operation cost, we randomly selected a cost. The random method is Gaussians, with a base cost of 0.8262 and deviation of 0.14711. The software/content cost is a randomly selected value ranging from 0.1058 to 2.1160 baht. The sending channel of our results is satellite. Figures 6 and 7 show the average actual costs of true tone. The average operation costs for CA, CP, and CO is 0.829624 baht. Figures 8 and 9 shows the monophonic actual cost. And the average operation cost for CA, CP, and CO is 0.821213 baht.

Figure 6 shows average costs of true tone content for all service models and the market price. The average costs of models 1-5 are 1.66, 6.66, 6.64, 20.02, and 19.95 respectively. The maximum actual costs are less than market about 0.6 times. Thus, we found the market price for true tone content overpriced.

Figure 7 shows the probability of the customer being willing to pay the cost price of true tone content. For Figure 7, we found customers almost willing to pay the cost price of model 1 and customers willing to pay about 60% of the market price.

Figure 8 shows the average costs for a monophonic through satellite compared with the market price. From this figure, we found the market

Figure 6. Average actual cost for a true tone content service

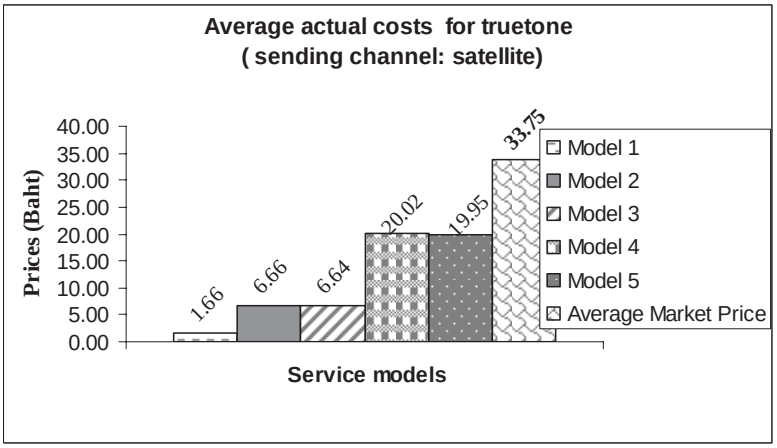


Figure 7. Probability of customer willing to pay the cost price of true tone content

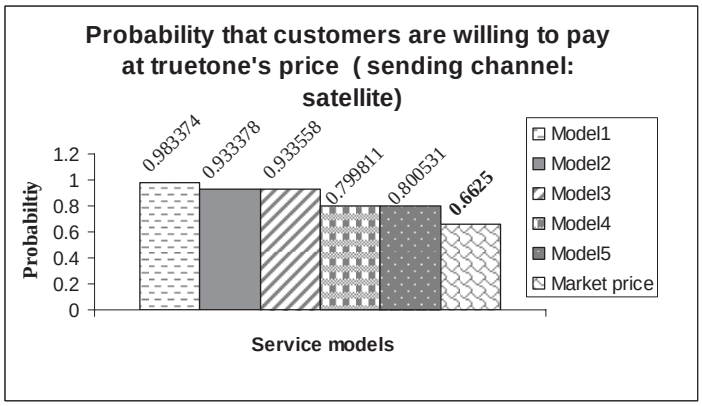


Figure 8. Average actual cost for a monophonic

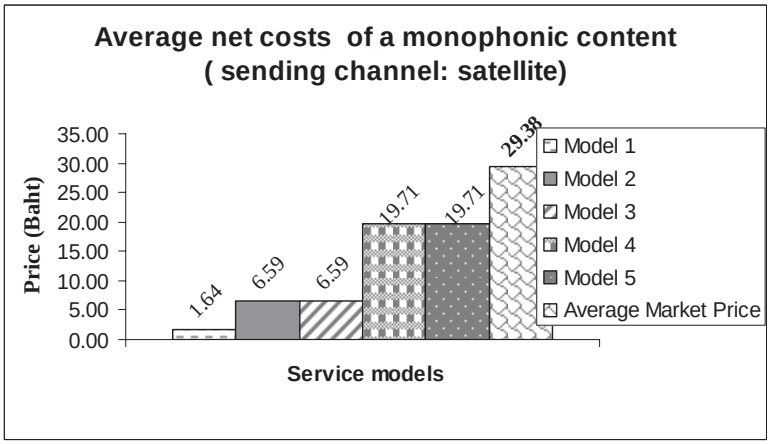
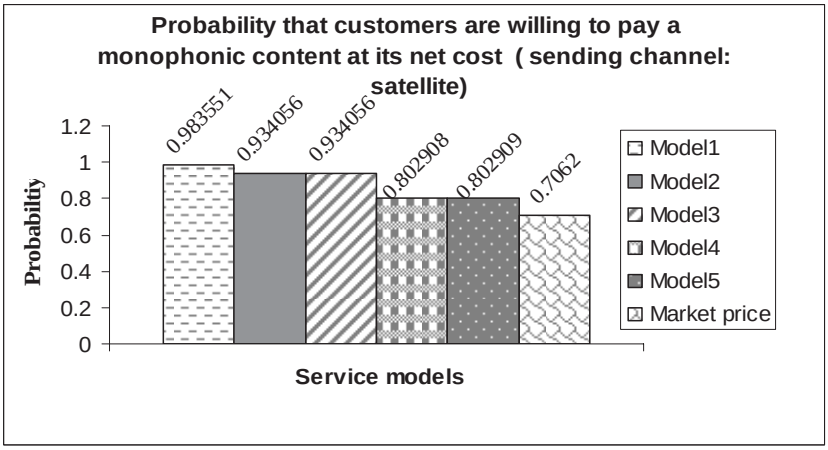


Figure 9. A probability that customers are willing to pay for monophonic content



price is more than the actual costs for all models. Market prices are very expensive. It is about 18 times more than model 1's actual cost.

Figure 9 shows the probability that the customers are willing to pay for actual costs. We found customers will gladly pay the cost prices for model 1, model 2, and model 3. Also, we found the probability of customers willing to pay market price is less than the probability of model 1 by nearly 30%.

SUMMARY

From our preliminary results using our formulas, the size of file, number of parties, and the sender affect the calculation of the actual cost. We presented an alternative method in pricing the mobile content services based on the actual cost. The method allows a provider to dynamically set its service price accordingly.

REFERENCES

Andreas S. (2001). *The digital content network receiver service market.* White Paper, HTRC Group, Canada. Retrieved March 17, 2005, from http://www.htrcgroup.com/pdffiles/dcnr.pdf

Bratsberg, H., & Wasenden, O. (2004, September). *Changing regulation – Impacts on mobile content distribution*. Retrieved March 16, 2005, from http://web.si.umich.edu/tprc/papers/2004/373/bratsberg_wasenden_tprc04_mobile_content_distribution_final.pdf

Cakewalk. (2005). *Desktop music handbook: Glossary of MIDI and digital audio terms.* Retrieved August 28, 2005, from http://www.cakewalk.com/tips/desktop-glossary.asp

ClearSky Mobile Media. (2005). Retrieved July 13, 2005, from http://www.clearskymobilemedia.com/carriersol/encont.asp

INET. (2005). *Always by your side*. Retrieved August 25, 2005, from http://www.inet.co.th

Internet KSC. (2005). Retrieved August 25, 2005, from http://www.ksc.net

Ji-NET. (2005). Retrieved August 25, 2005, from http://www.ji-net.com

Kivisaari, E., & Luukkainen, S. (2003, March). Content-based pricing of services in the mobile Internet. *Proceedings of the 7th IASTED International Conference on Internet and Multimedia Systems and Applications.* Retrieved March 15, 2003, from http://www.tml.tkk.fi/~sakaril/Content-based_pricing.pdf

LOXINFO. (2005). Retrieved August 25, 2005, from http://www.csloxinfo.co.th/

Nokia. (2005). Retrieved August 20, 2005 from http://www.nokia.co.th/nokia/0,,51297,00.html

Samsung. (2005). *Digital world*. Retrieved August 23, 2005, from http://product.samsung.com/cgi-bin/nabc/product/b2c_product_type.jsp?eUser=&prod_path=/Phones+and+Fax+Machines%2fWireless+Phones

Sanook.com. (2005). Retrieved June 10, 2005, from http://mobilemagic.sanook.com

Siam2you. (2005). Retrieved June 10, 2005, from http://www.siam2you.com

Smorodinsky. R. (2002). *Mobile entertainment – A value chain analysis and reference business scenario.* Retrieved from http://www.fing.org/ref/upload/GlobalCommunicationsrevised.pdf

Sonic Spot. (2005). *Glossary.* Retrieved September 1, 2005, from http://www.sonicspot.com/guide/glossary.html

Sony Ericsson. (2005). *Products.* Retrieved August 18, 2005, from http://www.sonyericsson.com/spg.jsp?cc=global&lc=en&ver=4001&template=pg1_1&zone=pp

Stiller, B., Reichl, P., & Leinen, S. (2001, March). Pricing and cost recovery for Internet services: Practical review, classification and application of relevant models. *NETNOMICS: Economic Research and Electronic Networking, 3*(1). Retrieved January 12, 2005, from http://userver.ftw.at/~reichl/publications/NETNOMICS00.pdf

True Internet. (2005). Retrieved October 7, 2005, from http://www.asianet.co.th/home.htm

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 220-226, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.25
A Technology Intervention Perspective of Mobile Marketing

Dennis Lee
The University of Queensland, Australia, & The Australasian CRC for Interaction Design, Australia

Ralf Muhlberger
The University of Queensland, Australia, & The Australasian CRC for Interaction Design, Australia

INTRODUCTION

In the last decade, the explosive growth and adoption of mobile phones has become commonplace in our everyday lives (Haghirian, Madlberger, & Tanuskova, 2005). In 1997, there were only 215 million people worldwide who used mobile phones as communication devices (Bauer, Barnes, Reichardt, & Neumann, 2005). Today, it is estimated that 2 billion people own a mobile phone worldwide and this number makes up a third of the entire human population (Wireless Intelligence, 2005).

Mobile phones are no longer thought of as mere personal communication tools (Cheong & Park, 2005; Ito & Okabe, 2005). They have become a fashion symbol for teenagers and young adults (Katz & Sugiyama, 2005). Personalised ring tones, colours, display logos and accessories are individualised accordingly to suit individuals' preferences (Bauer, Barnes, Reichardt, & Neumann, 2005). Furthermore, mobile phones are no longer just a platform for voice calls and sending and receiving text messages such as short messaging service (SMS). Photos, pictures and video clips can be attached as a multimedia message service (MMS) for communication purposes too (Okazaki, 2005a). With the recent introduction of 3G mobile technology, mobile phone users are able to perform more activities via their 3G enabled phone sets. They are able to browse the Internet fairly quickly, access online banking, play video games wirelessly, watch television programs, check for weather forecasts, allow instant messaging, and perform live video-conferencing (Okazaki, 2005b).

The rapid growth of the mobile industry has created a foundation for mobile commerce (m-

commerce). M-commerce facilitates electronic commerce via the use of mobile devices to communicate and conduct transactions through public and private networks (Balasubramanian, Peterson, & Jarvenpaa, 2002). The current emerging set of applications and services that m-commerce offers include mobile financial applications, mobile entertainment and services, product locating and shopping, wireless engineering, mobile auctions, wireless data centres and mobile advertising (Malloy, Varshney, & Snow, 2002). Commercial research has indicated that consumers' interest in m-commerce services and mobile payments have increased from 23% in 2001 to 39% in 2003 (Harris, Rettie, & Cheung, 2005). It is projected that by 2009 the global mobile commerce market will be worth at least US$40 billion (Juniper Research, 2004).

Considering the projected worth of mobile commerce and the number of mobile subscribers, mobile marketing is increasingly attractive, as companies can now directly convey their marketing efforts to reach their consumers without time or location barriers (Barnes, 2002). The potential of using the mobile medium to market is now more attractive than before (Karjaluoto, 2005), as it can assist companies in building stronger relationships with consumers (Barwise & Strong, 2002), and can be used as a promotional channel to reach consumers directly (Barnes, 2002; Kavassalis, Spyropoulou, Drossos, Mitrokostas, Gikas, & Hatzistamatiou, 2003; Okazaki, 2004) anywhere and anytime.

However, many aspects of mobile marketing are still in its infancy (Bauer, Barnes, Reichardt, & Neumann, 2005; Haghirian, Madlberger, & Tanuskova, 2005; Okazaki, 2004, 2005b; Tsang, Ho, & Liang, 2004). Research into mobile marketing is currently lacking, as this is a relatively new phenomenon. Very few studies have been conducted to demonstrate how the mobile phone channel can be successfully integrated into marketing activities of companies (Balasubramanian, Peterson, & Jarvenpaa, 2002; Haghirian, Madlberger, & Tanuskova 2005). Furthermore, no studies to date have compared the effectiveness of this mobile medium in delivering advertising and sales promotion with other more established media such as the print medium.

The fundamental question that remains unresolved is, "What is the difference between mobile marketing and traditional marketing?" Will this new form of marketing be effective? How will consumers respond to this form of marketing? What will be the benefit to marketers when consumers receive this type of advertising? These are just some of the issues that marketers are concerned with in order to evaluate the mobile channels for marketing purposes and are questions that are core to computer-supported collaborative work (CSCW) and technology intervention research.

MOBILE MARKETING

Mobile marketing via SMS-based advertising and sales promotions is now being carried out by several multinational corporations (MNCs) in Europe and the United States of America. MNCs are very cautious in integrating such a new medium into their marketing mix (Mayor, 2005; Okazaki, 2005a). This is mainly because marketers are not fully convinced of the value of mobile channels as a marketing tool (Haghirian, Madlberger, & Tanuskova, 2005). Marketers are unsure whether their marketing efforts will cause positive or negative impacts on their consumers.

Another issue is the difference in worldwide telecommunication networks and mobile handsets used in the last decade (Leppaniemi & Karjaluoto, 2005). The recent introduction of 3G mobile technology as a worldwide standard for telecommunication networks and mobile handsets has brought about a new level of investment safety for companies (Karjaluoto, 2005). Companies are beginning to test their marketing efforts via the mobile phone medium (Cheong & Park, 2005). This suggests the need for researchers to develop

theories and models to inform how mobile marketing can work effectively in the mobile phone context (Karjaluoto, 2005).

According to Tsang, Ho, and Liang (2004), mobile marketing can be classified as either permission-based, incentive-based or location-based. Permission-based marketing requires mobile users' prior approval before specific marketing messages can be sent (Barwise & Strong, 2003). By getting the permission of the mobile users, the factor of irritation may be reduced when users read the advertisement. Incentive-based marketing provides specific rewards to individuals who agree to receive promotions (Tsang, Ho, & Liang, 2004). For instance, mobile phone users may get free connection time from their mobile service providers for retrieving and reading advertisements. Location-based marketing targets mobile users in a certain location. The advantage of location-based marketing is that advertisements are sent to those individuals who are present or near the location (Barnes, 2003).

The incentive-based marketing approach is adopted because most consumers perceive the current mobile marketing as advertising, without making a distinction between sales promotions and advertising messages (Gogus, 2004). In other words, most consumers will generally term any marketing message received on their mobile phones as an advertisement, regardless of content (Gogus 2004). Moreover, the dominant form of mobile "advertising" appears to be in the form of promotion (Kavassalis, Spyropoulou, Drossos, Mitrokostas, Gikas, & Hatzistamatiou, 2003; Haghirian, Madlberger, & Tanuskova, 2005; Mayor, 2005; Okazaki, 2004; Tsang, Ho, & Liang, 2004).

In the marketing literature, a sales promotion can be defined as a more direct form of persuasion that may offer incentives to stimulate immediate purchase behaviour (Rossiter & Percy, 1998). Examples of sales promotional incentives include coupons, on-pack promotions, bonus packs, samples, premiums, and sweepstakes (Rossiter & Bellman, 2005; Shimp, 2003). Most of these promotional tools are based in print and termed as traditional promotional incentives (Belch & Belch, 2004).

On the other hand, advertising can be defined as a relatively indirect form of persuasion that may cause a favourable mental impression and then create an inducement toward a purchase response (Rossiter & Percy, 1998). Advertising is considered as the placement of a message to either increase product awareness, promote sales of goods and services, or just disseminate information (Leppaniemi & Karjaluoto, 2005). Advertisements may also include the element of sales promotion, a common example of which is in the form of coupons.

Coupons are considered to be some types of inducement that provide extra incentives to buy (Belch & Belch, 2004). Thus, in the context of mobile promotion, a mobile coupon is defined as an incentive that is paperless and electronic in nature (Wehmeyer & Müller-Lankenau, 2005). It is the fusion of the traditional print-based coupon with the mobile phone medium. A mobile coupon is delivered to a mobile phone handset as a message and is associated with mobile services and contents (Wehmeyer & Müller-Lankenau, 2005).

INTERACTION DESIGN AND THE LOCALES FRAMEWORK

The Locales Framework is a comprehensive theoretical CSCW and interaction design framework in the field of information and computer science (Fitzpatrick, 2003). According to Fitzpatrick, Kaplan, & Mansfield (1998), this research framework is an approach that allows for the creation of shared abstractions among stakeholders (e.g., companies, individuals, consumers, marketers), and also to narrow the gap between social and computing concerns with a common language. Understanding the social phenomenon and designing a relevant application that can fit the

social setting are the two important factors when applying the Locales Framework. It is the aim of Locales Framework analysis that more pragmatic design and systems applications are built to suit the social world (Fitzpatrick, 2003).

The Locales Framework is based on five aspects, each of which are interdependent and overlapping, as they share various concerns with one another and are used to approach the domain to be studied from different perspectives—rather than separating the domain into distinct subdomains to be studied independently.

The *locale foundation* aspect portrays the social world and the locale it uses for its interaction (Fitzpatrick, 1998, p. 91). The social world can be characterised by a number of issues such as collective goal, memberships, duration, structure, culture and roles. A locale is the primary unit of analysis in the Locales Framework. A locale consists of the site and means that a social world uses in its pursuit of the shared purposes. According to Fitzpatrick (2003), a site is a place the social world uses and means are the objects within this place. The social world needs sites and means to facilitate their shared interactions.

The *civic structure* aspect takes the locale of interest and considers its relationships and interactions with the wider community (Fitzpatrick, 1998, p. 92). In other words, it concerns the facilitation of interaction with the wider community within and beyond a person's known social worlds and locales. The interaction with a wider community can possibly relate to an environment that is physical, spatial, geographical, organizational, informational, professional, legislative, and so on.

The *individual view* aspect describes an individual's single perspectives on one social world as well as on multiple social worlds (Fitzpatrick, 1998, p. 115). A single perspective is how an individual sees one social world, and is dependent on the level of engagement with the centre of that world, whereas multiple view sets incorporate the individual's views of all the social worlds with which he or she is engaged. Individuals personalize their views to suit their tasks according to their current level of engagement.

The *interaction trajectory* aspect identifies the dynamic, temporal aspects of the social world in action (Fitzpatrick, 1998, p. 122). This aspect identifies the actual interactions individuals have over time within the setting and with each other. Moreover, this aspect is not only concerned with the current action, but also with the past and projected futures. Awareness of past actions and outcomes, present situations, and visions for the future are important for creating plans and strategies. An important consideration to understand this aspect is to look at what perspective or point of view is applied to any particular domain.

The *mutuality* aspect is a collaborative activity that draws specific attention to how the locale supports presence, and how awareness of that presence is supported for the achievement of shared activity. The mutuality aspect enables questions on who, what, when, where, why and how to be answered.

When the Locales Framework is used, it involves a two-phase approach. This is iterative in order to better understand the nature of the given (Fitzpatrick, 1998). The first phase is to understand the current locales of interest from the view of the interaction needs. This could involve using qualitative data collected through an ethnographic study or a one-to-one interview. Generally the data collected could then help to provide some relevant structure to designers when they engage in the design process of an application. It is argued that for designers who do not have any social science background, the Locales Framework could be applied as a sensitizing device to aid in formulating initial questions for the design process (Fitzpatrick, 1998).

The second phase in applying this framework is to evolve new locales. The goal is to discover more possibilities for the existing locale of interest in order to better support the activities that take place there and to explore possible newer locales that can evolve as a result. This phase is to identify

the advantages of any available medium, physical or computational, and the synergy among them, so that the needs of the social world are better meet. Specific questions that will help to drive this phase include: What interaction needs does the social world need that are lacking in this current locale? How can the existing locale be enhanced to support the aspects of the Locales Framework; namely, mutuality, individual views, civic structure and interaction trajectory? Can new technology be applied to the locale? Can new social worlds evolve if the resources are used in newer locales?

MOBILE MARKETING AS TECHNOLOGY INVENTION

Prior research has identified the importance of coupons in affecting consumers' cognitive, affective and conative behaviour during promotional campaigns (Raghubir, Inman, & Grande, 2004), but relatively little research has been conducted into the use of electronic coupons (Fortin, 2000; Suri & Swaminathan, 2004), particularly the form of mobile coupons (Okazaki, 2004). Most research in coupon studies is based mainly in the traditional medium of print (Coyle & Thorson, 2001; Liu & Shrum, 2002).

Much of the current literature that has been mentioned is adapted purely from a marketing perspective. Since mobile marketing involves people, technology and applications, mobile marketing should also be investigated from a human-computer interaction (HCI) and CSCW perspective. This will perhaps provide a better understanding of how and what is best for mobile marketing.

To better design mobile marketing strategies from a technological viewpoint, the use of the Locales Framework can be applied. The five interdependent characteristics of the Locales Framework guide study of the product or service to be marketed. An example may be a coffee shop:

Locale Foundation

The social world will be portrayed by consumers trying to buy beverages or food at the cafes and the locale is the cafe. The means in this case will be the chairs, tables, coffee machines, coffee counters, and the cashier's machine found within this site. The new technologies, that is, mobile phones and systems to send mobile coupons, are also included. More broadly the cafe may be situated in a locale such as a shopping centre or University that has its own means.

Civic Structure

The civic structure aspect considers issues such as the physical location of the café in its broader situation, store layout, and any competitors of café.

Individual View

In the case of mobile marketing, the individual that comes into the café will be a consumer and thus his or her task may be to purchase a cup of beverage for enjoyment. When they leave the café, their perspective may change to follow the priority that they may have to engage in. Perhaps the perspective may change to acquire knowledge and thus attend lesson at a lecture theatre or maybe need to catch a ride home by become a passenger when boarding a public transport such as a bus.

Interaction Trajectory

In this case study, the interaction trajectory aspect will determine the objective of the consumer coming to the café and how does the consumer interact with the surrounding environment. Despite receiving a discount coupon for cheaper beverages via the mobile phone, the consumer may come into the café with the purpose of meeting someone and not buy any beverages at all. To this consumer, the café has become a meeting venue

and not a place for consumption. The café may become a place for taking a coffee break with fellow colleagues, and therefore the consumer may take advantage to purchase a cup of coffee at the special price for enjoyment.

Mutuality

In this case scenario, the mutuality factor will look into how mobile marketing is supported and how applications can be created to support mobile marketing in the context of a café.

Several advantages of the Locales Framework, as according to Fitzpatrick (1998, pp. 152,153), are listed as follow:

- It provides a common tool for understanding and designing of a social problem.
- It has the potential to analyse issues from group to individual level, local setting to global context, and structure to process;
- It is independent of any one theoretical orientation when investigating into one phenomenon.
- It is a framework that is strong in identifying key elements of a collaborative environment but sufficiently generic, open and incomplete so as not to prescribe nor circumscribe all that is of interest.

Applying the Locales Framework approach to mobile marketing, we are able to build several "locales" (which can be defined as potential social world scenarios) in helping companies to consider before they actually implement their marketing plan. Moreover, companies that implement mobile marketing should consider the aspect of civic structure—the facilitation of interaction among various factors like physical, informative, geographical and technological parties. In the context of mobile marketing, companies should look at who their telecommunication service providers are, where their potential consumers are located, what applications should be used to generate response from consumers and what types of mobile phones should be able to receive mobile marketing. Companies need to understand that their potential consumers have many different perspectives and opinions, another aspect considered in the Locales Framework analysis.

As mentioned earlier, the inducement of using a coupon to induce potential consumers to respond to mobile advertisements is a possible suggestion. Furthermore, companies need to understand that the locale does not stay static, as it is always changing and evolving. Therefore companies need to involve the interaction trajectory aspect that identifies actual interactions individuals have over time within a given context setting and with each other. The last aspect on mutuality involves companies to consider how best mobile marketing can be supported in a given location and how mobile marketing can create awareness for the companies.

The Locales Framework does not attempt to account the findings of one particular phenomenon that is generalisable. It does, however, deliberately aim to characterise its findings that are open in many ways. In fact, one of its aims is to focus on providing an evolvable framework that can be made relevant for both understanding the social situation (in this case mobile marketing) and for improving better technology development to it. This approach, unlike many traditional marketing strategic research approaches, does not assume that the technology introduction has to accept static technologies, or unchanging user attitudes towards technology (rather than the product). A dynamic, multi-dimensional picture of clients and possible interactions allows more dynamic engagement models—supported by dynamic technology, not based on working around systems controlled by other developments.

FUTURE TRENDS

Mobile marketing still lacks research. However, approaching this area from a technological perspective we can suggest several possible outcomes that a marketing and CSCW combined approach indicate:

First, companies are increasingly able to understand potential consumers' attitudes and behaviours in the context of mobile marketing from a more holistic perspective. In particular, the adoption of Locales Framework can provide insights to companies on how to further improve their design and concept for mobile marketing strategies in a particular situation.

Second, there is a need to consider the aspect of interaction design in mobile marketing campaigns. Companies who intend to reach the target market effectively should consider factors like interactivity in their marketing materials, what types of technology (Wifi, Bluetooth, RFID, or global positioning system) can the companies adapt in mobile marketing and how best can the mobile medium fit in a given situation as well as to their potential consumers. The technology perspective of mobile marketing should be considered thoroughly.

Third, situational factors like time and location are important issues that any given companies who decide to use the mobile channel for marketing need to consider. At present, there are no concrete solutions and applications for companies to fully adapt when they design their marketing materials.

Fourth, marketing and technology are both at the forefronts of innovation and fashion. Technology introduction-based marketing methods may also drive technology R&D, when the integrated study approach suggests technological improvements.

Lastly, companies may need to consider social factors like culture, values and norms prior to the launch of a mobile marketing campaign. A single set of mobile marketing materials cannot be carried out in different places, as the social factors are often different. Thus, companies operating across many countries may just need to create a general set of guidelines for mobile marketing with the ability to be tailored to specific context situations. The adoption of the Locales Framework is a suitable tool to be considered for such a multi-level guidelines and customisation approach.

CONCLUSION

Using mobile phones as a medium for marketing is a new phenomenon. Companies need to understand the impact of this medium thoroughly before proceeding. Current research into mobile marketing begins with the traditional marketing perspectives and replaces existing media with mobile technology. Such an approach is based on a historical perspective that doesn't arise from the capabilities of human-computer interaction. The introduction of a theoretical-based research framework such as the Locales Framework is suitable in the investigation of this new medium for mobile marketing. A holistic perceptive of technology introduction in a business-to-client interaction to understand mobile marketing is described, with guidelines for supporting the development of mobile marketing strategies. Such a hybrid marketing/interaction design approach generates new possibilities for both technology development and client engagement that either approach individually would not.

ACKNOWLEDGMENTS

This work is supported by ACID (the Australasian CRC for Interaction Design) established and supported under the Cooperative Research Centres Programme through the Australian Government's Department of Education, Science and Training.

REFERENCES

Balasubramanian, S., Peterson, R. A., & Jarvenpaa, S. L. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of Academy of Marketing Science, 30*(4), 348-361.

Barnes, S. J. (2002). Wireless digital advertising: Nature and implications. *International Journal of Advertising, 21,* 399-420.

Barwise, P., & Strong, C. (2002). Permission-based mobile advertising. *Journal of Interactive Marketing, 16*(1), 14-24.

Bauer, H. H., Barnes, S. J., Reichardt, T., & Neumann, M. M. (2005). Driving consumer acceptance of mobile marketing: A theoretical framework and empirical study. *Journal of Electronic Commerce Research, 6*(3), 181-192.

Belch, G. E., & Belch, M. A. (2004). *Advertising and promotion: An integrated marketing communications perspective* (6th ed.). Boston: McGraw-Hill/Irwin.

Cheong, J. H., & Park, M-C. (2005). Mobile Internet acceptance in Korea. *Internet Research, 15*(2), 125-140.

Coyle, J. R., & Thorson, E. (2001). The effects of progressive levels of interactivity and vividness in Web marketing sites. *Journal of Advertising, 30*(3), 65-77.

Fortin, D. R. (2000, June). Clipping coupons in cyberspace: A proposed model of behavior for deal-prone consumers. *Psychology & Marketing, 17,* 515-534.

Fitzpatrick, G. (1998). *The locales framework: Understanding and designing for cooperative work*. PhD thesis. University of Queensland, Australia.

Fitzpatrick, G. (2003). *The locales framework: Understanding and designing for wicked problems.* Kluwer Academic Publishers.

Fitzpatrick, G., Kaplan, S. K., & Mansfield, T. (1998). *Applying the locales framework to understanding and designing.* Paper presented at OZCHI 1998, Australasian Computer Human Interaction Conference.

Gogus, C. (2004) Understanding young adults' participation in mobile sales promotions. *Paper presented at the 13th EDAMBA Summer School.* Soreze, France.

Haghirian, P., Madlberger, M., & Tanuskova, A. (2005). *Increasing advertising value of mobile marketing—An empirical study of antecedents.* Paper presented at 38th Hawaii International Conference on System Sciences HICSS-38. Hawaii, USA. IEEE Computer Society Press.

Harris, P., Rettie, R., & Cheung, E., (2005). Adoption and usage of m-commerce: A cross-cultural comparison of Hong Kong and the United Kingdom. *Journal of Electronic Commerce Research, 6*(3), 210-224.

Ito, M., & Okabe, D. (2005). Intimate connections: Contextualizing Japanese youth and mobile messaging. In R. Harper, L. Palen, & A. Taylor (Eds.), *The inside text: Social perspectives on SMS in the mobile age*. London: Kluwer.

Juniper Research. (2004). *M-commerce market to grow to $40bn by 2009.* Juniper Research. Retrieved November 20, 2006, from http://www.finextra.com/fullstory.asp?id=12605

Katz, J. E., & Sugiyama, S. (2005). Mobile phones as fashion statements: The co-creation of mobile communication's public meaning. In R. Ling & P. Pedersen (Eds.), *Mobile communications: Re-negotiation of the social sphere* (pp. 63-81). Surrey, UK: Springer.

Karjaluoto, H. (2005). *An investigation of third generation (3G) mobile technologies and services.* Paper presented at BAI2005 International Conference on Business and Information. Hong Kong.

Kavassalis, P., Spyropoulou, N., Drossos, D., Mitrokostas, E., Gikas, G., & Hatzistamatiou, A. (2003). Mobile permission marketing: Framing the market inquiry. *International Journal of Electronic Commerce, 8*(1), 55-79.

Leppaniemi, M., & Karjaluoto, H. (2005). Factors influencing consumers' willingness to accept mobile advertising: A conceptual model. *International Journal of Mobile Communications, 3*(3), 197-213.

Liu, Y., & Shrum, L. J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. *Journal of Advertising, 31*(4), 53-64.

Malloy, A. D., Varshney, U., & Snow, A. P. (2002). Supporting mobile commerce applications using dependable wireless networks. *Mobile Networks and Applications, 7,* 225-234.

Mayor, T. (2005). *The potential of mobile marketing is huge, but is there more to it than just fun and games?* AlertAds.com. Retrieved November 20, 2006, from http://alertads.com/mobile-marketing-is-huge.html

Okazaki, S. (2004). How do Japanese consumers perceive wireless ad? A multivariate analysis. *International Journal of Advertising, 23,* 429-454.

Okazaki, S. (2005a). Mobile advertising adoption by multinationals - senior executives' initial responses. *Internet Research, 15*(2), 160-180.

Okazaki, S. (2005b). New perspectives on m-commerce research. *Journal of Electronic Commerce Research, 6*(3), 160-164.

Raghubir, P. J., Inman, J., & Grande, H. (2004). The three faces of price promotions: Economic, informative and affective. *California Management Review, 46*(4), 1-19.

Rossiter, J. R., & Bellman, S. (2005). *Marketing communications: Theory and applications.* Pearson: Prentice Hall.

Rossiter, J. R., & Percy, L. (1998). *Advertising communications and promotion management* (2nd ed.). The McGraw-Hill Companies, Inc.

Shimp, T. A. (2003). *Advertising, promotion and supplemental aspect to integrated marketing communications* (6th ed.). Thomson: South Western.

Suri, R., Swaminathan, S., & Monroe, K. B. (2004). Price communications in online and print coupons: An empirical investigation. *Journal of Interactive Marketing, 18*(4).

Tsang, M. M., Ho, S. H., & Liang, T. P. (2004). Consumer attitudes toward mobile advertising: An empirical study. *International Journal of Electronic Commerce, 8*(3), 65-79.

Wireless Intelligence. (2005). Worldwide cellular connections exceeds 2 billion. *GSM Association Press Release 2005.* Retrieved November 20, 2006, from http://www.gsmworld.com/news/press_2005/press05_21.shtml

Wehmeyer, K., & Müller-Lankenau, C. (2005). *Mobile couponing: Measuring consumers' acceptance and preferences with a limit conjoint approach.* Paper presented at the 18th Bled eConference, eIntegration in Action. Slovenia.

KEY TERMS

Computer-Supported Collaborative Work (CSCW): Combines the understanding of the way people work in groups with the enabling technologies of computer networking and associated hardware, software, services and techniques (Wilson, 1991). CSCW also addresses how collaborative activities and coordination can be sup-

ported by means of computer systems (Carstensen & Schmidt, 2002). Moreover, CSCW involves incommensurate perspectives as well as incongruent strategies and discordant motives (Schmidt & Bannon, 1992) to gain a better understanding of collaborative efforts within organizations so that this understanding can be used to effectively design collaborative technology that can be best deployed within the organizations.

Human-Computer Interaction (HCI): The study of how people interact with computers and to what extent computers are or are not developed for successful interaction with human beings (ACM SIGCHI, 1996). In fact, HCI is a very broad discipline that encompasses different fields with different perspectives regarding computer development. For instance, HCI in psychology is concerned with the cognitive processes of humans and the behaviour of users, while HCI in computer science is concerned with the application design and engineering of the human interfaces. In sociology and anthropology, HCI is concerned with the interactions between technology, work and organization and the way that human systems and technical systems mutually adapt to each other.

Interaction Design (ID): The study of designing interactive products to support people in their everyday and working lives (Sharp, Rogers, & Preece, 2002). One of the objectives in ID is to produce usable products that are easy to learn, effective to use and also provide an enjoyable experience. Generally users are engaged in the design process.

Locales Framework, The: The Locales Framework is a theoretical-based research framework for interaction design, with a key focus on CSCW. It approaches the study of a certain context or domain with the aim to discover findings that are open in many ways. The general set of guidelines derived from a context or domain is known as the five interdependent aspects. They are: locale foundations, civic structure, individual views, interaction trajectory, and mutuality.

Mobile Advertising (M-Advertising): Advertising is a mass-mediated communication tool. Its aim is to communicate with the intended audience to buy into the desired message. In the context of mobile phones, mobile advertising aims to present the desired information to the consumers, hoping that consumers will react. Currently, most "advertising" contents found on mobile phones are considered as sales promotional materials, which aim to persuade consumers to buy the products.

Mobile Marketing (M-Marketing): Marketing a company's advertised and promotional materials via mobile phones through short message service (SMS) or multimedia messaging service (MMS) is known as mobile marketing. The mobile phone is to the adapted a marketing channel to reach consumers.

Technology Intervention: Technology intervention is the intentional introduction of a technology, or method, into a context to alter that environment. Technology intervention may be targeted at improving information flow, communication, or other types of awareness. In mobile marketing, mobile phones can be seen as a technology intervention in the company-client relationship. Interaction design specifically focuses how to use technology intervention to improve interactions.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 933-939, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.26
Definitions, Key Characteristics, and Generations of Mobile Games

Eui Jun Jeong
Michigan State University, USA

Dan J. Kim
University of Houston Clear Lake, USA

INTRODUCTION

In the emerging wireless environment of digital media communications represented as *ubiquitous* and *convergence*, rapid distribution of handheld mobile devices has brought the explosive growth of the mobile content market. Along with the development of the mobile content industry, mobile games supported by mobile features such as portability (mobility), accessibility (generality), and convenience (simplicity) have shown the highest growth rate in the world game market these days.

In-Stat/MDR (2004) and Ovum (2004) expect that the mobile games' annual growth rate between 2005 and 2009 will be around 50% in the United States and 30% in the world. According to KGDI (2005) and CESA (2005), compared to the rate of the whole game market (5%) of the world, it is about six times higher, and it exceeds the rate of video console (10%) and online games (25%). Mobile games thus are predicted to be one of the leading platforms in the world game market in 10 years' time. In addition, as the competition among game companies has been enhanced with the convergence of game platforms, mobile games are being regarded as a breakthrough for the presently stagnant game market, which has focused on heavy users.

However, due to the relative novelty of mobile games, there are a few visible barriers in the mobile game industry. First, definitions and terminologies and key characteristics related to mobile games are not clearly arranged as yet. Second, there is little research on the classification and development trends of mobile games. Therefore, this article is designed to contribute insights into these barriers in three ways. Firstly, the article

provides narrow and broad definitions of mobile games. Secondly, key characteristics, platforms, and service types of mobile games are discussed. Finally, following the broad definition of mobile games, this article classifies mobile games as one to fourth generations and one pre-generation. Characteristics and examples of each generation are also presented.

DEFINITIONS OF MOBILE GAMES

Each country and each game research institution has different definitions and terminologies. The definition of mobile games is important because the functions of mobile devices are being converged with those of other devices. Mobile games—more precisely, mobile network games—are narrowly defined as *games conducted in handheld devices with network functionality.* The two key elements of this definition are *portability* and *networkability.* In this definition, mobile games are generally referred to as the games played in handheld mobile devices such as cell phones and PDAs with wireless communication functionality. In terms of portability and networkability, the characteristics of mobile games are different from other device platforms such as PC and console games, which do not have both portability and wireless capability. For example, Game Boy (GB) with no communication functionality was only regarded as a portable console device. However, this concept has lost some of its ground in the market since the advent of new mobile game devices from portable consoles such as Play Station Portable (PSP) and Nintendo Dual Screens (NDS), as wireless networked games began to be serviced through the new mobile game devices.

Mobile games can be broadly defined as *embedded, downloaded, or networked games conducted in handheld devices such as mobile phones, portable consoles, and PDAs.* The key element of this concept is portability: all games in portable devices can be thought of as mobile games without regard to wireless functions. Therefore, this concept expands mobile games by including video games in portable consoles and embedded games in general portable devices such as PDAs, calculators, and dictionaries. As most game devices have been adopted with wireless networking functions, this definition becomes more powerful in game markets.

Recently, the narrow definition of mobile games has been generally used. However, since the meaning of *mobile* includes that of *portable and network (either wired or wireless function is embedded),* the broad definition of mobile games including portable game-dedicated devices such as GBs and PSPs should be used. This definition is more persuasive in the present and future game market. For instance, the competition between Nokia's N-gage (i.e., a cell phone integrating the functions of MP3 and games) and Sony's PSP (i.e., a portable game machine including functions of MP3 and networking) is for the preoccupation of a future mobile platform.

KEY CHARACTERISTICS, PLATFORMS, AND SERVICE TYPES

Characteristics and Limitations of Mobile Games

Mobile games are differentiated from other platform games such as console, PC, and arcade games in terms of their portability, accessibility, networkability, and simplicity. Owing to the *portability* (i.e., mobility), users can play games anytime. This characteristic has attracted many light users, who play simple games such as puzzle, card, or word games, because these games can be played in one's spare time in a short amount of time. Compared to players in other genres such as role playing games (RPGs) and simulation games that require a long time to play, light users vary broadly in terms of age, and many women players also belong to this group. This is one of the

strongest potentials of mobile games. The second characteristic of mobile games is *accessibility*. This can be defined as to the extent one can use a mobile device to play games at anytime and at anyplace. Console games are restricted to owners who have console machines and who want to enjoy games for a long time in a particular place. Likewise, most PC games and arcade games need to be somewhere in front of game devices with network facilities. However, mobile games—especially using mobile devices—are easy to access, because people almost always bring those devices anywhere and can download games anywhere as long as wireless networks are available. The third characteristic is *networkability*. Through wired or wireless connections, online games and console games are transplanted into mobile games to facilitate game usages. For example, some online games are linked to mobile games, so those games can be used both in PCs and mobile devices: game users can play the games with no limits in terms of location, machine, and time. Furthermore, mobile game users can play multi-user real-time games such as MMORPG (massively multiplayer online role-playing game) and real-time strategy (RTS) games. The final characteristic is their *simplicity* to use: mobile devices are simpler to handle than other platform machines. In addition, it is much easier to acquire the skills of the games and use them than those of other platforms.

Because of these characteristics, mobile games develop faster than other platform games. According to W2F (2003) and KGDI (2005), the development of a PC or console game usually takes at least two years to develop with more than 20 trained people and about $3 million. But in mobile games, about three to six months are spent with five people and less than $150,000. This is why the initial market entry barrier of mobile games is lower than that of other platform game markets. However, the average lifecycle of mobile games is less than six months, and value chains are more complex than those of other platform games.

Despite the major advantages of mobile games, there are drawbacks in some points. The most essential point is from not-unified platforms. With Internet and console games, converting of original games is not necessary, because the original games can be available in any PC via the Internet. However, mobile games should be converted to make them fit to other platforms, even in the same area. In other words, the conversion is necessary for service to be available in other mobile devices. The second is small screens and low capable devices. Although 3D networked games are being serviced, small screens and monotonous sounds are not sufficient to maximize the feelings of presence for users, and mobile game devices still do not have enough capacity to download high-capacity games through mobile networks.

Mobile Game Platforms

Mobile platforms function as game engines by running applications: a game engine is the core code handling the basic functionality of a game. Each mobile device has its own platform, so developers make games based on the formats of those platforms. With the development of platforms, downloaded, 3D games, and more advanced games are now serviced. These platforms are either freely opened or purchased with license fees. Platform holders have tried to expand their platforms, because the prevalence of their platforms implies a strong influence in mobile markets. These days, Java is the most influential platform both in mobile phone games and in handset manufacture. The Java 2 Micro Edition (J2ME) is a freeware version of Java; Execution Engine (ExEn) and Mophun are also freeware platforms distributed mainly in Europe. Brew is the licensed platform mainly used in the United States, Japan, and Korea. Different from mobile phone games, portable console games such as GB, N-Gage, PSP, and NDS have their own development tools for the platforms. Developers

who want to make mobile games in portable consoles should use such development kits with the charge of license fees. Since developers adopt more prevailing game kits for the better benefit of their games, the market prevalence of console platforms is parallel with the amount of license fees for portable console manufacturers.

Mobile Game Service Types

With the development of mobile service technologies, mobile game services have evolved from single/embedded to multiplayer networked games. Single/embedded are games with which just one player can play without network services. These embedded games are still used in many mobile devices as a service for device customers. Message-based are games using short message service (SMS). These types are played in wireless network environments through WAP (Wireless Application Protocol) browsing environments, but these games are shifting into multi-media message service (MMS) with high capabilities providing enhanced messaging services such as graphics, sound, animation, cartoons, and texts. Downloaded games have been developed with the advent of download platforms such as Java, ExEn, and Mophun. These games have been taken usually from mobile portals managed by mobile network operators, with charges based on both content and traffic fees. Networked games are the newest type of mobile games which are activated with the advent of the flat sum systems.

GENERATIONS OF MOBILE GAMES

From the broad definition of mobile games perspective, portable console machines were the first mobile devices that emerged in the 1970s. These games have been categorized as console games because most hit games were published by console game companies such as Nintendo, Sega, and Sony, and game users were the same as those of console games. However, they have expanded their ranges into color graphic games in the 1990s and mobile network games in the 2000s, so users of such games are no longer limited to young boys not yet in their teens. Following the broad definition of mobile games, this article includes portable console games as a part of mobile games. Portable console games, made before the advent of network portable console games around 2003, are regarded as "portable embedded games," which are categorized as the pre-generation of mobile games.

The first wireless mobile phone game, *Snake*, was serviced as a text-based (or early-graphic) game in 1997. However, today's state-of-the-art games are 3D, fully networked multi-user games with high definitions in wide color screens. As the development of mobile interfaces and network functionalities continues, mobile games can be divided into four generations and one pre-generation. These generations are categorized by stand-alone (off-line) or networked, text-based or graphic-based, and 2D or 3D graphics.

Pre-generation (Pre-G) refers to portable console games that are played in standalone portable devices. In the 1970s, these games were all embedded in only-one-game-use portable game machines such as *Auto Race, Merlin,* and *Missile Attack* by American vendors such as Mattel, Entex, and Tomy. However, in the 1980s, both embedded and cartridge usable games were pervaded with the initiatives of Japanese game companies such as Nintendo and Bandai. From 1989, portable console games were converged into the Nintendo Game Boy era with cartridge games. In the mid-1990s, these games were serviced with color graphic games. With various games usually transplanted from console games, these portable console games flourished with the development of console games.

The *first generation (1G)* refers to text-based mobile phone games like puzzle games. They had been usually serviced by wireless application protocol (WAP) from 1998, and most of

them are single-player embedded games. Some early-graphic games were embodied by white and black dots. These games spread until around 2001 when mobile platforms such as Java, Brew, and ExEn began to spread for the development of mobile graphic games. The *second generation (2G)* refers to graphic games. Developers transported popular games in PC or console games into mobile devices. At first, all graphic services were 2D white and black, but from around 2002, color phones rapidly spread with color graphic games, and some functions such as chatting and reviewing were added. With the prevalence of download platforms, downloaded games generally began to be provided by mobile portals. Traditional board games such as card and chess games were also translated into mobile graphic games in this generation with the concept of licensed games.

The *third generation (3G)* refers to networked games with simple network functionalities. Around 2003, most games were 2D graphic: 3D games were just a state of experiment. Network functionality was not fully serviced, because of the high cost of network use and low capabilities of mobile devices. However, owing to network capability, new games such as various simulations, multi-user role-playing games (MRPG), and location-based service (LBS) games were firstly developed in this period. Additionally, new mobile devices such as N-gage, PSP, and NDS had changed the traditional concepts of mobile games with the mixture of wireless and networked game services. These devices are estimated to have promoted the degree of mobile games as much as that of console games. With the prevalence of device convergences between mobile and console devices, from this generation there is no accurate difference among game genres. The *fourth generation (4G)* games refer to 3D and full networked games such as massively multi-user online role playing games (MMORPG). In addition, around 2004, full 3D mobile game services began to be serviced, and many developers joined the development of 3D network games. With the spread of new 3D graphic mobile phones and flat sum systems, 3D networked games are steadily gaining their shares in game markets. Table 1 illustrates the mobile game generations, key characteristics of each generation, and examples.

Table 1. Generations of mobile games

	Outset	**Characteristic Game Genres**	**Examples**
Pre-G	1970s	Portable console games Portable console color games (Embedded or cartridge games)	*Auto Race, Football* *Bomberman* *Pac Man, Tetris (Console)*
1G	1997/1998	Text-based games (early-graphic) WAP games	*Snake* *Dataclash, Gladiator*
2G	2001/2002	Downloaded games (Java, Brew) 2D color graphic games	*Tetris, Chess Mobile* *Samurai Romanesque*
3G	2003/2004	Portable console network games Half network games 3D graphic games	*Pokemon Ruby* *Badlands, Samgukji* *3D Pool*
4G	2004/2005	Full 3D graphic games Full 3D network games	*3D Golf, 3D Bass Fishing* *Homerun King Mobile*

CONCLUSION

With the expansion of a convergence and ubiquitous environment, the range of mobile games has grown to include all games available in handheld devices with portability. At first, mobile games were regarded as embedded single-user games. However, through the development of network and graphic technology, mobile games have been played as both full network games with multi-users and 3D graphic games with high-definition devices. So, many high-capability games such as MMORPG and multi-user simulation games have been adopted in mobile devices with high-speed network capability. In addition, with the convergence of game devices, the boundary between mobile phones and console devices has been eroded, while games in PC and console machines have been transformed into mobile games. Due to the accessibility, portability, and ease of use, mobile games have a wide range of users and do not impose limitations in age, sex, and social status. Traditionally game users were usually young males, but when it comes to mobile games, the game users are more diversified: not only young boys, but also elderly people and women are joining in on the new mobile gaming era. With the development of mobile technology, diversification of mobile content services, and generalization of mobile game users, mobile games will continue to gain more power within game markets.

Mobile games are summarized along with taxonomies. In addition, recent trends with game application areas will be discussed in the next article, "Mobile Games Part II: Taxonomies, Applications, and Trends."

REFERENCES

CESA. (2005). *2005 CESA game white paper.* Tokyo: Computer Entertainment Suppliers' Association.

DeMaria, R., & Wilson, J. L. (2002). *High score: The illustrated history of electronic games.* Berkeley: Osborne Media Group.

Entertainment Software Association. (2002). *Top ten industry facts.* Retrieved from www.theesa.com/pressro om.html

Ermi, L., & Mäyrä, F. (2005). Challenges for pervasive mobile game design: Examining players' emotional responses. *Proceedings of the ACM SIGCHI International Conference on Advances in Computer Entertainment Technology* (pp. 47-55), Valencia, Spain.

Hall, J. (2005). Future of games: Mobile gaming. In J. Raessens & J. Goldstein (Eds.) *Handbook of computer game studies* (pp. 47-55). Cambridge, MA: MIT Press.

In-Stat/MDR. (2004). Mobile gaming services in the U.S., 2004-2009. *In-Stat/MDR,* (August).

KGDI. (2005). *2005 game white paper.* Seoul: Korea Game Development & Promotion Institute.

Newman, J. (2004). *Videogames.* London: Routledge.

Nokia. (2003). *Introduction to mobile game development, Nokia Corporation*. Retrieved from www.forum.nokia.com/html_reader/main/1,,2768,00.html

Ovum. (2004, December). *Ovum forecasts global wireless market.* Ovum.

Ring, L. (2004). *The mobile connection: The cell phone's impact on society.* San Francisco: Morgan Kaufmann.

Taylor Nelson Sofres. (2002). *Wireless and Internet technology adoption by consumers around the world*. Retrieved from www.tnsofres.com/IndustryExpertise/IT/WirelessandInternetAdoptionbyConsumersAroudtheWorldA4.pdf

W2F. (2003, October). *Winning and losing in mobile games.* W2F Limited.

KEY TERMS

Device Platform: A device such as a cell phone, PDA, PC, or console machine through which games can be played.

Local-Based Service (LBS) Game: A mobile network game played within a local place around a telecommunication base with the information of a user's position.

Massively Multi-player Online Game (MMOG): A game where a huge number of users can play simultaneously based on their roles or missions.

Mobile Game Platform: Core code handling of the basic functionality of a game such as downloading, networking, or activating 3D graphics.

Mobile Game: An embedded, downloaded, or networked game conducted in a handheld device such as a mobile phone, portable console, or PDA.

Portable Console (Device): Handheld console machine such as PSP, NDS, and GBA with portable capabilities.

Role Playing Game (RPG): A game where a gamer takes a role and uses items in accomplishing missions or quests.

3D Network Game: A game played in connection with other users, using 3D graphics.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 184-189, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.27
Mobile Agents

Kamel Karoui
Institut National des Sciences Appliquées de Tunis, Tunisia

INTRODUCTION

The concept of mobile agent is not new; it comes from the idea of *OS process migration* firstly presented by *Xerox* in the 1980's. The term *mobile agent* was introduced by White & Miller (1994), which supported the mobility as a new feature in their programming language called *Telescript*.

This new research topic has emerged from a successful meeting of several sub-sciences: computer networks, software engineering, object-oriented programming, artificial intelligence, human-computer interaction, distributed and concurrent systems, mobile systems, telematics, computer-supported cooperative work, control systems, mining, decision support, information retrieval and management, and electronic commerce. It is also the fruit of exceptional advances in distributed systems field (Hirano 1997; Holder, Ben-Shaul, & Gazit 1999; Lange et al., 1999).

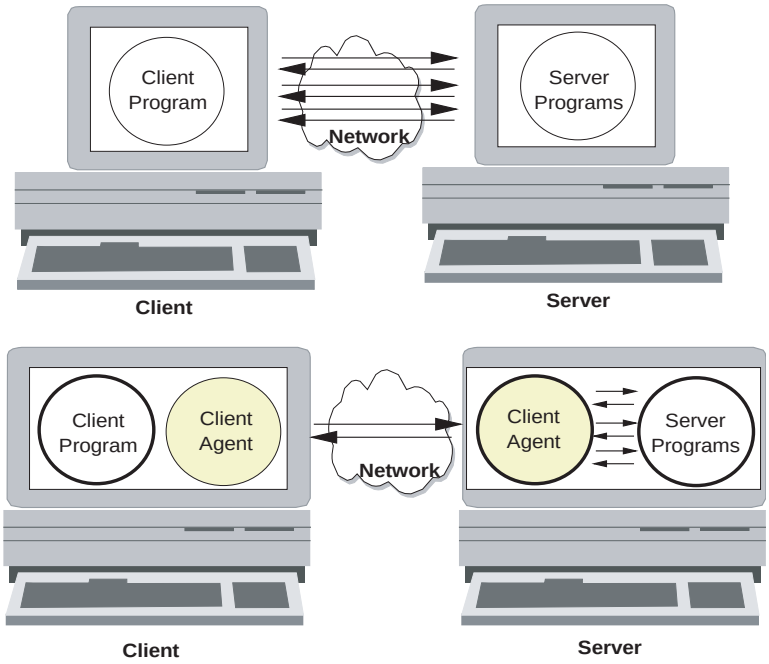
The main idea of the mobile agent technology is to replace the old approach of the client-server Remote Procedure Call (RPC) paradigm, by a new one consisting of transporting and executing programs around a network. The results of the programs execution are then returned back to the sending entity. Figure 1 illustrates this new approach.

Mobile agents are dynamic, non-deterministic, unpredictable, proactive, and autonomous entities. They can decide to exercise some degree of activities without being invoked by external entities. They can watch out for their own set of internal responsibilities. Agents can interact with their environment and other entities. They can support method invocation as well as more complex degree of interaction as for example the observable events reaction within their environment. They can decide to move from one server to another in order to accomplish the system global behavior.

BACKGROUND

As the information technology moves from a focus on the individual computer system to a situation in which the real power of computers is realized through distributed, open and dynamic systems, we are faced with new technological challenges. The characteristics of dynamic and open environments in which heterogeneous sys-

Figure 1. RPC vs. mobile agent approach



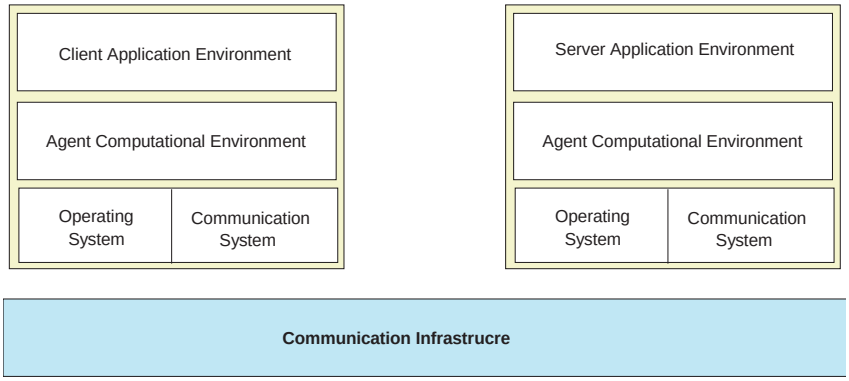
tems must interact require improvements on the traditional computing models and paradigms. It is clear that these new systems need some degree of intelligence, autonomy, mobility, and so on. The mobile agent concept is one of the new system environment that has emerged from this need. Several researches have proposed a definition of mobile agents (Bradshaw, Greaves, Holmback, Jansen, Karygiannis, Silverman, Suri, & Wong, 1999; Green & Somers, 1997; White 1997). Until now, there is neither standard nor a unique consensus on a unique definition. In general, a mobile agent can be defined using its basic attributes: the mobility, the intelligence and the interactivity. Based on these attributes, we can propose the following definition:

A mobile agent is a computational entity which acts on behalf of other entities in an intelligent way (autonomy, learning, reasoning, etc.). It performs its tasks in software open and distributed environment with some level of mobility, co-operation, proactivity, and/or reactiveness.

This attributes based definition gives an abstract view of what a mobile agent does, but it doesn't present how it does it. This definition doesn't mean that mobility, interactivity, and intelligence are the unique attributes of mobile agents. Effectively, a large list of other attributes exists such as: application field, communication, delegation, and so on.

This definition shows that a mobile agent doesn't exist without a software environment called a mobile agent environment (see Figure 2).

Figure 2. Mobile agent environment

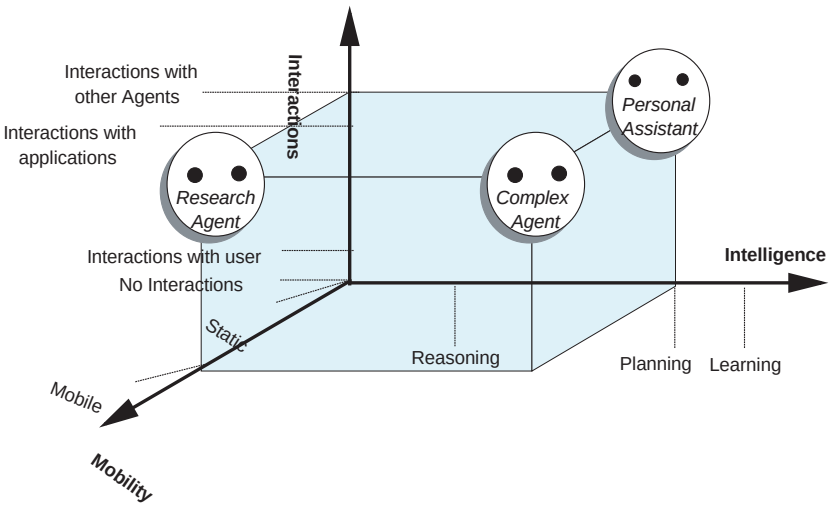


AGENT CLASSIFICATION

According to the literature (Frankllin & Graesser, 1996), agents, and especially mobile agents, can be classified using the three agent basic attributes depicted in Figure 3.

- The first agent attribute is mobility, so an agent can be static or mobile.

Figure 3. Agent classification



- The second attribute is intelligence; an agent can be characterized by its abilities of reasoning, planning, learning, and so on.
- Interaction is the third agent attribute. Agents can have different kinds of interactions. This category of agents contains the agents that: do not interact at all, interact with users, interact with applications, and interact with other agents.

There are of course many other classification methods (Frankllin & Graesser, 1996). For example, we can classify agents according to the task they perform, for example, information gathering agents or e-mail filtering agents.

MOBILE AGENT ADVANTAGES

Using mobile agents is not the unique way to solve some class of problems, alternative solutions exists. However, for some class of problems and applications, we believe that mobile agent technology is more adapted than classical methods. For example, in managing large scale intranet, where we must continuously, install, update, and customize software for different users without bringing the server down. In the following we present three types of application domains where it is better to use mobile agent technology:

- Data-intensive application where the data is remotely located. Here, agents are sent in order to process and retrieve data.
- Disconnected computing application where agents are launched by an appliance. For example, shipping an agent from a cellular phone to a remote server.
- Application where we need to extend the server behavior by sending agents that can represent permanently or not the server in different location (host or server).

In the following we present a list of the main advantages of mobile agent's technology:

- Efficiency: mobile agents consume fewer network resources.
- Reduction of the network traffic: mobile agents minimize the volume of interactions by moving and executing programs on special host servers.
- Asynchronous autonomous interactions: mobile agents can achieve tasks asynchronously and independently of the sending entity.
- Interaction with real-time entities: for critical systems (nuclear, medical, etc.) agents can be dispatched from a central site to control local real-time entities and process directives from the central controller.
- Dynamic adaptation: mobile agents can dynamically react to changes in its environment.
- Dealing with vast volumes of data: by moving the computational to the sites containing a large amount of data instead of moving data, we can reduce the network traffic.
- Robustness and fault tolerance: by its nature, a mobile agent is able to react to multiple situations, especially faulty ones. This ability makes the systems based on mobile agents fault tolerant.
- Support for heterogeneous environments: mobile agents are generally computer and network independent, this characteristic allows their use in a heterogeneous environment.

MOBILE AGENT DISADVANTAGES

In the following we present a list of the major problems for mobile agent approach:

- Security is one of the main concerns of the mobile agent approach. The issue is how

to protect agent from malicious hosts and inversely how to protect hosts from mobile agents. The main researchers' orientation is to isolate the agent execution environment from the host critical environment. This separation may limit the agent capabilities of accessing the desired data and from accomplishing its task.

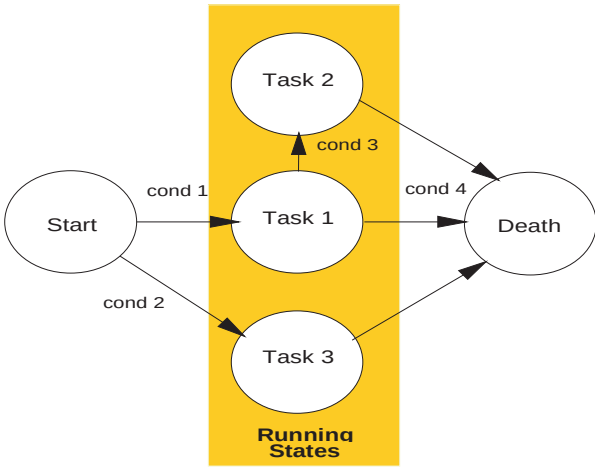
- Another big problem of the mobile agent approach is the lack of standardization. In the recent years, we have seen the development of many mobile agent systems based on several slightly different semantics for mobility, security, and communication. This will restrict the developers to small applications for particular software environments.
- Mobile agents are not the unique way to solve major class of problems, alternative solutions exists: messaging, simple datagram, sockets, RPC, conversations, and so on. There are neither measurement methods nor criteria that can help developer choose between those methods. Until now there is no killer application that uses the mobile agent approach.
- Mobile agents can achieve tasks asynchronously and independently of the sending entity. This can be an advantage for batch applications and disadvantage for interactive applications.

MOBILE AGENT MODELS

A successful mobile agent system should be designed based on the following six models. The implementation of these models depends on the agent construction tools.

- **Agent model:** It defines the intelligent part (autonomy, reasoning, learning, etc.) of the agent internal structure.
- **Computational model:** It defines how the agent executes its self when it is in its running states (see Figure 4). In general, this model

Figure 4. Agent life cycle model



is represented by a finite state machine or an extended finite state machine (Karoui, Dssouli, & Yevtushenko 1997).

- **Security model:** This model describes the different approach of the security part of the system. In general, there are two main security concerns, protection of hosts from malicious agents and protection of agents from malicious hosts.
- **Communication model:** It presents how the agents communicate and interacts with other agents of the system.
- **Navigation model:** This model deals with the mobility in the system. It describes how an agent is transported from one host to another.
- **Life-cycle model:** Each agent can be characterized by a life cycle. The life cycle starts from the agent creation state *Start*, and ends in the death state *Death*. The intermediate states depend on the nature of the mission. Those last states are called running states (see Figure 4).

AGENT CONSTRUCTION TOOLS

Several mobile agent construction tools have appeared since 1994. Most of them are built on top of the Java system or the Tcl/tk system (Morisson & Lehenbauer 1992). Table 1 provides a survey of some currently available agent construction tools. Although each of these tools supports different levels of functionality, they each attempt to address the same problem: namely, enabling portions of code to execute on different machines within a wide-area network. Many research groups are now focusing on Java as the development language of choice thanks to its portability and code mobility features.

Table 1. Mobile agent construction tools

<TBODY>**Product**	**Company**	**Lang.**	**Description**
AgenTalk	NTT/Ishida	LISP	Multiagent Coord.
Agentx	International Knowledge Systems	Java	Agent Development Environment
Aglets	IBM Japan	Java	Mobile Agents
Concordia	Mitsubishi Electric	Java	Mobile Agents
DirectIA SDK	MASA - Adaptive Objects	C++	Adaptive Agents
Gossip	Tryllian	Java	Mobile Agents
Grasshopper	IKV++	Java	Mobile Agents
iGEN™	CHI Systems	C/C++	Cognitive Agent
JACK Intelli Agents	Agent Oriented Software Pty. Ltd.	JACK	Agent Development Environment
JAM	Intelligent Reasoning Systems	Java	Agent Architecture
LiveAgent	Alcatel	Java	Internet Agent
AgentTcl	Dartmouth College	Tcl/tk	Mobile Agents
MS Agent	Microsoft Corp.	Active X	Interface creatures

One feature that all of these mobile agent construction tools have currently failed to address is in defining a domain of applicability; they all concentrate on the mobility of agents rather than the integration of agents with information resources.

MOBILE AGENT-BASED SYSTEM EXAMPLE

As an example of multi-agent and mobile agent systems, we present an application in telemedicine that we have developed in previous works (Karoui, Loukil, & Sounbati 2001; Karoui & Samouda 2001). The idea from proposing such system starts from the statistics about health care system of a small country. We have seen that this system suffers from two main weaknesses: insufficiency of specialists and bad distribution of the specialists over the country. Thus, we thought about a system which is able to provide, to a non-expert practitioner (Physician), the appropriate computerized or not help of a distant expert. The system is influenced by the following set of constraints and considerations:

1) Before asking for a help of a distant expert, the system should be able to proceed a multilevel automatic diagnose in order to refine, classify, and document the case.
2) The non-expert site of our system should be able to learn from previous experiences and the diagnosed cases by specialists.
3) The responses should not exceed a limit of time specified by the requestor on the basis of the case emergency.
4) The expert can refuse to respond to a query.
5) In order to facilitate and accelerate the expert diagnosis, the information related to a query and sent to the experts should be as complete as possible.
6) For security purposes, the system should ensure the authentication of both the requestor and the advisor, and also the integrity and confidentiality of the interchanged data.
7) The system should be easy to extend and to maintain.

Taking into account these requirements, we present here after how the system works. First of all, our system is composed of a set of medical sites; each of them has a server connected to a telemedicine network. This later can be either a private network or the Internet. In each medical site we have at least one physician able to collect patient symptoms. When a patient goes for a consultation, we cannot insure that in his local medical center there has the appropriate expert for his disease. In case of expert deficiency, the local physician collects the symptoms through a guided computerized user interface, and a multilevel diagnoses process. In the following, we explain the four-level diagnosis process which is composed of two human diagnoses (levels 1 and 4) and two computerized automatic diagnosis (levels 2 and 3).

1. **The first level diagnosis:** The physician who collects the symptoms can propose a diagnosis. This diagnosis will be verified by a computerized process called the second level diagnosis.
2. **The second level diagnosis:** The local system analyzes automatically the collected symptoms. If the system detect a disease, it automatically informs the physician giving him all the information used to reach such diagnosis (used rules and symptoms). This diagnostic may be different from the one given by the physician himself (*first level diagnosis).* The system then asks the physician if he wants to confirm this diagnosis by getting the advice of an expert. If yes, a request is sent to a set of experts chosen

automatically by the system. The request is represented by mobile agents sent to distant servers. The request contains all the information needed to get the right diagnosis.

3. **The third level diagnosis:** When the distant servers receive the request, each one of them verifies automatically the correctness of the information used in order to produce and send back to the requester (through the mobile agent) a computerized *third level diagnosis*. If this information is not correct (not complete or bad rules), the request is returned back to the sender asking the local system for more special information or symptoms about the case.
4. **The fourth level diagnosis:** If the information contained in the request is correct, but the expert server site cannot produce a computerized diagnostic (*third level diagnosis*). It is presented to a human expert who will analyze, give his diagnosis about the case and take the necessary actions.

For the system performance, we suppose that the non expert part learns (self learning) from its previous experience. So, for a given case, the system starts with a minimal amount of information about diseases, then from the multilevel diagnosis process (specially the third level diagnosis) the system will automatically update its diagnostic rules and databases.

CONCLUSION

Agent-oriented approach is becoming popular in the software development community. In the future, agent technology may become be a dominant approach. The agent-based way of thinking brings a useful and important perspective for system development.

Recent years have seen the development of many mobile agent systems based on several slightly different semantics for mobility, security, communication, and so on. We need now to start the process of choosing the best ideas from the huge number of the proposed approaches and identify the situations where those approaches are useful and may be applied. In order to achieve this goal, we need some quantitative measurements of each kind of mobility communication and security methods. This will automatically result in a kind of standardization.

REFERENCES

Bradshaw, J.M., Greaves, M., Holmback, H., Jansen, W., Karygiannis, T., Silverman, B., Suri, N., & Wong, A. (1999). Agents for the masses: Is it possible to make development of sophisticated agents simple enough to be practical? *IEEE Intelligent Systems.* 53-63.

Franklin, S. & Graesser, A. (1996). Is it an agent, or just a program?: A Taxonomy for Autonomous Agents. *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages.* Springer-Verlag.

Fuggetta, A., Picco, G.P., & Vigna, G. (1998). Understanding Code Mobility. *IEEE Transactions on Software Engineering, 24*(5).

Green, S. & Somers, F. (1997). Software Agents: A review. Retrieved August 5 1998, from *http://www.cs.tcd.ie/research_ groups/aig/iag/pubreview/*

Hirano, S. (1997). HORB: Distributed Execution of Java Programs, *Worldwide Computing and Its Applications'97, Springer Lecture Notes in Computer Science, 1274*, 29-42.

Holder, O., Ben-Shaul, I., & Gazit, H. (1999). System Support for Dynamic Layout of Distributed Application. *Proceedings of the 21 st International Conference on Software Engineering (ICSE'99),* 163- 173.

IBM (1998). Aglets software development kit. Retrieved June 4, 1999. From *http://www.trl.ibm.co.jp/aglets/*

Karoui, K., Dssouli, R., & Yevtushenko, N. (1997). Design For testability of communication protocols based on SDL. *Eighth SDL FORUM 97, Evry France.*

Karoui, K., Loukil, A., & Sonbaty, Y. (2001). Mobile agent hybrid route determination framework for health-care telemedicine systems, *ISC 2001,* Tampa Bay USA.

Karoui, K., Samouda, R., & Samouda, M. (2001). Framework for a telemedicine multilevel diagnose system. *IEEE EMBS'2001, Vol. 4,* Istanbul, Turkey, 3508-3512.

Kotz, D., Gray, R., Nog, S., Rus, D., Chawla, S., & Cybenko, G. (1997). Agent TCL: Targeting the needs of mobile computers. *IEEE Internet Computing, 1*(4).

Lange, D. et al. (1999). Seven good reasons for mobile agents. *Communications of the ACM, 42*(3), 88-89.

Lange, D. & Oshima, M. (1998). *Programming and deploying java mobile agents with aglets.* Addison -Wesley.

Morisson, B. & Lehenbauer, K. (1992). Tcl and Tk: Tools for the system administration.administration. *Proceedings of the Sixth System Administration Conference*,225-234.

White, J. (1997). *Mobile agent.* J.M. Bradshaw (Ed.), Software Agents. Cambridge, MA: The AAAI Press/The MIT Press, 437-472.

White, J.E. (1994). *Telescript technology: The foundation for the electronic marketplace.* Mountain View, CA: General Magic, Inc.

KEY TERMS

Agent: A computational entity which acts on behalf of other entities.

Agent Attributes: An agent can be classified using some of its characteristics called attributes. An agent has three basic attributes: mobility, intelligence, and interaction.

Client-Server Model: A client-server model defines a basis for communication between two programs called respectively the client and the server. The requesting program is a client and the service-providing program is the server.

Intelligent Agent: An agent who acts in an intelligent way (autonomy, learning, reasoning, etc.).

Mobile Agent: An intelligent agent who performs its tasks with some level of mobility, co-operation, proactivity, and/or reactiveness.

Multiagent System: A system composed of agents interacting together in order to achieve the system common task or behaviour.

RPC: Remote Procedure Call is one way of communication in a client server model. The client and the server are located in different computers in a network. An RPC is a synchronous operation requiring the requesting (client) to pass by value all the needed parameters to the server then the client is suspended until the server returns back the associated results.

This work was previously published in Encyclopedia of Multimedia Technology and Networking, edited by M. Pagani, pp. 608-614, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.28
Protection of Mobile Agent Data

Sheng-Uei Guan
Brunel University, UK

INTRODUCTION

One hindrance to the widespread adoption of mobile agent technology is the lack of security. Security will be the issue that has to be addressed carefully if a mobile agent is to be used in the field of electronic commerce. SAFER—or Secure Agent Fabrication, Evolution, and Roaming—is a mobile agent framework that is specially designed for the purpose of electronic commerce (Zhu, Guan, Yang, & Ko, 2000; Guan & Hua, 2003; Guan, Zhu, & Maung, 2004). Security has been a prime concern from the first day of our research (Guan & Yang, 1999, 2002; Yang & Guan, 2000). By building strong and efficient security mechanisms, SAFER aims to provide a trustworthy framework for mobile agents, increasing trust factors to end users by providing the ability to trust, predictable performance, and a communication channel (Patrick, 2002).

Agent integrity is one such area crucial to the success of agent technology (Wang, Guan, & Chan, 2002). Despite the various attempts in the literature, there is no satisfactory solution to the problem of data integrity so far. Some of the common weaknesses of the current schemes are vulnerabilities to revisit attack when an agent visits two or more collaborating malicious hosts during one roaming session and illegal modification (deletion/insertion) of agent data. Agent Monitoring Protocol (AMP) (Chionh, Guan, & Yang, 2001), an earlier proposal under SAFER to address agent data integrity, does address some of the weaknesses in the current literature. Unfortunately, the extensive use of PKI technology introduces too much overhead to the protocol. Also, AMP requires the agent to deposit its data collected to the agent owner/butler before it roams to another host. While this is a viable and secure approach, the proposed approach—Secure Agent Data Integrity Shield (SADIS)—will provide an alternative by allowing the agent to carry the data by itself without depositing it (or the data hash) onto the butler.

Besides addressing the common vulnerabilities of current literature (revisit attack and data modification attack), SADIS also strives to achieve maximum efficiency without compromising secu-

rity. It minimizes the use of PKI technology and relies on symmetric key encryption as much as possible. Moreover, the data encryption key and the communication session key are both derivable from a key seed that is unique to the agent's roaming session in the current host. As a result, the butler can derive the communication session key and data encryption key directly. Another feature in SADIS is strong security.

Most of the existing research focuses on detecting integrity compromise (Esparza, Muñoz, Soriano, & Forné, 2006) or on bypassing integrity attacks by requiring the existence of a cooperating agent that is carried out within a trusted platform (Ouardani, Pierre, & Boucheneb, 2006), but which neglected the need to identify the malicious host. With SADIS, the agent butler will not only be able to detect any compromise to data integrity, but to identify the malicious host effectively.

BACKGROUND

Agent data integrity has been a topic of active research in the literature for a while. SADIS addresses the problem of data integrity protection via a combination of techniques discussed by Borselius (2002): execution tracing, encrypted payload, environmental key generation, and undetachable signature.

One of the recent active research works is the security architecture by Borselius, Hur, Kaprynski, and Mitchell (2002). Their security architecture aims at defining a complete security architecture designed for mobile agent systems. It categorizes security services into the following: agent management and control, agent communications service, agent security service, agent mobility service, and agent logging service. SADIS addresses the agent communication service as well as agent security services (integrity protection), while previous research on SAFER addresses agent mobility service.

While many of the security services are still under active research, the security mechanisms for protecting agents against malicious hosts were described by Borselius, Mitchell, and Wilson (2001). Their paper proposes a threshold scheme to protect mobile agents. Under the mechanism, a group of agents is dispatched to carry out the task, each agent carrying a vote. Each agent is allowed to contact a merchant independently and gathers bids based on the given criteria. Each agent votes for the best bid (under a trading scenario) independently. If more than n out of m ($m > n$) agents vote for the transaction, the agent owner will agree to the transaction.

Such a mode of agent execution effectively simplifies agent roaming by allowing one agent to visit one merchant only. While the approach avoids the potential danger of having the agent compromised by the subsequent host, it does not employ a mechanism to protect the agent against the current host. Most important of all, the threshold mechanism's security is based on the probability that no more than n hosts out of m are malicious. In other words, the security is established based on probability. Different from this approach, SADIS's security is completely based on its own merits without making any assumption about probability of hosts being benign or malicious. This is because the author believes that in an e-commerce environment, security should not have any dependency on probability.

Other than the research by Borselius, there are related works in the area. One such work on agent protection is SOMA, or Secure and Open Mobile Agent, developed by Corradi, Cremonini, Montanari, and Stefanelli (1999). SOMA is a Java-based mobile agent framework that provides for scalability, openness, and security on the Internet. One of the research focuses of SOMA is to protect the mobile agent's data integrity. To achieve this, SOMA makes use of two mechanisms: Multi Hop (MH) Protocol and Trusted Third Party (TTP) Protocol. MH protocol works as follows. At each

intermediate site the mobile agent collects some data and appends them to the previous ones collected. Each site must provide a short proof of the agent computation, which is stored in the agent. Each proof is cryptographically linked with the ones computed at the previous sites. There is a chaining relation between proofs. When the agent moves back to the sender, the integrity of the chained cryptographic proofs is verified allowing the sender to detect any integrity violation.

The advantage of MH protocol is that it does not require any trusted third party or even the agent butler for its operation. This is a highly desirable feature for agent integrity protection protocol. Unfortunately, MH protocol does not hold well against revisit attack when the agent visits two or more collaborating malicious hosts during one roaming session (Chionh et al., 2001). Roth (2001) provides more detailed descriptions on potential flaws of the MH protocol.

Another agent system that addresses data integrity is Ajanta (Tripathi, 2002). Ajanta is a platform for agent-based application on the Internet developed in the University of Minnesota. It makes use of an append-only container for agent data integrity protection. The main objective is to allow the host to append new data to the container, but to prevent anyone from modifying the previous data without being detected. Similar to the MH protocol, such an append-only container suffers from revisit attack.

From these attacks on existing research, the importance of protecting agent itinerary is obvious. In SADIS, the agent's itinerary is implicitly updated in the agent butler during key seed negotiation. This prevents any party from modifying the itinerary recorded on the butler and guard against all itinerary-related attacks.

There is one recent research work on agent data integrity protection called One-Time Key Generation System (OKGS) researched at the Kwang-Ju Institute of Science and Technology, South Korea (Park, Lee, & Lee, 2002). OKGS does protect the agent data against a number of attack scenarios under revisit attack, such as data insertion attack and data modification attack to a certain extent. However, it does not protect the agent against deletion attack, as two collaborating malicious hosts can easily remove roaming records in between them.

Inspired by OKGS's innovative one-time encryption key concept, SADIS will extend this property to the communication between agent and butler as well. Not only the data encryption key is one time, but the communication session key is as well. Using efficient hash calculations, the dynamic communication session key can be derived separately by the agent butler and the agent with minimum overhead. Despite the fact that all keys are derived from the same session-based key seed, SADIS also ensures that there is little correlation between these keys. As a result, even if some of the keys are compromised, the key seed will still remain secret.

PROTECTION OF AGENT DATA INTEGRITY

SADIS is designed based on the SAFER framework. The proposal itself is based on a number of assumptions that were implemented under SAFER. Firstly, entities in SAFER, including agents, butlers, and hosts, should have a globally unique identification number (ID). This ID will be used to uniquely identify each entity. Secondly, each agent butler and host should have a digital certificate that is issued by a trusted CA under SAFER. These entities with digital certificates will be able to use the private key of its certificate to perform digital signatures and, if necessary, encryption. Thirdly, while the host may be malicious, the execution environment of mobile agents should be secure and the execution integrity of the agent can be maintained. This assumption is made because protecting the agent's execution environ-

ment is a completely separate area of research that is independent of this article. Without a secure execution environment and execution integrity, none of the agent data protection schemes will be effective. The last assumption is that entities involved are respecting and cooperating with the SADIS protocol. And finally, SADIS does not require the agent to have a pre-determined itinerary. The agent is able to decide independently which host is the next destination.

Key Seed Negotiation Protocol

When an agent first leaves the butler, the butler will generate a random initial key seed, encrypt it with the destination host's public key, and deposit it into the agent before sending the agent to the destination host. It should be noted that agent transmission is protected by the *SAFE* supervised agent transport protocol (Guan & Yang, 2002). Otherwise, a malicious host (man-in-the-middle) can perform an attack by replacing the encrypted key seed with a new key seed and encrypt it with the destination's public key. In this case, the agent and the destination host will not know the key seed has been manipulated. When the agent starts to communicate with the butler using the wrong key seed, the malicious host can intercept all the messages and re-encrypt them with the correct key derived from the correct key seed and forward them to the agent butler. In this way, a malicious host can compromise the whole protocol.

The key seed carried by the agent is session-based: it is valid until the agent leaves the current host. When the agent decides to leave the current host, it must determine the destination host and start the key seed negotiation process with the agent butler.

The key seed negotiation process is based on the Diffie-Hellman (DH) key exchange protocol (Diffie & Hellman, 1976) with a variation. The agent will first generate a private DH parameter *a* and its corresponding public parameter *x*. The value *x*, together with the ID of the destination host, will be encrypted using a communication session key and sent to the agent butler.

The agent butler will decrypt the message using the same communication session key (derivation of communication session key will be discussed later in the section). It too will generate its own DH private parameter *b* and its corresponding public parameter *y*. With the private parameter *b* and the public parameter *x* from the agent, the butler can derive the new key seed and use it for communications with the agent in the new host. Instead of sending the public parameter *y* to the agent as in normal DH key exchange, the agent butler will encrypt the value *y*, host ID, agent ID, and current timestamp with the destination host's public key to get message *M*. Message *M* will be sent to the agent after encrypting with the communication session key.

$$M = E(y + \text{host ID} + \text{agent ID} + \text{timestamp}, H_{pubKey})$$

At the same time, the agent butler updates the agent's itinerary and sends it to the agent. When the agent receives the double-encrypted DH public parameter *y*, it can decrypt with the communication session key.

Subsequently, the agent will store *M* into its data segment and requests the current host to send itself to the destination host using the agent transport protocol (Guan & Yang, 2002).

Upon arriving at the destination host, the agent will be activated. Before it resumes normal operation, the agent will request the new host to decrypt message *M*. If the host is the right destination host, it will be able to use the private key to decrypt message *M* and thus obtain the DH public parameter *y*. As a result, the decryption of message *M* not only completes the key seed negotiation process, but also serves as a means to authenticate the destination host. Once the message *M* is decrypted, the host will verify that the

agent ID in the decrypted message matches the incoming agent, and the host ID in the decrypted message matches that of the current host.

With the plain value of *y*, the agent can derive the key seed by using its previously generated private parameter *a*. With the new key seed derived, the key seed negotiation process is completed. The agent can resume normal operation in the new host.

Whenever the agent or the butler needs to communicate with the other, the sender will first derive a communication session key using the key seed and use this communication session key to encrypt the message. The receiver can make use of the same formula to derive the communication session key from the same key seed to decrypt the message.

The communication session key K_{CSK} is derived using the formula below:

$$K_{CSK} = \text{Hash}(\text{key\_seed} + \text{host ID} + \text{seqNo})$$

The sequence number is a running number that starts with 1 for each agent roaming session. Whenever the agent reaches a new host, the sequence number will be reset to 1. Given the varying communication session key, if one of the messages is somehow lost without being detected, the butler and agent will not be able to communicate afterwards. As a result, SADIS makes use of TCP/IP as a communication mechanism so that any loss of messages can be immediately detected by the sender. In the case of an unsuccessful message, the sender will send 'ping' messages to the recipient in plain format until the recipient or the communication channel recovers. Once the communication is re-established, the sender will resend the previous message (encrypted using the same communication session key).

When the host provides information to the agent, the agent will encrypt the information with a data encryption key K_{DEK}. The data encryption key is derived as follows:

$$K_{DEK} = \text{Hash}(\text{key\_seed} + \text{hostID})$$

Data Integrity Protection Protocol

The key seed negotiation protocol lays the necessary foundation for integrity protection by establishing a session-based key seed between the agent and its butler. Agent data integrity is protected through the use of this key seed and the digital certificates of the hosts. Our data Integrity Protection protocol is composed of two parts: chained signature generation and data integrity verification. Chained signature generation is performed before the agent leaves the current host. The agent gathers data provided by the current host d_i and construct D_i as follows:

$$D_i = E(d_i + ID_{host} + ID_{agent} + timestamp, k_{DEK})$$

or,

$$D_i = d_i + ID_{host} + ID_{agent} + \text{timestamp}$$

The inclusion of a host ID, agent ID, and timestamp is to protect the data from possible replay attack, especially when the information is not encrypted with the data encryption key. For example, if the agent ID is not included in the message, a malicious host can potentially replace the data provided for one agent with that provided for a bogus agent. Similarly, if a timestamp is not included into the message, earlier data provided to the same agent can be used at a later time to replace current data provided to the agent from the same host. The inclusion of the IDs of the parties involved and a timestamp essentially creates an unambiguous memorandum between the agent and the host.

After constructing D_i, the agent will request the host to perform a signature on the following:

$$c_i = \text{Sig}(D_i + c_{i-1} + ID_{host} + ID_{agent} + \text{timestamp}, k_{priv})$$

where c_0 is the digital signature on the agent code by its butler.

There are some advantages with the use of chained digital signature compared to the conventional signature approach. In the scenario when a malicious host attempts to modify the data from an innocent host *i* and somehow manages to produce a valid digital signature c_i, the data integrity would have been broken if the digital signature is independent and not chained to each other. The independent digital signature also opens the window for host *i* to modify data provided to the agent at a later time (one such scenario is the agent visits one of the host's collaborating partners later). Regardless of the message format used, so long as the messages are independent of each other, host *i* will have no problem reproducing a valid signature to the modified message. In this way, data integrity can be compromised. With chained digital signature, even if the malicious host (or host *i* itself) produces a valid digital signature after modifying the data, the new signature c_i' is unlikely to be the same as c_i. If the new signature is different from the original signature, as the previous signature is provided as input to the next signature, the subsequent signature verification will fail, thus detecting compromise to data integrity. The inclusion of a host ID, agent ID, and timestamp prevents anyone from performing a replay attack.

When the agent reaches a new destination, the host must perform an integrity check on the incoming agent. In the design of SADIS, even if the new destination host does not perform an immediate integrity check on the incoming agent, any compromise to the data integrity can still be detected when the agent returns to the butler. The drawback, however, is that the identity of the malicious host may not be established. One design focus of SADIS is not only to detect data integrity compromise, but more importantly, to identify malicious hosts. To achieve malicious host identification, it is an obligation for all hosts to verify the incoming agent's data integrity before activating the agent for execution. In the event of data integrity verification failure, the previous host will be identified as the malicious host.

Data integrity verification includes the verification of all the previous signatures. The verification of signature c_0 ensures agent code integrity, the verification of c_i ensures data provided by host h_i is intact. If any signature failed the verification, the agent is considered compromised.

While the process to verify all data integrity may seem to incur too much overhead and be somewhat redundant (e.g., why verify the integrity of d_1 in h_3 while host h_2 already verifies that), it is necessary to ensure the robustness of the protocol and to support the function of malicious host identification.

FUTURE TRENDS

Besides agent data integrity and agent transport security, there are other security concerns to be addressed in SAFER. One such concern is a mechanism to assess the agent's accumulated risk level as it roams. There have been some considerations for using the 'agent battery' concept to address this during the earlier stages of the research. Furthermore, in order to establish the identity of different agents from different agent communities, a certain level of certification by trusted third parties or an agent passport is required (Guan, Wang, & Ong, 2003). More research can be conducted in these areas.

CONCLUSION

In this article, a new data integrity protection protocol, SADIS, is proposed under the SAFER research initiative. Besides being secure against a variety of attacks and robust against vulnerabilities of related work in the literature, the research objectives of SADIS include efficiency. This is reflected in minimized use of PKI operations and

reduced message exchanges between the agent and the butler. The introduction of variation to DH key exchange and evolving communication session key further strengthened the security of the design. Unlike some existing literature, the data integrity protection protocol aims not only to detect data integrity compromise, but more importantly, to identify the malicious host.

With security, efficiency, and effectiveness as its main design focus, SADIS works with other security mechanisms under SAFER (e.g., Agent Transport Protocol) to provide mobile agents with a secure platform.

REFERENCES

Bellavista, P., Corradi, A., & Stefanelli, C. (2000). Protection and interoperability for mobile agents: A secure and open programming environment. *IEICE Transactions on Communications.*

Borselius, N. (2002). Mobile agent security. *Electronics & Communication Engineering Journal, 14*(5), 211-218.

Borselius, N., Hur, N., Kaprynski, M., & Mitchell, C.J. (2002). A security architecture for agent-based mobile systems. *Proceedings of the 3rd International Conference on Mobile Communications Technologies* (3G2002) (pp. 312-318), London.

Borselius, N., Mitchell, C.J., & Wilson, A.T. (2001). On mobile agent based transactions in moderately hostile environments. In B. De Decker, F. Piessens, J. Smits, & E. Van Herreweghen (Eds.), *Advances in Network and Distributed Systems Security Proceedings of the IFIP TC11 WG11.4 1st Annual Working Conference on Network Security* (pp. 173-186). Boston: Kluwer Academic.

Chionh, H.B., Guan, S.-U., & Yang, Y. (2001). Ensuring the protection of mobile agent integrity: The design of an agent monitoring protocol. *Proceedings of the IASTED International Conference on Advances in Communications* (pp. 96-99), Rhodes, Greece.

Corradi, A., Cremonini, M., Montanari, R., & Stefanelli, C. (1999). Mobile agents and security: Protocols for integrity. *Proceedings of the 2nd IFIP WG 6.1 International Working Conference on Distributed Applications and Interoperable Systems* (DAIS'99).

Diffie, W., & Hellman, M.E. (1976). New directions in cryptography. *IEEE Transactions on Information Theory, 22,* 644-654.

Esparza, O., Muñoz, J.L., Soriano, M., & Forné, J. (2006). Secure brokerage mechanisms for mobile electronic commerce. *Computer Communications, 29*(12), 2308-2321.

Guan, S.-U., & Hua, F. (2003). A multi-agent architecture for electronic payment. *International Journal of Information Technology and Decision Making, 2*(3), 497-522.

Guan, S.-U., Wang, T., & Ong, S.-H. (2003). Migration control for mobile agents based on passport and visa. *Future Generation Computer Systems,* 19(2), 173-186.

Guan, S.-U., & Yang, Y. (1999). SAFE: Secure-roaming agent for e-commerce. *Proceedings of the Computer & Industrial Engineering Conference '99* (pp. 33-37), Melbourne, Australia.

Guan, S.-U., & Yang, Y. (2002). SAFE: Secure agent roaming for e-commerce. *Computer & Industrial Engineering Journal, 42,* 481-493.

Guan, S.-U., Zhu, F., & Maung, M.T. (2004). A factory-based approach to support e-commerce agent fabrication. *Electronic Commerce and Research Applications, 3*(1), 39-53.

Ouardani, A., Pierre, S., & Boucheneb, H. (2006). A security protocol for mobile agents based upon the cooperation of sedentary agents. *Journal of Network and Computer Applications.*

Patrick, A.S. (2002). Building trustworthy software agents. *IEEE Journal of Internet Computing,* 46-53.

Park, J.Y., Lee, D.I., & Lee, H.H. (2002). One-time key generation system for agent data protection. *IEICE Transactions on Information and Systems,* 535-545.

Roth, V. (2001). On the robustness of some cryptographic protocols for mobile agent protection. *Proceedings of Mobile Agents 2001* (MA'01) (pp. 1-14).

Tripathi, A.R. (2002), Design of the Ajanta system for mobile agent programming. *Journal of Systems and Software, 62*(2), 123-140.

Wang, T., Guan, S.-U., & Chan, T.K. (2002). Integrity protection for code-on-demand mobile agents in e-commerce. *Journal of Systems and Software, 60*(3), 211-221.

Yang, Y., & Guan, S.-U. (2000). Intelligent mobile agents for e-commerce: Security issues and agent transport. In *Electronic commerce: Opportunities and challenges.* Hershey, PA: Idea Group.

Zhu, F., Guan, S.-U., Yang, Y., & Ko, C.C. (2000). SAFER e-commerce: Secure Agent Fabrication, Evolution and Roaming for e-commerce. In *Electronic commerce: Opportunities and challenges.* Hershey, PA: Idea Group.

KEY TERMS

Agent: A piece of software that acts to accomplish tasks on behalf of its user.

Cryptography: The art of protecting information by transforming it (*encrypting* it) into an unreadable format, called *cipher text*. Only those who possess a secret *key* can decipher (or *decrypt*) the message into *plain text*.

Flexibility: The ease with which a system or component can be modified for use in applications or environments other than those for which it was specifically designed.

Integrity: Regards the protection of data or program codes from being modified by unauthorized parties.

Mobile Agent: An agent that can move from machines to machines for the purpose of data collection or code execution. Also called a roaming agent.

Protocol: A convention or standard that controls or enables the connection, communication, and data transfer between two computing endpoints. Protocols may be implemented by hardware, software, or a combination of the two. At the lowest level, a protocol defines a hardware connection.

Security: The effort to create a secure computing platform, designed so that agents (users or programs) can only perform actions that have been allowed.

This work was previously published in Encyclopedia of Information Ethics and Security, edited by M. Quigley, pp. 556-562,

Chapter 1.29
Indexing Mobile Objects:
An Overview of Contemporary Solutions

Panayiotis Bozanis
University of Thessaly, Greece

ABSTRACT

Mobile computing emerged as a new application area due to recent advances in communication and positioning technology. As David Lomet (2002) notices, a substantial part of the conducted work refers to keeping track of the position of moving objects (automobiles, people, etc.) at any point in time. This information is very critical for decision making, and, since objects' locations may change with relatively high frequency, this calls for providing fast access to object location information, thus rendering the indexing of moving objects a very interesting as well as crucial part of the area. In this chapter we present an overview on advances made in databases during the last few years in the area of mobile object indexing, and discuss issues that remain open or, probably, are interesting for related applications.

INTRODUCTION

During the last years, a significant increase in the volume and the diversity of the data which are stored in database management systems has happened. Among them, spatio-temporal data is one of the most fast developing categories. This phenomenon can be easily explained since there is a flurry of application development concerning continuously evolving spatial objects in several areas. To name a few, mobile communication systems, military equipment in (digital) battlefields, air traffic, taxis, truck and boat fleets, and natural phenomena (e.g., hurricanes) all generate data whose spatial components are constantly changing.

In the standard database context, data remains unchanged unless an update is explicitly stated; for example, the phone number in an employee's record remains the same unless it is explicitly

updated. If this assumption was employed to continuously moving objects, then highly frequent updates should be performed. Otherwise, the database would be inaccurate and thus query outputs would be obsolete and unreliable.

In order to capture continuous movement and, additionally, spare unnecessary updates, it is widely accepted to store moving object positions as time-dependent functions, which results in updates triggered only by function parameter changes. For example, when objects follow linear movement, the parameters could be the position and the velocity vector of each object at the particular time the function (and therefore the object) is registered to the database. Usually, the moving objects are considered responsible for updating the database about alterations of their movement.

The following paragraphs present a comprehensive review on the various indexing proposals for accommodating moving objects in database systems, so that complex queries about their location in the past, the present, or the future can be served. The more elementary problem of location management, which asks for storing and querying the location of mobile objects based on the underlying network architecture, is surveyed in Pitoura and Samaras (2001). The works of Agarwal, Guibas, et al. (2002), Wolfson (2002), and Lomet (2002) discuss various aspects of modeling and manipulating motion, while the "lower level" subject of organizing (indexing) data for efficient broadcasting in wireless mobile computing is treated in Chen, Wu, and Yu (2003) and Shivakumar and Venkatasubramanian (1996); the interested reader could consult all these references for a wider introduction.

DEFINITIONS AND BACKGROUND

The indexes developed to accommodate moving objects can be classified into two broad categories:

(a) those optimizing queries about past states of movement, the so-called *historical queries*, and
(b) those designed to answer queries about future positions of the moving objects, which are termed *future* or *predictive queries.*

This categorization is not strict since there are structures enabling queries about both past and future positions. One can also group the indexes based upon whether the object trajectories are indexed or the objects themselves. However, in this study, we have chosen the classification according to whether the proposals are *practical* ones (that is, they have been actually implemented and experimentally investigated in realistic environments) or *theoretical* ones, which aim mainly at indicating the inherent complexity of the problem since their adoption to real applications is problematic because of the hidden constants or the involvement they exhibit, and, thus, their use is avoided.

Moving Object Representation

Since the size or the shape of a moving object is unimportant compared to the significance of its position as time evolves, its representation is directly related only to time and location. On the other hand, the way this spatiotemporal information is registered depends on the kind of the processing needed, namely, the postprocessing of recorded data and the exploration of current and future data location.

In the first case, object trajectories must be maintained. This means that one has the complete knowledge of location at any past time instance. Obviously, this is impossible, not only for storage limitations but also due to the nature of the underlying application; for instance, communication frameworks, like Global Positioning System (GPS) equipment, generate discrete location data. Therefore, trajectories must be calculated by

interpolating the sampled locations. Since linear interpolation is widely accepted for this task, each trajectory eventually turns into a polyline, that is, a sequence of connected line segments, in three-dimensional space. In conclusion, the problem of indexing trajectories reduces to indexing semantically related line segments. As we will see in the sequel, this extra information requires cautious extensions to spatial database indexes. The interested reader could find more about spatial indexes in Gaede and Günther (1998).

In the second case, the object position must be considered as a time function $x(t)$. Employing x, the application administrating the database can anticipate the future object locations, given that objects report any change in x parameter(s). Usually, $x(t)$ is modeled as a linear function of time, $x(t) = x(t_{ref}) + v(t - t_{ref})$, specified by two parameters: (a) the *reference position* $x(t_{ref})$ at a specific time t_{ref} and (b) the *velocity vector* v. This kind of representation is characterized as *parametric*, and its parameters define a dual to the (real) time-location space framework, which, as we elaborate later, was exploited for the development of new indexing methods.

Query Types

Location-Based Queries

The queries of this category can be further subdivided into *range queries* and *proximity* or *nearest-neighbor* (NN) ones. There are three different classes of range search queries: (a) *time-slice* or *snapshot query* (r, t), which specifies a hyper-rectangle r located at a time instant t, and asks for all moving objects that will be contained in r at that time, (b) *window query* (r, t), which requests reporting all objects crossing the hyper-rectangle r during the time interval $t = [t_s, t_e]$, and (c) *moving query* (r_1, r_2, t), where $t = [t_s, t_e]$, which specifies a $(d + 1)$-dimensional trapezoid τ by connecting r_1 at t_s and r_2 at t_e, and inquires all objects that will pass through τ.

On the other hand, a proximity query asks for specifying the nearest moving objects to a given location (spatial point) at time instance t or during a time interval t. There is also a generalization which asks for the *k nearest neighbors* (kNN). Sometimes, *reversed nearest neighbor* (RNN) searching is required, which asks for all objects having a given one as their nearest neighbor.

Trajectory-Based Queries

This type of queries concerns: (a) the topology of the trajectories, that is, information about the semantics of the movement, for example, objects entering, leaving, crossing, and bypassing a region during a given time instance or interval (for example, "Find all mobile phones entering a particular cell between 2 p.m. and 5 p.m. today."), and (b) derived or navigational information, for example, traveled distance, covered area, and velocity (for instance, "Report all objects whose traveled distance between 2 p.m. and 5 p.m. three days ago was smaller than 60 km/h.").

Continuous Queries

This kind of queries stands as a quite natural enhancement of the location-based ones by considering that the query range or point is also moving, for example, "Find all my nearest restaurants as I drive towards the current direction for the next 5 minutes." Despite its conceptual simplicity, these questions are not straightforward at all since they are equivalent to constantly posing location-based queries. As we will exhibit, current solutions capitalize on the semantics of the movement to achieve efficient processing.

Soundness-Enriched Queries

The members of this category are distinguished from the simple location-based counterparts since they demand answers enriched with validity-temporal or spatial-information. Specifically, they

additionally specify the future time *t* that the result expires and the change that will occur at time *t*. For instance, when one asks, "Report the nearest pharmacy to my position as I am moving now," the database will return the nearest pharmacy ID *i*, the future time *t* that *i* ceases to be the closest, and a new pharmacy *i'* that would be the next nearest at time *t*. Alternatively, the database can return a validity region *r* around the query position within which the answer remains valid. Returning to the previous example, *i* will be accompanied by the region that covers as the nearest.

This class of queries aims at reducing subsequent queries. To be explanatory, consider the scenario of a moving user posing a query to a server. The query and the server responses are delivered via a wireless network. Based on the fact that the mobility makes the validity of the answer highly volatile, the extra transmitted information could spare network bandwidth since the user will release new inquiries only when it is absolutely necessary.

PRACTICAL INDEXES

The indexes of this category can be roughly divided into two subcategories according to the time dimension: All these structures that are able to answer queries about the present and the future belong to the first group, while in the second one, one can find indexes which accommodate historical spatiotemporal data, and so, their main purpose is to respond to inquires about the past. We will see that this taxonomy is not strict-the structure in Sun, Papadias, Tao, and Liu (2004) indicates this fact.

Querying the Present and the Future

The members of this grouping can be further classified based on three types of supporting inquiries: range, nearest neighbor, and soundness enriched.

Structures Supporting Range Queries

Tayeb, Ulusoy, and Wolfson (1998) presented one of the earliest works on indexing mobile objects. Using linear functions to approximate object movement, they reduced the problem to indexing lines in the *xt*-plane. The transformed data set is stored in a periodically generated bucket Point Region (PR) quadtree (Samet, 1990), which stores a line segment in every quadrant of the underlying space that it crosses, partially or fully. The authors provided algorithms for answering snapshot and window queries. In short, this index suffers from the need of continuous rebuilding and the rather high space requirements.

Kollios, Gunopoulos, and Tsotras (1999b) suggested solutions for one- and two-dimensional range searching: When the objects are moving on the line, then either they could be indexed in the *xt*-plane using standard spatial solutions like R-trees (Guttman, 1984), or the problem could be solved by simplex range searching in the two-dimensional dual space. In the last case, when Hough-X dual space (Jagadish, 1990) is used, the employment of a dynamic version of external partition trees (Agarwal, Arge, Erickson, Franciosa, & Vitter, 2000) guarantees linear space complexity and $O(n^{1/2+\varepsilon} + k)$ worst-case query performance. On the other hand, the alternative of employing Hough-Y dual space (Jagadish, 1990) permits a practical approximation algorithm with linear space and expected logarithmic query time by segmenting the plane into *c* horizontal stripes and indexing them with *c* independent B+-trees in the same way Kollios, Gunopoulos, and Tsotras (1999a) did. When one accepts exact answers for a specified time period *T*, then the authors provide a solution of linear space complexity and worst-case logarithmic query performance by storing the relative ordering of the objects. Specifically, since between two crosses of objects the relative order of objects remains unchanged, they discover all objects crossing during *T* and store the orderings in an external version of persistent lists (Driscoll,

Sarnak, Sleator, & Tarjan, 1989). Finally, Kollios et al. considered also (a) the extension to a 1.5-dimensional space (that is, movement into routes) by, first, indexing the trajectories and then, on each trajectory, the moving points, and (b) the two-dimensional version of the problem by mapping each trajectory to a point in a four-dimensional space and indexing the resulting points with external partition trees.

On the other hand, Chon, Agrawal, and El Abbadi (2001) conducted preliminary investigation in which two parameters are sufficient for indexing moving objects. Adopting the convention that object movement can be described by four independent parameters, namely the velocity, the starting time, and the starting and ending positions of the movement, they examined all six different combinations when only two parameters can vary, the other two being constant, and concluded that the most promising dimensionality reduction results if one varies only the starting time and the ending position. The validity of the conclusion depends on performance study conducted using the SS-tree (White & Jain, 1996) as the underlying spatial index, and comparing the results with the dual space transformation of Kollios et al. (1999b).

In Šaltenis, Jensen, Leutenegger, and Lopez (2000), TPR-tree, a time-parameterized (TP) version of R*-trees (Beckmann, Kriegel, Schneider, & Seeger, 1990) for objects moving with constant velocities in one-, two-, and three-dimensional space, was introduced, which became the de facto spatial index for future queries. Therefore, we will be more analytical in our discussion. So, TPR-tree can be defined as a balanced, multiway tree. The moving points are accommodated in the leaf nodes while the internal nodes store pairs of a pointer to a subtree T and a TP bounding rectangle (TPBR) R_T, augmented with a velocity vector in such a way, it can bound the positions of all moving objects or other bounding rectangles in T. The TPBRs are defined in a conservative manner: If t_{ref} is the reference time and S is the set of all involved objects, then in each dimension x_i, the lower bound is set to be the minimum x_i-coordinate value in S at time t_{ref}, moving with the minimum observed velocity in S, and the upper bound is defined to be the maximum x_i-coordinate value in S at time t_{ref}, moving with the maximum observed velocity in S. Please notice that TPR-tree indexes the future trajectories of moving points as infinite lines. TPBRs never shrink and, since they may mistakenly grow, the authors suggested algorithms that keep the index tuned for H time units; after that, a global reorganization (or rebuilding) of the structure is necessary. Toward this end, the insertion and bulk-loading algorithms of the R*-tree are generalized so that their respective objective functions are time parameterized in the following way. Let A be an objective function and $A(t)$ its generalized counterpart with the involved metrics, like perimeters or overlap, being time dependent. Then, they minimize the integral

$$\int_{t_{ref}}^{t_{ref}+H} A(t)dt.$$

Three types of queries are supported: (a) time-slice queries, (b) window queries, and (c) moving queries. Type a calls for calculating the bounding rectangles of the index at time t before the intersection test is performed. Types b and c are served based on a very simple observation: the extents of a moving TPBR and the moving query should intersect in each dimension at a time point. The authors provide analytical formulae that compute such time intervals, thus avoiding the time-expensive, generic polyhedron intersection tests. In conclusion, one could say that the TPR-tree is very practical and, in the sense of avoiding the usage of dual space or the reduction of higher dimensional spaces, is a straightforward solution tested for uniformly generated one-, two-, and three-dimensional workloads and various values of the validation parameter H.

TPR-Tree algorithms extended in Benetis, Jensen, Karćiaukas, and Šaltenis (2002) so that they could answer two-dimensional NN and

RNN queries for a query point q during a time interval t. For the first case, they employ the standard approach of prune and depth-first search in R-tree-like indexes, like, for example, (Roussopoulos, Kelly, & Vincent, 1995) maintaining a list of intervals whose union equals to t, each associated with a point (or points) which is (are) the closest to q among the examined so far. The authors provide a metric M so that parts of the tree, on the average closest to q, are to be visited first; in this way, TPBRs with no chance of enclosing closer than the current closest set of descendant points are early pruned. The RNN queries are served based on the following fact: If we divide the space around the query point q into six equal sectors s_i by straight lines intersected at q, then there exist *at most* six RNN points, at most, one in each s_i. So, we first find the NN point(s) of q for each sector, which is a candidate for being the RNN point of q. If there exist two or more of them in a sector, then there are no tentative RNN points for q. Then each candidate point is checked for whether it has q as the NN. In order to spare disk access, all RNN candidate points are tested in one traversal of the index.

Šaltenis and Jensen (2002) proposed R^{EXP}-tree for indexing the current and anticipated future positions of moving objects, based on the assumption that objects' positions expire after a time period; this is quite natural to the context of location-based services where objects that have not reported their position within a given time period t_{exp} are assumed to be uninterested in the service, and so they are declared expired and are removed from the data set. Also, the authors introduced a new parameter, the query window length W, which is an upper bound on how far from their issue time queries are expected to refer to. This extra knowledge of the expiration time is used to derive better (tighter) TPBRs. In the one-dimensional case, let S be the set of involved moving objects (points or intervals), t_{exp} be the maximum expiration time observed in S, and $h = \min\{H, t_{exp}\}$. Then it is proven that the best TPBR enclosing S is the trapezoid with upper and lower bounds containing the edges of the convex hull of S, which intersect the line $t = t_{ref} + h/2$. Extending this observation to the d-dimensional case, the authors suggested either (straightforward) independent computation in each dimension, or the processing, in random order, of each dimension so that the computation in the ith dimension considers the processing made for the first i - 1 dimensions. The latter solution produces tighter TPBRs based on analytical formulas. As for the update operations, R^{EXP}-tree uses integrals of the form

$$\int_{t_{ref}}^{t_{ref}+\min\{H,t_{exp}\}} A(t)dt,$$

while employing a lazy removal of expired objects (points or TPBRs) only after an update operation discovers an expired entry. Compared to TPR-trees, where the objects are assumed not to expire, R^{EXP}-trees exhibit better experimental performance on artificially generated index workloads of factor 2 or more without any degradation of the update time.

Procopiuc, Agarwal, and Har-Peled (2002) introduced the Spatiotemporal Self-Adjusting R-tree (STAR) for two-dimensional moving points as an improvement over TPR-trees (Šaltenis et al., 2000) which self-adjusts whenever the query performance deteriorates without user interference; actually, the user specifies the parameter determining the quality or space consumption and self-adjusting time or query performance trade-offs. Using the result of Agarwal and Har-Peled (2001) for approximating the extent of moving points, they store the Minimum Bounding Rectangles (MBRs) as sequences (to be accurate, chains) of points so that every MBR, at any time instance, is described by interpolating along two axes. Additionally, employing a priority queue, they "refresh" the approximations as time evolves in order to keep them valid, and redistribute the

children of a node *v* when the children of *v* overlap too much. In order to exhibit the advantages of the STAR technique, the authors provide experimental evidence on both synthetic and realistic data sets, following the method used for experimentation in Šaltenis et al. (2000). Their main findings can be summarized as follows: (a) A speedup of 2-3 with respect to TPR-tree was achieved, (b) the deterioration of the scheme over time was proven to be not too much, and (c) the proposed approximations and heuristics actually work well. Overall, one can argue that the STAR proposal is a good example of how one can incorporate theoretical (geometrical) results into a practical mobile index in order to achieve good query time performance.

TPR-trees were recently improved in Tao, Papadias, and Sun (2003) by introducing TPR*-trees, which exhibit new insertion and deletion algorithms. In order to achieve this, the authors suggested a cost model for the original TPR-tree which emphasizes the factors that influence its performance. To be more specific, Tao et al. observed that the probability that a node is intersected by a query window *q* during a time interval t_q depends on the area $A_{SR}(o', t_q)$ swept by its extension-according to *q* characteristics-on each axis MBR *o'* during t_q. This metric is quite different from the integral metrics of TPR-trees, which do not differentiate between static and moving MBRs of the same area. So, the main goal of TPR*-tree's update algorithms is minimizing the quantity

$$C(q) = \sum_{\forall \text{node with MBR } o} A_{SR}(o', t_q).$$

In order to make the quantity work for every query *q*, the authors suggest optimizing TPR*-trees for the static point query q_0 for the time horizon of *H* time units, a choice which is fully justified by a thorough experimentation. Additionally, the insertion and deletion algorithms employ some more-elaborated decision making processes for insertion path selection, node reinsertion, and children node redistribution, which pay off a lot in terms of search performance. Finally, Tao et al. conducted extensive experiments that proved the superiority of TPR*-trees over the TPR-trees under all conditions; the average query cost is almost five times less and the average update cost is nearly constant while the TPR*-tree remains effective as time evolves. In conclusion, based on the previous discussion, the TPR*-tree can be characterized as the "state-of-the-art" index for serving range queries about the present and the future on moving objects. This fact, none the less, does not invalidate the practicality of the TPR-trees when one affords some administration cost on index maintenance in order to be able to answer two-dimensional NN and RNN queries for a query point *q* during a time interval *t*.

Structures Supporting Nearest Neighbor Queries

The members of this subcategory are additionally distinguished as follows.

Simple Proximity Indexes

Kollios et al. (1999a) suggested practical solutions for dealing with the (conceptually) simple problem of locating the nearest moving neighbor of a static query in the plane. The authors present performance studies for (a) indexing in the *xt*-plane, which is equivalent to indexing line segments or lines (i.e., the trajectories of the objects) with standard spatial indexes, like, for example, R-trees (Guttman, 1984) or R*-trees (Beckmann et al., 1990), so that one has to find the closest trajectory, and (b) segmenting the plane into *c* horizontal stripes and indexing in the dual Hough-Y space, with the employment of *c* B+-trees (Comer, 1979), one for each of the *c* horizontal stripes. The generality of the approach also permits finding the NN (a) within a specified time interval-since the query point becomes a line

segment, and (b) for restricted data movement in fixed line segments (routes) or, as it is called, movement in the 1.5-dimensional space; one first indexes line segments in spatial indexes, and, after locating the nearest route, then employs an NN search on the objects of the identified route. In short, this work, besides being the first one, can be characterized as preliminary since the cases studied are very restrictive.

On the other hand, Aggarwal and Agrawal (2003) introduced methods for indexing a special case of moving objects with nonlinear trajectories in arbitrary dimensions. Specifically, the functional representation $F(\Theta, t)$ of the trajectory, where $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is its associated parametric representation, is said to satisfy the convex hull property when, for any set of n trajectories $F(\Theta^1, t)$, $F(\Theta^2, t)$, …, $F(\Theta^n, t)$, the following holds: If Q' lies inside the convex hull of $\Theta^1, \Theta^2, \ldots, \Theta^n$, then $F(\Theta', t)$ lies inside the convex hull of $F(\Theta^1, t)$, $F(\Theta^2, t)$, …, $F(\Theta^n, t)$. Although the convex hull property is a property of the particular parametric representation of a trajectory and not a property of the trajectory itself, the authors demonstrated that very interesting categories of movement satisfy the property: d-dimensional trajectories with constant velocity, d-dimensional parabolic trajectories, elliptic orbits, and trajectories that accept approximate Tailor expansion. Since the convex hull property relates the locality in parametric space to the locality of the positions of objects, Aggarwal and Agrawal suggested the indexing of parametric representations of their trajectories using common multidimensional indexes like R*-trees (Beckman et al., 1990). Actually, they exhibited their method for solving the NN search problem by introducing a branch-and-bound, best-first algorithm for pruning and searching the underlying index which was investigated for linear and parabolic trajectories in three- and two-dimensional space, respectively. In conclusion, this work is quite interesting as being the first one extending, in a nontrivial way, the class of indexed trajectories for NN inquiries.

Continuous Proximity Indexes

Song and Roussopoulos (2001b) studied the kNN problem for a moving query point and static (i.e., not moving) data points, suggesting algorithms that extend static kNN ones, which is equivalent to the continuous kNN problem. The intuition behind the solutions presented there is that, when the query point moves to a new position, then some part of the previous answer must also belong to the new one. So, a series of conditions were proved that state when the previous answer set, or a part of it, remains valid as the query point moves to a new location, which search bound is appropriate to reinitiate the branch and bound at the new position, and how a prefetched-in-main-memory NNs set of cardinality $M > k$ is changing with query movement. The proposed algorithms were experimentally examined, and their simplicity renders them easy to implement, a fact that one should consider when using off-the-shelf R-trees.

Continuous NN queries were treated from a different perspective in Ishikawa, Kitagawa, and Kawashima (2002). Based on the observation that ellipsoid areas around moving query points are better than circular ones when the movement is conducted in a semiconstrained manner-for example, consider cars moving on city roads-they proposed an incremental ellipsoid query-generation algorithm which is built up on current, past, and future trajectory positions, and carefully selected metrics. The authors suggested indexing the static data set with a "standard" spatial structure, like, for example, R*-trees, but they do not reuse the previous answer for evaluating the next query position; instead, they search the spatial index from scratch, anticipating that page caching will spare some I/O cost. In a nutshell, this method needs elaborate tuning to be effective since the incremental update procedures are quite costly.

Continuous kNN search in a static data set has been studied by Tao, Papadias, and Shen

(2002). They actually followed the approach of Tao and Papadias (2002) which, as we will see in the sequel, is based on influence points and TP NN search. However, the authors tried to avoid the performance of an NN search from scratch each time the result expires. Instead, they suggest algorithms for evaluating the answers within a single traversal of the input-point set. Assuming the accommodation of the input set in an R-tree $\mathcal{T}$, the authors observed that when the query point is moving on a line segment l, the result consists of a list of l of NN points p_i; p_is partition l into a number of disjoint subsegments l_i, each one of them having p_i as the NN. So, Tao et al. proposed a bound-and-bound traversal of T that employs heuristic node-pruning rules which capitalize on the fact that a new point belongs to l as long as it "cancels" a current member(s) of l. The authors also provided conditions for guiding the search while pruning nodes during continuous kNN, and suggested cost models for node access estimation in case of uniform distributions; for the general case, they recommend the use of histograms. The extensive experimental evaluation of the methods strongly suggests the usefulness of the proposal when, of course, the data set is a static one.

Raptopoulou, Papadopoulos, and Manolopoulos (2003) proposed an algorithm for answering continuous kNN queries on moving points stored in a TPR-tree. Since the squared Euclidean distance between two moving points is described by a parabolic function of time, the authors observed that the kNN points during a time period t_q can be determined by the k levels of the arrangement of the squared distance functions of moving points with respect to the moving query during t_q. So they suggested a two-phase algorithm. First, the underlying TPR-tree is searched in a depth-first manner according to the minimum distance metric:

$$mindist(x, y) = \sqrt{\sum_j |x_j - y_j|^2}$$

between the moving query and the bounding rectangles at a starting time of t_q. When m ($m \geq k$) moving points are located, then, by considering their arrangement, the kNN points for the entire t_q are determined. Next, the second phase commences, which retraverses the TPR-tree, employing two pruning heuristics, and refines the answer as long as relevant, "promising" subtree branches still exist. The conducted experiments advocate the applicability of the method in case one has adopted the TPR-tree for indexing moving points and needs to serve continuous kNN query requests.

Finally, Iwerks, Samet, and Smith (2003) presented algorithms for answering continuous kNN queries on a constantly moving point set whose members can change either location or velocity. The proposed methods refer to static query points, but they can be applied to moving ones as well. Their solution capitalizes on the observation that answering continuous queries about objects within distance D to the query point is much easier to maintain than a kNN one. So, the authors suggested one to first filter points with a continuous within-distance query in order to reduce the number of points considered for the kNN query. By calculating the time instances when points change their distance to the query point or when points change their order with the current kNN point, and by ranking these two kinds of events with a priority queue, Iwerks et al. presented algorithms for query maintenance through the proper use of the queue, which is experimentally investigated. In general, the approach is quite interesting, extending naturally the repertoire of mobile NN queries; but, as Iwerks et al. noticed, further research is needed for fine-tuning.

Structures Supporting Soundness-Enriched Inquires

This class of indexes returns, along with the answer to either range or proximity queries, either tempo-

ral or spatial validity information, which specifies the conditions under which the answer remains in effect. In general, the solutions presented so far are relatively new, but also quite promising for further development and research.

Time-Parameterized Queries

Tao and Papadias (2002) introduced TP queries which return (a) the objects that satisfy the (spatial) conditions of the query, (b) the expiration time t_{exp} of the validity of the answer, and (c) the change that invalidates the answer at time t_{exp}. TP queries are very useful in contexts like location-based applications, air-traffic control, and so forth. In order to solve the problem, the authors introduced the concept of the influence time t^o_{inf} associated with each moving object *o*, which indicates the time *o* influences the result. In this way, finding the expiration time of the answer reduces to discovering the minimum influence time of all objects which, in turn, is equivalent to NN search with distance metric being the influence time. The above observation is valid for the treatment of window, kNN, and join queries. In the first case, the influence time of an object or its MBR equals the minimum intersection time along all dimensions. The authors provided also analytical formulae for evaluating the t^o_{inf} of an object *o*, which is used in a standard branch-and-bound traversal of the index accommodating the object set. The same rationale was also followed for the kNN queries; however, due to relatively increased complexity of the intersection conditions, simpler approaches were proposed which, none the less, behave excellent in practice. Now, in the join case, the influence time is defined for pairs of objects (o_1, o_2) as the minimum time when o_1 and o_2 either stop or start satisfying the join conditions. Concluding their work, Tao and Papadias also exhibited how TP queries can be used to answer continuous *spatiotemporal* and *earliest event* queries. The first type refers to the case when a time interval is defined during which one should provide results as they are generated. It follows promptly that this can be served by continuous TP queries at times when the current result expires. The second type asks for the evaluation of the earliest time in the future a specified event could take place; for example, in the scene of moving point objects and query point *q*, find the first time *q* catches a point. By surrounding *q* with a time-varying radius cycle, this query reduces to TP by evaluating the earliest time the circle contains a point, which, in turn, equals to determining the smallest radius of such a circle.

Validity Queries

In Zheng and Lee (2001), the problem of enabling the mobile clients to determine the validity of query results based on their current location was considered. This approach aims at reducing the network traffic by reducing the number of client queries: The server, additionally to the query result, returns a validity region *r* within which there is no alteration of the answer, and so, clients can issue subsequent queries only when they leave *r*. Please note that this treatment departs from the "standard" future prediction queries where the time, and *not* the query location, defines the answer. This view permits the following approach: Since the data set is static, its Voronoi diagram is constructed. This diagram is used to locate the NN of a moving query with respect to its current position by determining the Voronoi cell *c* which encloses it. Additionally, the maximum circle around a query point which does not cross any bounding edge of *c* is calculated and returned to the user as the safer lower bound of the validity of the query result.

Zhang, Zhu, Papadias, Tao, and Lee (2003) also dealt with validity NN, kNN, and window queries. For the first two cases, observing that the validity region is the Voronoi and order-*k* Voronoi cell, respectively, they suggested algorithms that implicitly calculate the respective cell-that is, without calculating the whole Voronoi or the

order-*k* Voronoi diagram-by, first, finding the NN(s), and then issuing TP NN or kNN queries (Tao & Papadias, 2002) for locating the points defining the border of the cell. The third case of the validity region of a window query is reduced to finding the maximal rectangle around the focus (center) of the window, where the result remains unchanged, which is then sharpened by removing the parts of it that would force the query containing other points not in the reported answer. Zhang et al. proved that these calculations require posing one standard window query, one "holey" window query, and just a few main-memory TP window queries.

Querying the Past

Historical databases on moving objects accommodate spatiotemporal information which can be processed to either *report* or *enumerate* all objects satisfying certain spatial and temporal conditions. In the first case, the corresponding indexes are characterized as *reporting*, while, in the second one, they are termed as *aggregating* or *enumerating*. Here we must note that an aggregating index does not simply count objects; it must be able to provide whatever summarized data, like, for example, average statistics, the related application needs.

Reporting Indexes

In Pfoser, Jensen, and Theodoridis (2000), spatiotemporal queries on mobile object trajectories without capabilities in future prediction queries are treated. The work is based on the observation that the simple use of R-tree (Guttman, 1984), and therefore the employment of MBR approximations, for storing the line segments of trajectories just leave a lot of "dead" space. They proposed two solutions, the STR-tree and the TB-tree. The first one consists of an extension of R-trees so that lines belonging to the same trajectory are kept close together while the main dimension is time. This is accomplished by trying to store new segments to the leaf accommodating their predecessors, and, when this is impossible, it performs leaf split, putting together more recent segments into a new node only if there exists a vacant predecessor, at most, p - 1 levels up; otherwise, new segments are inserted to leaves lying to the right of the search path. This modification permits the combined queries since, at the first stage, the execution of the R-tree range search algorithm identifies the segment trajectories belonging to the initial range and, at the second stage, recursive range queries with endpoints of discovered segments retrieve the final answer.

On the other hand, TB-tree departs from the underlying rationale of R-trees, which assumes independency among geometries, and focuses on spatiotemporal queries. For this, it cuts every trajectory into pieces, each piece is stored in a leaf, and the various leaves are connected to form a linearly linked list. In order to insert a new segment, one first locates the leaf accommodating its predecessor as STR-tree does. If there does not exist any space, then a new leaf is created and becomes the rightmost one of the index. This approach means that a trajectory is allocated to a set of leaves, and the leaves are organized in a tree hierarchy so that combined queries in the beginning locate the relevant leaves, and then follow list pointers as long as the second range constraints are met. Experimental evaluation of the index proved its efficiency compared to the R-tree. Further tuning of the TB-tree performance in case of it being constrained by the presence of infrastructure, like roads, lakes, and so forth, was achieved in Pfoser and Jensen (2001) by suggesting segmentation of the original query range in a set of subranges so that dead space can be eliminated.

The case of indexing moving objects with potential changing extended for historical queries without future prediction capability (snapshot and small-range queries) was treated in Kollios, Gunopoulos, Tsotras, Delis, and Hadjieleftheriou

(2001) and Hadjieleftheriou, Kollios, Tsotras, and Gunopoulos (2002). The authors followed the approach of approximating the object movement with an MBR, but, in order to eliminate empty space and overlap, they suggested the usage of artificial splits and insertion of the pieces into a partially persistent R-tree (Kumar, Tsotras, & Faloutsos, 1998). The first work refers to the case of linear functions of time. It presents a greedy algorithm which, given the preferred number of splits is proportional to the number of objects, calculates split points which minimize the overall empty space, capitalizing on the monotonicity of such a kind of objects, which guarantees that the savings in empty space decrease as the number of splits increases. This property does not apply to the case of general time functions, so the second paper suggested (a) an optimal, straightforward, dynamic programming algorithm, quadratic to the number of splits, and (b) two greedy solutions of linear-to-the-number-of-splits and subquadratic-to-the-number-of-objects complexity, aiming at finding the next object to split according to the maximum possible volume reduction.

The treatment of mobile objects in a spatio-temporal context was also addressed in Porkaew, Lazaridis, and Mehrotta (2001). Assuming the use of the "ubiquitous" R-tree (Guttman, 1984) as the underlying index, the authors proposed algorithms and intersection conditions for various combinations of spatial and temporal range, NN, and kNN queries when the trajectories are approximated by MBRs, and when the movement is described in location-velocity, multidimensional, parametric space.

Song and Roussopoulos (2001a, 2003), based on the fact that indexing static objects has been extensively investigated, suggested the zoning-based, index-updating police. The space is partitioned into regions, the so-called *zones*. Each zone is related to the objects it contains through a bucket. In that way, objects trigger updates only when they move into a different zone while range queries are reduced to querying the buckets whose zones intersect the query region. The generality of the approach gives the freedom to choose any index for bucket manipulation while the authors introduced a new index, the SEB-tree, for indexing the objects that left the bucket. SEB-tree exploits the fact that "absent" objects can be represented by two-dimensional points with increasing coordinate values. This property permits the segmentation of the plane into rectangular regions indexed by a B-tree.

Lazaridis, Porkaew, and Mehrotra (2002), assuming the approximation of object trajectories by MBRs which are stored in R-trees, presented algorithms for moving window queries with known (predictive) and unknown (nonpredictive) moving patterns. In short, the authors suggested the use of a priority queue and specific intersection conditions which guide and prune the search of the underlying R-tree in a way similar to the one in Roussopoulos et al. (1995) that serves standard kNN queries in static points.

Aggregating or Enumerating Indexes

Papadias, Tao, Kalnis, and Zhang (2002) introduced indexes for answering aggregate queries in spatiotemporal databases, that is, queries that ask for summarized data over two-dimensional regions satisfying specific spatiotemporal conditions. The first index, aggregate RB-tree (aRB-tree), refers to the case of fixed spatial dimensions where one needs to maintain historical summary data. The aRB-tree consists of the generalization of aggregate R-trees (aR-trees) of Papadias, Kalnis, Zhang, and Tao (2002) to three-dimensional space (two spatial dimensions and one time dimension). The main block is an R-tree indexing spatial regions in the form of MBRs. Each MBR region stores accumulated data for the entire region and historical summarized data for the time dimension in a B-tree. In this way, the spatial part of the query guides the search using the R-tree part, and, in each intersecting node, the corresponding B-tree is accessed for

the temporal part. When the regions are dynamic, the authors proposed two solutions. The first one, aggregate historical RB-tree (aHRB-tree), uses the node copying technique (Driscoll et al., 1989) for the main R-tree in order to avoid unnecessary replication of unaffected regions. The second one, the aggregate 3DRB-tree (a3DRB-tree), can be used when all region changes are known in advance; in such case, each version of a region is stored in the form of a three-dimensional box as a distinct entry in a 3DR-tree (Papadias, Kalnis, et al., 2002), resulting, thus, in space reduction compared to the aHRB-tree.

The shortcoming of the above work is that it does not support distinct counting of the objects; if an object remains in the query region for several time stamps during the query interval, then it will also participate in the result several times. As the distinct counting property is very crucial in many decision-making queries, like, for example, traffic analysis and mobile phone users' statistics, Tao, Kollios, Considine, Li, and Papadias (2004) suggested algorithms that do not use direct counting of the objects, as the volume of the data or legal issues about personal data may prohibit this approach. Instead, they are using FM sketches (Flajolet & Martin, 1985), as one nowadays finds in many data stream processing algorithms. So, they replace the summary B-tree of the aRB-tree proposal with a "sketch" B-tree, where the aggregate information of the leaves is maintained in an FM structure, while each intermediate node sketch is formed by or forming the sketches of all the sketches in its subtree. In this way, the resultant structure becomes an *approximate*, distinct counting index of bounded probability failure, while the search algorithm of aRB-tree remains the same. The authors actually proposed three heuristic rules that exploit the pruning capabilities of sketches additively to the spatial and temporal conditions of the query. Additionally, they discussed how their scheme can be extended to support distinct sum queries, how their sketch B-tree can be applied to a hierarchy of increasing-resolution regular grids, and in which way sketches can be used in mining spatiotemporal association rules.

A first approach to support approximate queries about the past, the present, and the future was recently reported in Sun et al. (2004). In order to achieve their goals, the authors presented an adaptive multidimensional histogram (AMH) for answering present-time queries. AMH is based on a regular grid partition of the plane into cells. Each cell maintains its frequency, that is, the number of enclosing objects. The cells are distributed into a (upper bounded) number of buckets with the aid of a binary partition tree. Each bucket contains its area, average, and average squared frequency of the respective cells. Employing continuous bucket merging and splitting during idle CPU cycles, AMH reorganizes itself, in an interruptive way, to follow the data distribution and, thus, it can successfully serve present-time queries by retrieving the buckets intersected by the query window. The outdated bucket versions are stored in a main-memory index for answering historical queries, which is either a packed B-tree or a three-dimensional R-tree, depending upon the update and query performance tradeoff one accepts. The historical index gradually emigrates to disk. Finally, based on the observation that overall data distribution varies slowly and smoothly despite the abrupt nature of the individual objects' velocities, the authors applied exponential smoothing, a time-series forecasting method, for serving queries about the future with the use of recent history data.

THEORETICAL SOLUTIONS

Basch, Guibas, and Hershberger (1999) provided a theoretical framework for dealing with moving objects by introducing the concept of *kinetic framework*, which refers to storing only a combinatorial instance of the data set at any time. That is, in spite of the continuous character of the movement, the data structure changes only

on certain discrete *kinetic events* which depend strictly on combinatorial properties. The crux of the framework is that since the way the points move is known, one can calculate when kinetic events will take place and manipulate them with a priority queue, the so-called *event queue*.

A series of theoretical, computational-geometry main-memory results followed the above paradigm. To name a few; Agarwal, Gao, and Guibas (2002) proposed algorithms for "kinetizing" kd-trees; Karavelas and Guibas (2001) coped with kinetic spanners; Czumaj and Sohler (2001) approximate kinetic data structures like binary search trees, range trees, and heaps; Kaplan, Tarjan, and Tsioutsiouliklis (2001) designed kinetic heaps; Agarwal and Har-Peled (2001) presented approximate algorithms for maintaining descriptors of the extent of moving points; Agarwal, Guibas, Murali, and Vitter (2000) considered the cylindrical binary space partitions; Agarwal, Basch, de Berg, Guibas, and Hershberger (2000) proved lower bounds of kinetic planar subdivisions; and Agarwal, Eppstein, Guibas, and Henzinger (1998) studied kinetic minimum spanning trees. In the following lines, we will present secondary-memory theoretical solutions, that is, theoretical indexing schemes for data residing in a database.

Indexes Supporting Range Queries

Agarwal, Arge, and Erickson (2003) proposed four mainly theoretical indexing schemes for moving points in the plane. The first one improves upon the approach of Kollios et al. (1999b); instead of mapping points to four-dimensional space and using one-level partition trees, they project the points to the *xt*-planes and to *yt*-planes, employ transformations to dual *xt*-planes and to *yt*-planes, and the resulting point sets are stored to a two-level external partition tree so that the answer set is included in two stripes of the dual planes. This solution has linear space and $O(n^{1/2+\varepsilon} + k)$, k output size, and $O(\log^2 n)$ amortized update complexity. The structure can also serve range queries for a time interval t with the same complexity bounds. The authors also suggested two indexes based on the kinetic framework of Basch et al. (1999) for queries referring to the present time or arriving in a strict chronological order. In case of one-dimensional linear moving points, they stored the points ordered across the line in a kinetic B-tree which is updated only when two points interchange positions; these moments are handled by an external priority queue. The solution has linear space and logarithmic query and update complexity, and it can be combined with partition trees to achieve a trade-off between the query time and the number of kinetic events. In order to extend the approach to two dimensions, the authors stored the points into a kinetic version of range tree whose function is governed by a global event queue and two kinetic B-trees, one for each coordinate. This approach results in increases in space and in update time by an $O(\log_B n/\log_B \log_B n)$ factor. The fourth solution offers approximate results to NN searching by replacing the Euclidean metric with a polyhedral one, whose unit ball is a regular m-gon, and storing the points in a three-level quasi-linear space index with $O(n^{1/2+\varepsilon}/\sqrt{\delta})$ query time, $0 < \varepsilon, \delta < 1$. The first two levels are partition trees on dual *xt*- and *yt*-planes while the third one consists of lists containing the lower envelope of the trajectories.

Agarwal, Arge, and Vahrenhold (2001) provide kinetic solutions for answering window and moving window queries for one- and two-dimensional moving points with time complexities that depend on the number of events $\phi(t_q)$ that occurred between the current time t_{now} and the query time t_q. For the case of moving points on the line, the *yt*-plane is segmented into a logarithmic number of vertical strips or slabs so that the number of events in slab i is $O(nB^i)$, B being the page capacity, and, in each slab, the part of the arrangement that belongs to it is represented by $O(n/B^i)$ carefully selected *y*-ordered levels which are stored into a persistent B-tree. In this way, a window query

with $nB^{i-1} < \phi(t_q) \leq nB^i$ reduces to first, locating the respective window, and then searching the right version of the B-tree in $O(\log_B n + k/B + B^{i-1})$ time, k being the output size. The update cost of an event is bounded by $O(\log^3 n)$ and the space complexity is $O(n/B\log_B n)$. The authors extended their approach to the case of moving points in the plane. They first divided the *xyt*-space into a logarithmic number of slabs along the time axis so that the number of events in slab i is $O(nB^i)$. In each slab, the projections of the arrangement in the *xt*- and *yt*-planes are represented in an analogy to the one-dimensional case manner, that is, by carefully selected levels of arrangements. This construction results in $O(\log^3 n)$ update cost of an event, $O(n/B\log_B n)$ space complexity, and

$$O(\sqrt{n/B^i}\,(B^{i-1} + \log_B n) + k/B)$$

moving window query cost.

Indexes Supporting Soundness-Enriched Queries

Tao, Mamoulis, and Papadias (2003) presented the first theoretical bounds on validity information of various types of range query, or of NN-search query answers which refer to their expiration time t and their change at time t to remain valid. More specifically, when the query's length and movement are chosen from a constant number of combinations and the point set is static, the query cost is logarithmic and the space is linear by subdividing the plane into disjoint areas which are stored in a persistent B-tree. When the point set is static, the query's length is arbitrary, but the movement is axis parallel; then, by combining a primary B-tree of logarithmic fan-out with an external priority search tree (Arge, Samoladas, & Vitter, 1999), the query cost is $O(\log_B{}^2(n/B)/\log_B\log_B(n/B))$ and the space is $O(n/B\log_B(n/B)/\log_B\log_B(n/B))$. For the general case of static point set and queries with arbitrary length and movement, a slightly modified version of partition trees (Agarwal, Arge, et al., 2000) permits linear space and $O((n/B)^{1/2+\varepsilon})$ query cost, whereas a double transformation to the slope-rank space allows the application of simplified external range trees so that a validity query costs $O(\log_B{}^2(n/B)/\log_B\log_B(n/B))$ time and $O(n^2/B\log_B(n/B)/\log_B\log_B(n/B))$ space. When the data points are dynamic whereas the query is static, the authors proposed plane sweeping during which the ordering of the trajectories is stored in a persistent, aggregate B-tree which guarantees logarithmic query cost and $O(n^2/B\log_B(n/B))$ space. The general case of both dynamic data points and query is only considered in one-dimensional space by accommodating the cells of the arrangement currently intersected by the query (the zone) in a B-tree and maintaining it with an external priority queue, resulting in linear space consumption and logarithmic query time. On the other hand, for the NN search queries, when the points are statically lying in the plane, the careful use of Voronoi diagrams solves the problem with linear space and logarithmic query cost, whereas, in case of moving points on the line, the storage of the zone permits linear space and logarithmic query complexity.

CONCLUSION AND FUTURE DIRECTION

A broad category of applications, emerging from the technological advances that occurred in the last years in telecommunications and hardware, constitute the framework known as mobile computing. A key issue in such a framework is the fast and on-time access to a constantly changing data set. This fact indicates the crucial role of indexing moving objects. In this study we provided a comprehensive review of contemporary solutions in this area of research from the database (application) perspective, which, none the less, presents some very interesting topics for further consideration. First, it would be very helpful of the indexes to provide results that capture the

uncertainty associated with the location of moving objects due to network delays and the continuous character of motion. Trajcevski, Wolson, Zhang, and Chamberlain (2002) deal with the issue of modeling and querying about uncertainty, but, there is a lot to be done. It would be also very interesting to efficiently cope with nonlinear trajectories since the scope and the range of indexable moving objects will be significantly extended. Another appealing subject, especially for extending mobile application capabilities, is the design of indexing structures capable of serving "mixed" queries concerning the past and the future of movement; the approach in Sun et al. (2004) can be considered as a first attempt toward this end. The incremental valuation of validity queries is very intriguing, as well. Finally, from an engineering perspective, it would be very helpful (a) to test all indexes with real data sets, as, until now, every experimental investigation is conducted with "semireal" ones, where the movement component is actually generated, and (b) to design efficient updating algorithms for the indexes, different from the usual "deletion and reinsertion" practice, accepting perhaps a tradeoff between either the query time or the accuracy of the result and the update time.

REFERENCES

Aggarwal, C. C., & Agrawal, D. (2003). On nearest neighbor indexing of nonlinear trajectories. In *Proceedings of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, June 9 -12, 2003, San Diego, California, USA (pp. 252-259). New York: ACM.

Agarwal, P. K., & Har-Peled, S. (2001). Maintaining approximate extent measures of moving points. In *Proceedings of the 12th Annual Symposium on Discrete Algorithms*, January 7-9, 2001, Washington, DC (pp. 148-157). New York: ACM/SIAM.

Agarwal, P. K., Arge, L., & Erickson, J. (2003). Indexing moving points. *Journal of Computer and System Sciences, 66* (1), 207-243.

Agarwal, P. K., Arge, L., & Vahrenhold, J. (2001). Time responsive external data structures for moving points. In F. K. H. A. Dehne, J.-R. Sack, & R. Tamassia (Eds.), *Algorithms and Data Structures, Seventh International Workshop, WADS 2001, Providence, RI, USA, August 8-10, 2001, Proceedings, LNCS 2125* (pp. 50-61). Berlin, Germany: Springer.

Agarwal, P. K., Arge, L., Erickson, J., Franciosa, P. G., & Vitter, J. S. (2000). Efficient searching with linear constraints. *Journal of Computer and System Sciences, 61*(2), 194-216.

Agarwal, P. K., Basch, J., de Berg, M., Guibas, L. J., & Hershberger, J. (2000). Lower bounds for kinetic planar subdivisions. *Discrete & Computational Geometry, 24*(4), 721-733.

Agarwal, P. K., Eppstein, D., Guibas, L. J., & Henzinger, M. R. (1998). Parametric and kinetic minimum spanning trees. In *39th Annual Symposium on Foundations of Computer Science, FOCS '98*, November 8-11, 1998, Palo Alto, California, USA (pp. 596-605). New York: IEEE Computer Society.

Agarwal, P. K., Gao, J., & Guibas, L. J. (2002). Time responsive external data structures for moving points. In R. H. Mohring & R. Raman (Eds.), *Algorithms: ESA 2002, 10th Annual European Symposium*, Rome, Italy, September 17-21, 2002, Proceedings, LNCS 2461 (pp. 5-16). Berlin, Germany: Springer.

Agarwal, P. K., Guibas, L. J., Edelsbrunner, H., Erickson, J., Isard, M., Har-Peled, S., et al. (2002). Algorithmic issues in modeling motion. *ACM Computing Surveys, 34*(4), 550-572.

Agarwal, P. K., Guibas, L. J., Murali, T. M., & Vitter, J. S. (2000). Cylindrical static and kinetic

binary space partitions. *Computational Geometry, 16*(2), 103-127.

Arge, L., Samoladas, V., & Vitter, J. S. (1999). On two-dimensional indexability and optimal range search indexing. *Proceedings of the 18th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems,* May 3-June 2, 1999, Philadelphia, Pennsylvania (pp. 346-357). New York: ACM Press.

Basch, J., Guibas, L. J., & Hershberger, J. (1999). Data structures for mobile data. *Journal of Algorithms, 31*(1), 1-28.

Beckmann, N., Kriegel, H.-P., Schneider, R., & Seeger, B. (1990). The R*-tree: An efficient and robust access method for points and rectangles. In H. Garcia-Molina & H. V. Jagadish (Eds.), *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data,* Atlantic City, NJ, May 23-25, 1990 (pp. 322-331). New York: ACM.

Benetis, R., Jensen, S., Karčiauskis, G., & Šaltenis, S. (2002). Nearest neighbor and reverse nearest neighbor queries for moving objects. In M. A. Nascimento, M. T. Ozsu, & O. R. Zaiane (Eds.), *International Database Engineering & Applications Symposium, IDEAS'02,* July 17-19, 2002, Edmonton, Canada, Proceedings (pp. 44-53). New York: IEEE Computer Society.

Chen, M.-S., Wu, K.-L., & Yu, P. S. (2003, January/February). Optimizing index allocation for sequential data broadcasting in wireless mobile computing. *IEEE Transactions on Knowledge and Data Engineering, 15*(1), 161-173.

Chon, H. D., Agrawal, D., & El Abbadi, A. (2001). Storage and retrieval of moving objects. In K.-L. Tan, M. J. Franklin, & J. C. S. Lui (Eds.), *Mobile Data Management, Second International Conference, MDM 2001,* Hong Kong, China, January 8-10, 2001, Proceedings, LNCS 1987 (pp. 173-184). Berlin, Germany: Springer.

Comer, D. (1979). The ubiquitous B-tree. *ACM Computing Surveys, 11*(2), 121-137.

Czumaj, A., & Sohler, C. (2001). Soft kinetic data structures. In *Proceedings of the 12th Annual Symposium on Discrete Algorithms,* January 7-9, 2001, Washington, DC (pp. 865-872). New York: ACM/SIAM.

Driscoll, J. R., Sarnak, N., Sleator, D. D., & Tarjan, R. E. (1989). Making data structures persistent. *Journal of Computer and System Sciences, 38*(1), 86-124.

Flajolet, P., & Martin, G. N. (1985). Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences, 31*(2),182-209.

Gaede, V., & Günther, O. (1998). Multidimensional access methods. *ACM Computing Surveys, 30*(2), 170-231.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In B. Yormark (Ed.), *SIGMOD'84, Proceedings of Annual Meeting,* Boston, Massachusetts, June 18-21, 1984 (pp. 47-57). New York: ACM Press.

Hadjieleftheriou, M., Kollios, G., Tsotras, V. J., & Gunopoulos, D. (2002). Efficient indexing of spatiotemporal objects. In C. S. Jensen, K. G. Jeffery, J. Pokorny, S. Šaltenis, E. Bertino, K. Bohm, & M. Jarke (Eds.), *Advances in Database Technology EDBT 2002, Eighth International Conference on Extending Database Technology,* Prague, Czech Republic, March 25-27, Proceedings, LNCS 2287 (pp. 251-268). Berlin, Germany: Springer.

Ishikawa, Y., Kitagawa, H., & Kawashima, T. (2002). Continual neighborhood tracking for moving objects using adaptive distances. In M. A. Nascimento, M. T. Ozsu, & O. R. Zaiane (Eds.), *International Database Engineering & Applications Symposium, IDEAS'02,* July 17-19, 2002, Edmonton, Canada, Proceedings (pp. 54-63). New York: IEEE Computer Society.

Iwerks, G. S., Samet, H., & Smith, K. (2003). Continuous *k*-nearest neighbor queries for continuously moving points with updates. In J. C. Freytag, P. C. Lockemann, S. Abiteboul, M. J. Carey, P. G. Selinger, & A. Heuer (Eds.), *VLDB 2003, Proceedings of 29th International Conference on Very Large Data Bases,* September 9-12, 2003, Berlin, Germany (pp. 512-523). St. Louis, MO: Morgan Kaufmann.

Jagadish, H. V. (1990). On indexing line segments. In *VLDB 1990, Proceedings of 16th International Conference on Very Large Data Bases,* August 1990, Brisbane, Queensland, Australia (pp. 614-625). St. Louis, MO: Morgan Kaufmann.

Kaplan, H., Tarjan, R. E., & Tsioutsiouliklis, K. (2001). Faster kinetic heaps and their use in broadcast scheduling. In *Proceedings of the 12th Annual Symposium on Discrete Algorithms,* January 7-9, 2001, Washington, DC, (pp. 836-844). New York: ACM/SIAM.

Karavelas, M. I., & Guibas, L. J. (2001). Static and kinetic geometric spanners with applications. In *Proceedings of the 12th Annual Symposium on Discrete Algorithms,* January 7-9, 2001, Washington, DC (pp. 168-176). New York: ACM/SIAM.

Kollios, G., Gunopoulos, D., & Tsotras, V. J. (1999a). Nearest neighbor queries in a mobile environment. In M. H. Bohlen, C. S. Jensen, & M. Scholl (Eds.), *Spatio-Temporal Database Management, International Workshop STDBM'99,* Edinburgh, Scotland, September 10-11, 1999, Proceedings, LNCS, 1678 (pp. 119-134). Berlin, Germany: Springer.

Kollios, G., Gunopoulos, D., & Tsotras, V. J. (1999b). On indexing mobile objects. In *Proceedings of the 18th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems,* May 31-June 2, 1999, Philadelphia, Pennsylvania (pp. 261-272). New York: ACM Press.

Kollios, G., Gunopoulos, D., Tsotras, V. J., Delis, A., & Hadjieleftheriou, M. (2001). Indexing animated objects using spatio-temporal access methods. *IEEE Transactions on Knowledge and Data Engineering, 13*(5), 742-777.

Kumar, A., Tsotras, V. J., & Faloutsos, C. (1998). Designing access methods for bitemporal databases. *IEEE Transactions on Knowledge and Data Engineering, 10*(1), 1-20.

Lazaridis, I., Porkaew, K., & Mehrotra, S. (2002). Dynamic queries over mobile objects. In C. S. Jensen, K. G. Jeffery, J. Pokorny, S. Šaltenis, E. Bertino, K. Bohm, & M. Jarke (Eds.), *Advances in Database Technology EDBT 2002, Eighth International Conference on Extending Database Technology,* Prague, Czech Republic, March 25-27, Proceedings, LNCS 2287 (pp. 269-286). Berlin, Germany: Springer.

Lomet, D. (2002). Letter from the editor-in-chief [Special issue]. *Bulletin of the Technical Committee on Data Engineering, 25*(2), 1.

Papadias, D., Kalnis, P., Zhang, J., & Tao, Y. (2001). Efficient OLAP operations in spatial data warehouses. In C. S. Jensen, M. Schneider, B. Seeger, & V. J. Tsotras (Eds.), *Advances in Spatial and Temporal Databases, Seventh International Symposium, SSTD 2001,* Redondo Beach, California, USA, July 12-15, 2001, Proceedings, LNCS 2121 (pp. 443-459). Berlin, Germany: Springer.

Papadias, D., Tao, Y., Kalnis, P., & Zhang, J. (2002). Indexing spatio-temporal data warehouses. In *Proceedings of the 18th International Conference on Data Engineering,* February 26- March 1, 2002, San Jose, California, USA (pp. 166-175). New York: IEEE Computer Society.

Pfoser, D., & Jensen, C. S. (2001). Querying the trajectories of on-line mobile objects. In *Proceedings of the Second ACM International Workshop on Data Engineering for Wireless and Mobile Access, MobiDE,* May 20, 2001, Santa Barbara, California, USA (pp. 66-73). New York: ACM.

Pfoser, D., Jensen, C. S., & Theodoridis, Y. (2000). Novel approaches to the indexing of moving object trajectories. In A. El Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, & K.-Y. Whang (Eds.), *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases,* September 10-14, 2000, Cairo, Egypt (pp. 395-406). St. Louis, MO: Morgan Kaufmann.

Pitoura, E., & Samaras, G. (2001, July/August). Locating objects in mobile computing. *IEEE Transactions on Knowledge and Data Engineering, 13*(4), 571-592.

Porkaew, K., Lazaridis, I., & Mehrotta, S. (2001). Querying mobile objects in spatio-temporal databases. In C. S. Jensen, M. Schneider, B. Seeger, & V. J. Tsotras (Eds.), *Advances in Spatial and Temporal Databases, Seventh International Symposium, SSTD 2001,* Redondo Beach, California, USA, July 12-15, 2001, *Proceedings, LNCS 2121* (pp. 59-78). Berlin, Germany: Springer.

Procopiuc, C. M., Agarwal, P. K., & Har-Peled, S. (2002). STAR-tree: An efficient self-adjusting index for moving objects. In D. M. Mount & C. Stein (Eds.), *Algorithm Engineering and Experiments, Fourth International Workshop, ALENEX 2002,* San Francisco, California, USA, January 4-5, 2002, Revised Papers, LNCS 2409 (pp. 178-193). Berlin, Germany: Springer.

Raptopoulou, K., Papadopoulos, A., & Manolopoulos, Y. (2003). Fast nearest-neighbor query processing in moving-objects databases. *GeoInformatica, 7*(2), 113-137.

Roussopoulos, N., Kelly, S., & Vincent, F. (1995). Nearest neighbor queries. In M. J. Carey & D. A. Schneider (Eds.), *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data,* San Jose, California, May 22-25, 1995 (pp. 71-79). New York: ACM Press.

Šaltenis, S., & Jensen, C. S. (2002). Indexing of moving objects for location-based services. In *Proceedings of the 18th International Conference on Data Engineering,* February 26-March 1, 2002, San Jose, CA (pp. 463-472). New York: IEEE Computer Society.

Šaltenis, S., Jensen, C. S., Leutenegger, S. T., & Lopez, M. A. (2000). Indexing the positions of continuously moving objects. In W. Chen, J. F. Naughton, & P. A. Bernstein (Eds.), *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data,* May 16-18, 2000, Dallas, Texas, USA (pp. 331-342). New York: ACM.

Samet, H. (1990). *The design and analysis of spatial data structures.* Reading, MA: Addison-Wesley.

Shivakumar, N., & Venkatasubramanian, S. (1996). Efficient indexing for broadcast based wireless systems. *Mobile Networks and Applications, 1,* 433-446.

Song, Z., & Roussopoulos, N. (2001a). Hashing moving objects. In K.-L. Tan, M. J. Franklin, & J. C. S. Lui (Eds.), *Mobile Data Management, Second International Conference, MDM 2001,* Hong Kong, China, January 8-10, 2001, Proceedings, LNCS 1987 (pp. 161-172). Berlin, Germany: Springer.

Song, Z., & Roussopoulos, N. (2001b). *k*-nearest neighbor for moving query point. In C. S. Jensen, M. Schneider, B. Seeger, & V. J. Tsotras (Eds.), *Advances in Spatial and Temporal Databases, Seventh International Symposium, SSTD 2001,* Redondo Beach, CA, USA, July 12-15, 2001, Proceedings, LNCS 2121 (pp. 79-96). Berlin, Germany: Springer.

Song, Z., & Roussopoulos, N. (2003). *k*-nearest neighbor for moving query point. In M.-S. Chen, P. K. Chrysanthis, M. Sloman, & A. B. Zaslavsky

(Eds.), *Mobile Data Management, Fourth International Conference, MDM 2003*, Melbourne, Australia, January 21-24, 2003, Proceedings, LNCS 2574 (pp. 340-344). Berlin, Germany: Springer.

Sun, J., Papadias, D., Tao, Y., & Liu, B. (2004). *Querying about the past, the present and the future in spatio-temporal databases.* Paper accepted for presentation at the 20th IEEE International Conference on Data Engineering (ICDE), Boston, MA, March 30-April 2, 2004. Retrieved October 10, 2003, from http://www.cs.ust.hk/~dimitris/publications.html

Tao, Y., & Papadias, D. (2002). Time-parameterized queries in spatio-temporal databases. In M. J. Franklin, B. Moon, & A. Ailamaki (Eds.), *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, Madison*, Wisconsin, June 3-6, 2002 (pp. 334-345). New York: ACM.

Tao, Y., Kollios, G., Considine, J., Li, F., & Papadias, D. (2004). *Spatio-temporal aggregation using sketches.* Paper accepted for presentation at the 20th IEEE International Conference on Data Engineering (ICDE), Boston, MA, March 30-April 2, 2004. Retrieved October 10, 2003, from http://www.cs.ust.hk/~dimitris/publications.html

Tao, Y., Mamoulis, N., & Papadias, D. (2003). Validity information retrieval for spatio-temporal queries: Theoretical performance bounds. In T. Hadzilacos, Y. Manolopoulos, & J. F. Roddick (Eds.), *Advances in Spatial and Temporal Databases, Eighth International Symposium, SSTD 2003*, Santorini Island, Greece, July 24-27, 2003, Proceedings, LNCS 2750 (pp. 159-178). Berlin, Germany: Springer.

Tao, Y., Papadias, D., & Shen, Q. (2002). Continuous nearest neighbor search. In *VLDB 2002, Proceedings of 28th International Conference on Very Large Data Bases*, August 20-23, 2002, Hong Kong, China (pp. 287-298). St. Louis, MO: Morgan Kaufmann.

Tao, Y., Papadias, D., & Sun, Q. (2003). The TPR*-tree: An optimized spatio-temporal access method for predictive queries. In J. C. Freytag, P. C. Lockemann, S. Abiteboul, M. J. Carey, P. G. Selinger, & A. Heuer (Eds.), *VLDB 2003, Proceedings of 29th International Conference on Very Large Data Bases*, September 9-12, 2003, Berlin, Germany (pp. 790-801). St. Louis, MO: Morgan Kaufmann.

Tayeb, J., Ulusoy, Ö., & Wolfson, O. (1998). A quadtree-based dynamic attribute indexing method. *The Computer Journal*, *41*(3), 185-200.

Trajcevski, G., Wolson, O., Zhang, F., & Chamberlain, S. (2002). The geometry of uncertainty in moving objects databases. In C. S. Jensen, K. G. Jeffery, J. Pokorny, S. Šaltenis, E. Bertino, K. Bohm, & M. Jarke (Eds.), *Advances in Database Technology EDBT 2002, Eighth International Conference on Extending Database Technology,* Prague, Czech Republic, March 25-27, Proceedings, LNCS 2287 (pp. 233-250). Berlin, Germany: Springer.

White, D. A., & Jain, R. (1996). Similarity indexing with the ss-tree. In S. Y. W. Su (Ed.), *Proceedings of the 12th International Conference on Data Engineering*, February 26-March 1, 1996, New Orleans, Louisiana (pp. 516-523). New York: IEEE Computer Society.

Wolfson, O. (2002). Moving objects information management: The database challenge. In A. Y. Halevy & A. Gal (Eds.), *Next Generation Information Technologies and Systems, Fifth International Workshop, NGITS 2002*, Caesarea, Israel, June 24-25, 2002, Proceedings, LNCS 2382 (pp. 75-89). Berlin, Germany: Springer.

Zhang, J., Zhu, M., Papadias, D., Tao, Y., & Lee, D. L. (2003). Location-based spatial queries. In A. Y. Halevy, Z. G. Ives, & A.-H. Doan (Eds.), *Proceedings of the 2003 ACM SIGMOD Interna-*

tional Conference on Management of Data, San Diego, California, USA, June 9-12, 2003 (pp. 443-454). New York: ACM.

Zheng, B., & Lee, D. L. (2001). Semantic caching in location-dependent query processing. In C. S. Jensen, M. Schneider, B. Seeger, & V. J. Tsotras (Eds.), *Advances in Spatial and Temporal Databases, Seventh International Symposium, SSTD 2001, Redondo Beach, CA, USA, July 12-15, 2001, Proceedings, LNCS 2121 (pp. 97-113). Berlin, Germany: Springer.*

This work was previously published in Wireless Information Highways, edited by D. Katsaros, A. Nanopoulos, and Y. Manalopoulos, pp. 315-338, copyright 2005 by IRM Press (an imprint of IGI Global).

Chapter 1.30
Database Queries in Mobile Environments

N. Marsit
IRIT—Paul Sabatier University, France

A. Hameurlain
IRIT—Paul Sabatier University, France

Z. Mammeri
IRIT—Paul Sabatier University, France

F. Morvan
IRIT—Paul Sabatier University, France

ABSTRACT

The technological evolution of networks together with the development of positioning systems has contributed to the emergence of numerous location-based services. Services related to this expanding area will become of major technical as well as economical interest in the coming few years. This aroused a great deal of interest from the scientific community at large and specifically from those studying these services and their diverse requirements and constraints. One of the direct consequences in the database field is the appearance of new types of queries (mobile queries issued from mobile terminals and/or requesting information associated with moving objects such as vehicles). Our objective in this chapter is to present a comprehensive survey of the field of research work related to mobile queries, with particular attention to the location issue.

INTRODUCTION

Mobile units are obviously on the rise. Thanks to the technological progress realized in this domain, mobile terminals and units have become successful and widely used by the general public. At the beginning of the wireless revolution, the main objective of wireless networks was to enable mo-

bile units to communicate. Nowadays, networks support various new services and applications. In fact, the significant technological evolutions added to the development of positioning systems such as GPSs have contributed to the emergence of a large number of location-based services and applications (e.g., a mobile user asking for data related to his location such as the closest hotel). These types of applications are about to become the major focus of economical interest in the next few years. The location-based service benefits are expected to exceed $40 billion in 2006, while they were estimated to approximately $1 billion in 2000 (Mokhtar & Su, 2004). This has aroused the interest of a great part of the scientific community devoted to research and development in this area. One of the direct consequences in the database field is the appearance of new types of queries. In mobile environments, there are two basic categories of queries (Marsit, Hameurlain, Mammeri, & Morvan, 2005). The first one includes queries issued from mobile terminals and querying data related to fixed objects (such as hotels, gas stations, hospitals)—for example, "select the closest restaurants." The second category includes the queries issued from mobile or fixed terminals and querying data related to moving objects (such as vehicles, helicopters, boats, people). For example, "select all ambulances that will be at 2 km from the hospital within 10 minutes." Within these two categories we can distinguish different types of queries according to their location dependence, to the association of spatial and temporal dimensions and to the evaluation mode (continuous or not).

The main objective of this chapter is to review work related to mobile queries (i.e., queries issued by mobile terminals and/or querying data related to moving objects). We start by classifying the different types of mobile queries. This step is central because it allows us to highlight the constraints of each type of query and to identify their underlying problems. In the field of mobility, various topics and problems were addressed by several different research communities. We only present the problems generated by the requirements of the new types of mobile queries. Finally, we point out the still open problems and identify the new challenges related to query processing in mobile environments.

QUERY CLASSIFICATION

Context

In mobile environments, entities can be either fixed or mobile. Hence, defining what mobility means is an essential requirement:

- **Mobile Client:** The query is submitted by a mobile terminal, here called mobile client.
- **Mobile Server:** The query or part of the query is processed at one or several mobile servers.
- **Moving Object:** Data targeted by the query can represent, in databases, moving objects (e.g., vehicles).

In this context we can highlight some query classification criteria. The first is the mobility constraint which allows distinguishing two basic forms of queries: (1) queries submitted by mobile terminals and querying data related to fixed objects (e.g., hotels), and (2) queries submitted by mobile or fixed terminals and querying data related to moving objects.

Notice that the mobility of servers does not add additional types of queries. Nevertheless, it may have an impact on query execution models since other problems have to be considered (e.g., network connection, server localization, etc.) (Holliday, Agrawal, & Abbadi, 2002).

For both query categories mentioned above, other types of queries could be distinguished. In fact, a second criterion, location constraint, brings out three types of queries: *Non-Location-Related Query*, *Location-Aware Query*, and *Location-Dependent Query*. The third criterion

depends on the association of the spatial and temporal dimensions. Finally, according to the query evaluation mode, we distinguish another type called continuous queries.

Location-Dependent Queries

Seydim, Dunham, and Kumar (2001a) present a framework for the location relatedness in queries. They consider that attributes, relations, operators, and simple predicates can be either location-related or not location-related:

- An attribute is considered as location-related if its domain is location-related. For instance, in the relation Hotel (id-Hotel, name, occupancy, city, street), the attributes city and street are location-related attributes.
- A location-related relation contains at least one location-related attribute. Otherwise, it is considered a non-location-related relation. (e.g., the relation Hotel is location-related because it contains the location-related attributes city and street).
- An operator is location-related if at least one of its operands is location-related.
- A simple location-related predicate is a simple predicate where the operator is location-related and the operands are defined in location-related domains.

Based on these definitions we can describe the types of queries mentioned above.

Non-Location-Related Queries (NLRQs)

If all the predicates and attributes used in a query are non-location-related, then it is called a non-location-related query (Seydim et al., 2001a)—for example, select availability of hotel with identifier 10. NLRQs are considered in many cases as traditional queries. However, the issuer of the query is mobile. So, peculiar problems specific to mobility like location management have to be addressed. These problems are discussed later.

Location-Aware Queries (LAQs)

If a query has at least one location-related simple predicate or one location-related attribute, then it is called Location-Aware Query—for example, select the names and the availabilities of the hotels in Toulouse. To express such a query, we need the attribute city, which is location-related. This is sufficient to consider the query as LAQ. Other examples may highlight the need for a new special location-related operator such as "*close to*" to express proximity—for example, select restaurants, close to the hotel, whose identifier is 10. Thus, we underline the need to define new methods to evaluate this type of operator. Moreover, several studies considered the problem of defining new location-related operators and efficient methods to process them (Seydim et al., 2001a).

Location-Dependent Queries (LDQs)

If the query results depend on the location of the query issuer (the mobile client), then the query is called a Location-Dependent Query—for example, select the closest hospital. Note that LDQ processing brings new challenges to the database community. First, new operators that take into account the notion of proximity (*close, closest to*) and orientation of the mobile client (*straight ahead*) have to be defined. Second, another step is required to bind the location of the mobile client to the query. The query issuer location could be recovered from different sources (wireless network operator's databases, positioning system such as GPS, etc.). Another problem occurs when the mobile client location is determined with a granularity that differs from the one of the data stored in the database. Indeed, a process of changing the granularity of the mobile client location given by the location service to the appropriate granularity required by the application has to be considered.

Moving Object Database Queries (MODQs)

This type includes queries issued by mobile or fixed terminals and querying moving object databases (i.e., databases in which data represent moving objects such as vehicles or planes). This type of query has appeared with the emergence of numerous applications requiring moving object data storage capability (Seydim et al., 2001a; Sistla, Wolfson, Chamberlain, & Dao, 1997). Such applications owe their popularity to the increasing development of positioning systems and technologies, allowing real-time tracking of moving objects—for example, select all taxis that are now in Wilson Square. For such a query, the size and the shape of the object are not important. Generally, the position of the moving object is the most requested information. Hence, several problems arise such as modeling and querying moving objects with rapidly changing locations, tracking and updating the location of moving objects, and managing the uncertainty on location due to imprecision of sensor technology and to the continuous movement of moving objects.

Spatio-Temporal Queries

The spatio-temporal type includes queries combining space dimension with time dimension (Erwig, Güting, Schneider, & Vazirgiannis, 1999). In the literature, this type of query is often associated with moving objects. Indeed, numerous applications do not only focus on the location of an object, but also on its position at a given time or on its trajectory during a certain time interval. The time notion generally implies the past, the present, and the future. Thus, we can distinguish between two types of spatio-temporal queries:

- The first type considers trajectories describing a time history of object movement—for example, select all moving objects with trajectories included within the area 'R' between 04:00 p.m. and 05:00 p.m..
- The second type focuses on the current position of the moving object and possibly its future position—for example, *select all ambulances that will be 2 km away from the hospital in less than 10 min*.

The spatio-temporal queries raise problems at several levels. Representing continuously changing data (e.g., position) has been one of the major points of interest during the past few years. Complex structures have to be managed to represent moving object trajectories. In addition, effective methods must be established to represent the moving object movement and to predict its future positions. The extension of traditional languages, such as SQL, enables support of querying spatial or temporal data. However, they have to be further extended to really represent the strong relationship between space and time. Finally, the notion of uncertainty, whether in the data model or in the proposed extensions of languages, has to be dealt with. We present in the fourth section different proposals developed in this research area.

Continuous Queries

A continuous query (CQ) allows users to get changing results from a database without having to issue the query repeatedly (Chen, Dewitt, Tian, & Wang, 2000). Assume that a driver is asking for a selection of hotels within 5 km from his position. If this query is submitted as non-continuous, then the results are sent back immediately after the query has been processed. If the same query is submitted as continuous, then the set of selected hotels varies "continuously" with the movement of the user. Notice that continuous queries require considerable modifications in the query evaluation algorithms. In fact, issues such as "when" and "how often" the continuous query should be re-evaluated have to be addressed. Also, the pos-

sibility of partial or incremental re-evaluation has to be investigated (Gök & Ulusoy, 2000).

Discussion

In this section, we presented the types of query which appear to be most crucial in mobile environments. However, we must insist here on two essential points: first, certain types of queries previously presented form non-separate groups. LDQ, MODQ, and spatio-temporal queries can be submitted as continuous queries. Nevertheless, with our classification we covered the types of mobile queries most often encountered in the literature, and we focused on their characteristics. In Table 1, we summarize six types of queries presented in this chapter. We observe that the first three types (NLRQ, LAQ, and LDQ) involve only the mobility of the client, whereas the other types involve moving objects and/or the mobility of clients. The fundamental difference between LDQ and LAQ is that for LDQ the notion of location is implicitly involved, while it is explicitly specified for LAQ. So, the LDQ type can be considered as part of the LAQ type. Spatio-temporal queries add a temporal dimension to the spatial dimension. Finally, note that CQ is a transverse type which can be associated to the other types of queries.

A second essential point concerns the common problems met for the various types of queries. Indeed, certain problems can be recurring in all query types (e.g., the management of mobile localization). Then, certain types of queries share very close problems such as data modeling of moving objects and representation of spatio-temporal data on moving objects. Moreover, considerable effort was devoted to model moving objects in the final objective to support spatio-temporal queries (Forlizzi, Güting, Nardelli, & Schneider, 2000; Grumbach, Rigaux, & Segoun, 2000; Sistla et al., 1997).

QUERYING FIXED-OBJECT DATABASES

The mobility domain is very large, and it would be over ambitions to present an exhaustive state of

Table 1. Summary of the mobile query types

Query type	Mobile client	Moving object	Spatial dimension	Temporal dimension
NLRQ *(Non Location-Related Queries)*	Yes	No	No	No
LAQ *(Location Aware Queries)*	Yes	No	Location (explicit)	No
LDQ *(Location-Dependent Queries)*	Yes	No	Location of client (implicit)	No
MODQ *(Moving Object Database Queries)*	Yes/No	Yes	Yes	No
Spatio-Temporal Queries	Yes/No	Yes	Yes	Yes
CQ *(Continuous Queries)*	Yes/No	Yes/No	Yes/No	Yes/No

the art on all work related to this field. We rather focus on the problems related to query processing in mobile environments. At the beginning of the previous section, we classified the mobile queries into two basic categories: those which are issued from mobile terminals and which are querying data on fixed objects, and those which are querying data on moving objects. In this section we describe problems of the first query class and give some methods proposed to cope with them. In the next section, we consider related work on the second category of queries.

Localization of Mobile Units

Since the beginning of the 1990s, localization of mobile units is one of the major concerns of researchers interested in mobility. In fact, while a mobile user submits a query, the system should be able to locate him in order to send back the answer. For these reasons, the localization of mobile units is an important issue related to query processing in mobile environments.

Location Models

The location model for mobile units is closely dependent on sensor systems used to detect the location. There are two basic models for representing the location: symbolic models and geometric models (Lee, Lee, Xu, & Zheng, 2002). The use of symbolic or geometric models depends on the application to be developed because they meet different requirements in terms of location representation and required precision.

In the symbolic model, the location is represented by entities from the real world such as streets, cities, and zip code. In this model, location can also be represented by elements defined in any particular systems, for instance, cells in a cellular system. The location information in a symbolic model is well structured and easy to manage. Their granularity is well suited for several location-based applications and services because their representation is often based on relationships between entities (e.g., a street is in a city, a city is in a zip code). However, this granularity is not very fine and it often depends on applications or on used systems.

In the geometric model, the location is represented by n-dimensional coordinates (generally 2 or 3). This model can give good accuracy and is compatible with heterogeneous systems. However, it can be costly in terms of data volume and sometimes the location information needs to be translated into a level understandable by the application.

Location Management

It is important to know the current location of mobile units. This information is usually stored at specific network sites. The main issue here is to find a compromise between the update cost and the lookup cost of location information. In fact, to reduce the cost of lookup, it is necessary to increase the number of sites where this information is stored. Hence, the availability of the location information is improved but the cost of updates becomes higher. On the other hand, if the frequency of the update is reduced, then the precision of the location information is compromised. In this area, various strategies that balance the cost of lookups against the cost of updates were proposed. The main relevant approaches are based on two types of location database architectures: two-tier schemes and hierarchic schemes. In Pitoura and Samaras (2001), a comprehensive survey of locating objects in mobile environments is proposed.

Location-Dependent Query Processing

The proposed techniques of localization were generally dedicated to wireless network operators. They allow mobile users to communicate or to get the answer of a query. With the development of

location-based services and applications, several new requirements arose in terms of data querying. In fact, the appearance of new types of queries like LDQ changed the problem of localization, since the location of a mobile client is involved even in query processing. This led to much research on LDQ processing.

Binding Location to LDQ

To better analyze the problem of LDQ query processing, let us take the example: select the closest hospital. The idea presented in Seydim et al. (2001a, 2001b) is to translate the query into select the closest hospital to my current position. A question arises here: how to bind the mobile client location to query, especially when this information must be given by network operator databases or by particular sensor network technologies such as GPSs? The proposed solution is based on the use of a particular service called *location service*. A mobile client identifier is sent to the location service which returns its actual location (Leonhardi & Kubach, 1999; Seydim et al., 2001a, 2001b). This service allows developing applications regardless of operator or sensor system. Location information is generally gathered from different sources: positioning system (e.g., GPS, GSM) and network operator databases. Thus, with the help of location service, our query becomes select hospitals closest to position X (where X is the value corresponding to the mobile client location sent back by the location service). This step is called *location binding* (Seydim et al., 2001a).

Matching Location Granularity Level

The location granularity sent by the location service and the location granularity required by the application may not be the same. Assume that the position X of the example studied above is sent according to the geometric model (e.g., latitude/longitude) and the hospital locations are represented in the database by zip code or by city name. In this case, a process of translating the query location granularity into the appropriate location granularity as needed by the application is necessary. This process is called *location leveling* (Seydim et al., 2001b). The problem of granularity mismatch may persist even when a unique location model is used. For example, cell and city (symbolic model) do not have the same level of granularity. The first solutions proposed to solve this problem were based on a location translation mechanism (ESRI, 2000; SignalSoft, 2000). These assumed that the set of location granularity needed by applications and those needed by queries are known in advance. So, translations are well defined and explicitly specified by mapping functions. Seydim et al. (2001b) proposed a general process leading to the coordination of dynamic location granularity levels. This process is based on metadata describing the hierarchy and relationship between different location granularities possibly used by applications and queries. Assume for example that the given position X is cell number 3. While consulting the hierarchy stored in metadata, one can find that the cell number 3 is in the zip code 31000. Thus, our query is finally transformed into select hospital within zip code 31000.

Processing of Location-Related Operators

Another aspect to which many researchers paid particular attention concerns processing of location-related operators. Indeed, the location constraint in queries implies necessarily the use of this type of operators. Although spatial operators (e.g., *Distance, Intersect, Contains, Within*) form an important part of location-related operators, they are not sufficient to deal with all the requirements of LDQ. So, new operators have to be introduced to express proximity (e.g., *closest to*) and mobile client orientation (e.g., *straight ahead*). Some of these operators can have different semantic interpretations according to the application requirements. In Seydim et al. (2001a),

the operators *closest to* and *straight ahead* are treated in spatial manner. Thus, for the *closest to* operator, the authors define an area around the query issuer to access the related data. This area may be a circle, a half circle, and so on. For the *straight ahead* operator, they define a window to select an area ahead of the direction of the user. These operators are not always interpreted in spatial manner. In fact, *closest to* may mean *object from the same city or the same zip code or the same cell or neighbor cell.*

To close this section, let us recall that in this chapter we were only interested in two aspects (i.e., localization of mobile terminals and query processing) of work related to processing of queries issued by mobile terminals. Obviously, other research directions were developed to solve problems of cache coherency, of transaction management (Serrano-Alvarado, Roncancio, & Adiba, 2001), and of data access in mobile environments (Birman et al., 1999).

QUERYING MOVING-OBJECT DATABASES

Much effort has been devoted to solving problems related to processing queries issued by mobile terminals and querying data related to fixed objects and more particularly LDQ. However, fast evolution of technologies providing increasingly precise information on location and movement of objects, like GPSs, encouraged the development of new types of applications. Indeed, the past few years have witnessed the emergence of a wide range of complex applications managing moving objects (e.g., fleet management, air traffic management, road traffic management, emergencies). These new applications have generated new constraints and new requirements. Hence, the database community is facing new challenges. In this section, we expose a part of the most representative topics of research related to querying moving-object databases. We concentrate on the problems of modeling, querying, language extensions, and uncertainty management. Obviously, other problems were dealt with in the literature such as indexing spatio-temporal data. Although these aspects are also important, we will not present them in this chapter.

Modeling of Moving Objects

Representing the continuous movement of objects in databases is one of the central problems of modeling moving objects. First, the computer systems are not able to handle infinite sets. Indeed, classical DBMSs consider data unchanged until they are explicitly updated. This is not sufficient to represent continuously changing information such as the position of a moving object. Second, positioning systems return locations of objects in a discrete manner. This problem is similar to the one of modeling spatio-temporal data. In fact, a spatio-temporal object is defined as an object whose shape or position varies as a function of time. The shape of moving objects is often ignored since they are generally assimilated to moving points. This assumption simplifies the problem compared to the more general one of representing spatio-temporal data. Approaches proposed to represent the spatio-temporal behavior of moving objects follow two main directions: the first centers on representing a history of the movement and trajectories of the moving objects; the second focuses on modeling the current location of a moving object and its possible future position. In the first direction, we distinguish two approaches: one which proposes to extend the existing systems with abstract data types, while the second is based on constraint databases. The second direction relies on a dynamic attributes modeling approach.

Abstract Data Types

The idea of extending the classical DBMS features by Abstract Data Type (ADT) appeared near the

end of the 1980s (Güting, 1989). Nearly 10 years later, Erwig et al., 1999) suggested exploiting the ADT concept to define spatio-temporal ADT. This work was pursued with the aim of introducing a system of ADT with suitable operations into the DBMS. This way, it is possible to better represent spatio-temporal data and to extend query languages (Forlizzi et al., 2000; Güting et al., 2000). Erwig et al. (1999) conclude that two levels of abstraction are necessary. The first level, called abstract model, is relatively simple (Güting et al., 2000). It allows handling infinite sets without considering details of their representation. However, the implementation of this model is not obvious. The second level (a discrete model) makes it possible to implement the types introduced in the abstract model by associating each data type of the abstract model with "discrete" types whose domains are defined according to a finite representation.

The abstract model relies on base types *(int, real, string, bool),* spatial types of 2D dimension *(point, points, line),* and temporal types *(intime).* One of the most important types of constructors is *moving.* It enables construction of types whose values change dynamically. For example, a moving(point) value is a function defined from time into point values. So, with this constructor one can represent the movement of objects. One can also represent a moving region by moving (region). Moreover, several operations on these types can be defined such as *at, mdistance, trajectory,* and so forth.

Further comprehensive explanations can be found in several papers (Erwig et al., 1999; Forlizzi et al., 2000; Güting et al., 2000).

Constraint Database Model

The constraint database model was initially proposed by Kanellakis, Kuper, and Revesz (1990, 1995). Even if the paradigm presented allows the representation of any type of data, this idea especially seduced the spatial database community. The principle is to consider the spatial objects as infinite sets of points satisfying first-order logic formulae (these formulae correspond to constraints). The constraint representation can be viewed as an extension of finite relational representations. In fact, a relation is considered as a first-order logic formula applied to infinite sets. Grumbach et al. (2000) relied on constraint database to represent and manipulate multidimensional data. They illustrated their model with spatio-temporal applications manipulating time and spatial objects. They considered space and time as natural components of 3D point sets. Thus, the spatio-temporal data are seen as mathematical objects which can be analyzed with already known tools (Mokhtar, Su, & Ibarra, 2002). So the location of moving objects can be modeled by a linear function mapping time to n-dimensional space. A trajectory is generally approximated by a succession of segments connected pairwise in a three-dimensionsional space. Each segment is represented as a conjunction of linear constraints using time variables and coordinate variables. The complete trajectory is represented by the disjunction of all its linear constraints.

Let us retain that the abstract representation of this model is completely independent of the implementation and query processing algorithms. In addition to this model, query processing has a reasonable complexity and query languages are simple. In fact, users can use standard languages because they do not have to care about internal mechanisms (data structures, operations) (Grumbach et al., 2000). However, this model is not suited for spatio-temporal queries which involve future locations of objects. Little work tried to extend this model in order to support this type of query.

Model Based on Dynamic Attributes

Sistla et al. (1997) proposed a model called MOST (Moving Objects Spatio-Temporal). This model was designed to be implemented in a software layer on the top of existing DBMSs. A proto-

type was performed within the framework of the DOMINO project (Wolfson, Sistla, Xu, Zhou, & Chamberlain, 1999b). The objective of this work is to extend classical DBMS features in order to represent in an effective way the current location of moving objects and to enable prediction of their future positions. The originality of this work lies in the introduction of the notion of dynamic attributes. The values of dynamic attributes change over time according to a given function without having to explicitly update them. Thus, within the MOST model, object attributes can be either static or dynamic. Let us take the object airplane as example. It includes static attributes such as the identifier and the company name, dynamic attributes such as the position represented by coordinate (X, Y, Z). Each dynamic attribute is represented by three sub-attributes. For example X is represented by X.value (value of X at the last update time), X.updatetime (the last update time), X.function (function of a variable t. at t=0, X.function=0). This function has to represent the motion vector of the object (e.g., speed). At each time t+updatetime, the value of X is given by computing X.value+X.function (t). Hence, with the MOST model, we can implicitly deduce the future positions of moving objects and answer queries which concern the future state of the database. In Sistla et al. (1997), the FTL (Future Temporal Logic) language is proposed in order to express this type of query.

Languages for Querying Moving Object Databases

The expression of moving objects database queries is closely tied to the chosen model. Therefore, several proposals of language extensions were introduced according to the underlying model of data. In the following, we will focus on language extensions proposed for each approach.

Querying in Abstract Data Type Model

Within the abstract model, semantics of operations and functions are defined in a relatively simple way. Indeed, they can handle infinite sets without considering their physical representation. Many operators are defined in Forlizzi et al. (2000) and Güting et al. (2000). These operators can be incorporated into languages intended for users (e.g., extended SQL). To illustrate this, let us take the following example:

airplane (id : string, company : string, pos : mpoint) is a relation. mpoint is moving(point). pos is a spatio-temporal attribute, which represents the location of the plane as a function of time. Assume that we have two operators, trajectory and length, defined as follows:

trajectory: mpoint $\rightarrow$ line, returns a line which is the projection of moving point in plan.
Length: line $\rightarrow$ real, determines the line length.

Now, we can express the query such as select Air France planes having a trajectory longer than 3000 miles:

```
select id
from airplane
Where company =''Air France'' and length(
trajectory(pos))>3000;
```

Obviously, in this framework further operators are defined in order to express more spatio-temporal queries.

Querying in Constraint Database Model

The particularity of this model lies in the fact that it hides from users the complexity of the data model. Indeed, this model allows handling infinite sets of points, forming tuples of relations, without caring about their finite representations

(managed by the model). Therefore, the user can simply use a standard language like SQL. Assume that the relation trajectory(X, Y, Z) represents the position of an object at any time; the user does not care about the finite representation of this abstract relation (infinite). Thus, to ask a query such as select position of the plane at time t1, we have only to use SQL language as follows:

```
Select X, Y, Z from trajectory where t = t1;
```

Naturally, the evaluation of these queries requires algorithms based on finite representation. Grumbach et al. (2000) propose a method exploiting the interpolated form of data (to obtain one of the variables describing an object as a function of the other variables). The query evaluation is then reduced to a small number of operators applied to a sub-space formed by key attributes (in the constraint model, all the attributes can be expressed as a function of key attributes of the infinite relations). Thus, the selection operation can be evaluated by a simple conjunction formula. Other language extensions and other evaluation methods were proposed in the framework of the constraint database model (Mokhtar et al., 2002).

Querying in MOST Model

Sistla et al. (1997) developed the MOST model, which allows the management of the current position of a mobile object and to anticipate its future movement. In this framework, an extension of standard languages such as SQL or OQL by temporal logic predicates was proposed. This language, called FTL, is based on two basic temporal operators: *Until* and *Nexttime*. *f until g* is satisfied if and only if the predicate *g* is true in this state or *g* is satisfied in the future and until then *f* is satisfied. *Nexttime f* is satisfied *if* the formulae *f* is satisfied at the next state of the history. From these two basic operators, other temporal operators can be expressed in FTL: *Eventually f = true Until f* ; *Always f* = ¬*Eventually* ¬*f* ; *Eventually_within_c g* (i.e., *g* will be satisfied within *c* time units).

Given that the language and the model are designed to be implemented on top of DBMS, the FTL language is assumed to support non-temporal operators of the underlying languages. Thus, by means of this language, we can express the spatio-temporal queries in a simple and intuitive way.

Example: Select all the planes which enter a region P around the airport within 30 min.

Assume the relation airplane (id, company, X_pos). Assume that the *X_pos* attribute gives the geographical coordinates of the plane and that a polygon P is defined as a spatial object. Assume also that a spatial operator Inside (X_pos, P) returns true if the first argument X_pos is inside a given area P. Else, it returns false. We can now express the query using the temporal operator already defined in FTL:

```
Eventually_within_30 inside(X_pos, P);
```

In Sistla et al. (1997), an algorithm to evaluate queries in the MOST model is presented. These queries are given in the form of conjunctive formulae. The proposed method of evaluation may support continuous spatio-temporal queries.

Update Policy and Uncertainty Management

The uncertainty aspect is very important in the field of moving object databases. Indeed the current location of moving objects is known only with limited accuracy. This is due to the continuous nature and the unpredictable behavior of the object movement, to the imprecision of po-

sitioning systems and to the network delays. We can consider the uncertainty notion at different levels. First, concerning the update of moving objects' current locations, an important issue is: "How often and when is updating the current location of a moving object needed in order to have an acceptable uncertainty without affecting the system performance?" Second, it is necessary to take into account the uncertainty notion when querying the moving object location or trajectory. In the following, we present work that deals with these two aspects.

Moving Object Location Update

The first proposed approaches were based on periodic updates of moving object locations (e.g., every 2 km or every 2 minutes). These traditional approaches were used in commercial transportation (Wolfson, 2003). Their advantage is that they enable knowledge about the error bound in the query answers since the uncertainty is known in advance and it corresponds to the distance between two updates. However, if more accuracy on location is required, many more updates are needed. Thus, other update policies were proposed (Wolfson et al., 1999b). In this work the authors extend the MOST model by proposing an update policy based on deviation and uncertainty. The basic idea is to update the position of a moving object only when the distance between its current location and the database location exceeds a certain threshold. This distance is called deviation. It is computed by the moving object itself (e.g., equipped by GPS) or by sensors network. The choice of the suited uncertainty threshold is a very important issue. Indeed, the purpose is to find the compromise between the uncertainty cost and the update cost. So, three methods for determining the adequate uncertainty threshold were studied. The first one considers a fixed uncertainty threshold during the entire moving object trip. The second one changes the uncertainty threshold for each update in order to minimize the update cost, the deviation cost, and the uncertainty cost. The third one deals with the disconnection detection. More details on these approaches are given in Wolfson et al. (1999a). Another recent approach was proposed (Wolfson, 2003). It is also based on deviation but it considers unknown itinerary contrary to hypotheses of Wolfson et al. However, the moving object is supposed to move on the same street between two updates.

Querying with Uncertainty

It is very important to deal with uncertainty in the data model, in the query language and execution model. In Wolfson et al. (1999a), the answers to queries involving current or future location of moving objects imply an uncertainty area (e.g., a circular area around the computed position having the uncertainty threshold as radius). In the case of DBMS representing the trajectory of a moving object, the problem of uncertainty management is more complex. Pfoser and Jensen (1999) have focused on the uncertainty due to sampling and to the imprecision of the GPS system. They propose a method that enables the quantification of uncertainty in moving object data modeling. Their approach enables limiting the uncertainty on the past movement of the moving object, but the error becomes higher when we are close to the present moment. Another relevant work considering the uncertainty on the trajectories of moving objects is Trajcevski, Wolfson, Zhang, and Chamberlain (2002). The authors propose to model the moving object trajectory as a cylindrical volume in 3D space.

OPEN PROBLEMS

In this section, we identify some open problems that seem to us to be the most relevant, and new challenges related to query processing in mobile environments.

New Scenarios

We have already classified the different query types related to mobility. Next, a very important question has to be asked about this subject: Do the query types studied meet all the requirements of the location-based applications? To our knowledge, the scenarios proposed so far do not take into account the 'criticality' aspect required by certain location-based applications. Assume, for instance, that a captain is looking for "the rescue helicopters able to reach his boat within 10 minutes." Suppose also that this captain requires an answer in two minutes, another decision being taken otherwise. This query can be considered as a spatio-temporal query with real-time constraints. Indeed, we would like in that case for the query processing and result transmission to respect a particular deadline (two minutes). We can also imagine spatial constraints. A driver, for instance, could require obtaining an answer to an LDQ before reaching the next crossing or not farther than three miles from his current location. Thus, temporal deadlines or spatial constraints could both be specified in an explicit manner for different query types. In other cases, we could draw implicit constraints after the semantic analysis of a query. For example, let us consider spatio-temporal queries involving the future position of a moving object: select the ambulances which will be within 10 minutes of the hospital. In this case, we can consider the 10-minute delay as a deadline. Indeed, we can wonder what is the utility that will have the answer after that deadline?

We believe that new scenarios could open opportunities for new research that would propose methods to still open issues of processing mobile queries including real-time constraints. One of the interesting ways to do this would be to find a compromise between respecting the constraints and having a sufficient accuracy of the information returned by the system. A user would sometimes prefer to have a result, even partial, that respects the deadline, rather than a complete result given after the specified deadline. In the example of the captain, he can accept a partial result (not all helicopters and not necessarily the closest) in order to have the answer in the two-minute delay. To our knowledge, such problems have not yet been addressed.

Unification of Approaches

There are various possibilities and ways of use, and the application domains are very wide. In fact, some applications provide location-dependent services to customers (e.g., local yellow pages). Others are tracking oriented (fleet management, transportation, emergency) or navigation-assistance oriented (e.g., inform a subscriber how to go from point A to point B). These applications have different requirements and different constraints. Consequently, they have been developed in several directions in order to fit differing location-based application requirements. Currently the trend is to develop tools and techniques dedicated to each application field. Regarding querying in mobile environments, the research was often led with the purpose of meeting the requirements and handling the constraints of particular types of queries. This has unfortunately led to a spread of energy and duplicate work. We believe that it is important, today, to generalize the approaches and to develop them in a unified framework in order to fit the common requirements of different types of applications. Hence, it could be easier to implement the approaches and even to standardize them. This effort has to be made at several levels (location management, data modeling, querying). We think that a relevant research direction would be to thoroughly study the constraints associated to each type of query, with the aim of highlighting the common ones (e.g., the location, the movement, the temporal dimension). Such work would enable a common basis to all query types to be highlighted.

Standard Query Languages

For several types of queries studied in this chapter, we noted that querying involves specifying new operators. For spatio-temporal queries, in each proposed data model, new operators have been defined. Also for LDQ, we have seen that new operators taking into account proximity and orientation of the mobile client have been introduced. We think that it becomes necessary to enhance the "expressiveness" of the queries by standardizing these various extensions. Indeed, we must study formally the different proposals to integrate them in standard language intended for users (like SQL or OQL).

Mobile Queries Scheduling

In mobile environments, the determination of the queries scheduling, at the server side, requires some special considerations. Indeed, the answer to a query, submitted by a mobile client, could be sent back to him/her at a location that he/she would have already left. This problem requires, at best, re-routing the results to the new location of the client. This could be sufficient for NLRQ or LAQ, but not at all for LDQ. In fact, by changing location, the mobile client could receive invalid results, because answers to an LDQ are dependant on the client location. Hence, he/she would either reject the results or ask the server to process the query again. This could affect the system performance. Thus, we believe that an interesting direction to investigate would be to find a query processing order, taking into account the types of queries, their constraints, or the degree of mobility of the client submitting them (high or low mobility). We think that future work could contribute to addressing this issue. Indeed, we can envisage the development of optimal scheduling algorithms of mobile queries, taking into account new constraints of mobility and different requirements of every type of query. This issue was also discussed in Lee et al. (2002).

Unpredictable Object Movement Modeling

We have outlined the fact that spatio-temporal queries may involve the anticipation of the future movement of objects. To process these queries, many authors assume that the movement of objects has certain regularity. In fact, the motion vector is determined by the positioning system. The position or the trajectories of mobile objects are often estimated by interpolations or extrapolations. Generally, these approaches are addressing applications that are moving object tracking oriented and navigation-assistance oriented. However, other applications that are information service oriented may require location prediction capability. Indeed, the integration of GPS modules in mobile phones and PDAs is increasingly spreading. Thus, new location-dependent commercial services may involve future positions of users having unpredictable movement. Indeed, the movement of a pedestrian using a mobile phone does not possess the same properties of regularity as other moving objects such as vehicles or planes.

CONCLUSION

In this chapter, we first proposed a classification of mobile queries into two categories: queries issued from mobile terminals and querying data related to fixed objects, and queries issued from mobile or fixed terminals and querying moving object databases.

These queries generated multiple problems, which created new fields of work. We focused on the research areas that were developed with the aim of resolving the problems related to query processing in mobile environments. We have chosen to present separately the proposed approaches within each of the two basic categories of query. For the first category we described the problems of location management. We also dedicated an important part to describing work related to lo-

cation-dependent query processing. Concerning the second category of queries, we presented the most relevant work related to moving object databases modeling, especially their positions and their trajectories. We briefly described the work done with the aim of extending query languages in order to handle the spatial dimension and the temporal dimension. Finally, we were interested with the proposals made to face the problems of uncertainty management. Although progress is being made, there are still open issues which are challenging. In the last section, we presented some new challenges facing researchers before the use of complex mobile queries becomes an industrial and commercial reality.

REFERENCES

Birman, K. P., Hayden, M., Özkasap, O., Xiao, Z., Budiu, M., & Minsky, Y. (1999). Bimodal multicast. *ACM Transactions on Computer Systems, 17*(2), 41-88.

Chen, J., Dewitt, D. J., Tian, F., & Wang, Y. (2000, May). NiagaraCQ: A scalable continuous query system for Internet databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, TX (pp. 379-390).

Erwig, M., Güting, R. H., Schneider, M., & Vazirgiannis, M. (1999). Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica, 3*(3), 269-296.

ESRI. (2000). *About GIS: How GIS works.* Retrieved August 8, 2000, from http://www.esri.com/library/gis/abtgis/giswrk.html

Forlizzi, F., Güting, R.H., Nardelli, E., & Schneider, M. (2000, May). A data model and data structures for moving objects databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, TX (pp. 319-330).

Gök, H. G., & Ulusoy, O. (2000). Transmission of continuous query results in mobile computing systems. *Information Science, 125*(1-4), 37-63.

Grumbach, S., Rigaux, P., & Segoun, L. (2000, September). Manipulating interpolated data is easier than you thought. *Proceedings of the International Conference on Very Large Data Bases*, Cairo, Egypt (pp. 156-165).

Güting, R. H. (1989, August). Gral: An extensible relational database system for geometric applications. *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, Amsterdam, The Netherlands (pp. 33-44).

Güting, R. H., Bohlen, M. H., Erwig, M., Jensen, C. S., Lorentzos N. A., Schneider M., et al. (2000). A foundation for representing and querying moving objects. *ACM Transactions on Database Systems, 25*(1), 1-42.

Holliday, J., Agrawal, D., & Abbadi, A. E. (2002). Disconnection modes for mobile databases. *Wireless Networks, 8*(4), 391-402.

Kanellakis, P. C., Kuper, G. M., & Revesz, P. Z. (1990, April). Constraint query languages. *Proceedings of the Symposium on Principles of Database Systems*, Nashville, TN (pp. 299-313).

Kanellakis, P. C., Kuper, G. M., & Revesz, P. Z. (1995). Constraint query languages. *Journal of Computer and System Science, 51*(1), 26-52.

Leonhardi, A., & Kubach, U. (1999, October). An architecture for a universal distributed location service. *Proceedings of the European Wireless '99 Conference*, Munich, Germany (pp. 351-355).

Lee, D. L., Lee, W. C., Xu, J., & Zheng, B. (2002). Data management in location-dependent information services. *IEEE Pervasive Computing, 1*(3), 65-72.

Marsit, N., Hameurlain, A., Mammeri, Z., & Morvan, F. (2005, February). Query processing in mobile environments: A survey and open prob-

lems. *Proceedings of the International Conference on Distributed Frameworks for Multimedia Applications,* Besançon, France (pp. 150-157).

Mokhtar, H., & Su, J. (2004, January). Universal trajectory queries for moving object databases. *Proceedings of the International Conference on Mobile Data Management,* Berkeley, CA (pp. 133-145).

Mokhtar, H., Su, J., & Ibarra, O. H. (2002, June). On moving object queries. *Proceedings of the Symposium on Principles of Database Systems,* Madison, WI (pp. 188-198).

Pfoser, D., & Jensen, C. S. (1999, July). Capturing the uncertainty of moving-object representations. *Proceedings of the International Symposium Advances in Spatial Databases,* Hong Kong, China (pp. 111-132).

Pitoura, E., & Samaras, G. (2001). Locating objects in mobile computing. *IEEE Transactions on Knowledge and Data Engineering, 13*(4), 571-592.

Serrano-Alvarado, P., Roncancio, C., & Adiba, M. E. (2001, September). Analyzing mobile transaction supports for DBMS. *Proceedings of the International Workshop on Database and Expert Systems Applications (DEXA),* Munich, Germany (pp. 595-600).

Seydim, A. Y., Dunham, M. H., & Kumar, V. (2001a, May). Location-dependent query processing. *Proceedings of the International Workshop on Data Engineering for Wireless and Mobile Access,* Santa Barbara, CA (pp. 47-53).

Seydim, A. Y., Dunham, M. H., & Kumar, V. (2001b, September). An architecture for location-dependent query processing. *Proceedings of the International Workshop on Database and Expert Systems Applications (DEXA),* Munich, Germany (pp. 549-555).

SignalSoft. (2000). *SignalSoft Corporation—Wireless location services.* Retrieved December 2, 2000, from http://www.signalsoft corp.com

Sistla, A. P., Wolfson, O., Chamberlain, S., & Dao, S. (1997, April 7-11). Modeling and querying moving objects. *Proceedings of the International Conference on Data Engineering,* Birmingham, UK (pp. 422-432).

Trajcevski, G., Wolfson, O., Zhang, F., & Chamberlain, S. (2002, March). The geometry of uncertainty in moving objects databases. *Proceedings of Advances in Database Technology—EDBT, International Conference on Extending Database Technology,* Prague, Czech Republic (pp. 233-250).

Wolfson, O. (2003, July). Accuracy and resource consumption in tracking and location prediction. *Proceedings of the International Symposium on Advances in Spatial and Temporal Databases,* Santorini Island, Greece (pp. 325-343).

Wolfson, O., Jiang, L., Sistla, A. P., Chamberlain, S., Rishe, N., & Minglin, D. (1999). Databases for tracking mobile units in real time. *Proceedings of Database Theory—ICDT, International Conference,* Jerusalem, Israel (pp. 169-186).

Wolfson, O., Sistla, A.P., Xu, B., Zhou, J., & Chamberlain, S. (1999, June). DOMINO: Databases fOr MovINg Objects tracking. *Proceedings of the ACM SIGMOD International Conference on Management of Data,* Philadelphia (pp. 547-554).

This work was previously published in the Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 267-284, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.31
A Taxonomy of Database Operations on Mobile Devices

Say Ying Lim
Monash University, Australia

David Taniar
Monash University, Australia

Bala Srinivasan
Monash University, Australia

ABSTRACT

In this chapter, we present an extensive study of database operations on mobile devices which provides an understanding and direction for processing data locally on mobile devices. Generally, it is not efficient to download everything from the remote databases and display on a small screen. Also in a mobile environment, where users move when issuing queries to the servers, location has become a crucial aspect. Our taxonomy of database operations on mobile devices mainly consists of on-mobile join operations and on-mobile location dependent operations. For the on-mobile join operation, we include pre- and post-processing whereas for on-mobile location dependent operations, we focus on set operations arise from location-dependent queries.

INTRODUCTION

In these days, mobile technology has been increasingly in demand and is widely used to allow people to be connected wirelessly without having to worry about the distance barrier (Myers, 2003; Kapp, 2002). Mobile technologies can be seen as new resources for accomplishing various everyday activities that are carried out on the move. The direction of the mobile technology industry is beginning to emerge as more mobile users have been evolved. The emergence of this new technology provides the ability for users to access information anytime, anywhere (Lee, Zhu, & Hu, 2005; Seydim, Dunham, & Kumar, 2001). Quick and easy access of information at anytime anywhere is now becoming more and more popular.

People have tremendous capabilities for utilizing mobile devices in various innovative ways for various purposes. Mobile devices are capable to process and retrieve data from multiple remote databases (Lo, Mamoulis, Cheung, Ho, & Kalnis, 2003; Malladi & Davis, 2002). This allows mobile users who wish to collect data from different remote databases by sending queries to the servers and then be able to process the multiple information gathered from these sources locally on the mobile devices (Mamoulis, Kalnis, Bakiras, Li, 2003; Ozakar, Morvan, & Hameurlain, 2005). By processing the data locally, mobile users would have more control on to what they actually want as the final results of the query. They can therefore choose to query information from different servers and join them to be process locally according to their requirements. Also, by being able to obtain specific information over several different sites would help bring optimum results to mobile users queries. This is because different sites may give different insights on a particular thing and with this different insights being join together the return would be more complete. Also processing that is done locally would helps reduce communication cost which is cost of sending the query to and from the servers (Lee & Chen, 2002; Lo et al, 2003).

Example 1: A Japanese tourist while traveling to Malaysia wants to know the available vegetarian restaurants in Malaysia. He looks for restaurants recommended by both the Malaysian Tourist Office and Malaysian Vegetarian Community. First, using his wireless PDA, he would download information broadcast from the Malaysian Tourist Office. Then, he would download the information provided by the second organization mentioned previously. Once he obtains the two lists from the two information providers, he may perform an operation on his mobile device that joins the contents from the two relations that may not be collaborative to each other. This illustrates the importance of assembling information obtained from various non-collaborative sources in a mobile device.

This chapter proposes a framework of the various kinds of join queries for mobile devices for the benefits of the mobile users that may want to retrieve information from several different non-collaborative sites. Our query taxonomy concentrates on various database operations, including not only join, but as well as location-dependent information processing, which are performed on mobile devices.

The main difference between this chapter and other mobile query processing papers is that the query processing proposed here is carried out locally on mobile devices, and not in the server. Our approach is whereby the mobile users gather information from multiple servers and process them locally on a mobile device. This study is important, not only due to the need for local processing, but also due to reducing communication costs as well as giving the mobile users more control on what information they want to assemble. Frequent disconnections and low bandwidth also play a major motivation to our work which focuses on local processing.

The rest of this chapter is organized as follows. In the next section, we will briefly explain the background knowledge of mobile database technology, related work, as well as the issues and constraints imposed by mobile devices. We will then present a taxonomy of various database operations on mobile devices, including join operation in the client-side and describes how location-dependent affects information gathering processing scheme on mobile devices. Last but not least, we will discuss the future trend which includes the potential applications for database processing on mobile devices.

PRELIMINARIES

As the preliminary of our work, we will briefly discuss the general background of mobile database

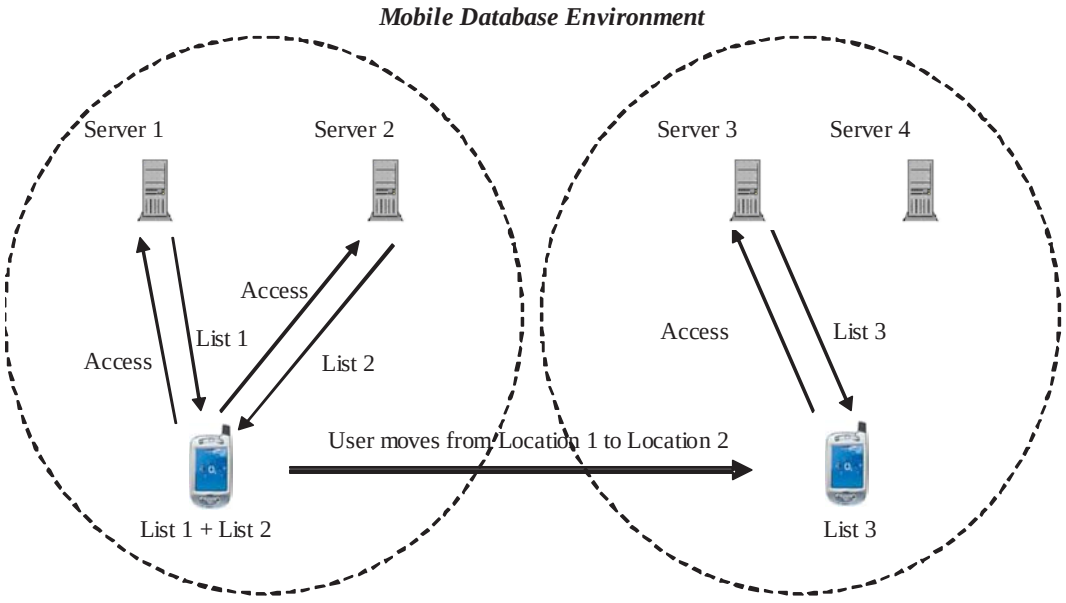
environment which includes some basic knowledge behind a mobile environment. Next, we will discuss related work of mobile query processing done by other researchers. Lastly, we will also cover the issues and complexity of local mobile database operations.

Mobile Database Environment: A Background

Mobile devices are defined as electronic equipments which operate without cables for the purposes of communication, data processing, and exchange, which can be carried by its user and which can receive, send, or transmit information anywhere, anytime due to its mobility and portability (Myers, 2003). In particular, mobile devices include mobile phones, personal digital assistants (PDA), laptops that can be connected to network and mixes of these such as PDA-mobile phones that add mobile phone to the functionality of a PDA. This chapter is concerned with devices categorized as PDA-mobile phones or PDAs.

Generally, mobile users with their mobile devices and servers that store data are involved in a typical mobile environment (Lee, Zhu, & Hu, 2005; Madria, Bhargava, Pitoura, & Kumar, 2000; Wolfson, 2002). Each of these mobile users communicates with a single server or multiple servers that may or may not be collaborative with one another. However, communication between mobile users and servers are required in order to carry out any transaction and information retrieval. Basically, the servers are more or less static and do not move, whereas the mobile users can move from one place to another and are therefore dynamic. Nevertheless, mobile users have to be within specific region to be able to received signal in order to connect to the servers (Goh & Taniar, 2005; Jayaputera & Taniar, 2005). Figure 1 illustrates a scenario of a mobile database environment.

Figure 1. A mobile database environment



It can be seen from Figure 1 that mobile user 1 when within a specific location is able to access servers 1 and 2. By downloading from both servers, the data will be stored in the mobile device which can be manipulated later locally. And if mobile user 1 moves to a different location, the server to access maybe the same but the list downloaded would be different since this mobile client is located in a different location now. The user might also be able to access to a different server that is not available in his pervious location before he moves.

Due to the dynamic nature of this mobile environment, mobile devices face several limitations (Paulson, 2003; Trivedi, Dharmaraja, & Ma, 2002). These include limited processing capacity as well as storage capacity. Moreover, limited bandwidth is an issue because this wireless bandwidth is smaller compared with the fixed network. This leads to poor connection and frequent disconnection. Another major issue would be the small display which causes limitations in the visualizations. Therefore, it is important to comprehensively study how database operations may be carried out locally on mobile devices.

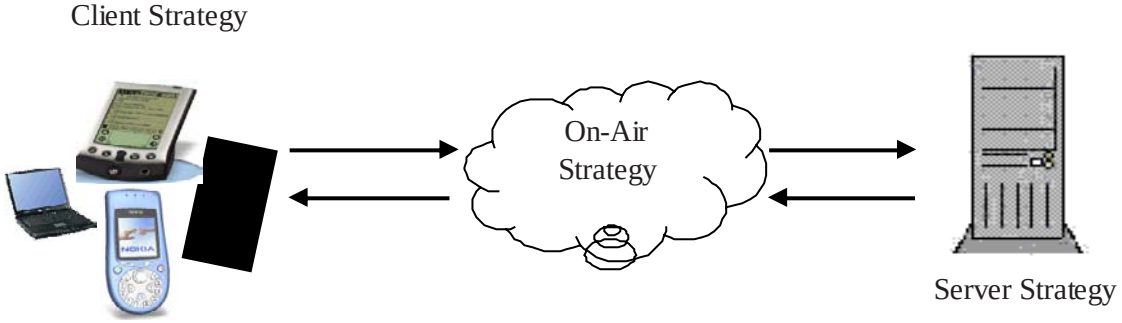
Mobile Query Processing: Related Work

As a result of the desire to process queries between servers that might not be collaborative, traditional join query techniques might not be applicable (Lo et al, 2003). Recent related work done by others in the field of mobile database queries includes processing query via *server strategy*, *on-air strategy* and *client strategy* (Waluyo, Srinivasan, & Taniar, 2005b). Figure 2 gives an illustration of the three strategies of query processing on a mobile environment.

In general, the *server strategy* is referring to mobile users sending a query to the server for processing and then the results are returned to the user (Seydim, Dunham, & Kumar, 2001; Waluyo, Srinivasan, & Taniar, 2005b). Issues, such as location-dependent, take into account since different location will be accessing different servers, and subsequently it relates to the processing by the server and the return of the results based on the new location of the mobile user (Jayaputera & Taniar, 2005). Our approach differs from this strategy in the sense that we focus on how to process the already downloaded data on a mobile device and manipulate the data locally to return satisfactorily results taken into account the limitations of mobile devices.

As for the *on-air strategy* which is also known as the *broadcasting strategy* is basically the server broadcasts data to the air and mobile users tune into a channel to download the necessary data (Tran, Hua, & Jiang, 2001; Triantafillou, Harpantidou, & Paterakis, 2001). This broadcasting technique broadcasts a set of database items to

Figure 2. Mobile query processing strategies



the air to a large number of mobile users over a single channel or multiple channels (Huang & Chen, 2003; Prabhakara, Hua, & Jiang, 2000; Waluyo, Srinivasan, & Taniar, 2005a, 2005c). This strategy greatly deals with problem of channel distortion and fault transmission. With the set of data on the air, mobile users can tune into one or more channel to get the data. This, subsequently, improves query performance. This also differs from our approach in the sense that our focus is not how the mobile users download the data in terms of whether it is downloaded from data on the air or whether downloaded from data in the server, but rather how we process the downloaded data locally on mobile devices.

The *client strategy* is whereby the mobile user downloads multiple lists of data from the server and processes them locally on their mobile device (Lo et al, 2003; Ozakar, Morvan, & Hameurlein, 2005). This strategy deals with processing locally on the mobile devices itself, such as when data are downloaded from remote databases and need to be process to return a join result. Downloading both non-collaborative relations entirely may not be a good method due to the limitations of mobile devices which have limited memory space to hold large volume of data and small display which limits the visualization (Lo et al, 2003). Thus efficient space management of output contents has to be taken into account. In addition, this strategy also relates to maintaining cached data in the local storage, since efficient cache management is critical in mobile query processing (Cao, 2003; Elmagarmid, Jing, Helal, & Lee, 2003; Xu, Hu, Lee, & Lee, 2004; Zheng, Xu, & Lee, 2002). This approach is similar to our work in terms of processing data that are downloaded from remote databases locally and readily for further processing.

The related work intends to concentrate on using different strategies, such as via server or on air to download data and how to perform join queries locally on mobile devices taking into account the mobile devices limitations. However our approach focus on using a combination of various possible join queries that is to be carried out locally to attend to the major issues such as the limited memory and limited screen space of mobile devices. We also incorporate the location-dependent aspects in the local processing.

Issues and Complexity of Local Mobile Database Operations

Our database wireless environment consists of PDAs (personal digital assistant), wireless network connections, and changing user environment (e.g., car, street, building site). This arises some issues and complexity of the mobile operations. And also secondly, the limited screen space is another constraint. If the results of the join are too long, then it is cumbersome to be shown on the small mobile device screen. The visualization is thus limited by the small screen of the mobile devices. Figure 3 shows an illustration of how join results are displayed on a PDA.

Figure 3. Join display on a PDA

Processors may also be overloading with time consuming joins especially those that involve thousands of records from many different servers, and completion time will be expected to be longer.

Another issue to be taken account is by having a complex join that involves large amount of data, the consequences would lead to increase communication cost. One must keep in mind that using mobile devices, our aim is to minimize the communication cost with is the cost to ship query and results from database site to requested site.

The previous limitations such as small displays, low bandwidth, low processor power, and operating memory are dramatically limiting the quality of obtaining more resourceful information. The problem of keeping mobile users on the satisfactory level becomes a big challenge. Due to the previously mentioned hardware limitations and changing user environment, the limitations must be drastically overcome and adapted to the mobile environment capabilities. As a result, it is extremely important to study comprehensive database operations that are performed on mobile devices taking into account all the issues and complexities. By minimizing and overcoming these limitations it can further help to boost the number of mobile users in the near future.

TAXONOMY OF DATABASE OPERATIONS ON MOBILE DEVICES

This chapter proposes a taxonomy of database operations on mobile devices. These operations give flexibility to mobile users in retrieving information from remote databases and processing them locally on their mobile devices. This is important because users may want to have more control over the lists of data that are downloaded from multiple servers. They may be interested in only a selection of specific information that can only be derived by processing the data that are obtained from different servers, and this processing should be done locally when all the data have been downloaded from the respective servers. As a result, one of the reasons for presenting the taxonomy of database operations on mobile results is because there is a need to process data locally based on user requests. And since it is quite a complex task that requires more processing from the mobile device itself, it is important to study and further investigate. It also indicates some implications of the various choices one may make when making a query.

We classify database operations on mobile devices into two main groups: (*1*) *on-mobile join processing*, and (*2*) *on-mobile location-dependent information processing.*

On-Mobile Join Processing

It is basically a process of combining data from one relation with another relation. In a mobile environment, joins are used to bring together information from two or more different information that is stored in non-collaborative servers or remote databases. It joins multiple data from different servers into a single output to be displayed on the mobile device. In on-mobile join, due to a small visualization screen, mobile users who are joining information from various servers normally require some pre- and post-processing.

Consider Example 1 presented earlier. It shows how a join operation is needed to be performed on a mobile device as the mobile user downloads information from two different sources which are not collaborative between each other and wants to assemble information through a join operation on his mobile device. This example illustrates a simple on-mobile join case.

On-Mobile Location-Dependent Information Processing

The emerging growth of the use of intelligent mobile devices (e.g., mobile phones and PDAs) opens up a whole new world of possibilities which

includes delivering information to the mobile devices that are customized and tailored according to their current location. The intention is to take into account location dependent factors which allow mobile users to query information without facing location problem. Data that are downloaded from different location would be different and there is a need to bring together these data according to user request who may want to synchronize the data that are downloaded from different location to be consolidated into a single output.

Example 2: A property investor while driving his car downloads a list of nearby apartments for sale from a real-estate agent. As he moves, he downloads the requested information again from the same real-estate agent. Because his position has changed since he first enquires, the two lists of apartments for sale would be different due to the relative location when this investor was inquiring the information. Based on these two lists, the investor would probably like to perform an operation on his mobile device to show only those apartments exist in the latest list, and not in the first list. This kind of list operation is often known as a "difference" or "minus" or "exclude" operation, and this is incurred due to information which is location-dependent and is very much relevant in a mobile environment.

Each of the previous classifications will be further explained into more detail in the succeeding sections.

ON-MOBILE JOIN OPERATIONS

Joins are used in queries to explain how different tables are related (Mamoulis, Kalnis, & Bakiras, 2003; Ozakar, Morvan, & Hameurlain, 2005). In a mobile environment, joins are useful especially when you want to bring together information from two or more different information that is stored in non-collaborative servers. Basically, it is an operation that provides access to data from two tables at the same time from different remote databases. This relational computing feature consolidates multiple data from different servers for use in a single output on the mobile devices.

Based on the limitations of mobile devices which are the limited mount of memory and small screen space, it is important to take into account the output results to ensure that it is not too large. And furthermore, sometimes user may want to join items together from different databases but they do not want to see everything. They may only want to see certain related information that satisfies their criteria. Due to this user's demand, a join alone is not sufficient because it does not limit the conditions based on user's requirements. The idea of this is basically to ensure mobile users has the ability to reduce the query results with maximum return of satisfaction because with the pre and post-processing, the output results will greatly reduce base on the user's requirements without having to sacrifice any possible wanted information. There will also be more potential of data manipulation that a mobile user can perform.

Therefore we will need to combine a pre-processing which is executed before mobile join and/or a post-processing which is executed after the mobile join. Figure 4 shows an illustration of the combination of pre and post-processing with the mobile join.

Figure 4. On-mobile join taxonomy

Join Operations

Generally, there are various kinds of joins available (Elmasri & Navathe, 2003). However, when using joins in a mobile environment, we would like to particularly focus on two types of joins which is equi-join and anti-join. Whenever there are two relations from different servers that wanted to be joined together into a single relation, this is known as equi or simple join. What it actually does is basically combining data from relation one with data from relation two.

Referring to Example 1 presented earlier, which shows an equi-join, which joins the relations from the first server (i.e., Malaysian Tourist Office) with the second server (i.e., Malaysian Vegetarian Community) to have a more complete output based on user requirements. The contents of the two relations which are hosted by the two different servers that is needed to be joined can be seen on Figure 5.

An *anti-join* is a form of join with reverse logic (Elmasri & Navathe, 2003). Instead of returning rows when there is a match (according to the join predicate) between the left and right side, an anti-join returns those rows from the left side of the predicate for which there is no match on the right. However one of the limitations of using anti-join is that the columns involved in the anti-join must both have not null constraints (Kifer, Bernstein, & Lewis, 2006).

Example 3: A tourist who visits Australia uses his mobile device to issue a query on current local events held in Australia. There is a server holds all types of events happened all year in 2005. The tourist may want to know if a particular event is a remake in the past years and is only interested in non-remake events. So if the list obtained from Current Local Events list matches with events in Past Events list, then he will not be interested and hence it is not needed to display as output on his mobile device.

Figure 5. An equi-join between two relations

Name	Address	Category	Rating
Restaurant A	Address 1	Chinese	Excellent
Restaurant B	Address 2	Vietnamese	Satisfactory
Restaurant C	Address 3	Thai	Excellent
Restaurant D	Address 4	Thai	Satisfactory
--------	--------	-------	---------

Server 1 : Malaysian Tourist Office

Name	Address
Restaurant A	Address 1
Restaurant F	Address 6
Restaurant X	Address 24
Restaurant G	Address 7
---------	----------

Server 2 : Malaysian Vegetarian Community

Example 3 shows an example of the opposite of an equi-join. The tourist only wants to collect information that is not matched with the previous list. In other words, when you get the match, then you do not want it.

Nevertheless, if join is done alone, it may raise issues and complexity especially when applying to a mobile device that has a limited memory capacity and a limited screen space. Therefore, in a mobile device environment, it is likely that we impose pre and post-processing to make on-mobile join more efficient and cost effective.

Pre-Processing Operations

Pre-processing is an operation that is being carried out before the actual join between two or more relations in the non-collaborative servers are carried out (in this context, we then also call it a *pre-join* operation). The importance of the existence of pre-processing in a mobile environment is because mobile users might not be interested in all the data from the server that he wants to download from. The mobile users may only be interested in a selection of specific data from one of the server and another selection of data from another server. Therefore, pre-processing is needed to get the specific selection from each of the servers before being downloaded into the mobile device to be further processed. This also leads to reducing communication cost since less data is needed to download from each server and also helps to discard unwanted data from being downloaded into the mobile devices.

Filtering is a well-known operation of pre-processing. It is similar to the selection operation in relational algebra (Elmasri & Navathe, 2003). Filter is best applied before join because it will helps reduce size of the relations before join between relations occurs. Basically it is being used when the user only needs selective rows of items so that only those requested are being process to be joined. This is extremely handy for use in a mobile environment because this helps to limits the number of rows being process which in return helps to reduce the communication cost since the data being process has been reduced. Filtering can be done in several different ways. Figure 6 shows illustration of pre-processing whereby two lists of data from two different servers that are filtered by the respective server before they are downloaded into the mobile device.

Example 4: A student is in the city centre and wants to know which of the bookshops in the city centre sell networking books. So using his mobile device, he looks for the books recommended by two of the nearest bookshops based on his current location which are called bookshop1 and bookshop2. The student's query would first scans through all the books and filters out only those that he is interested which in this case is networking books, and then joins together the relation from both bookshop1 and bookshop2.

Filtering one particular type of item can be expressed as in terms of a table of books titles. In this case, the user may be only interested in networking book, so filter comes in to ensure only networking books are being processed.

Filtering a selection group of items can be expressed as in terms having a large list of data and you want to select out only those that are base on the list which contains a specific amount of data, such as top 10 list and so on.

Example 5: A customer is interested in buying a notebook during his visit to a computer fair. However, he is only interested in the top 10 best selling based in Japan and he wants to know the specifications of the notebook from the top 10 list. And because he is in a computer fair in Singapore, so he uses his mobile device to make a query to get the ten notebooks from the top 10 Japan list and then joins with the respective vendors to get the details of the specifications. This type of filter gets the top ten records, instead of a specific one like in the previous example.

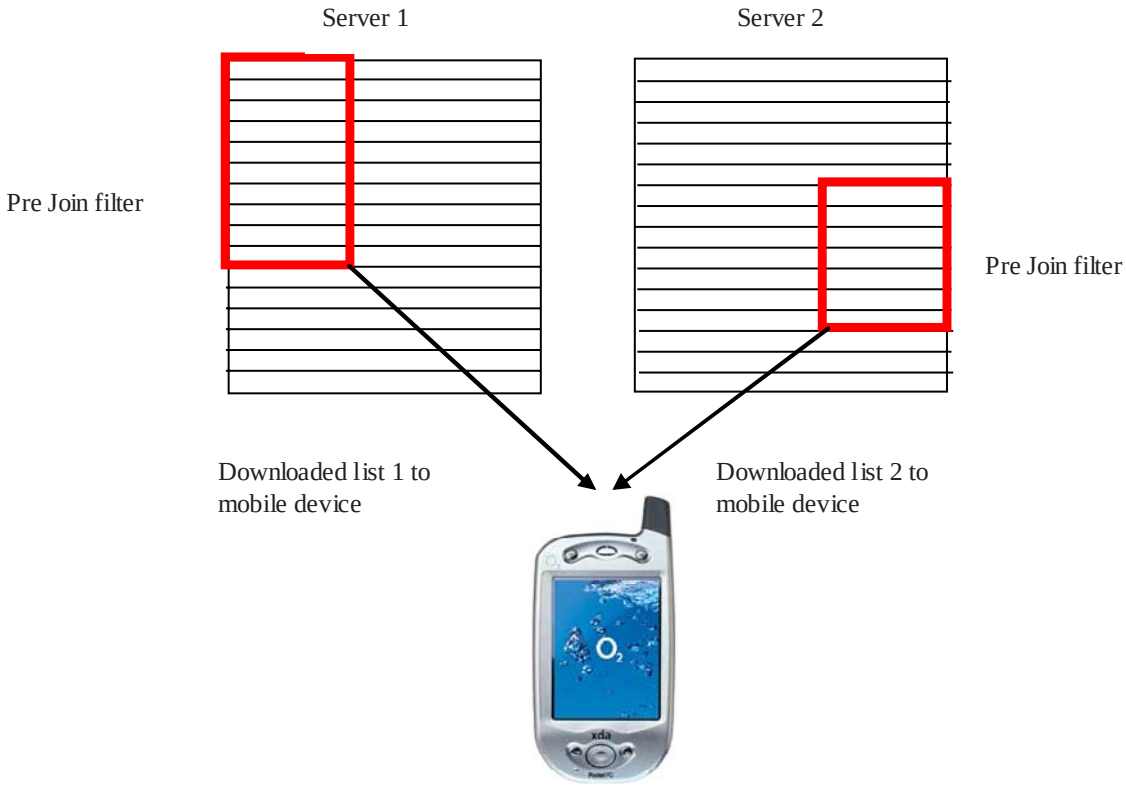
From Examples 4 and 5, we use pre-processing because the first list of data has to be filtered first before joining to get the matching with the second list of data.

Post-Processing Operations

Post-processing is an operation that is being carried out after the actual join (in this context, we then also call it a *post-join* operation). It is when the successive rows output from one step which is the pre-processing and then join with the other relation are then fed into the next step that is a post-join. The importance of the existence of post-processing in a mobile environment is because after mobile joins are carried out which combines lists from several remote databases, the results maybe too large and may contain some data that are neither needed nor interested by the users. So with post-processing comes into operation, the results of the output can further be reduced and manipulated in a way that it shows the results in which the user is interested. Therefore, post-processing operation is important because it is the final step that is being taken to produce the users the outputs that meets their requirements.

In general, there is a range of different post-processing operations that is available. However, in this chapter, we would like to focus only on aggregation, sorting, and projection that are to be used in a mobile environment.

Figure 6. Filtering



Aggregation

Aggregation is a process of grouping distinct data (Taniar, Jiang, Liu, & Leung, 2002). The aggregated data set has a smaller number of data elements than the input data set which therefore helps reduce the output results to meet the limitation of the mobile device of smaller memory capacity. This also appears to be one of the ways for speeding up query performance due to facts are summed up for selected dimensions from the original fact table. The resulting aggregate table will have fewer rows, thus making queries that can use them go faster. Positioning, count, and calculations are commonly used to implement the aggregation concepts.

Positioning aggregation gives the return of a particular position or ranking after joins are completed (Tan, Taniar, & Lu, 2004). Fundamentally, after joining required information from several remote databases, the user may want to know a particular location of a point base on the new joined list of data. Positioning can be relevant and useful in a mobile environment especially when a mobile user who has two lists of data on hand and wants to know the position of a particular item in the list base on the previous list of data.

Example 6: A music fan who attends the Annual Grammy Award event is interested in knowing what the ranking of the songs that won the best romantic song in the top 100 best songs list. So using his mobile device, he first gets that particular song he is interested in and then joins with the top 100 best songs list to get the position of that romantic song that won the best award.

From Example 6, it shows an example of post-processing, because getting the position of the song that has won a Grammy Award from the top 100 best songs list can only be obtained after the join between the two lists is performed.

Count aggregation is an aggregate function which returns the number of rows of a query or some part of a query (Elmasri & Navathe, 2003). Count can be used to return a single count of the rows a query selects, or the rows for each group in a query. This is relevant for a mobile environment especially when a mobile user, for instance, is interested in knowing the number of petrol kiosks in his nearby location.

Example 7: Referring to Example 6 on the Grammy Award Event, in this example the mobile user wants to know the number of awards previously won which is obtained from the idol biography server who is a current winner in the Grammy Award. So using his mobile device, he first gets the name of his idol he is interested in and then joins with the idol biography server site to get the number of awards previously won and return the number of count of all awards he/she has won.

From Example 7, the post-processing shows that the return of the specific numeric value which is the count of the previously won awards, is also only obtainable after the join between the two lists to the final value.

Calculation aggregation is a process of mathematical or logical methods and problem solving that involves numbers (Elmasri & Navathe, 2003). This is relevant for a mobile environment especially when a mobile user who is on the road wants to calculate distance or an exact amount of the two geographical coordinates between two different lists of data.

Example 8: A tourist who was stranded in the city and wants to get home but do not know which public transport and where to take them. He wants to know which is the nearest available transportation and how far it is from its current standing position. He only wants the nearest available with its timetable. So using his mobile device, he gets a list of all surrounding transportation available but narrows down based on the shortest distance calculated by kilometers and then joins both relations together so that both the timetable information and the map getting there for that transportation are available. As a result of looking for the shortest distance, calculations are needed in order to get the numeric value.

From Example 8, post-processing is carried after joining two different lists from different sources and if the user wants to make calculation on specific thing such as the distance, it can only be calculated when the query joins together with the type of transportation selected with the other list which shows the tourist current coordinate location.

Sorting

Sorting is another type of post-processing operation, which sorts the query results (Taniar & Rahayu, 2002). It can help user to minimize looking at the unwanted output. Therefore, mobile users might use sorting techniques after performing the mobile join to sort the data possibly based on the importance of user desire. This means that the more important or most close related to user desire conditions would be listed at the top in a descending order. This makes it more convenient for the mobile user to choose what they would like to see first since the more important items have been placed on top. Another possible reason for using this technique is because the mobile device screen is small and the screen itself it might not cover everything on a single page. So by sorting the data then the user can save time looking further at other pages since the user can probably have found what he wants at the top of the list.

Example 9: By referring to previous Example 1 on vegetarian restaurants, the mobile user is only interested in high rating vegetarian restaurants. So in this case, sorting comes into consideration because there is no point to list vegetarian restaurants that is low ratings since the tourist is not interested at all.

From Example 9, sorting is classified as post-processing because it is done when you have got the final list that has been joined. Sorting basically reorders the list in terms of user preference.

Projection

Projection is defined as the list of attributes, which a user wants to display as a result of the execution of the query (Elmasri & Navathe, 2003). One of the main reasons that projection is important in a mobile environment is because of the limitation of mobile device which has small screen that may not be able to display all the results of the data at once. Hence, with projection, those more irrelevant data without ignoring user requirements will be further discarded and so less number of items would be produced and displayed on the limited screen space of a mobile device.

Example 10: By referring to previous Example 5 regarding enquiring the top 10 notebooks, the user may only want to know which of the top 10 notebooks in Japan that has DVD-RW. Generally, the top 10 list only contains names of the notebook and may not show the specification. Hence in order to see the specification, it can only be obtained by making another query to a second list which contains detail of the specification.

From Example 10, projection is a sub class of post-processing in the sense that the user only wants specific information after the join which get every details of the other specifications.

Figure 7. Ratio between PDA screen and join results

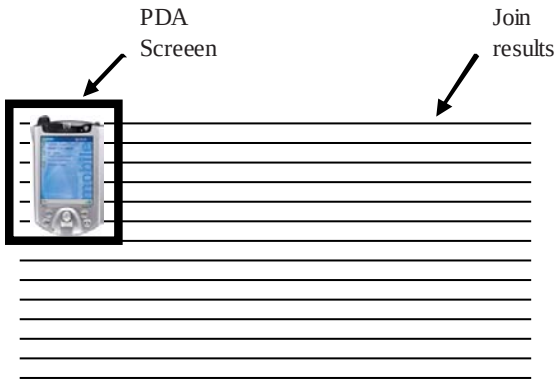


Figure 7 shows an illustration of how aggregation, projection, and sorting are important in a mobile device after performing a typical join which has returned a large amount of data. As can be seen, the screen of a mobile device is too small and may affect the viewing results of a typical join situation which has produced too many join results.

ON-MOBILE LOCATION-DEPENDENT OPERATIONS

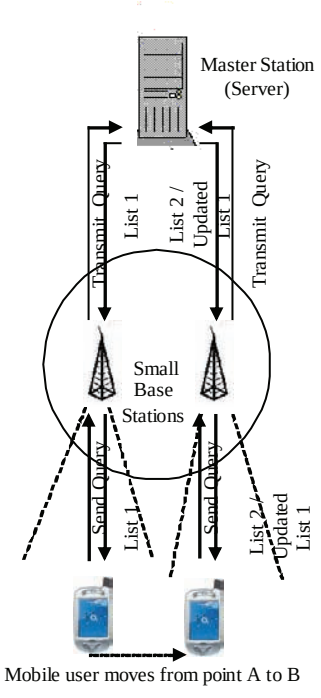
Location-dependent processing is of interest in a number of applications, especially those that involves geographical information systems (Cai & Hua, 2002; Cheverst, Davies, Mitchell, 2000; Jung, You, Lee, & Kim, 2002; Tsalgatidou, Veijalainen, Markkula, Katasonov, & Hadjiefthymiades, 2003). An example query might be "to find the nearest petrol kiosk" or "find the three nearest vegetarian restaurants" queries that are issued from mobile users. As the mobile users move around, the query results may change and would therefore depend on the location of the issuer. This means that if a user sends a query and then changes his/her location, the answer of that query has to be based on the location of the user issuing the query (Seydim, Dunham, & Kumar, 2001; Waluyo, Srinivasan, & Taniar, 2005a).

Figure 8 shows a general illustration of how general mobile location dependent processing is carried out in a typical mobile environment (Jayaputera & Taniar, 2005). The query is first transmitted from a mobile user to the small base station which will send it to the master station to get the required downloaded list and sent back. Then as the user moves from point *A* to point *B* the query will be transmitted to a different small base station that is within the current location of the user. Then again, this query is send to the master station to get relevant data to be downloaded or update if the data already exist in the mobile device and sent back.

In order to provide powerful functions in a mobile environment, we have to let mobile users to query information without facing the location problem. This involves data acquirement and manipulation from multiple lists over remote databases (Liberatore, 2002). We will explain the type of operations that can be carried out to synchronize different lists that a mobile user downloads due to his moving position to a new location. Hence, the list the mobile user downloaded is actually location dependent which depends on where is his current location and will change if he/she moves. Since this operation is performed locally on a mobile device, we call it "on-mobile location-dependent operations."

On-mobile location dependent operations have been becoming a growing trend due to the constant behavior of mobile users who move around. In this section, we look at examples of location

Figure 8. A typical location-dependent query



dependent processing utilizing traditional set operations commonly used in relational algebra and other set operations. It involves the circumstances when mobile users are in the situation where they download a list when in a certain location and then they move around and download another list in their new current location. Or another circumstance might be mobile user might already have a list in his mobile device but moves and require to download the same list again but from different location. In any case, there is a need to synchronize these lists that has been downloaded from different location.

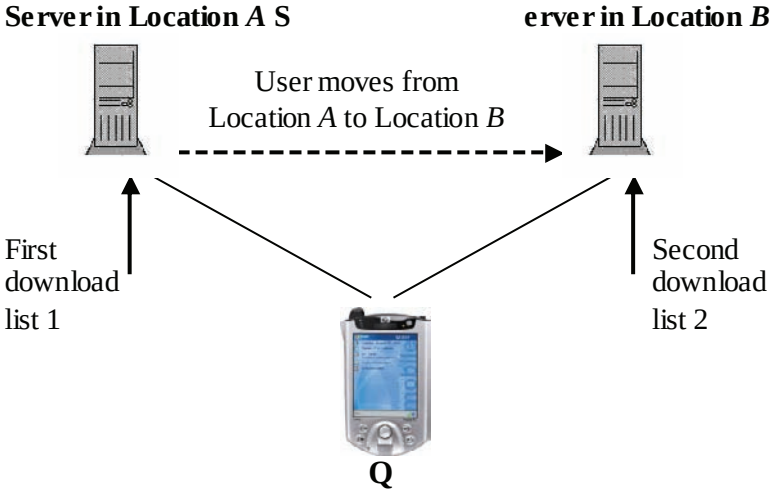
Figure 9 shows an example of how location dependent play a role when a mobile user who is on the highway going from location *A* to location *B* and wants to find the nearest available petrol kiosk. First, the mobile user establishes contact with server located at location *A* and downloads the first list which contains petrol kiosk around location *A*. As he moves and comes nearby to new location *B* he downloads another new list and this time the list is different from the previously downloaded because the location has been changed and therefore only contains petrol kiosk around location *B*. These two lists represent possible solutions for the mobile user. Through a local list processing, it can determine by comparing both the lists, which is indeed its nearest gas station based on current location.

Traditional Relational Algebra Set Operations

In a mobile environment, mobile users would possibly face a situation when he/she is required to download a list of data from one location and then download again another list of data from the same source but from different location. So, the relevance of using set operations to on-mobile location dependent processing is that both involve more than one relation. Due to the possible situation that mobile users face concerning downloading different list of data from similar source but different location, the needs of processing the two lists of data into a single list is highly desirable, particularly in this mobile environment.

Therefore, relational algebra set operations can be used for list processing on mobile devices which involves processing the data that are obtained from the same source but different locations. Different types of traditional relational algebra set opera-

Figure 9. On-mobile location-dependent operations



tions that can be used include union, intersection and difference (Elmasri & Navathe, 2003).

Union Set Operation

Union operation combines the results of two or more independent queries into a single output (Kifer, Bernstein, & Lewis, 2006). By default, no duplicate records are returned when you use a union operation. Given that the union operation discards duplicate record, this type of set operation is therefore handy when processing user query that requires only distinct results that are obtained by combining two similar kinds of lists. For instance, when a mobile user needs to download data from the same source but different location, and wishes to get only distinct results. This operation can help bring together all possible output downloaded from same source but different location into a single output list of result.

However, the limitation is that the mobile user that access queries in a union operation must ensure the relations are union compatible. For achieving union compatible in mobile environment, a user must ensure the lists are downloaded from the same source. This means that the user may download from one source and then moves to a new location and download again but from the same source. Then only the user can perform a union operation on the mobile device. However the contents may be different between the two lists of data downloaded from different location although the same source. This is because in a location dependent processing when the user moves to a new location, the data downloaded is different from the data downloaded in the previous location.

Nevertheless, if both lists are too large then using union operation by itself may not be substantial. This brings in post-processing operation. Post-processing are processing that are further executed after a typical on-mobile join operation is being carried out.

Example 11: A tourist currently visiting Melbourne wants to know places of interest and downloads a list of interesting places in Melbourne from tourist attraction site and stores in his mobile device. Then he visits Sydney and again downloads another list of interesting places from tourist attraction site but this time it shows places in Sydney. He wants to perform a join that shows only the places regardless of the states but in terms of the types of places such as whether it is a historical building, zoo, religions centre and so on.

Example 11 demonstrates a union operation whereby the query combine all data from the first relation which contains places in Melbourne together with places in Sydney that are downloaded from similar source but the list are different because they are in different location. And since they are similar source, the number of fields is basically the same and so union operator is relevant. In this example, the results of the union operation are further post-processed to do the grouping based on type of places.

Intersection Set Operation

Given collections R1 and R2, the set of elements that is contained in both R1 and R2 are basically called *intersection*. It only returns results that appear in both R1 and R2. The intersection set operation is handy in a mobile environment when the user would like to know only information that has common attribute that exist in both relations that he/she has downloaded when moving from one place to another. An intersection of two lists basically gives the information that appears in both lists (Elmasri & Navathe, 2003). However, a post-processing operation might be highly desirable if the current output result is too large. With the post-processing, it can further reduce the final results by manipulating the multiple list of data in a way that shows only results in which the user is interested.

Example 12: A group of student in Location *A* wants to know where is the nearest McDonalds and using the mobile device they downloaded a list of McDonalds locations which shows all available McDonalds in surrounding location. As they travel further until they arrive in Location *B*, they download another McDonalds lists again and realize the list is somewhat different since they have move from *A* to *B*. Therefore based on these two lists, the student wants to display only those McDonalds that provide drive through service regardless of whether it is in *A* or *B*.

Example 12 demonstrates an intersection operation because what the students are interested is based on both the downloaded lists as well as they want to know which McDonalds has the common field of providing drive through service. The drive-through service can also be thought as part of the post-processing.

Difference Set Operation

Difference set operation is also sometimes known as minus or excludes operation (Elmasri & Navathe, 2003). Given collections R1 and R2, the set of elements that is contained in R1 and not in R2 or vice versa is called difference. Therefore, the output results return only results that appear in R1 that does not appear in R2. The difference set operation may come into benefit especially when the mobile user would like to find certain information that is unique and only appears in one relation and not both from the downloaded list of data, and in the context of location-dependent the information requested must come from one location only.

Example 13: A student wants to know what movie is currently showing in a shopping complex that houses a number of cinemas. He downloads a list when he is at the complex. Then he goes to another shopping complex and wants to know the movies currently showing there. So now the new list is downloaded which contain movies in his new location. The student then wants to know which movies are only showing in this current location and not shown in the previous location.

Example 13 demonstrates a difference in operation because having two different lists downloaded from the two shopping complex, the student only wants the query to return movies that show in either one of the cinemas only and not both.

Other Set Operations

Besides the traditional relational algebra set operations, there are different types of set operations that maybe applicable for location dependent processing on mobile devices. An example of this is a list comparison operation that maybe useful in local mobile device processing between two list of data that is downloaded from the same source. Mobile users are often on the move — moving from one place to another. However, they may typically send query to similar source in different locations. With the implementation of comparison operation in the mobile device, a mobile user can now obtain a view side by side and weight against each other between the two lists of data that is downloaded from similar source but different location. This is useful when mobile user want to compare between the two different lists together.

Example 14: In the city market, a user has downloaded a list of current vegetables prices and keeps then in her mobile device. Then she went to a countryside market and downloaded another list of vegetables prices. With these two lists, she wants to make a comparison and show which vegetables type is cheaper in which market.

From Example 14, it is known that the first list which contains the city price list has been downloaded and kept in the mobile device locally. And then the user further downloads a new list when she is in the country which contains a different list of prices. With these two different lists on hand that contain common items, the mobile user wants her mobile device to locally process

these two lists by making a comparison result and then show which of the two list has cheaper price for the respectively vegetables items.

FUTURE TRENDS

Database operations on mobile devices are indeed a potential area for further investigation, because accessing and downloading multiple data anywhere and anytime from multiple remote databases and process them locally through mobile devices is becoming an important emerging element for mobile users who want to have more control over the final output. Also, location dependent processing has becoming more important in playing a role on operations on mobile devices (Goh, & Taniar, 2005; Kubach & Rothernel, 2001; Lee, Xu, Zheng, & Lee, 2002; Ren & Dunham, 2000). The future remains positive but there are some issues need to be addressed. Hence, this section discusses some future trend of database operations on mobile devices in terms of various perspectives, including query processing perspective, user application perspective, technological perspective, as well as security and privacy perspective. Each of the perspectives gives different view of the future work in the area of mobile database processing and applications.

Query Processing Perspective

From the query processing perspective, the most important element is to help reduce the communication cost, which occurs due to data transfer between to and from the servers and mobile devices (Xu, Zheng, Lee, & Lee, 2003). These also includes are location dependent processing, future processing that takes into consideration various screen types and storage capacity.

The need for collecting information from multiple remote databases and processing locally becomes apparent especially when mobile users collect information from several non-collaborative remote databases. Therefore, it is of great magnitude to investigating the optimization of database processing on mobile devices, because it helps addresses issue of communication cost. It would also be of a great interest to be able to work on optimizing processing of the database operations to make the processing more efficient and cost effective.

For location dependent processing, whenever mobile users move from one location to another location, the downloaded data would be different even though the query is direct to similar source. And because of this, whenever the downloaded data differ as the users move to a new location, the database server must be intelligent enough to inform that existing list contains different information and prompt if user wants to download a new list.

There are various types of mobile devices available in the market today. Some of them may have bigger screen and some of them may have smaller screen. Therefore, in the future the processing must be able to be personalized or to be adopted to any screen types or sizes. The same goes for storage space. Some mobile phones may have just built in limited memory, whereas PDAs may allow expansion of storage capacity through the use of storage card. So, future intelligent query processing must be able to adapt to any storage requirement such as when downloading list of data to limited build in memory, the data size is reduced to a different format that can adapt to the storage requirement. As we notice, one of the major limitations of mobile devices is the limited storage capacity. Thus, filtering possible irrelevant data from mobile users before being downloaded would most likely help the storage limitation in terms of having irrelevant data automatically filtered out before being downloaded into a mobile device. This also helps in increasing the speed of returning downloaded list of data to the mobile devices.

User Application Perspective

User application perspective looks at the type of future applications that may be developed taking into account the current limitations of mobile devices and its environment processing capabilities. This includes developing future applications taking into account location dependent technology, communication bandwidth, and different capabilities of mobile devices.

There are numerous opportunities for future development of applications especially those that incorporate the need for extensive location dependent processing (Goh & Taniar, 2005). In this case, we would like to explain an example of a particular application that uses location dependent technology. Essentially, there is a need for constant monitoring movement of people because it may be useful in locating missing persons. Therefore, operators are required to provide police with information allowing them to locate an individual's mobile device in order to retrieve the persons that were reported as missing. This can be made possible by inserting tracking software according to user agreement (Wolfson, 2002).

Although, communication bandwidth is still relatively small at the moment, but as more and more demand towards the use of mobile devices, there has been a trend in 3G communication to provide a wider bandwidth (Kapp, 2002; Lee, Leong, & Si, 2002; Myers & Beigl, 2003). This makes it available for mobile users to be able to do more things with their mobile devices such as downloading video and so on. Therefore, future applications can make use of a faster bandwidth and query processing can be easier.

Despite the fact that processing capabilities of mobile devices varies such as small mobile phone which does not have processing capabilities to PDAs which has bigger memory and processors, and so, future applications must be able to distinguish these and program applications that has the option of whether it is to be loaded into mobile phones or PDAs.

Technological Perspective

Technological perspective looks at how technology plays a role for future development of better and more powerful mobile devices. This may includes producing mobile devices that are capable to handle massive amount of data and devices that are able to have combined voice and data capabilities (Myers & Beigl, 2003).

Another case from a technological point of view is that when operationally active, mobile users will often handle large amount of data in real time which may cause overload processing. Hence, this requires hardware that is capable of processing these data with minimum usage of processing power. The processing power required increases as the number of servers and data downloaded by the user increases. Therefore, strategies would be to further develop hardware that capable to process faster.

There are some users who prefer to listen than reading from a mobile device especially the user is driving from point *A* to *B* and is querying directions. This is practical since the screen display of a mobile device is so small and it may require constant scrolling up down and left right to get see the map from one point to another point on the mobile device. It would be proficient if there is a convergence towards voice and data combination whereby the mobile device are voice enabled in the sense that as the user drives the mobile device read out the direction to the user.

Security and Privacy Perspective

Security and privacy perspective arises due to more and more mobile users from all over the world accessing data from remote servers wirelessly through an open mobile environment. As a result, mobile users are often vulnerable to issues such as possible interference from others in this open network. This exists largely due to the need for protecting human rights by allowing them to remain anonymous, and allowing the user to be

able to do things freely with minimal interference from others. Therefore, security and privacy issue remain important factors (Lee et al, 2002).

Hence, it is important to have the option for enabling the user to remain anonymous and unknown of their choice and behavior unless required by legal system. This also includes higher security levels whenever accessing the open network wirelessly. This issue could potentially be addressed by means of privacy preserving methods, such as user personal information are carefully being protected and when the user are connected to the network, identify the user with a nickname rather than the real name.

CONCLUSION

In this chapter, we have presented a comprehensive taxonomy of database operations on mobile devices. The decision of choosing the right usage of operations to minimize results without neglecting user requirements is essential especially when processing queries locally on mobile devices from multiple list of remote database by taking into account considerations of the issues and complexity of mobile operations. And, this chapter also covers issues on location-dependent queries processing in mobile database environment. As the wireless and mobile communication of mobile users has increased, location has become a very important constraint. Lists of data from different locations would be different and there is a need to bring together these data according to user requirements who may want need these two separate lists of data to be synchronized into a single output.

REFERENCES

Cai, Y., & Hua, K. A. (2002). An adaptive query management technique for real-time monitoring of spatial regions in mobile database systems. *Proceedings of the 21st IEEE International Conference on Performance, Computing, and Communications* (pp. 259-266).

Cao, G. (2003). A scalable low-latency cache invalidation strategy for mobile environments. *IEEE Transactions on Knowledge and Data Engineering, 15*(5), 1251-1265.

Cheverst, K., Davies, N., Mitchell, K., & Friday, A. (2000). Experiences of developing and deploying a context-aware tourist guide. *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking* (pp. 20-31).

Elmargamid, A., Jing, J., Helal, A., & Lee, C. (2003). Scalable cache invalidation algorithms for mobile data access. *IEEE Transactions on Knowledge and Data Engineering, 15*(6), 1498-1511.

Elmasri, R., & Navathe, S. B. (2003). *Fundamentals of database systems* (4th ed.). Reading, MA: Addison Wesley.

Goh, J., & Taniar, D. (2005, Jan-Mar). Mining parallel pattern from mobile users. *International Journal of Business Data Communications and Networking, 1*(1), 50-76.

Huang, J. L., & Chen, M. S. (2003) Broadcast program generation for unordered queries with data replication. *Proceedings of the 8th ACM Symposium on Applied Computing* (pp. 866-870).

Jayaputera, J., & Taniar, D. (2005). Data retrieval for location-dependent query in a multicell wireless environment. *Mobile Information Systems, IOS Press, 1*(2), 91-108.

Jung, II, D., You, Y. H., Lee, J. J., & Kim, K. (2002). Broadcasting and caching policies for location-dependent queries in urban areas. *Proceedings of the 2nd International Workshop on Mobile Commerce* (pp. 54-59).

Kapp, S. (2002). 802.11: Leaving the wire behind. *IEEE Internet Computing, 6*(1).

Kifer, M., Bernstein, A., & Lewis, P. M. (2006). *Database systems: An application-oriented approach* (2nd ed.). Addison Wesley.

Kubach, U., & Rothermel, K. (2001). A map-based hoarding mechanism for location- dependent information. *Proceedings of the 2nd International Conference on Mobile Data Management* (pp. 145-157).

Lee, C. H., & Chen, M. S. (2002). Processing distributed mobile queries with interleaved remote mobile joins. *IEEE Tran. on Computers, 51*(10), 1182-1195.

Lee, D. K., Xu, J., Zheng, B., & Lee, W. C. (2002, July-Sept.). Data management in location-dependent information services. *IEEE Pervasive Computing, 2*(3), 65-72.

Lee, D. K., Zhu, M., & Hu, H. (2005). When location-based services meet databases. *Mobile Information Systems, 1*(2), 81-90.

Lee, K. C. K., Leong, H. V., & Si, A. (2002). Semantic data access in an asymmetric mobile environment. *Proceedings of the 3rd Mobile Data Management* (pp. 94-101).

Liberatore, V. (2002). Multicast scheduling for list requests. *Proceedings of IEEE INFOCOM Conference* (pp. 1129-1137).

Lo, E., Mamoulis, N., Cheung, D. W., Ho, W. S., & Kalnis, P. (2003). Processing ad-hoc joins on mobile devices. *Database and Expert Systems Applications, Lecure Notes in Computer Science, 3180,* 611-621.

Madria, S. K., Bhargava, B., Pitoura, E., & Kumar, V. (2000). Data organisation for location-dependent queries in mobile computing. *Proceedings of ADBIS-DASFAA* (pp. 142-156).

Malladi, R., & Davis, K. C. (2002). Applying multiple query optimization in mobile databases. *Proceedings of the 36th Hawaii International Conference on System Sciences* (pp. 294-303).

Mamoulis, N., Kalnis, P., Bakiras, S., & Li, X. (2003). Optimization of spatial joins on mobile devices. *Proceedings of the SSTD.*

Myers, B. A., & Beigl, M. (2003). Handheld computing. *IEEE Computer Magazine, 36*(9), 27-29.

Ozakar, B., Morvan, F., & Hameurlain, A. (2005). Mobile join operators for restricted sources. *Mobile Information Systems, 1*(3).

Paulson, L. D. (2003). Will fuel cells replace batteries in mobile devices? *IEEE Computer Magazine, 36*(11), 10-12.

Prabhakara, K., Hua, K. A., & Jiang, N. (2000). Multi-level multi-channel air cache designs for broadcasting in a mobile environment. *Proceedings of the IEEE International Conference on Data Engineering (ICDE'00)* (pp. 167-176).

Ren, Q., & Dunham, M. H. (1999). Using clustering for effective management of a semantic cache in mobile computing. *Proceedings of the ACM International Workshop on Data Engineering for Wireless and Mobile Access* (pp. 94-101).

Ren, Q., & Dunham, M. H. (2000). Using semantic caching to manage location-dependent data in mobile computing. *Proceedings of the 6th International Conference on Mobile Computing and Networking* (pp. 210-221). 2000.

Seydim, A. Y., Dunham, M. H., & Kumar, V. (2001). Location-dependent query processing. *Proceedings of the 2nd International Workshop on Data Engineering on Mobile and Wireless Access (MobiDE'01)* (pp. 47-53).

Tan, R. B. N., Taniar, D., & Lu, G. J. (2004, Sept.). A taxonomy for data cube query. *International Journal of Computers and Their Applications, 11*(3), 171-185.

Taniar, D., & Rahayu, J. W. (2002). Parallel database sorting. *Information Sciences, 146*(1-4), 171-219.

Taniar, D., Jiang, Y., Liu, K. H., & Leung, C. H. C. (2002). Parallel aggregate-join query processing. *Informatica: An International Journal of Computing and Informatics, 26*(3), 321-332.

Tran, D. A., Hua, K. A., & Jiang, N. (2001). A generalized design for broadcasting on multiple physical-channel air-cache. *Proceedings of the ACM SIGAPP Symposium on Applied Computing (SAC'01)* (pp. 387-392).

Triantafillou, P., Harpantidou, R., & Paterakis, M. (2001). High performance data broadcasting: A comprehensive systems perspective. *Proceedings of the 2nd International Conference on Mobile Data Management (MDM 2001)* (pp. 79-90).

Trivedi, K. S., Dharmaraja, S., & Ma, X. (2002). Analytic modelling of handoffs in wireless cellular networks. *Information Sciences, 148*(1-4), 155-166.

Tsalgatidou, A., Veijalainen, J., Markkula, J., Katasonov, A., & Hadjiefthymiades, S. (2003). Mobile e-commerce and location-based services: Technology and requirements. *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Services* (pp. 1-14).

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005a). Indexing schemes for multi channel data broadcasting in mobile databases. *International Journal of Wireless and Mobile Computing.* To appear Mar/Apr.

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005b, Mar.). Research on location-dependent queries in mobile databases. *International Journal of Computer Systems Science & Engineering, 20*(3), 77-93.

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005c). Global indexing scheme for location-dependent queries in multi-channels broadcast environment. *Proceedings of the 19th IEEE International Conference on Advanced Information Networking and Applications, Volume 1, AINA 2005,* (pp. 1011-1016).

Wolfson, O. (2002). Moving objects information management: The database challenge. *Proceedings of the 5th Workshop on Next Generation Information Technology and Systems (NGITS)* (pp. 75-89).

Xu, J., Hu, Q., Lee, W. C., & Lee, D. L. (2004). Performance evaluation of an optimal cache replacement policy for wireless data dissemination. *IEEE Transaction on Knowledge and Data Engineering (TKDE), 16*(1), 125-139.

Xu, J., Zheng, B., Lee, W. C., & Lee, D. L. (2003). Energy efficient index for querying location-dependent data in mobile broadcast environments. *Proceedings of the 19th IEEE International Conference on Data Engineering (ICDE '03)* (pp. 239-250).

Zheng, B., Xu, J., Lee, D. L. (2002). Cache invalidation and replacement strategies for location-dependent data in mobile environments. *IEEE Transactions on Computers, 51*(10), 1141-1153.

KEY TERMS

Location-Dependent Information Processing: Information processing whereby the information requested is based on the current location of the user.

Mobile Database: Databases which are available for access by users using a wireless media through a wireless medium.

Mobile Query Processing: Join processing carried out in a mobile device.

On-Mobile Location-Dependent Information Processing: Location-dependent information processing carried out in a mobile device.

Post-Join: Database operations which are performed after the join operations are completed. These operations are normally carried out to further filter the information obtained from the join.

Pre-Join: Database operations which are carried out before the actual join operations are performed. A pre-join operation is commonly done to reduce the number of records being processed in the join.

This work was previously published in the Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 49-70, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 1.32
Addressing the Credibility of Mobile Applications

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

Mobile access has opened new vistas for various sectors of society including businesses. The ability that anyone using (virtually) any device could be reached anytime and anywhere presents a tremendous commercial potential. Indeed, the number of mobile applications has seen a tremendous growth in the last few years.

In retrospect, the fact that almost *anyone* can set up a mobile application claiming to offer products and services raises the question of credibility from a consumer's viewpoint. The obligation of establishing credibility is essential for an organization's reputation (Gibson, 2002) and for building consumers' trust (Kamthan, 1999). If not addressed, there is a potential for lost consumer confidence, thus significantly reducing the advantages and opportunities the mobile Web as a medium offers. If a mobile application is not seen as credible, we face the inevitable consequence of a product, however functionally superior it might be, rendered socially isolated.

The rest of the article is organized as follows. We first provide the motivational background necessary for later discussion. This is followed by introduction of a framework within which different types of credibility in the context of mobile applications can be systematically addressed and thereby improved. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

In this section, we present the fundamental concepts underlying credibility, and present the motivation and related work for addressing credibility within the context of mobile applications.

Basic Credibility Concepts

For the purposes of this article, we will consider credibility to be synonymous to (and therefore

interchangeable with) believability (Hovland, Janis, & Kelley, 1953). We follow the terminology of Fogg and Tseng (1999), and view credibility and trust as being slightly different. Since trust indicates a *positive* belief about a person, object, or process, we do not consider credibility and trust to be synonymous.

It has been pointed out in various studies (Fogg, 2003; Metzger, 2005) that credibility consists of two primary dimensions, namely *trustworthiness* and *expertise* of the source of some information. Trustworthiness is defined by the terms such as well-intentioned, truthful, unbiased, and so on. The trustworthiness dimension of credibility captures the *perceived* goodness or morality of the source. Expertise is defined by terms such as knowledgeable, experienced, competent, and so on. The expertise dimension of credibility captures the *perceived* knowledge and skill of the source. Together, they suggest that "highly credible" mobile applications will be perceived to have high levels of *both* trustworthiness and expertise.

We note that trustworthiness and expertise are at such a high level of abstraction that direct treatment of any of them is difficult. Therefore, in order to improve credibility, we need to find quantifiable attributes that can improve each of these dimensions.

A Classification of Credibility

The following taxonomy helps associating the concept of credibility with a specific user class in context of a mobile application. A user could consider a mobile application to be credible based upon direct interaction with the application (*active credibility*), or consider it to be credible in absence of any direct interaction but based on certain pre-determined notions (*passive credibility*). Based on the classification of credibility in computer use (Fogg & Tseng, 1999) and adapting them to the domain of mobile applications, we can decompose these further.

There can be two types of *active credibility:* (1) *surface credibility,* which describes how much the user believes the mobile application is based on simple inspection; and (2) *experienced credibility,* which describes how much the user believes the mobile application is based on first-hand experience in the past.

There can be two types of *passive credibility:* (1) *presumed credibility,* which describes how much the user believes the mobile application because of general assumptions that the user holds; and (2) *reputed credibility,* which describes how much the user believes the mobile application because of a reference from a third party.

Finally, credibility is not absolute with respect to users and with respect to the application itself (Metzger, Flanagin, Eyal, Lemus, & McCann, 2003). Also, credibility can be associated with a whole mobile application or a part of a mobile application. For example, a user may question the credibility of information on a specific product displayed in a mobile application. We contend that for a mobile application to be labeled non-credible, there must exist at least a part of it that is labeled non-credible based on the above classification by at least one user.

The Origins and Significance of the Problem of Mobile Credibility

The credibility of mobile applications deserves special attention for the following reasons:

- **Delivery Context:** Mobile applications are different from the desktop or Web environments (Paavilainen, 2002) where context-awareness (Sadeh, Chan, Van, Kwon, & Takizawa, 2003) is a unique challenge. The delivery context in a changing environment of mobile markup languages, variations in user agents, and constrained capabilities of mobile devices presents unique challenges towards active credibility.

- **Legal Context:** Since the stakeholders of a mobile application need not be co-located (different jurisdictions in the same country or in different countries), the laws that govern the provider and the user may be different. Also, the possibilities of fraud such as computer domain name impersonation (commonly known as "pharming") or user identity theft (commonly known as "phishing") with little legal repercussions for the perpetrators is relatively high in a networked environment. These possibilities can impact negatively on presumed credibility.
- **User Context:** Users may deploy mobile devices with varying configurations, and in the event of problems with a mobile service, may first question the provider rather than the device that they own. In order for providers of mobile portals to deliver user-specific information and services, they need to know details about the user (such as profile information, location, and so on). This creates the classical dichotomy between personalization and privacy, and striking a balance between the two is a constant struggle for businesses (Kasanoff, 2002). The benefits of respecting one can adversely affect the other, thereby impacting their credibility in the view of their customers. Furthermore, the absence of a human component from non-proximity or "facelessness" of the provider can shake customer confidence and create negative perceptions in a time of crisis such as denial of service or user agent crash. These instances can lead to a negative passive credibility.

Initiatives for Improving Mobile Credibility

There have been initiatives to address the credibility of Web applications such as a user survey to identify the characteristics that users consider necessary for a Web application to be credible (Fogg et al., 2001) and a set of guidelines (Fogg, 2003) for addressing *surface, experienced, presumed,* and *reputed credibility* of Web applications.

However, these efforts are limited by one or more of the following issues. The approach towards ensuring and/or evaluating credibility is not systematic, the proposed means for ensuring credibility is singular (only guidelines), and the issue of feasibility of the means is not addressed. Moreover, these guidelines are not specific to mobility, are not prioritized and the possibility that they can contradict each other is not considered, can be open to broad interpretation, and are stated at such a high level that they may be difficult to realize by a novice user.

ADDRESSING THE CREDIBILITY OF MOBILE APPLICATIONS

In this section, we consider approaches for understanding and improving active credibility of mobile applications.

A Framework for Addressing Active Credibility of Mobile Applications

To systematically address the active credibility of mobile applications, we take the following steps:

1. View credibility as a qualitative aspect and address it indirectly via quantitative means.
2. Select a theoretical basis for communication of information (semiotics), and place credibility in its setting.
3. Address semiotic quality in a practical manner.

Based on this and using the primary dimensions that affect credibility, we propose a framework for active credibility of mobile applications

Table 1. A semiotic framework for active credibility of mobile applications

Semiotic Level	Quality Attributes	Means for Credibility Assurance and Evaluation	Decision Support
Social	**Credibility**	• "Expert" Knowledge (Principles, Guidelines, Patterns) • Inspections • Testing • Metrics • Tools	Feasibility
	Aesthetics, Legality, Privacy, Security, (Provider) Transparency [T5;E]		
Pragmatic	Accessibility, Usability [T4;E]		
	Interoperability, Portability, Reliability, Robustness [T3;E]		
Semantic	Completeness and Validity [T2;I]		
Syntactic	Correctness [T1;I]		

(see Table 1). The external attributes (denoted by E) are extrinsic to the software product and are directly a user's concern, while internal attributes (denoted by I) are intrinsic to the software product and are directly an engineer's concern. Since not all attributes corresponding to a semiotic level are at the same echelon, the different tiers are denoted by "Tn."

We now describe each of the components of the framework in detail.

Semiotic Levels

The first column of Table 1 addresses semiotic levels. Semiotics (Stamper, 1992) is concerned with the use of symbols to convey knowledge.

From a semiotics perspective, a representation can be viewed on six interrelated levels: physical, empirical, syntactic, semantic, pragmatic, and social, each depending on the previous one in that order. The physical and empirical levels are concerned with the physical representation of signs in hardware and communication properties of signs, and are not of direct concern here. The syntactic level is responsible for the formal or structural relations between signs. The semantic level is responsible for the relationship of signs to what they stand for. The pragmatic level is responsible for the relation of signs to interpreters. The social level is responsible for the manifestation of social interaction with respect to signs.

Quality Attributes

The second column of Table 1 draws the relationship between semiotic levels and corresponding quality attributes.

Credibility belongs to the social level and depends on the layers beneath it. The external quality attributes *legality, privacy, security,* and *(provider) transparency* also at the social level depend upon the external quality attributes *accessibility* and *usability* at the pragmatic level, which in turn depend upon the external quality attributes *interoperability, performance, portability, reliability,* and *robustness* also at the pragmatic level. (We note here that although *accessibility* and *usability* do overlap in their design

and implementation, they are not identical in their goals for their user groups.)

We discuss in some detail only the entries in the social level. Aesthetics is close to human senses and perception, and plays a crucial role in making a mobile application "salient" to its customers beyond simply the functionality it offers. It is critical that the mobile application be legal (e.g., is legal in the jurisdiction it operates and all components it makes use of are legal); takes steps to respect a user's privacy (e.g., does not use or share user-supplied information outside the permitted realm); and be secure (e.g., in situations where financial transactions are made). The provider must take all steps to be transparent with respect to the user (e.g., not include misleading information such as the features of products or services offered, clearly label promotional content, make policies regarding returning/exchanging products open, and so on).

The internal quality attributes for syntactic and semantic levels are inspired by Lindland, Sindre, and Sølvberg (1994) and Fenton and Pfleeger (1997). At the semantic level, we are only concerned with the conformance of the mobile application to the domain(s) it represents (that is, semantic correctness or completeness) and vice versa (that is, semantic validity). At the syntactic level the interest is in conformance with respect to the languages used to produce the mobile application (that is, syntactic correctness).

The definitions of each of these attributes can vary in the literature, and therefore it is important that they be adopted and followed consistently. For example, the definition of usability varies significantly across ISO/IEC Standard 9126 and ISO Standard 9241 with respect to the perspective taken in their formulation.

Means for Credibility Assurance and Evaluation

The third column of Table 1 lists (in no particular order, by no means complete, and not necessarily mutually exclusive) the means for assuring and evaluating active credibility.

- **"Expert" Body of Knowledge:** The three types of knowledge that we are interested in are principles, guidelines, and patterns. Following the basic principles (Ghezzi, Jazayeri, & Mandrioli, 2003) underlying a mobile application enables a provider to improve quality attributes related to (T1-T3) of the framework. The guidelines encourage the use of conventions and good practice, and could also serve as a checklist with respect to which an application could be heuristically or otherwise evaluated. There are guidelines available for addressing accessibility (Chisholm, Vanderheiden, & Jacobs, 1999; Ahonen, 2003), security (McGraw & Felten, 1998), and usability (Bertini, Catarci, Kimani, & Dix, 2005) of mobile applications. However, guidelines tend to be more useful for those with an expert knowledge than for a novice to whom they may seem rather general to be of much practical use. Patterns are reusable entities of knowledge and experience aggregated by experts over years of "best practices" in solving recurring problems in a domain including that in mobile applications (Roth, 2001, 2002). They are relatively more structured compared to guidelines and provide better opportunities for sharing and reuse. There is, however, a lack of patterns that clearly address quality concerns in mobile applications. Also, there is a cost of adaptation of patterns to new contexts.
- **Inspections:** Inspections (Wiegers, 2002) are a rigorous form of auditing based upon peer review that can address quality concerns at both technical and social levels (T1-T5), and help improve the credibility of mobile applications. Inspections could, for example, decide what information is and is not considered "promotional," help

improve the labels used to provide cues to a user (say, in a navigation system), and assess the readability of documents. Still, inspections do involve an initial cost overhead from training each participant in the structured review process, and the logistics of checklists, forms, and reports.

- **Testing:** Some form of testing is usually an integral part of most development models of mobile applications (Nguyen, Johnson, & Hackett, 2003). There are test suites and test harnesses for many of the languages commonly used for representation of information in mobile applications. However, due to its very nature, testing addresses quality concerns of only some of the technical and social levels (T1, subset of T2, T3, T4, subset of T5). Therefore, testing *complements* but does not replace inspections. Accessibility or usability testing that requires hiring real users, infrastructure with video monitoring, and subsequent analysis of data can prove to be prohibitive for small-to-medium-size enterprises.
- **Metrics:** In a resource-constrained environment of mobile devices, efficient use of time and space is critical. Metrics (Fenton & Pfleeger, 1997) provide a quantitative means for making qualitative judgments about quality concerns at technical levels. For example, metrics for a document or image size can help compare and make a choice between two designs, or metrics for structural complexity could help determine the number of steps required in navigation, which in turn could be used to estimate user effort. However, well-tested metrics for mobile applications are currently lacking. We also note that a dedicated use of metrics on a large scale usually requires tool support.
- **Tools:** Tools that have help improve quality concerns at technical and social levels. For example, tools can help engineers detect security breaches, report violations of accessibility or usability guidelines, find non-conformance to markup or scripting language syntax, suggest image sizes favorable to the small devices, or detect broken links. However, at times, tools cannot address some of the technical quality concerns (like complete semantic correctness of the application with respect to the application domain), as well as certain social quality concerns (like provider intent or user bias). Therefore, the use of tools as means for automatic quality assurance or evaluation should be kept in perspective.

Decision Support

A mobile application project must take a variety of constraints into account: organizational constraints of time and resources (personnel, infrastructure, budget, and so on) and external forces (market value, competitors, and so on). These compel providers to make quality-related decisions that, apart from being sensitive to credibility, must also be feasible.

For example, the provider of a mobile application should carry out intensive accessibility and usability evaluations, but ultimately that application must be delivered on a timely basis. Also, the impossibility of complete testing is well known.

Indeed, the last column of Table 1 acknowledges that with respect to any assurance and/or evaluation, and includes feasibility as an all-encompassing consideration on the layers to make the framework practical. There are well-known techniques such as analytical hierarchy process (AHP) and quality function deployment (QFD) for carrying out feasibility analysis, and further discussion of this aspect is beyond the scope of this article.

Limitations of Addressing Credibility

We note here that credibility, as is reflected by its primary dimensions, is a socio-cognitive concern that is not always amenable to a purely technological treatment. However, by decomposing it into quantifiable elements and approaching them in a systematic and feasible manner, we can make improvements towards its establishment.

We assert that the quality attributes we mention in pragmatic and social levels are necessary but make no claim of their sufficiency. Indeed, as we move from bottom to top, the framework gets less technically oriented and more human oriented. Therefore, finding sufficient conditions for establishing credibility is likely to be an open question, and it may be virtually impossible to provide complete guarantees for credibility.

FUTURE TRENDS

In the previous section, we discussed active credibility; the issue of passive credibility poses special challenges and is a potential area of future research. We now briefly look at the case of reputed credibility.

In case of Web applications, there have been two notable initiatives in the direction of addressing reputed credibility, namely WebTrust and TRUSTe. In response to the concerns related to for business-to-consumer electronic commerce and to increase consumer confidence, the American Institute of Certified Public Accountants (AICPA) and Canadian Institute of Chartered Accountants (CICA) have developed WebTrust Principles and Criteria and the related WebTrust seal of assurance. Independent and objective certified public accountant or chartered accountants, who are licensed by the AICPA or CICA, can provide assurance services to evaluate and test whether a particular Web application meets these principles and criteria. The TRUSTe program enables companies to develop privacy statements that reflect the information gathering and dissemination practices of their Web application. The program is equipped with the TRUSTe "trustmark" seal that takes users directly to a provider's privacy statement. The trustmark is awarded only to those that adhere to TRUSTe's established privacy principles and agree to comply with ongoing TRUSTe oversight and resolution process. Admittedly, not in the realm of pure academia, having similar quality assurance and evaluation programs for mobile applications, and perhaps even the use of ISO 9001:2000 as a basis for a certification, would be of interest.

A natural extension of the preceding discussion on credibility could be in the context of the next generation of mobile applications such as semantic mobile applications (Alesso & Smith, 2002) and mobile Web services (Salmre, 2005). For example, ontological representation of information can present certain human-centric challenges (Kamthan & Pai, 2006) that need to be overcome for it to be a credible knowledge base.

Finally, viewing a mobile application as an information system, it would of interest to draw connections between credibility and ethics (Johnson, 1997; Tavani, 2004).

CONCLUSION

Although there have been significant advances towards enabling the technological infrastructure (Coyle, 2001) for mobile access in the past decade, there is much to be done in addressing the social challenges. Addressing credibility of mobile applications in a systematic manner is one step in that direction.

The organizations that value credibility of their mobile applications need to take two aspects into consideration: (1) take a *systematic* approach to

the development of the mobile applications, and (2) consider credibility as a first-class concern *throughout* the process. The former need to particularly include support for modeling a user's environment (context, task, and device) (Gandon & Sadeh, 2004) and mobile user interface engineering. The latter implies that credibility is viewed as a *mandatory* non-functional requirement during the analysis phase and treated as a central design concern in the synthesis phase.

In a user-centric approach to engineering, mobile applications belong to an *ecosystem* that includes both the people and the product. If the success of a mobile application is measured by use of its services, then establishing credibility with the users is critical for the providers. By making efforts towards improving the criteria that directly or indirectly affect credibility, the providers can meet user expectations and change the user perceptions in their favor.

REFERENCES

Ahonen, M. (2003, September 19). Accessibility challenges with mobile lifelong learning tools and related collaboration. *Proceedings of the Workshop on Ubiquitous and Mobile Computing for Educational Communities (UMOCEC 2003),* Amsterdam, The Netherlands.

Alesso, H. P., & Smith, C. F. (2002). *The intelligent wireless Web.* Boston: Addison-Wesley.

Bertini, E., Catarci, T., Kimani, S., & Dix, A. (2005). A review of standard usability principles in the context of mobile computing. *Studies in Communication Sciences, 1*(5), 111-126.

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *Web content accessibility guidelines 1.0.* W3C Recommendation, World Wide Web Consortium (W3C).

Coyle, F. (2001). *Wireless Web: A manager's guide.* Boston: Addison-Wesley.

Fenton, N. E., & Pfleeger, S. L. (1997). *Software metrics: A rigorous & practical approach.* International Thomson Computer Press.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do.* San Francisco: Morgan Kaufmann.

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., et al. (2001, March 31-April 5). What makes Web sites credible?: A report on a large quantitative study. *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference,* Seattle, WA.

Fogg, B. J., & Tseng, S. (1999, May 15-20). The elements of computer credibility. *Proceedings of the ACM CHI 99 Conference on Human Factors in Computing Systems,* Pittsburgh, PA.

Gandon, F. L., & Sadeh, N. M. (2004, June 1-3). Context-awareness, privacy and mobile access: A Web semantic and multiagent approach. *Proceedings of the 1st French-Speaking Conference on Mobility and Ubiquity Computing,* Nice, France (pp. 123-130).

Gibson, D. A. (2002). *Communities and reputation on the Web.* PhD Thesis, University of California, USA.

Ghezzi, C., Jazayeri, M., & Mandrioli, D. (2003). *Fundamentals of software engineering* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hovland, C. I., Janis, I. L., & Kelley, J. J. (1953). *Communication and persuasion.* New Haven, CT: Yale University Press.

Johnson, D. G. (1997). Ethics online. *Communications of the ACM, 40*(1), 60-65.

Lindland, O. I., Sindre, G., & Sølvberg, A. (1994). Understanding quality in conceptual modeling. *IEEE Software, 11*(2), 42-49.

Kamthan, P. (1999). *E-commerce on the WWW: A matter of trust.* Internet Related Technologies (IRT.ORG).

Kamthan, P., & Pai, H.-I. (2006, May 21-24). Human-centric challenges in ontology engineering for the semantic Web: A perspective from patterns ontology. *Proceedings of the 17th Annual Information Resources Management Association International Conference (IRMA 2006),* Washington, DC.

Kasanoff, B. (2002). *Making it personal: How to profit from personalization without invading privacy.* New York: John Wiley & Sons.

McGraw, G., & Felten, E. W. (1998). Mobile code and security. *IEEE Internet Computing, 2*(6).

Metzger, M. J. (2005, April 11-13). Understanding how Internet users make sense of credibility: A review of the state of our knowledge and recommendations for theory, policy, and practice, *Proceedings of the Internet Credibility and the User Symposium,* Seattle, WA.

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D., & McCann, R. (2003). Bringing the concept of credibility into the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication Yearbook, 27,* 293-335.

Nguyen, H. Q., Johnson, R., & Hackett, M. (2003). *Testing applications on the Web: Test planning for mobile and Internet-based systems* (2nd ed.). New York: John Wiley & Sons.

Paavilainen, J. (2002). *Mobile business strategies: Understanding the technologies and opportunities.* Boston: Addison-Wesley.

Roth, J. (2001, September 10). Patterns of mobile interaction. *Proceedings of the 3rd International Workshop on Human Computer Interaction with Mobile Devices (Mobile HCI 2001),* Lille, France.

Roth, J. (2002). Patterns of mobile interaction. *Personal and Ubiquitous Computing, 6*(4), 282-289.

Sadeh, N. M., Chan, T.-C., Van, L., Kwon, O., & Takizawa, K. (2003, June 9-12). A semantic Web environment for context-aware m-commerce. *Proceedings of the 4th ACM Conference on Electronic Commerce,* San Diego, CA (pp. 268-269).

Salmre, I. (2005). *Writing mobile code: Essential software engineering for building mobile applications.* Boston: Addison-Wesley.

Stamper, R. (1992, October 5-8). Signs, organizations, norms and information systems. *Proceedings of the 3rd Australian Conference on Information Systems,* Wollongong, Australia.

Tavani, H. T. (2004). *Ethics and technology: Ethical issues in an age of information and communication technology.* New York: John Wiley & Sons.

Wiegers, K. (2002). *Peer reviews in software: A practical guide.* Boston: Addison-Wesley.

KEY TERMS

Delivery Context: A set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user, and other aspects of the context into which a resource is to be delivered.

Mobile Web Engineering: A discipline concerned with the establishment and use of sound scientific, engineering, and management principles, and disciplined and systematic approaches to the successful development, deployment, and maintenance of high-quality mobile Web applications.

Personalization: A strategy that enables delivery that is customized to the user and user's environment.

Quality: The totality of features and characteristics of a product or a service that bear on its ability to satisfy stated or implied needs.

Semantic Web: An extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

Semiotics: The field of study of signs and their representations.

User Profile: A information container describing user needs, goals, and preferences.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 25-31, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Section II
Development and Design Methodologies

This section provides in-depth coverage of conceptual architectures, frameworks and methodologies related to the design and implementation of mobile systems and technologies. Throughout these contributions, research fundamentals in the discipline are presented and discussed. From broad examinations to specific discussions on particular frameworks and infrastructures, the research found within this section spans the discipline while also offering detailed, specific discussions. Basic designs, as well as abstract developments, are explained within these chapters, and frameworks for designing successful mobile applications, interfaces, and agents are discussed.

Chapter 2.1
Developing Smart Client Mobile Applications

Jason Gan
University of Technology, Australia

ABSTRACT

This chapter examines the convergence of mobile technologies based on smart client architecture. To improve the usability and accessibility of mobile applications and services, the smart client architecture extends the capabilities of the mobile computing platform with support for mutimodal interfaces, smart client database and synchronization, presence awareness, location awareness and identity management. Its broad impact on business communication and productivity is highlighted as a tangible benefit.

INTRODUCTION

In the highly competitive mobile market, a key differentiator is provided by improving the user experience of the mobile application. To improve the user experience, common usability and accessibility problems in mobile applications can be mitigated by multimodal interfaces and smart client architecture. For instance, the provision of multimodal interfaces for browser-based applications can help to overcome the limitations of small viewing areas and input options, whereas smart clients based on rich application interfaces can be utilized to push processing load onto the mobile device. Furthermore, smart clients that enable presence, context sensitivity, location awareness, and real-time collaboration promise a new paradigm for mobile communications, delivering far richer, dynamic user experiences.

SMART CLIENT

As mobile enabling technologies advance in capability, affordability, and availability, users expect improved design of mobile devices that will leverage the advances and convergence in technology and the Internet to deliver richer applications and value-added mobile services (a.k.a., m-services). A key enabling technology for delivering on the promise of mobile applications with high levels of functionality, performance, flexibility, and integration is the *smart client.* This is a type of

application model that bridges the gap between the thick and thin client models, providing the responsiveness and adaptability of a thick client model with the manageability of a thin client.

Dave Hill, from the Microsoft .NET Enterprise Architecture team, defines five characteristics of a smart client application (2004):

1. **Utilizes Local Resources:** Smart clients exploit local resources such as hardware for storage, processing, or data capture to deliver a richer user experience.
2. **Connected:** Smart clients are ready to connect and exchange data with various systems across the enterprise.
3. **Off-Line Capable:** Off-line capability using local caching and processing enable operation during periods of disconnection or intermittent network connectivity. Smart clients can send data in the background, resulting in greater responsiveness in the user interface.
4. **Intelligent Install and Update:** The smart client interface allows the remote update of the smartphone software to repair bugs, change characteristics, or incorporate new features.
5. **Client Device Flexibility:** Smart client applications support multiple versions that target specific device type and functionality.

The smart client architecture supports multimodality, data integration, Bluetooth interoperability, presence awareness, location awareness, and identity management. Each of these features extends serviceable functionality to the mobile application, from voice-activated commands to authentication and non-repudiation services. Moreover, the integration of serviceable functionality promises to deliver rich user experiences. For example, the voice-activated smart wireless device will automatically connect, authenticate, and show the identity and location of the receiver.

The convergence of presence, location, and identity management is an emergent technology integrating the services that support a secure mobile network and an online community environment with applications that facilitate information retrieval, communications, dating, gambling, financial management, trading, paying bills, games, and entertainment. As the mobile market is highly competitive and dynamic, and driven by the mass market demand for high-performance m-applications, the impact of technology convergence highlights the need for common industry standards.

INDUSTRY STANDARDS FOR SMART CLIENT MOBILE APPLICATIONS

As the specifications for smart client mobile application interfaces are complex, it is necessary for wireless developers to adhere to industry standards. The Mobile Industry Processor Interface (MIPI) Alliance is a non-profit organization that spearheaded the initiative of industry specifications for smartphones and application-rich mobile devices.

Wireless and embedded software developers have a choice between Microsoft .NET and Sun Java development frameworks and runtime environments for designing and delivering next-generation mobile applications.

The Microsoft .NET Compact Framework is a subset of the developer software for PCs and servers that allows powerful .NET applications to run on handheld computers, and specifically supports: Pocket PC, embedded solutions running on Windows CE .NET for smart mobile devices, and Microsoft Smartphone 2002. The .NET Compact Framework can be extended to support additional mobile device interfaces. The inclusion of SQL Server CE 2.0 provides developers with a powerful, local relational database for creating dynamic, client-side mobile applications with

database replication that enables remote devices to edit data in parallel. SQL Server 2000 supports two methods of replicating information from a back-end master to a remote client database: Remote Data Access (RDA) and merge replication (Thews, 2003).

The Sun Java platform includes the Micro Edition specification for building rich and smart client applications for mobile and embedded devices.

SMART CLIENT AND MULTIMODALITY

Smart client mobile applications equipped with multimodality and speech capabilities represent the next generation in portable office communications and wearable computer technology. Multimodality is defined as the optional presentation of the same information content in more than one sensory mode (European Communications Standards Institute, 2003). The multimodal interface on a multiple context-aware device enables interaction through different communication channels (a.k.a., modalities) to overcome the inherent input/output limitations of mobile devices. For example, a user might use voice-only commands rather than the standard keypad input, while audio output can deliver additional information that cannot fit on the screen. The multimodal information is processed by the interpreter component of a multimodal host server.

Multimodal integration is the combination of different modalities to form a flexible user interface. There are seven types of modalities: visual, auditory, tactile, olfactory, gustatory, vestibular, and proprioception. Of these, visual and auditory are the most commonly used in m-applications. However, users who are blind and/or mute stand to benefit from the availability of tactile feedback and gesture modalities that address their physical disabilities.

SMART CLIENT DATABASE AND SYNCHRONIZATION

Reliable, secure, and immediate access to enterprise data irrespective of time and place is paramount in a time-critical business application, and it is especially critical in completing an online session—for example, when you have to close a bid on time. The problem with wireless technology is that data access and connectivity are intermittently connected and unreliable due to dropped connections, coverage issues, low bandwidth, and high latency. Smart client off-line access to data using a localized mobile database and synchronization with a central database (a store-and-forward mechanism) can help reduce and eliminate the performance bottlenecks related to slow and unreliable networks, and thus improve the user experience.

Besides improving the user experience, mobile applications can also affect the work environment. In particular, the collaborations of disparate teams stand to benefit from presence awareness.

PRESENCE AWARENESS

A study on disparate software project teams discussed the potential impact on productivity and the bottom line from leveraging presence and collaborative technologies in multi-site environments (Herbsleb, Mockus, Finholt, & Grinter, 2001, p. 9).

Presence is defined as a collection of real-time data describing the ability and willingness of a user to communicate across specific media and devices (Schneyderman, 2004). Presence awareness, a vital property of instant messaging applications, allows users to know when other users in a community are online and willing to exchange messages, what devices can be used for communications,

and the real-time status of these devices. This can result in time and cost savings and improved productivity in many enterprise environments. For example, customer service environments stand to benefit from decreased operational costs and the ability to connect knowledge experts in real time across geographical locations and time zones. The adoption of presence technology into the workplace enables disparate workers to connect and communicate more efficiently by overcoming the lack of context and absence of informal communication (Herbsleb, Atkins, Boyer, Handel, & Finholt, 2002). A presence-based publish-and-subscribe channel can be deployed for discovery of available managers in workflow collaborations and for push content delivery.

It is important for developers to address the security and privacy concerns regarding presence technology. Building security and privacy controls into the presence-enabled device allows such features as access control, visibility, message blocking, and message encryption. Privacy policies restrict communication channels and prevent unsolicited conversations or uninvited listening. Biometric scanners built into the presence-equipped device help to prevent such exploits as identity masquerading and spoofing. In addition, each receiving device would be configured to transmit its location and identity information to the mobile network.

LOCATION AWARENESS AND PERVASIVE COMPUTING

Location awareness in smart mobile applications is provided by mobile positioning technology such as the Global Positioning System (GPS). The GPS is a satellite navigation system that allows a mobile device with a GPS receiver to pinpoint a location on Earth by measuring the distances from a number of satellites simultaneously. Integration with GPS technology provides a mobile device with a pervasive-computing interface to location-based services (LBSs) such as emergency assistance and personal navigation. Assisted GPS (AGPS) describes a mobile positioning system that consists of the integrated GPS receiver and network resources such as an assistance server and reference network. The assistance server communicates with the GPS receiver via the cellular link and accesses data from the reference network, resulting in boosted performance of the receiver.

IDENTITY MANAGEMENT

Smart client mobile applications may require that access privileges to specific resources are granted only to users who have the right to use those resources. In an e-business system, access to certain resources is restricted to those who can supply the proper identity credentials by a process called authentication. Identity management is the federation of trusted endpoints that are able to authenticate each other's presence at these endpoints with the intention of having a secure transaction (Smith, 2002). The banking and finance industry is a strong driver for this technology to reduce online fraud. Moreover, businesses and customers stand to benefit from expedited billing and payments. As identity and trust form the basis of successful e- and m-business, there is a strong demand from business and government sectors for efficient mobile integration of identity management services that will provide the end user with a more satisfactory online experience, enriched with higher levels of personalization, security, and control over identity information.

CONCLUSION AND FUTURE DIRECTION

The smart client provides an application model for developing richer, more responsive, and more usable mobile applications. Key areas of

technological development based on the smart client are multimodality, presence awareness, location awareness, and identity management. These technologies describe a future to be defined by voice-activated systems, real-time collaboration, location-based services for information or emergencies, and mobile integration of identity services. Developing smart mobile applications that leverage the advances and convergence in technology is a step closer to realizing the full potential of a wireless Internet.

REFERENCES

Antonopoulos, A. M. (2004, May 10). Location and presence take identity management to the next level. *Network World Data Center Newsletter.* Retrieved December 28, 2004, from http://www.nwfusion.com/newsletters/datacenter/2004/1004datacenter1.html

European Communications Standards Institute. (2003). Human factors (HF); multimodal interaction, communication and navigation guidelines. *ETSI EG 202 191, 1.1.1*(August).

Herbsleb, J. D., Atkins, D. L., Boyer, D. G., Handel, M., & Finholt, T. A. (2002). Introducing instant messaging and chat in the workplace. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our World, Changing Ourselves.* Minneapolis, MN. New York: ACM Press.

Herbsleb, J. D., Mockus, A., Finholt, T. A., & Grinter, R. E. (2001, May 12-19). An empirical study of global software development. *Proceedings of the 23rd International Conference on Software Engineering* (ICSE'01), Toronto, Canada. Retrieved from http://www-2.cs.cmu.edu/~jdh/collaboratory/research_papers/ICSE_01_final(2).pdf

Hill, D. (2004). *What is a smart client anyway?* Retrieved December 12, 2004, from http://weblogs.asp.net/dphill/articles/66300.aspx

Schneyderman, A. (2004). *Presence in mobile VoIP networks.* Retrieved from http://www.tmcnet.com/voip/0904/featureshney derman.htm

Smith, R. (2002). *Identity management—give me liberty or give me passport?* Retrieved December 28, 2004, from http://www.giac.org/practical/GSEC/Robert_Smith_GSEC.pdf

Thews, D. (2003, October). Create mobile database apps. *Visual Studio Magazine.* Retrieved January 23, 2005, from http://www.ftponline.com/vsm/2003_10/magazine/features/thews/

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 504-508 , copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.2
Engineering Wireless Mobile Applications

Qusay H. Mahmoud
University of Guelph, Canada

Zakaria Maamar
Zayed University, UAE

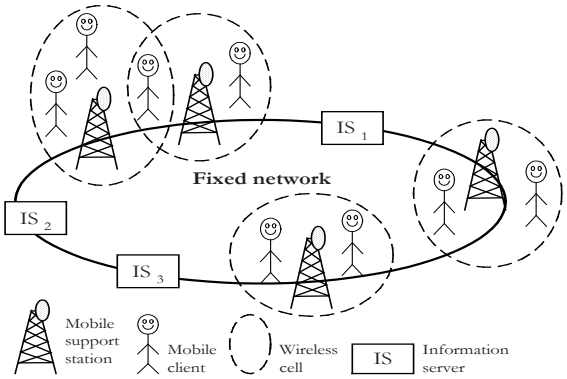
ABSTRACT

Conventional desktop software applications are usually designed, built, and tested on a platform similar to the one on which they will be deployed and run. Wireless mobile application development, on the other hand, is more challenging because applications are developed on one platform (like UNIX or Windows) and deployed on a totally different platform like a cellular phone. While wireless applications can be much smaller than conventional desktop applications, developers should think in small terms of the devices on which the applications will run and the environment in which they will operate instead of the amount of code to be written. This paper presents a systematic approach to engineering wireless application and offers practical guidelines for testing them. What is unique about this approach is that it takes into account the special features of the new medium (mobile devices and wireless networks), the operational environment, and the multiplicity of user backgrounds; all of which pose new challenges to wireless application development.

INTRODUCTION

The general mobile computing model in a wireless environment consists of two distinct sets of entities (Figure 1): Mobile Clients (MCs) and fixed hosts. Some of the fixed hosts, called Mobile Support Stations (MSSs), are enhanced with wireless interfaces. An MSS can communicate with the MCs within its radio coverage area called wireless cell. An MC can communicate with a fixed host/server via an MSS over a wireless channel. The wireless channel is logically separated into two sub-channels: an uplink channel and a downlink channel. The uplink channel is used by MCs to submit queries to the server via an MSS, whereas the downlink channel is used by MSSs to disseminate information or to forward the responses from the server to a target client.

Figure 1. Mobile computing model



Each cell has an identifier (CID) for identification purposes. A CID is periodically broadcasted to all the MCs residing in a corresponding cell.

A wireless mobile application is defined as a software application, a wireless service or a mobile service that can be either pushed to users' handheld wireless devices or downloaded and installed, over the air, on these devices.[1] Such applications must work within the daunting constraints of the devices themselves:

- **Memory:** Wireless devices such as cellular phones and two-way pagers have limited amounts of memory, obliging developers to consider memory management most carefully when designing application objects.
- **Processing power:** Wireless devices also have limited processing power (16-bit processors are typical).
- **Input**: Input capabilities are limited. Most cell phones provide only a one-hand keypad with twelve buttons: the ten numerals, an asterisk (*), and a pound sign (#).
- **Screen:** The display might be as small as 96 pixels wide by 54 pixels high and 1 bit deep (black and white). The amount of information that can be squeezed into such a tight screen is severely limited.

In addition, the wireless environment imposes further constraints: (1) wireless networks are unreliable and expensive, and bandwidth is low; (2) they tend to experience more network errors than wired networks; and (3) the very mobility of wireless devices increases the risk that a connection will be lost or degraded. In order to design and build reliable wireless applications, designers need to keep these constraints in mind and ask themselves, what impact do wireless devices with limited resources have on application design?

The motivation for this paper is provided in part by the above characteristics that form some of the foundations for pervasive computing environments. Such characteristics pose several challenges in designing wireless mobile applications for mobile computing. This paper provides a detailed treatment of the impact of these characteristics on engineering wireless mobile applications and presents a systematic approach for designing them. In addition, it offers practical design techniques for wireless application design and development.

WIRELESS APPLICATIONS

Wireless applications can be classified into two streams (Beaulieu, 2002; Burkhardt, Henn, Hepper, Rintdorff, & Schack, 2002):

1. **Browser-based:** Applications developed using a markup language. This is similar to the current desktop browser model where the device is equipped with a browser. The Wireless Application Protocol or WAP (http://www.openmobile alliance.org) follows this approach (Open Mobile Alliance, 2005).
2. **Native applications:** Compiled applications where the device has a runtime environment to execute applications. Highly interactive wireless applications are only possible with the latter model. Interactive applications, such as mobile computer games, are a good example. Such applications can be developed using the fast growing Java 2 Micro Edition

(J2ME) platform (http://www.java.sun.com/j2me), and they are known as MIDlets.

Another stream is the hybrid application model that aims at incorporating the best aspects of both streams above. The browser is used to allow the user to enter URLs to download native applications from remote servers, and the runtime environment is used to let these applications run on the device.

WAP Might be Dead, but What Did We Learn?

WAP and J2ME MIDP solve similar problems but each can learn a couple of things from the other. There are special features that are available in WAP but not in MIDP and *vice versa*. These features are summarized as follows:

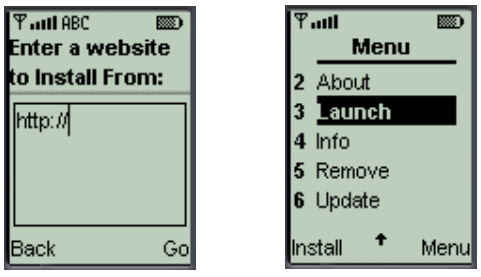
- MIDP provides a low-level graphics APIs that enable the programmer to have control over every pixel of the device's display. This is important for entertainment applications (such as games) in a wireless environment.
- MIDP is the way to go for games. The nature of MIDlets (they exist on the device until they are explicitly removed) allows users to run them even when the server becomes unavailable (support for disconnected operations).
- WML provides tags and possible presentation attributes, but it doesn't define an interaction model. For example, WML defines a SELECT tag for providing a list. Some WAP-enabled devices interpret the SELECT tag as a popup menu list while others interpret it as a menu that can be used for navigation. Therefore, there is no standard interaction model defined for this element. If a developer uses it, the application may run well on some devices and poorly on others. MIDlets, on the other hand, provide a clearly defined standard for interaction using commands.

A Micro Browser is Needed

MIDlets combine excellent online and off-line capabilities that are useful for the wireless environment, which suffers from low bandwidth and network disconnection. Integrating WAP and MIDP opens up possibilities for new wireless applications and over the air distribution models. Therefore, WAP and MIDP shouldn't be viewed as competing but rather as complementing technologies. In order to facilitate downloading wireless applications over the air, there is a need for some kind of an environment on the handset that allows the user to enter a URL for a MIDlet Suite, for example. This environment could very well be a WAP browser as shown in Figure 2.

Similar to Java Applets that are integrated into HTML, MIDlets can be integrated into a WML or an XHTML page. Such a page can then be called from a WAP browser, and the embedded MIDlet gets downloaded and installed on the device. In order to enable this, a WAP browser is needed on the device. Another alternative approach for over-the-air provisioning is the use of a Short Message Service (SMS) which has been done by Siemens where the installation of MIDlets is accomplished by sending a corresponding SMS. If the SMS contains a URL to a Java Application

Figure 2. Combining WAP and J2ME



Descriptor (JAD) file specifying a MIDlet Suite, then the recipient can install the application simply by confirming the SMS.

DESIGN CHALLENGES AND POSSIBLE SOLUTIONS

In this paper, we are more concerned with native interactive applications that can be developed using the J2ME platform or a similar technology. J2ME-based wireless applications can be classified into local (stand-alone) and network applications. Local applications perform all their operations on a handheld wireless device and need no access to external data sources through a wireless network. Examples include calculators and single-player games. Network applications, on the other hand, consist of some components running on a wireless device and others running on a network, and thus depend on access to external resources. An example would be an e-mail application with a client residing on a wireless phone interacting with a Simple Mail Transfer Protocol (SMTP) server to send/receive e-mail messages. A major difference between local and networked applications is in the way they are tested. Local applications are easier to test than network applications. For example, a calculator application can run on a wireless device even when it is not connected to any network, but an e-mail client will not work without a connection to e-mail servers.

Challenges

The constraints discussed earlier pose several crucial challenges, which must be faced in order for wireless applications to function correctly in the target environment.

- **Transmission errors:** Messages sent over wireless links are exposed to interference (and varying delays) that can alter the content received by the user, the target device, or the server. Applications must be prepared to handle these problems. Transmission errors may occur at any point in a wireless transaction and at any point during the sending or receiving of a message. They can occur after a request has been initiated, in the middle of the transaction, or after a reply has been sent. While wireless network protocols may be able to detect and correct some errors, error-handling strategies that address all kinds of transmission errors that are likely to occur are still needed.
- **Message latency:** Message latency, or the time it takes to deliver a message, is primarily affected by the nature of each system that handles the message, and by the processing time needed and delays that may occur at each node from origin to destination. Message latency should be taken into account and users of wireless applications should be kept informed of processing delays. It is especially important to remember that a message may be delivered to a user long after the time it is sent. A long delay might be due to coverage problems or transmission errors, or the user's device might be switched off or have a dead battery. Some systems keep trying, at different times, to transmit the message until it is delivered. Other systems store the message then deliver it when the device is reconnected to the network. Therefore, it is important to design applications that avoid sending stale information, or at least to make sure that users are aware that it is not up-to-date. Imagine the possible consequences of sending a stock quote that is three days old without warning the user!
- **Security:** Any information transmitted over wireless links is subject to interception. Some of that information could be sensitive, like credit card numbers and other personal information. The solution needed really depends on the level of sensitivity. To provide

a complete end-to-end security solution, you must implement it on both ends, the client and the server, and assure yourself that intermediary systems are secure as well.

Possible Solutions

Here are some practical hints useful to consider when developing mobile applications.

- **Understand the environment.** Do some research upfront. As with developing any other software application, we must understand the needs of the potential users and the requirements imposed by all networks and systems the service will rely on.
- **Choose an appropriate architecture.** The architecture of the mobile application is very important. No optimization techniques will make up for an ill-considered architecture. The two most important design goals should be to minimize the amount of data transmitted over the wireless link, and to anticipate errors and handle them intelligently.
- **Partition the application.** Think carefully when deciding which operations should be performed on the server and which on the handheld device. Downloadable wireless applications allow locating much of an application's functionality of the device; it can retrieve data from the server efficiently, then perform calculations and display information locally. This approach can dramatically reduce costly interaction over the wireless link, but it is feasible only if the device can handle the processing that the application needs to perform.
- **Use compact data representation.** Data can be represented in many forms, some more compact than others. Consider the available representations and select the one that requires fewer bits to be transmitted. For example, numbers will usually be much more compact if transmitted in binary rather than string forms.
- **Manage message latency.** In some applications, it may be possible to do other work while a message is being processed. If the delay is appreciable — and especially if the information is likely to go stale — it is important to keep the user informed of progress. Design the user interface of your applications to handle message latency appropriately.
- **Simplify the interface.** Keep the application's interface simple enough that the user seldom needs to refer to a user manual to perform a task. To do so: reduce the amount of information displayed on the device; make input sequences concise so the user can accomplish tasks with the minimum number of button clicks; and offer the user selection lists.

AD-HOC DEVELOPMENT PROCESS

An ad-hoc development process for wireless applications comprises three steps:

1. Write the application. Several Integrated Development Environments (IDEs) are available for developing Java-based wireless applications, for example, Sun's J2ME Wireless Toolkit, and Metrowerks CodeWarrior.
2. Test the application in an emulation environment. Once the application compiles nicely, it can be tested in an emulator.
3. Download the application to a physical device and test it. Once the application's performance is satisfactory on one or more emulators, it can be downloaded to a real device and tested there. If it is a network application, it is tested on a live wireless network to ensure that its performance is acceptable.

It is clear that many important software engineering activities are missing from this ad-hoc development process. For example, there is no formal requirements analysis phase, and so following an ad-hoc development process may lead to building a product different from the one customers want. Also, testing an application without knowing its requirements is not an easy task. In addition, issues related to the operating environment such as network bandwidth should be considered during the design so that the performance of the application will be satisfactory.

WIRELESS SOFTWARE ENGINEERING

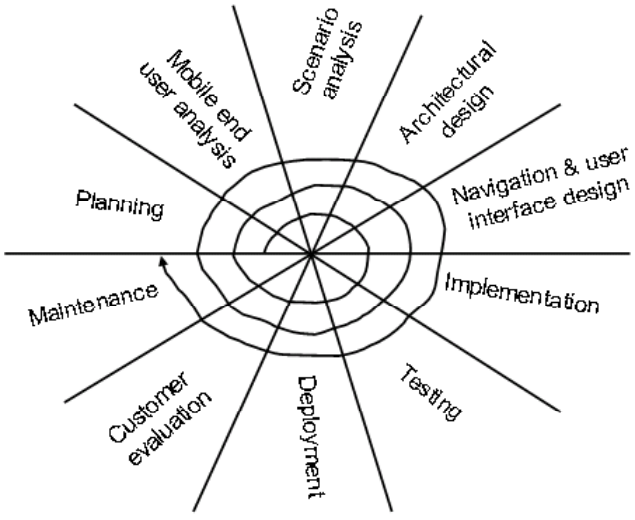
While wireless application development might appear to have less need for the coordination that a process provides, aspects of development, testing, evaluation, deployment, and maintenance of a wireless application have to be integrated in the design process throughout the full development life cycle. We have put forward a systematic approach to developing wireless applications, which is compatible with the Rational Unified Process or RUP (Jacobsen, Booch, & Rumbaugh, 2000) in the sense that it is iterative and responsibility-driven. We have developed this systematic approach based on our experience designing and building wireless applications. We recognized that the development of a wireless application is not a one-shot task, and testing wireless applications is more challenging than testing conventional desktop software applications; therefore, an ad-hoc development process cannot be used.

Development Activities

Our software engineering approach to wireless application development consists of a set of manageable activities that, if followed properly, leads to reliable and maintainable wireless applications. The activities of our approach are shown in Figure 3.

Planning. This iterative process begins with a planning phase, which is an activity that identifies the objectives of the wireless application and specifies the scope of the first increment. In addition, the costs of the overall project are estimated, the risks are evaluated, and a tentative schedule is set.

Figure 3. Wireless application development activities



Mobile user analysis. First, we must understand the audience of the application and the environment in which it will operate. As an example, if the application is a wireless network-aware application such as a multi-player game, the study will include the users of the application and how they plan to use it. The output at the end of this phase is a wireless application plan document that serves as the mobile end-user requirement.

Scenario analysis. This phase is similar to conventional software requirements analysis, and therefore concepts and principles of requirements analysis can be applied here (Pressman, 2005). In this phase, the mobile end user, an interaction designer, and a developer sit together to come up with a complete scenario analysis model that takes into account the following types of scenario analysis:

- **Screen and interaction analysis:** The basic unit of interaction between the user and the mobile device is the screen, which is an object that encapsulates device-specific graphic user input. Therefore, the content to be displayed on the screen is identified. Content may include text fields, menus, lists, and graphics. Interaction analysis specifies how the user interacts with the application. In order to find out how the user will interact with the application, UML (Booch et al., 2000) use cases are developed.
- **Usage analysis:** The use case model developed during screen and interaction analysis is mainly related to how users interact with the application through the screen. The whole functionality of the application should be described with use cases.
- **Environment analysis:** The environment in which the application will operate should be described in detail. This includes the different wireless networks and back-end systems used. In addition, target mobile devices such as cellular phones and PDAs on which the application will run should be described in detail.

The output of this phase is an information analysis model document produced by the interaction designer and developer that outlines the functional requirements of the application and the constraints of the environment. This document is reviewed by developers and other stakeholders and modified as required.

Architectural design. This phase is concerned with the overall architecture (or structure) of the wireless application. Architecture is very important for any application, and no optimization techniques will make up for an ill-considered architecture. Design patterns can be used in this phase to reuse experience in order to come up with an extensible, high-performance architecture. Some of the most important design goals should be to minimize the amount of data transmitted over the wireless link, and to anticipate errors and handle them intelligently. Other design and architecture issues include:

- **Application partitioning.** Designers need to think carefully when deciding which operations should be performed on the server and which on the wireless device. J2ME allows designers to locate much of an application's functionality on the device; it can retrieve data from the server efficiently, then perform calculations and display information locally. This approach can dramatically reduce costly interaction over the wireless link, but it is feasible only if the device can handle the processing your application needs to perform.
- **Message latency.** In some applications, it may be possible to do other work while a message is being processed. If the delay is appreciable — and especially if the information is likely to go stale — it is important to keep the user informed of progress.

The outcome of the architectural design phase is a design document that details the system architecture.

Navigation and user interface design. Once the application architecture has been established and its components identified, the interaction designer prepares screen mockups and navigation paths that show how the user moves from one screen to another to access services. Figure 4 shows a simple example where the user will have to login before she is able to check her messages.

The user interface is the face of the application to users. A poorly designed user-interface will scare the user away, and a well-designed user interface will give a good first impression and improves the user's perception of the services offered by the application. The user interface must be well-structured and easy to use. Here are some guidelines that can help in designing simple yet effective user interfaces for mobile devices with tiny screens.

- Keep the application's interface simple enough that the user seldom needs to refer to a user manual to perform a task.
- Reduce the amount of information displayed on the device.
- Make input sequences concise so the user can accomplish tasks with the minimum number of button clicks.
- Offer the user selection lists.
- Do not depend on any specific screen size.

The output of this phase is a user manual that describes the screen mockups and the navigational paths.

Implementation. In this phase development tools are used to implement the wireless application. There are several tools available for building wireless applications such as Sun's J2ME Wireless Toolkit. We would recommend using a tool that allows installing the application in various emulation environments. Conventional implementation strategies and techniques such as coding standards and code reviews can be used in this phase.

Testing. Software testing is a systemic process to find differences between the expected behavior of the system specified in the software requirements document and its observed behavior. In other words, it is an activity for finding errors in the software system and fixing them so users can be confident that they can depend on the software. Errors in software are generally introduced by people involved in software development (including analysts, architects, designers, programmers, and the testers themselves). Examples of errors include mismatch between requirements and implementation.

Many developers view the subject of software testing as "not fashionable," and, as a result, too few of them really understand the job software testers do. Testing is an iterative process and should start from the beginning of the project. Software developers need to get used to the idea

Figure 4. Screen mockups

of designing software with testing in mind. Some of the new software development methodologies such as eXtreme Programming (XP) (Beck, 1999) stress incremental development and testing. XP is ideally suited for some types of applications, depending on their size, scope, and nature. User interface design, for example, benefits highly from rapid prototyping and testing usability with actual users.

Wireless applications, like all other types of software, must be tested to ensure functionality and usability under all working conditions. Testing is even more important in the wireless world because working conditions vary a lot more than they do for most software. For example, wireless applications are developed on high-end desktop machines but deployed on handheld wireless devices with very different characteristics.

One way to make testing simple is to design applications with testing in mind. Organizing the system in a certain way can make it much easier to test. Another implication is that the system must have enough functionality and enough output information to distinguish among the system's different functional features. In our approach, and similar to many others, the system's functional requirements (features that the system must provide) are described using the Unified Modeling Language (Booch et al., 2000) to create a use-case model, then detailing the use cases in a consistent written form. Documenting the various uses of the system in this way simplifies the task of testing the system by allowing the tester to generate test scenarios from the use cases. The scenarios represent all expected paths users will traverse when they use the features that the system must provide.

Deployment. Deploying and running applications in an emulation environment is a very good way to test the logic and flow of your application generally, but you will not be certain it will satisfy users until you test it on a real physical device connected to a wireless network. Your application's performance may be stunning in the emulator, which has all the processing power and memory of your desktop machine at its command, but will it perform well on the handheld device, with its limited memory and processing power, low bandwidth, and other constraints? In this phase, the application is deployed on a live network and evaluated.

Customer Evaluation. Once the application has been deployed, it is ready to be downloaded by users for evaluation and usage. In this phase, users start using the deployed application and report any problems they may experience to the service provider.

Maintenance. Software maintenance encompasses four activities: error correction, adaptation, enhancement, and reengineering (Pressman, 2005). The application will evolve over time as errors are fixed and customers request new features. In this phase, users report errors to and request new features from the service provider, and developers fix errors and enhance the application.

TESTING ISSUES AND TESTING ACTIVITIES

The wide variety of mobile devices such as wireless phones and PDAs results in each device running a different implementation of the J2ME environment. Varying display sizes add to the complexity of the testing process. In addition, some vendors provide proprietary API extensions. As an example, some J2ME vendors may support only the HTTP protocol, which the MIDP 1.0 specification requires, while others support TCP sockets and UDP datagrams, which are optional. Here are some guidelines for testing wireless applications.

Implementation Validation. Ensuring that the application does what it is supposed to be

is an iterative process that you must go through during the implementation phase of the project. Part of the validation process can be done in an emulation environment such as the J2ME Wireless Toolkit (Sun Microsystems, 2005), which provides several phone skills and standard input mechanisms. The toolkit's emulation environment does not support all devices and platform extensions, but it allows for the application to look appealing and to offer a user-friendly interface on a wide range of devices. Once the application has been tested on an emulator, you can move on to the next step and test it on a real device, and in a live network.

Usability Testing. In usability testing, the focus is on the external interface and the relationships among the screens of the application. As an example, consider an e-mail application that supports entry and validation of a user name and password, enables the user to read, compose, and send messages, and allows maintenance of related settings, using the screens shown in Figure 3, among others.

In this example, start the test at the Login window. Enter a user name and a password and press the soft button labeled Login. Enter a valid user name and password. The application should display the main menu. Does it? The main menu should display a SignOut button. Does it? Press the SignOut button. Does the application return to the Login screen? Write yourself a note to raise the question, "Why does the user 'log' in but 'sign' out?" Now enter an invalid user name or password. The program should display a meaningful message box with an OK button. Does it? Press the OK button. Does the application return to the Login screen?

You need to test the GUI navigation of the entire system, making notes about usability along the way. If, for example, the user must traverse several screens to perform a function that's likely to be very popular, you may wish to consider moving that particular function up the screen layers. Some of the questions you should ask during usability testing include: is the navigation depth (the number of screens the user must go through) appropriate for each particular function, does the application minimize text entry (painful on a wireless phone) or should it provide more selection menus, can screens of all supported devices display the content without truncating it, and if you expect to deploy the application on foreign devices, does it support international character sets?

Network Performance Testing. The goal of this type of testing is to verify that the application performs well in the hardest of conditions (for example, when the battery is low or the phone is passing through a tunnel). Testing performance in an emulated wireless network is very important. The drawback with testing in a live wireless network is that so many factors affect the performance of the network itself that you cannot repeat the exact test scenarios. In an emulated network environment, it is easy to record the result of a test and repeat it later, after you have modified the application, to verify that the performance of the application has improved.

Server-Side Testing. It is very likely that wireless applications communicate with server-side applications. If your application communicates with servers you control, you have a free hand to test both ends of the application. If it communicates with servers beyond your control (such as quotes.yahoo.com), you just need to find the prerequisites of use and make the best of them. You can test server-side applications that communicate over HTTP connections using testing frameworks such as HttpUnit (http://httpunit.sourceforge.net), and measure a Web site's performance using httperf (http://citeseer.nj.nec.com/mosberger98httperf.html), a tool designed for measuring the performance of Web servers.

Testing Checklists

Here we provide checklists that are useful when testing your application, in both emulation and live environments. These checklists include tests that are usually performed by certification programs offered by Nokia and Motorola (Motorola Application Certification Program).

Navigation Checklist. Here are some items to check for when testing the navigational paths of wireless applications:

- **Successful startup and exit**: Verify that your application starts up properly and its entry point is consistent. Also make sure that the application exits properly.
- **Application name**: Make sure your application displays a name in the title bar.
- **Keep the user informed**: If your application does not start up within a few seconds, it should alert the user. For large applications, it is a good idea to have a progress bar.
- **Readable text**: Ensure that all kinds of content are readable on both grayscale and color devices. Also make sure the text does not contain any misspelled words.
- **Repainting screens**: Verify that screens are properly painted and that the application does not cause unnecessary screen repaints.
- **Soft buttons**: Verify that the functionality of soft buttons is consistent throughout the application. Verify that the whole layout of screens and buttons is consistent.
- **Screen navigation**: Verify that the most commonly used screens are easily accessible.
- **Portability**: Verify that the application will have the same friendly user interface on all devices it is likely to be deployed on.

Network Checklist. Some of the items that should be inspected when testing wireless applications are:

- **Sending/Receiving data:** For network-aware applications, verify that the application sends and receives data properly.
- **Name resolution:** Ensure that the application resolves IP addresses correctly, and sends and receives data properly.
- **Sensitive data:** When transmitting sensitive data over the network, verify that the data is being masked or encrypted.
- **Error handling:** Make sure that error messages concerning network error conditions (such as no network coverage) are displayed properly, and that when an error message box is dismissed, the application regains control.
- **Interruptions:** Verify that, when the device receives system alerts, SMS messages, and so on while the application is running, messages are properly displayed. Also make sure that when the message box is dismissed the application continues to function properly.

PROVISIONING WIRELESS APPLICATIONS

Developers usually build, test, and evaluate an application on a platform similar to the one on which it will be deployed and ran. Development of wireless applications is more challenging because they typically are developed on one platform (such as Solaris or MS Windows) but deployed on a totally different one (such as a cell phone or PDA). One consequence is that, while emulators enable developers to do some of their testing on the development platform, ultimately they must test and evaluate the application in the very different environment of a live wireless network.

Wireless applications fall into two broad categories:

- **Local applications** perform all their operations on a handheld wireless device and need no access to external data sources through a

wireless network. Examples include calculators and single-player games.

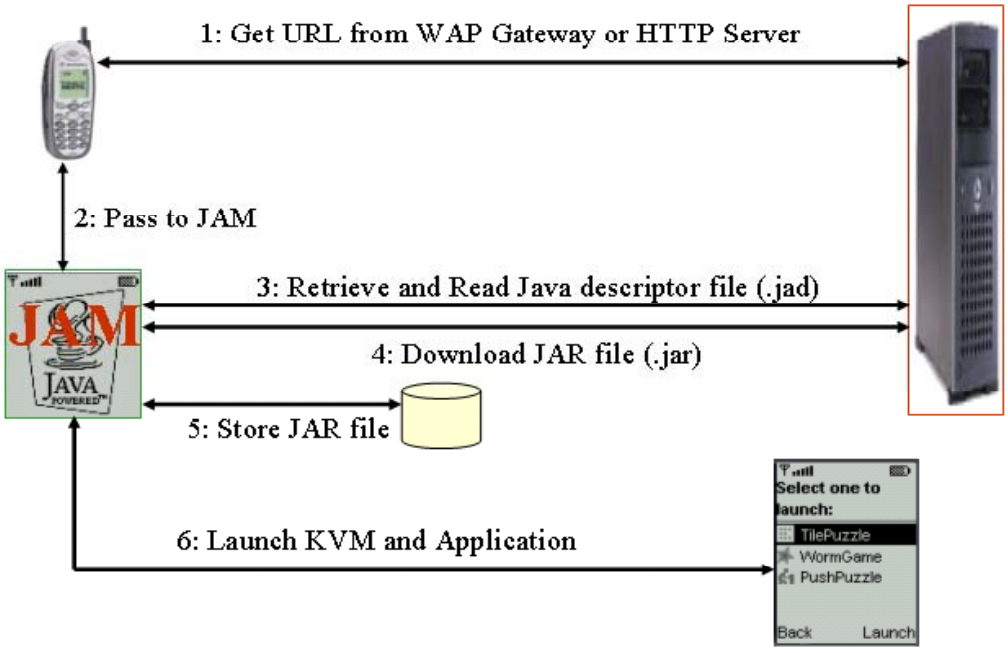
- **Network applications** consist of some components running on a wireless device and others running on a network, and thus depend on access to external resources. An example would be an e-mail application, with a client residing on a wireless phone that interacts with an SMTP server to send messages.

Although these two types of applications are different, they are deployed in the same way. The big difference shows up later: Local applications are easier to test than network applications. For example, a calculator application can run on a wireless phone even when it is not connected to any network, but an e-mail client won't work without a connection to the SMTP server that actually transmits the messages.

Over the Air Provisioning

For some time, wireless portals in Europe such as Midletcentral have allowed customers to download applications directly to their phones, over the air. Over-the-air provisioning of wireless applications (OTA) is finally available in North America. Nextel customers, for example, can download network-aware wireless applications without an update data cable. OTA is the deployment of wireless Java applications (*MIDlet suites*) from the Internet to wireless devices over a wireless network. Users need not connect their devices to the desktop with a data cable or visit a service center to install or upgrade software. To take advantage of OTA, you must equip your handheld device with a mechanism to discover MIDlet suites available for download, using the device's browser (such as a WAP browser) or a resident application written specifically to identify downloadable MIDlet suites. The process of downloading MIDlets over the air is illustrated in Figure 5.

Figure 5. Over-the-air provisioning



RELATED WORK

The explosive growth of the wireless mobile application market raises new engineering challenges (Morisio & Oivo, 2003); what is the impact of the wireless Internet on engineering wireless mobile applications for the new wireless infrastructure and wireless handheld devices? Due to the limited experience with wireless technologies and developing wireless applications, little work has been in the area of wireless software engineering. We found a special issue in the *IEEE Transactions on Software Engineering* on "Software Engineering for the Wireless Internet" (Morisio & Oivo, 2003). However, out of the six papers accepted in the special issue only two papers deal with the development process. Ocampo, Boggio, Munch, and Palladino (2003) provided an initial reference process for developing wireless Internet applications, which does not differ significantly from traditional iterative process models but includes domain-specific guidance on the level of engineering processes. Satoh (2003) developed a framework for building and testing networked applications for mobile computing. The framework is aimed to emulate the physical mobility of portable computing devices through the logical mobility of applications designed to run on them; an agent-based emulator is used to perform application-level emulation of its target device.

More recently, Chen (2004) proposed a methodology to help enterprises develop business strategies and architectures for mobile applications. It is an attempt to formulate a life cycle approach to assisting enterprises in planning and developing enterprise-wide mobile strategies and applications. This methodology is more concerned with business strategies rather than technical details, and thus it is targeted at managers rather than developers. And finally, Nikkanen (2004) presented the development work of a browser-agnostic mobile e-mail application. It reports on experiences porting a legacy WAP product to a new XHTML-based browser application and offers guidelines for developing mobile applications.

Our work is different in the sense that we provide a detailed treatment of the impact of the characteristics of mobile devices and the wireless environment on engineering wireless mobile applications; we discuss the challenges and offer practical solutions for developing mobile applications. We present a systematic approach for designing wireless mobile application. Our approach is iterative just like in Ocampo et al. (2003), but differs in the sense that our process has more focus on requirements elicitation and more importantly scenario analysis. We do not provide a testing framework, but our testing strategy and checklist is more practical than using just an emulated environment. Finally, unlike the work reported in Chen (2004), our methodology is targeted at developers and researchers rather than managers. And, unlike the work in Nikkanen (2004), our guidelines and systematic approach is not limited to WAP-based applications, but can be applied to engineering any wireless application.

CONCLUSION AND FUTURE WORK

As the wireless Internet becomes a reality and software developers become comfortable with the methods and processes required to build software, we recognize that the methods developed for conventional systems are not optimal for wireless applications. In particular, wireless application development doesn't always fit into the development model originated to cope with conventional large software systems. Most wireless application systems will be smaller than any medium-size project; however, a software development method can be just as critical to a small software project success as it is to that of a large one. In this paper, we have presented and discussed a systematic approach to wireless application development, and presented practi-

cal guidelines for testing wireless applications. The proposed approach takes into account the special features of the wireless environment. We have successfully used the approach presented to develop various wireless applications ranging from a stock portfolio management application to a mobile agent platform for mobile devices (Mahmoud, 2002). Our future work includes evaluating the effectiveness of the proposed methodology, documenting wireless software design patterns, and building tools to automate the task of testing wireless applications.

There are several interesting research problems in the emerging area of wireless mobile applications and services. Some of these research issues include: novel mobile services in the area of m-commerce and health care; security and privacy issues; mobile agents for mobile services; discovery and interaction of mobile services; enabling roaming of applications and profiles between different wireless standards; and location-aware and context-aware mobile services. We are currently addressing some of these research problems, and research results will be presented in future articles.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for the many helpful suggestions for improving this paper. The first author was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant No. 045635.

REFERENCES

Beaulieu, M. (2002). *Wireless Internet applications and architecture*. Boston: Addison-Wesley.

Beck, K. (1999). *Extreme programming explained: Embrace change*. Addison-Wesley.

Booch, G., Rumbaugh, J., & Jacobsen, I. (2000). *The Unified Modeling Language user guide*. Boston: Addison-Wesley.

Burkhardt, J, Henn, H., Hepper, S., Rintdorff, K., & Schack, T. (2002). *Pervasive computing technology and architecture of mobile Internet applications*. London: Addison-Wesley.

Chen, M. (2004). A methodology for building mobile computing applications. *International Journal of Electronic Business, 2*(3), 229-243.

Jacobsen, I., Booch, G., & Rumbaugh, J. (2000). *The unified software development process*. Boston: Addison-Wesley.

Httperf. Retrieved January 13, 2005, from http://www.hpl.hp.com/research/linux/httperf

HttpUnit. Retrieved January 13, 2005, from http://httpunit.sourceforge.net

Mahmoud, Q. (2002). MobiAgent: An agent-based approach to the wireless Internet. *Journal of Internet Computing, special issue on Wireless Internet, 3*(2), 157-162.

Morisio, M., & Oivo, M. (2003). Software engineering for the wireless Internet [Guest Editor's Introduction]. *IEEE Transactions on Software Engineering, 29*(12), 1057-1058.

Motorola Application Certification Program. (n.d.). Retrieved February 10, 2005, from http://qpqa.com/motorola/iden

Nikkanen, M. (2004). User-centered development of a browser-agnostic mobile e-mail application. In *Proceedings of the Third Nordic Conference on Human-Computer Interaction*, Tampere, Finland (pp. 53-56). New York: ACM Press.

Ocampo, A., Boggio, D., Munch, J., & Palladino, G. (2003). Towards a reference process for develop-

ing wireless Internet services. *IEEE Transactions on Software Engineering, 29*(12), 1122-1134.

Open Mobile Alliance. (2005). Retrieved from March 15, 2005, http://www.openmobilealliance.org

Pressman, R. S. (2005). *Software engineering: A practitioner's approach* (6th ed.). New York: McGraw Hill.

Satoh, I. (2003). A testing framework for mobile computing software. *IEEE Transactions on Software Engineering, 29*(12), 1112-1121.

Sun Microsystems J2ME. (2005). Retrieved from http://java.sun.com/j2me

Sun Microsystems J2ME Wireless Toolkit. (2005). Retrieved from http://java.sun.com/products/j2mewtoolkit

ENDNOTE

[1] We use the terms wireless application and mobile application interchangeably throughout this article.

This work was previously published in International Journal of Information Technology and Web Engineering, Vol. 1, Issue 1, edited by G. Alkhatib and D. Rine, pp. 59-75, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.3
Conceptual Framework for Mobile-Based Application in Healthcare

Matthew W. Guah
School of Business Economics, Erasmus University Rotterdam, The Netherlands

ABSTRACT

The significance of aligning IT with corporate strategy is widely recognised, but the lack of an appropriate framework often prevents practitioners from integrating emerging Internet technologies (like Web services and mobile technologies) within organisations' strategies effectively. This chapter introduces a framework that addresses the issue of deploying Web services strategically within a mobile-based healthcare setting. A framework is developed to match potential benefits of Web services with corporate strategy in four business dimensions: innovation, internal healthcare process, patients' pathway, and management of the healthcare institution. The author argues that the strategic benefits of implementing Web services in a healthcare organisation can only be realized if the Web-services initiatives are planned and implemented within the framework of an IT strategy that is designed to support the business strategy of that healthcare organisation.The chapter will use case studies to answer several questions relating to wireless and mobile technologies and how they offer vast opportunity to enhance Web services. It also investigates what challenges are faced if this solution is to be delivered successfully in healthcare. The healthcare industry globally, with specific emphasis on the USA and United Kingdom, has been extremely slow in adopting emerging technologies that focus on better practice management and administrative needs. The chapter elaborates on certain emerging information technologies that are currently available to aid the smooth process of implementing mobile-based technologies into healthcare industry.

INTRODUCTION

This chapter is based on research—using a longitudinal case study—into the National Programme for Information Technology (NPfIT). NPfIT is an initiative that has been budgeted to cost the UK government £6.3 billion for the purpose of improving the information systems in the National Health Service (NHS), with emphasis on IT infrastructure and the creation of a nationwide patient database.

The significance of aligning IT with corporate strategy in healthcare organisations is widely recognised, but the lack of an appropriate framework often prevents medical practitioners from integrating emerging Internet technologies (like Web services and mobile technologies) within healthcare organisations' strategies effectively. This chapter introduces a framework that addresses the issue of deploying Web services strategically within a mobile-based healthcare setting. A framework is developed to match potential benefits of Web services with corporate strategy in four business dimensions: innovation, internal healthcare process, patients' pathway, and the management of healthcare institution. The author argues that the strategic benefits of implementing Web services in a healthcare organisation can only be realized if the Web-services initiatives are planned and implemented within the framework of an IT strategy that is designed to support the business strategy of that healthcare organisation.

The chapter will also consider certain essential issues regarding the deployment of any mobile data solution (i.e., reliability, efficiency, and security) in the healthcare industry and how such deployment can support healthcare professionals in saving patients' lives. Using case studies, the chapter will answer the following questions:

- Wireless and mobile technologies offer vast opportunities to enhance services, but what challenges are faced if this solution is to be delivered successfully in healthcare?
- Why has the global healthcare industry, with specific emphasis on the USA and United Kingdom, been extremely slow in adopting technologies that focus on better practice management and administrative needs?
- How complacent can IS strategists be to the productivity paradox in the wake of HIPAA (Health Insurance Portability and Accountability Act in USA) and NPfIT (in UK)?
- What emerging information technologies are there to aid the smooth process of implementing mobile-based technologies into the healthcare industry?

The existing economics and IS literature on information-technology adoption often considers network externalities as one of the main factors that affect adoption decisions (Brown & Venkatesh, 2003). It is generally assumed that potential adopters achieve a certain level of expectations about network externalities when they have to decide whether to adopt a particular technology. However, there has been little discussion on how the potential adopters reach their expectations. This chapter attempts to fill a gap in the literature on the adoption of mobile healthcare technology by offering an optimal control perspective motivated by the rational expectations hypothesis and exploring the process dynamics associated with the actions of decision makers in the healthcare industry. They must adjust their expectations about the benefits of a mobile healthcare technology over time due to bounded rationality. The model posed in this chapter addresses mobile healthcare technologies that exhibit strong network externalities. It stresses adaptive learning to show why different healthcare organisations that initially have heterogeneous expectations about the potential value of a mobile healthcare technology eventually are able to arrive at contemporaneous decisions to adopt the same technology, creating

the desired network externalities. This further allows these organisations to become catalysts to facilitate processes that lead to healthcare industry-wide adoption.

BACKGROUND

The NHS has been responsible for the provision of healthcare and services in the United Kingdom for the past 56 years on the basis of being free for all at the point of delivery. The traditional perception of the NHS is one of a healthcare system organised as a professional guild, with unlimited finance from the government. This type of NHS is experiencing an irrevocable change as taxpayers are no longer complaisant and paternalistic employers are reacting against inflating costs and escalating complaints from the patients. The employer is reacting to the continuous massive flow of subsidies for inefficient physician practices, fragmented delivery systems, and cost-unconscious consumer demand. The patients are increasingly assertive as to their preferences and few have expressed their willingness to make additional contributions for particular health benefits and medical interventions.

Web services are technologies with roots in the Application Service Provision (ASP) business model that are used mostly to automate linkages among applications (Hagel, 2002). They are generally anticipated to make critical system connections not only possible but also easy and cheap (Kreger, 2003; Sleeper & Robins, 2001). One of the perceived benefits of Web services is that organisations would be able to concentrate on their core competencies (Perseid Software Limited, 2003). Service providers argued that the remote delivery of software applications would release managers from the perennial problems of running in-house IT departments, allowing more time to develop IT and e-business strategy rather than the day-to-day operations (Currie, Desai, & Kahn, 2004). This justification has been used in traditional forms of outsourcing over many years (Willcocks & Lacity, 1998).

The NHS is experiencing massive changes in the structure of information systems provision markets and organisations. The local service provision (LSP) and national service provision (NSP) models in use by the NPfIT are in a state of ferment. The payment methods borrow from both capitation and fee for service, and methods of utilisation management are compromised between arm's-length review and full delegation (Guah & Currie, 2006). LSP and NSP consist of large and more complex entities. These are the result of merger, acquisition, and product diversification. The service providers involved have had to take on a visible feature of ceaseless acquisition and divestiture, integration and outsourcing, and combination and recombination. Providers of medical systems, hospital administration systems, and health plans are coming together and then coming apart. They are substituting contracts for joint ownership, creating diversified conglomerates and refocused facilities, and experimenting with ever-new structures of ownership, finance, governance, and management (Robinson, 2000). These would give the NHSIA (National Health Service Information Authority) the benefits not only of a middle ground between the extremes of vertical integration and spot contracting, but also a balance of coordinated and autonomous adaptation in the face of its ever-new challenges.

The general assumption is that expenditures in the nation's health will outpace the overall growth in the economy (Collins, 2003; Pencheon, 1998). This is reflected in the percentage of the GDP (gross domestic product) of USA (13%), Germany (10.7%), France (9.6%), and the United Kingdom (7.6%) being devoted to the total cost of healthcare resources (Brown, 2002). Unlike the United Kingdom, however, some of these countries are faced with limitations in social willingness to pay. It has been documented that millions of U.S. residents currently lack the most basic insurance coverage (Institute of Medicine, 2002).

Response to Emerging Technologies in the NHS

Over the years, nontechnologists in the NHS have managed to muddle through one powerful new system after another. Generational strategy is one continuously being used to deal with some of the pressures induced by IS. Adopting such innovations as PCs (personal computers) and the Internet requires the personal and organisational costs of unfreezing deeply ingrained old habits. Many workforces ignore, deny, or deal awkwardly with such technologies.

Srinivasan, Lilien, and Rangaswamy (2004) found several reasons why an organisation should respond to new technology development. Two major reasons are listed below:

- Technological change is a principal driver of competition. This is principally because it destroys monopolies, creates new industries, and renders products and markets obsolete.
- Additional sources—both within and outside the organisation or industry—are increasingly complementing in-house technology development efforts.

A common response to new systems is the "not invented here" (NIH) syndrome (Collins, 2003; Guah & Currie, 2004; Haines, 2002). This often leads to certain organisations rejecting a perfectly useful system based on an implicit assumption that the system does not fully recognise or accommodate their own needs and idiosyncrasies (Brown & Venkatesh, 2003; Davis, 1989). Davis sees this as a likely result of a decline in communication with external sources. NIH syndrome could also result from competences that can be proven to be outdated and inefficient in comparison to an existing technology. One Trust, which places a central role in the direction of regional IS strategy, had to reject a system promoted by the Department of Health because the system was not as familiar as another bespoke system (Haines).

The common characteristics of new systems in the NHS are uniformity in products and prices in the face of great variability in consumer preferences and the actual costs of providing service (Collins, 2003). This one-size-fits-all approach usually leads to services that are of excessive costs for some users and insufficient quality for others, impeding the use of price flexibility to enhance capacity utilisation (Robinson, 2000). Also of concern is a combination of overcapacity and low load factors in some regional trusts with undercapacity and shortages elsewhere. Concerns are growing in the NHS that this may generate cross-subsidies from trusts for which the cost of service will be low to trusts for which the cost of service will be high (McGauran, 2002). Additionally, deregulation of healthcare costs has spurred an outpouring of new services. Consequently, several of these services are the following (Collins; Pencheon, 1998):

- A different cost structure.
- An impact on IS budgets.
- A better match between supply and demand.

Incomplete information has been a fascinating attribute of the NHS's unusual system's organisational and normative characteristics. The asymmetry of NHS information between patients and medical practitioners has changed in an exogenous fashion over its 56 years. The amount of healthcare information available to patients is usually the result rather than the cause of changes in the economic and political environment (Robinson, 2000).

PROJECT DESCRIPTION: NATIONAL PROGRAMME FOR INFORMATION TECHNOLOGY

The NPfIT is an initiative by the National Health Service Information Authority, born as a result of several plans to devise a workable IS strategy

for the NHS (NHSIA, 2003; Wanless, 2002). The NPfIT was designed to connect the capabilities of modern IT to the delivery of the NHS plan devised in 1998. The core of this strategy is to take greater control of the specification, procurement, resource management, performance management, and delivery of the information and IT agenda (NHSIA).

The NPfIT is an essential element in delivering the NHS plan. It has created £6 billion information infrastructure, which could improve patient care by increasing the efficiency and effectiveness of clinicians and other NHS staff. The intention of the plan is to address the following (http://www.npfit.nhs.uk):

- Create an NHS Care Records Service to improve the sharing of consenting patients' records across the NHS.
- Make it easier and faster for GPs (general practitioners) and other primary care staff to book hospital appointments for patients.
- Provide a system for electronic transmission of prescriptions.
- Ensure that the IT infrastructure can meet NHS needs now and in the future.

The decision to implement a national programme for IT into the NHS system complexity is only the first step in the IS modernisation journey for a multifaceted organisation. There are many examples of new technologies disrupting organisational routines and relationships in the NHS (Atkinson & Peel, 1998; Majeed, 2003; Metters, Abrams, Greenfield, Parmar, & Venn, 1997). These usually require both medical professionals and NHS regional trusts managers to relearn how to work together. Orlikowski (1993) and Edmondson (2003) suggest that one technology can be seen differently by two groups of people in an organisation. Findings from Barney and Griffin (1992) and Orlikowski have showed how this could result in the elicitation of different responses for members of that organisation.

Scope of Project Work

The chapter takes a more in-depth look at the role of the NHSIA (seen as the project leader for the NPfIT), currently the most visible spokesperson and translator for the potential implications of the resulting new technologies. Research has shown NHS IS staff to pay particular attention to what the NHSIA says and does in regard to information systems (Collins, 2003; Ferlie & Shortell, 2001). This research builds on a framework that identifies the key dimensions of the NHSIA tactic that is situation specific for NPfIT assumptions. The work looks at the NHSIA goal and roles for the NPfIT, as well as the role of the private-sector service providers in the implementation of NPfIT.

Here are a few objectives the NPfIT hopes to accomplish:

- To have a series of tightly specified and priced framework contracts on a short list (of about five) primary service providers (PSPs) who can work at the regional and local Strategic Health Authority (StHA) level. This should enforce the integration and implementation partnership—at a national level—during all aspects of the NPfIT project. Each PSP will have an aligned consortium of service providers and vendors for the integrated care resource service element of the NPfIT, and will be mandated to work with the domain PSP for electronic booking, the infrastructure providers, and healthcare providers. StHA PSPs may not make their products exclusive or mandatory to their StHA.
- To create priced packages of national services and applications that the PSPs and StHAs can together implement locally. This activity will include managing the creation of a single Human Resource Information Systems (HRIS) and other national services to access and move health-record information as required.

- To create service-level agreements for the national services and other services out of the scope of the PSP consortium that the PSPs can work toward in providing an integrated service to the StHA
- To develop and maintain the national standards and specifications that all vendors must use. It is also anticipated to create the national business cases required for the Department of Health governance (required by the National Treasury), and to support the local decision-making business cases required at the StHA level.
- To procure, under national contracts, a backbone network infrastructure

Such an arrangement provides the greatest clarity in respect to the appropriate allocation of responsibilities and should be well understood in the public and private sectors (see Table 1). Services will be procured on a long-term basis so the combination of local and central funding will be required for at least 5, and preferably 10, years at guaranteed levels.

THE RESEARCH STUDY

This study intended to address the gap in the existing literature with regard to the complex issues surrounding the adoption of mobile technology in

Table 1. PSP implementation timetable (as of July 2002)

Activity/Output	***Target Date***
Agree on procurement strategy (Department of Health (DoH) & local health authorities)	End Jul 2002
Service requirement finalized and approved	End Sep 2002
Outline business case developed and approved	End Sep 2002
Official Journal of European Communities (OJEC) advert	Oct 2002
Procurement of systems and implementation services for electronic booking begins	Oct 2002
National long list of PSPs created	Dec 2002
Invitation to negotiate issued	Jan 2003
National short list of PSPs created	Apr 2003
First local health authorities begin detailed planning with PSPs	Aug 2003
PSP framework contract finalized	Oct 2003
Infrastructure provider(s) contract agreed	Oct 2003
First local health authorities begin implementation	Nov 2003
Infrastructure migration begins	Mar 2004

healthcare. I define mobile healthcare as the use of all kinds of wireless devices (cell phones, personal digital assistants, mobile e-mail devices, handheld computers, etc.) to provide health information and patient-care records to healthcare practitioners, patients, and their caregivers, employers, and employees of health service providers and public regulars of healthcare and services.

The findings reported in this chapter are part of a larger 5-year research study that was developed to investigate the deployment, hosting, and integration of the ASP and Web-services technologies from both a supply-side and demand-side perspective. The overall research was in two phases. The first phase, comprising of a pilot study, was conducted in the USA and United Kingdom (Currie et al., 2004). An exploratory-descriptive case-study methodology (Yin, 1994) was used to investigate 28 ASP vendors and seven customer sites in the United Kingdom. The dual focus upon supply side and demand side was critical for obtaining a balanced view between vendor aspirations about the value of their business models and customer experiences, which may suggest a less optimistic picture. The unit of analysis was the business model (Amit & Zott, 2001), not the firm or industry level, so a case-study methodology was anticipated to provide a rich data set for analysing firm activities and behaviour (Currie et al.).

The result from the pilot study led to the funding of two additional research studies by the Engineering and Physical Sciences Research Council (EPSRC) and Economic and Social Research Council (ESRC) respectively. Industrial collaborators were selected for the roles of technology partners, service providers, and potential or existing customers. These studies were concerned with identifying sources of value creation from the ASP business model and Web-service technologies in different vertical sectors (including health).

Research Methodology

The research followed a number of stages involving the use of both qualitative and quantitative data-collection techniques and approaches (Walsham, 1993). A questionnaire survey was distributed by e-mail to businesses and healthcare organisations all over the United Kingdom. These organisations were listed on a national database maintained by the NHSIA, plus those maintained by the university. To ensure relevant managers and practitioners responded, the covering note clearly stated the purpose of the questionnaire and requested that it be passed on to the person(s) with responsibility for managing healthcare e-business strategy. Scales to address the research questions were not available from the literature, so the questionnaire was developed based on the theory of strategic value (Banker & Kauffman, 1988). It included a checklist, open-ended questions, and a section seeking organisational data. Research questions under Part 1 required respondents to answer yes or no if the application of Internet technologies in healthcare were bringing value to patients. Data in Part 2 of the questionnaire were collected by open-ended questions seeking respondents' views on the best approach to healthcare performance improvement and Web-service value creation. This line of questioning was used to increase the reliability of data since all respondents were asked the same questions, but some added additional information. The purpose was to impose uniformity across the sample of representation rather than to replicate the data obtained from each participant (Yin, 1994).

PATIENT-INFORMATION MANAGEMENT

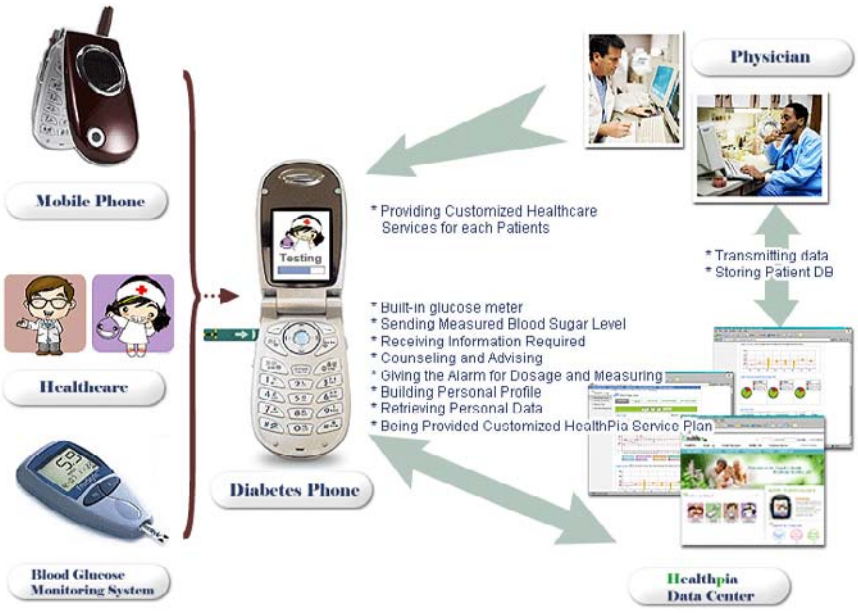
Healthcare organisations are showing a clear interest in accelerating the transformation of clinical care through the routine use of appropriate

emerging technologies by clinicians when diagnosing problems and subsequently planning and administering patient care. To support such noble efforts toward delivering better healthcare to the public, President George Bush and other national leaders have publicly called for the development of a national health information infrastructure. The U.S. government (in 2005) has over the past 2 years published plans for all Americans to have an electronic health record by 2014. Similar to the NPfIT in the United Kingdom, the plans call for a hierarchical set of local, regional, state, and national networks that facilitate peer-to-peer sharing of patient records.

When considering the deployment of any mobile data solution, reliability, efficiency, and security are essential, none more so than in the emergency services if lives are to be saved. To support such communication of critical and personal information, there has been an increased demand for the creation of electronic methods for storing and tracking clinical information (see Figure 1). This requires the solution for some fundamental architectural problems within the healthcare environment: scalability, reliability, recoverability, interchangeable vocabularies, and integration:

- Most service providers can support several thousands of simultaneous log-ons. Many are finding it difficult to demonstrate scalability in thin-client, rules-based order entry or structured clinical information.
- Service providers need to show evidence that they have appropriate schedules for downtime because healthcare organisations require reliability of 99.999% due to the critical nature of the information in the healthcare industry.
- There has to be a recognisable solution for a very quick data-recoverable plan in the event that downtime or a system failure occurs. Healthcare organisations have to ensure there are fault-proof backup plans to provide medical practitioners with information that is only available in an electronic form in the event that systems become suddenly unavailable. Certain work processes in the

Figure 1. Flexible and independent patient care (http://www.healthpia.us/services.asp)



healthcare industry (including emergency care, scheduling, registration, order entry, and clinical procedure recording) would need to continue seamlessly, even with a primary-system interruption (see Figure 1). After there has been an interruption, recovery must be complete with no loss of information. Backup, therefore, must prevent any IT failure from making care to the patient impossible. Where mobile healthcare technology is involved, adequate hardware, infrastructure, and tested processes should exist as part of a complete implementation to guarantee this recoverability.

Due to a previous lack of harmonized acquisition, healthcare organisations in both the United Kingdom and USA are frequently challenged by a variety of code sets and files that have proliferated across various healthcare institutions. HIPAA attachment transactions (in the United States) and the NPfIT (in the United Kingdom) are beginning to dictate that the future exchange of patient information be carried out electronically between healthcare organisations. To facilitate this portable and interoperable mobile healthcare technology, certain local vocabularies need to be replaced and the use of government-specified code sets should be synchronized. The way forward is to maintain current systems and historic data through mapping infrastructures that manage the correct translation, giving semantic meaning to patient data with the hope that one day soon there will be a complete migration to common vocabularies.

Many applications currently in the NHS have been designed with the assumption that that the approach and architecture does not need to coexist and interoperate. While some of these may support integration with other applications that also have significantly different philosophical stances, they do not fully recognise that the need for a healthcare organisation to implement a total solution involves the practicalities of many different dimensions of time, scope, economics, and service providers' organisational politics. Toward this goal, all interoperability for a mobile healthcare technology should require that all features and functions work across all applications. The NPfIT project is proving that all service providers have to significantly alter their current approach to internal and external integration, security, and nomenclatures during the life of the project.

Case 1: MotoHealth

Motorola, along with its partners, initiated a telemedicine service at Harvard Teaching Hospital called MotoHealth (http://www.motorola.com/mediacenter/news/detial) late 2004. The Motorola solution uses mobile phones to help healthcare professionals to monitor chronically ill patients during their normal daily routines. This product was designed to meet the customer's convenience, and as a discreet way of monitoring patients in the mobile environment, can replace in-home hospital and home monitoring devices. As a result, it gives the patient more independence to continue daily activities virtually anywhere they like. This method to providing healthcare pushes healthcare and services out of high-cost health institutions. It enables the patient's body to become the point of care and the mobile healthcare technology becomes the bridge to the patient's body, thus, enabling the delivery of care, educational advice, and support remotely and transparently.

This case has proven that when a mobile healthcare technology is implemented as part of a comprehensive healthcare program, it can give healthcare providers useful daily updates on a patient's physiological levels such as blood pressure, glucose level, and weight control. Such a method of healthcare facilitates proactive treatment action, resulting in fewer hospitalizations and visits to emergency rooms, potentially lowering the increasing demand on the costs of providing healthcare and services to the public.

Policy Issues

Arguably, the most viable techniques for successful mobile healthcare technology implementation are practical guidelines and good management practices. Policies established by a healthcare organisation are the first steps toward this goal. There are, however, few steps for establishing a policy for mobile healthcare technology:

Guidelines must be developed for the acquisition of mobile healthcare technologies. This would balance the need to encourage innovative applications against wasteful spending, which can be seen by certain members of the staff as duplication of effort. This in turn makes it the responsibility of every medical practitioner who may need a particular type of application to strictly adhere to this policy.

There must be regular inventory taken. This helps to identify and evaluate all installed hardware and software before setting or changing the standards. These would certainly affect policy decisions and future acquisitions. While standards can sometimes be looked upon as restrictive in the IT industry, medical practitioners actually see these to offer benefits for the care providers and the patients alike. Nearly all healthcare organisations have standards that cover many aspects of their operations within the healthcare process. Generally, standards in the NHS are recorded in formal standards and procedures manuals, but in certain cases, we came across informal handwritten notes (i.e., "this is the way we do things here") that are also considered to be standards.

Ruyter, Wetzels, and Kleijnen (2001) show how organisations implementing e-commerce first learn to exploit the Internet for information transfer before supporting transactions, and then finally use it for commercial trading and collaboration among various actors. Considering mobile healthcare is still in its infancy, borrowing from the e-commerce experience will mean that healthcare organisations will probably adopt wireless e-business methodologies first to support their existing healthcare processes and improve efficiency before they come up with new business models to transform the competitive landscape in the healthcare industry.

In the case of the NHS, wireless enterprise implementation issues frequently extend well beyond the technology domain and into areas of medical practices and organisational culture (see Figure 2). Nearly all the regional healthcare trusts that are actively pursuing wireless enterprise strategies at the moment are handcrafting solutions around their own local IT infrastructures

Figure 2. Home visits and general-practice consultation (http://www.bmj.com, 2004)

and their own homegrown healthcare processes partly because there are very few packaged mobile healthcare solutions on the market.

The focus, currently, is on accessing information via wireless mobile healthcare messaging. However, the future should hold more applications like mobile access, telemedicine, and alerts for facilitating better disease management and controls. Given the emerging state of mobile technology and its potential impact on healthcare, mobile healthcare can be seen as truly radical. Mobile healthcare has the potential to remake this entire industry and obsolete established strategies. Most healthcare organisations in Western Europe and North America feel they must participate in this emerging healthcare technology in order to survive the increasing demand to service a continuously evolving patient environment.

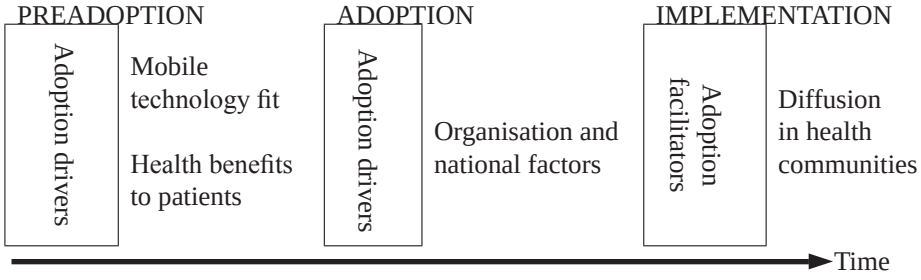
The research found two reasons why healthcare organisations are beginning to pay keen attention to mobile healthcare:

- These organisations are being defensive. There is a general belief that newcomers in the healthcare-provision market may be plotting to use new functionalities available through the use of mobile healthcare to attack the incumbents' core providers.
- Converse to the previous reason is the understanding that mobile healthcare could realize its potential. If this happens successfully, mobile healthcare would be too attractive a proposition to ignore, and joining in at a later date may prove too expensive.

The author poses a conceptual framework for the research upon which this chapter is based, showing the different stages of the technology-adoption process for healthcare organisations as well as the main factors operating at each stage (see Figure 3). In the preadoption stage, healthcare organisations take an internal perspective and analyse the fitness of the mobile healthcare technology for the contemplated task as well as the value of this technology. These should be the drivers of the adoption decision. The phase after that shows the healthcare organisation analysing whether organisational and environmental factors are favourable for continuing with this novel technology. This may uncover inhibitors that could slow down the adoption process. In situations where the healthcare organisation decides to implement the mobile healthcare technology in the next stage, it should find adoption facilitators that can help in the diffusion of mobile technology within the healthcare environment. Readers should note that implementation is beyond the scope of this framework.

This framework borrows from the technology-organisation-environment framework put forward by Tornatzky and Fleischer (1990) as well as the deinstitutionalisation framework by Tolbert and Zucker (1994). Tornatzky and Fleischer's framework identifies three aspects of an organisation's

Figure 3. Conceptual framework of mobile healthcare technology adoption



context that influence the process by which it adopts and implements innovation: technological context, organisational context, and environmental context. The omission of environmental context here is due to the fact that this chapter is about a single industry in which the environment is held constant.

The author considers innovation in healthcare as an idea, medical practice, or any material artefact in the healthcare process that is perceived to be new by the relevant unit of adoption in medical treatment. The relative advantage, compatibility, complexity, trial, and observation of such innovation can usually be used to determine the tendency for its adoption in the healthcare environment.

By relative advantage, the author means the degree to which an innovation within the health industry is perceived as better than the idea it supersedes. Compatibility, on the other hand, is seen as the degree to which an innovation within the health industry is perceived as being consistent with the existing values given by a particular healthcare community, or with past experiences and needs of potential adopters within the healthcare process. Compatibility of mobile healthcare technology can also be explained in terms of a combination of what healthcare practitioners feel or think about a particular innovation. This would also involve a critical look at the practical and operational compatibility with what healthcare practitioners are doing in the ongoing healthcare process. This interpretation of compatible innovation is in conjunction with perceived usefulness, the degree to which an end user believes a certain system can help perform a certain medical task. Complexity is the degree to which an innovation within the health industry is seen to be difficult to understand and use within the healthcare industry.

The trial of an innovation within the healthcare industry is an important part of evaluating new technologies within this critical industry. It is the degree to which an innovation within the health industry is experimented with on a limited basis. Given an opportunity to experiment with a new mobile healthcare technology before decisions are made about the adoption is an important benefit, especially for emerging technologies. This is an industry where practitioners take very highly the availability of information, while learning from experiences with previously disappointing IT projects.

Observation is a reliable means by which the healthcare industry evaluates innovations. This process identifies the degree to which the performance of a mobile healthcare technology and related benefits to the patients are visible to the medical practitioners and not only the service providers.

The determinants of mobile healthcare technology adoption are the benefits to the patients vs. the cost of such adoption. Most often, the NHS measures this in terms of the difference in costs for the shift from a previous technology to a mobile healthcare technology. Also worth considering are several factors that are important to the health service, such as the improvements made to the healthcare process as a result of a mobile healthcare technology after its introduction. There might even be a discovery of new uses for the mobile healthcare technology and the development of certain complementary inputs.

Hartmann and Sifonis (2000) relates to some of these features of mobile healthcare technology application by the identification of four dimensions within an organisation: leadership, governance, technology, and operational competencies. By leadership, they referred to the process of managing the initiatives and how the host organisation should stay motivated throughout the adoption process. By governance, they referred to the process of organising the innovation as regards the structure and operating procedures. Technology is where Hartmann and Sifonis looked at the organisation's ability to rapidly develop and implement new applications. They finally explained operational competencies as the way the host organisation manages the coordination of

the relationship between leadership, governance, and technology as well as exploiting the available resources.

Levy and Powell (2005), on the other hand, presented evidence—from their study of small and medium-sized businesses in the United Kingdom—that the adoption of emerging technologies posits a similar framework as adoption related to the readiness of organisations taking into consideration the perceived ease of use and perceived usefulness. The readiness of the NHS to adopt mobile healthcare technology can be determined by reviewing the financial and technological resources available as well as various other factors dealing with the compatibility and consistency of emerging technologies with organisational culture and values.

Case 2: Pervasive Monitoring System

Oracle, along with its London partners, piloted a wireless sensor interface technology platform in mid-2005 (http://www.toumaz.com/news.php?act). It used advanced transactional database capabilities and offered the potential to transform the treatment and management of chronic diseases for millions of patients. This project was meant to bring the economies of scale of semiconductors into the healthcare industry with its advantages of real-time, personalised care and the delivery of some form of breakthrough. The system was based on a low-cost, disposable, integrated sensor interface chip.

Due to the chip's ultra low power and very small battery size, it could be worn on the body with complete freedom of movement, or it could be implanted. Such a mobile healthcare technology is compatible with a wide range of sensors (see Figure 4) and can therefore be configured to detect vital signs such as ECG (electrocardiogram), blood oxygen and glucose, body temperature, and mobility. The device can also dynamically process and filter event data (including irregularities in heartbeat or blood pressure) and send the details to a mobile phone or PC via an ultra low-power, short-range radio telemetry link.

Further improvements to this kind of mobile healthcare technology could enhance the quality and efficiency of the healthcare patients of the NHS receive in the future. It could permit the following in future healthcare:

- Provide more timely and personalised care.

Figure 4. Mobile healthcare technology with built-in sensor (http://www.healthpia.us)

- Deliver unprecedented freedom, flexibility, and control for patients.
- Include exciting possibilities for medical practitioners to ultimately offer consumer items for which selection is based on quality and efficiency.

CONCLUSION

This chapter has provided examples of mobile healthcare technologies (case studies) for which the successful delivery of mobile solutions can help with certain kinds of emergency services. When considering the deployment of any mobile data solution, reliability, efficiency, and security are essential, none more so than in the emergency services if lives are to be saved. Wireless and mobile technology offers the opportunity to vastly enhance services, but there are still challenges to be faced if a cost-efficient solution is to be delivered.

Although some of these initiatives described as e-health can deliver certain benefits (including increased productivity and effectiveness of healthcare personnel and improved delivery of information and services), they will be faced with a number of challenges. Service providers in the private sector are looking to government for more leadership on identification issues and, as such, mobile healthcare technology should be a welcomed measure.

IT is seen as a key driver in the delivery of an efficient public sector, but how can departments justify further expenditure and eventually provide a clear road map to return on investment whilst delivering what has been promised? Also, what are the key short-term issues and, more importantly, the solutions that government departments can focus on? This panel will examine the savings that ICT investment is expected to deliver in the public sector, and how to serve more people by making things more efficient.

The United Kingdom has certainly increased its uptake on open-source software since the Office of Government Commerce's (OGC's) announcement that open source is a viable desktop alternative for the majority of users. However, many government organisations have chosen to remain with their existing proprietary software. This panel will examine the advantages and the drawbacks of both software solutions.

While this chapter is not intended to give specific guidelines for using mobile healthcare technologies, the author finds it useful to mention the following two points:

- Have clear objectives. Mobile healthcare technology is only a means to an end. It is advisable for managers of healthcare organisations to not be dazzled by the technology.
- Mobile healthcare technology is a unique medium, requiring management to capitalize on its uniqueness. The information being transmitted by mobile healthcare technology is the same, but is just delivered in a different way.

A conceptual framework has been posed from the research upon which this chapter is based, showing the different stages of the mobile healthcare technology-adoption process for healthcare organisations as well as the main factors operating at each stage.

In conclusion, the author has argued that Web services can aid the strategic planning of a healthcare organisation and can be used for competitive advantage. Web services can also contribute to improving our understanding and management of the critical issues surrounding mobile-based healthcare. Such understanding not only avoids disastrous consequences during the adoption of information systems, but also proves essential in supporting healthcare professionals to effectively manage the current trend of rapid increase in healthcare costs.

REFERENCES

Amit, R., & Zott, C. (2001). Value creation in e-business. *Strategic Management Journal, 22*,493-520.

Atkinson, C. J., & Peel, V. J. (1998). Transforming a hospital through growing, not building, an electronic patient record system. *Methods of Information in Medicine, 37*, 285-293.

Banker, R., & Kauffman, R. (1988). Strategic contributions of information technology: An empirical study of ATM networks. In *Proceedings of the Nineth International Conference on Information Systems*, Minneapolis.

Barney, J. B., & Griffin, R. W. (1992). *The management of organizations: Strategy, structure, behavior.* Boston: Houghton Mifflin.

Brown, S. (2001). NHS finance: The issue explained. *The Guardian*, 30 May (pp. 21).

Brown, S. A., & Venkatesh, V. (2003). Bringing non-adopter along: The challenge facing the PC industry. *Communications of the ACM, 46*(4), 76-80.

Collins, T. (2003). Doctors attack health IT codes. *ComputerWeekly*, February 6, (pp. 21).

Currie, W., Desai, B., & Khan, N. (2004). Customer evaluation of application services provisioning in five vertical sectors. *Journal of Information Technology, 19*, 39-58.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *Management Information Systems Quarterly, 13*(4), 982-1003.

Edmondson, A. C. (2003). Framing for learning: Lessons in successful technology implementation. *California Management Review, 45*(2), 34-54.

Ferlie, E. B., & Shortell, S. M. (2001). Improving the quality of healthcare in the United Kingdom and the United States: A framework for change. *Milbank Quarterly, 79*, 281-315.

Guah, M. W., & Currie, W. L. (2004). Application service provision: A technology and working tool for healthcare organisation in the knowledge age. *International Journal of Healthcare Technology and Management, 6*(1/2), 84-98.

Guah, M. W., & Currie, W. L. (2006). *Internet strategy: The road to Web services solutions.* PA: IRM Press.

Hagel, J. (2002). *Out of the box: Strategies for achieving profits today and growth tomorrow through Web services.* Boston: Harvard Business School Press.

Haines, M. (2002). *Knowledge management in the NHS: Platform for change*. Department of Health. Retrieved November 2002 from http://www.healthknowledge.org.uk

Hartmann, A., & Sifonis, J. (2000). *NetReady-strategies for the success in the e-economy.* New York: McGraw-Hill.

Institute of Medicine. (2002). *Crossing the quality chasm: A new health system for the 21st century.* Washington, DC: Committee on Quality Healthcare in America, National Academy Press.

Kreger, H. (2003). Fulfilling the Web services promise. *Communications of the ACM, 46*(6), 29-34.

Levy, M., & Powell, P. (2005). *Strategies for growth in SMEs: The role of information and information systems*. Oxford: Butterworth Heinemann.

Majeed, A. (2003). Ten ways to improve information technology in the NHS. *British Medical Journal, 326*, 202-206.

McGauran, A. (2002). Foundation hospitals: Freeing the best or dividing the NHS? *British Medical Journal, 324*(1), 1298.

Metters, J., Abrams, M., Greenfield, P. R., Parmar, J. M., & Venn, C. E. (1997). *Report to the Secretary of State for Health of the professional committee on the appeal of Mr. D. R. Walker under paragraph 190 of the terms and conditions of service of hospital medical and dental staff (England and Wales).* London: Department of Health.

National Health Service Information Authority (NHSIA). (2003). *Annual operating plan: To be the national provider of information and infrastructure services.* London: UK, Department of Health.

Orlikowski, W.J. (1993). CASE tools as organizational change: Investigating incremental and radical changes in systems development. *MIS Quarterly, 17*(3).

Orlikowski, W. J., & Tyre, M. J. (1994). Windows of opportunity: Temporal patterns of technological adaptation in organisations. *Organisation Science,* May, 98-118.

Pencheon, D. (1998). Matching demand and supply fairly and efficiently. *British Medical Journal, 316,* 1665-1667.

Perseid Software Limited. (2003). *The strategic value of Web services for healthcare and the life sciences.* Retrieved August 2003 from http://www.perseudsiftware.com

Robinson, J. C. (2000). Deregulation and regulatory backlash in healthcare. *California Management Review, 43*(1), 13-33.

Ruyter, K. D., Wetzels, M., & Kleijnen, M. (2001). Customer adoption of e-service: An experimental study. *International Journal of Service Industry Management, 2,* 184-206.

Sleeper, B., & Robins, B. (2001). *Defining Web services.* Retrieved April 2002 from http://www.stencilgroup.com

Srinivasan, R., Lilien, G. L., & Rangaswamy, A. (2004). Technological opportunism and radical technology adoption: An application to e-business. *Journal of Marketing, 66*(3), 47-60.

Tolbert & Zucker. (1994). *Institutional Analysis of Organizations: Legitmate but not Institutionalized*. Institute for Social Science Research working paper, University of California, Los Angeles, Vol. 6, No. 5.

Tornatzky, L. G., & Fleischer, M. (1990). *The process of technology innovation*. Lexington: Lexington Books.

Walsham, G. (1993). *Interpreting information systems in organisations*. Chichester, United Kingdom: Wiley.

Wanless, D. (2002). *Securing our future health: Taking a long-term view. Final report of an independent review of the long-term resource requirement for the NHS*. London: Department of Health.

Willcocks, L., & Lacity, M. (1998). *Strategic sourcing of information systems.* John Wiley & Sons, New York.

Yin, R. K. (1994). *Case study research: Design and methods.* CA: Sage Publications.

This work was previously published in Web and Mobile-Based Applications for Healthcare Management, edited by L. Al-Hakim, pp. 355-375, copyright 2007 by IRM Press (an imprint of IGI Global).

Chapter 2.4
Design of an Enhanced 3G-Based Mobile Healthcare System

Julián Fernández Navajas
University of Zaragoza, Spain

José Ruiz Mas
University of Zaragoza, Spain

Antonio Valdovinos Bardají
University of Zaragoza, Spain

Eduardo Antonio Viruete Navarro
University of Zaragoza, Spain

Robert S. H. Istepanian
Kingston University, UK

Carolina Hernández Ramos
University of Zaragoza, Spain

José García Moros
University of Zaragoza, Spain

Álvaro Alesanco Iglesias
University of Zaragoza, Spain

ABSTRACT

An enhanced mobile healthcare multi-collaborative system operating over Third Generation (3G) mobile networks is presented. This chapter describes the design and use of this system in different medical and critical emergency scenarios provided with universal mobile telecommunications system (UMTS) accesses. In these environments, it is designed to communicate healthcare personnel with medical specialists in a remote hospital. The system architecture is based on advanced signalling protocols that allow multimedia multi-collaborative conferences in IPv4/IPv6 3G scenarios. The system offers real-time transmission of medical data and videoconference, together with other non real-time services. It has been optimized specifically to operate over 3G mobile networks using the most appropriate codecs. Evaluation results show a reliable performance over IPv4 UMTS accesses (64 Kbps in the uplink). In the future, advances in m-Health systems will make easier for mobile patients to interactively get the medical attention and advice they need.

INTRODUCTION

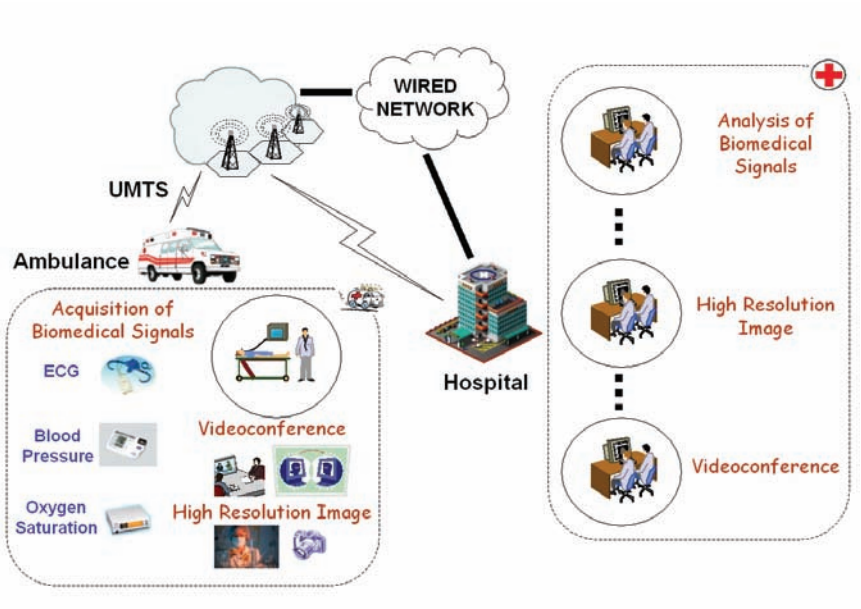
Mobile health (m-health) is an emerging area of telemedicine in which the recent development in mobile networks and telemedicine applications converge. m-health involves the exploitation of mobile telecommunication and multimedia technologies and their integration into new mobile healthcare delivery systems (Istepanian & Lacal, 2003). Wireless and mobile networks have brought about new possibilities in the field of telemedicine thanks to the wide coverage provided by cellular networks and the possibility of serving moving vehicles. One of the first wireless telemedical systems that utilized second-generation (2G) global system for mobile communications (GSM) networks addressed the electrocardiogram (ECG) transmission issues (Istepanian, 2001a). In recent years, several m-health and wireless telemedical systems based on GSM were reported (Istepanian, 2001b), allowing the accomplishment of remote diagnosis in mobile environments, as well as communication to geographic zones inaccessible by wired networks. The recent developments in digital mobile telephonic technologies (and their impact on mobility issues in different telemedical and telecare applications) are clearly reflected in the fast growing commercial domain of mobile telemedical services. A comprehensive review of wireless telemedicine applications and more recent advances on m-health systems is presented in Istepanian, Laxminarayan, and Pattichis (2005).

However, 2G-based systems lack the necessary bandwidth to transmit bandwidth-demanding medical data. The third-generation (3G) universal mobile telecommunications system (UMTS) overcomes limitations of first and second mobile network generations supporting a large variety of services with different quality of service (QoS) requirements. However, this fact makes network design and management much more complex. New applications require networks to be able to handle services with variable traffic conditions keeping the efficiency in the network resources utilization. The UMTS air interface is able to cope with variable and asymmetric bit rates, up to 2 Mbps and 384 kbps in indoor and outdoor environments, respectively, with different QoS requirements such as multimedia services with bandwidth on demand (Laiho, Wacker, & Novosad, 2000). In this kind of scenario, the emergence of 3G mobile wireless networks will permit to extend the use of m-health applications thanks to the provided higher transmission rates and flexibility over previous mobile technologies.

UMTS introduces the IP multimedia core network subsystem (IMS) (3GPP, 2005a), an IPv6 network domain designed to provide appropriate support for real-time multimedia services, independence from the access technologies and flexibility via a separation of access, transport and control. The fundamental reason for using IPv6 is the exhaustion of IPv4 addresses. Support for IPv4 is optional, but since network components require backward compatibility, it is clear that a dual stack configuration (IPv4 and IPv6) must be provided. The IMS uses the session initiation protocol (SIP) as signalling and session control protocol (Rosenberg et al., 2002). SIP allows operators to integrate real-time multimedia services over multiple access technologies such as general packet radio service (GPRS), UMTS or, ultimately, other wireless or even fixed network technologies (interworking multimedia domains). This chapter presents a 3G-based m-health system designed for different critical and emergency medical scenarios, as shown in Figure 1. Several medical specialists in the hospital take part in a multipoint conference with the ambulance personnel, receiving compressed and coded biomedical information from the patient, making it possible for them to assist in the diagnosis prior to its reception.

The 3G system software architecture includes intelligent modules such as information compression and coding, and QoS control to significantly improve transmission efficiency, thus optimizing the use of the scarce and variable wireless channel

Figure 1. Typical medical mobility scenario



bandwidth compared to previous systems (Chu & Ganz, 2004; Curry & Harrop, 1998). Finally, unlike Chu and Ganz (2004), this m-health system follows a multi-collaborative design which supports IPv6/IPv4 interworking, uses SIP as the service control protocol and integrates new real-time multimedia features intended for 3G wireless networks.

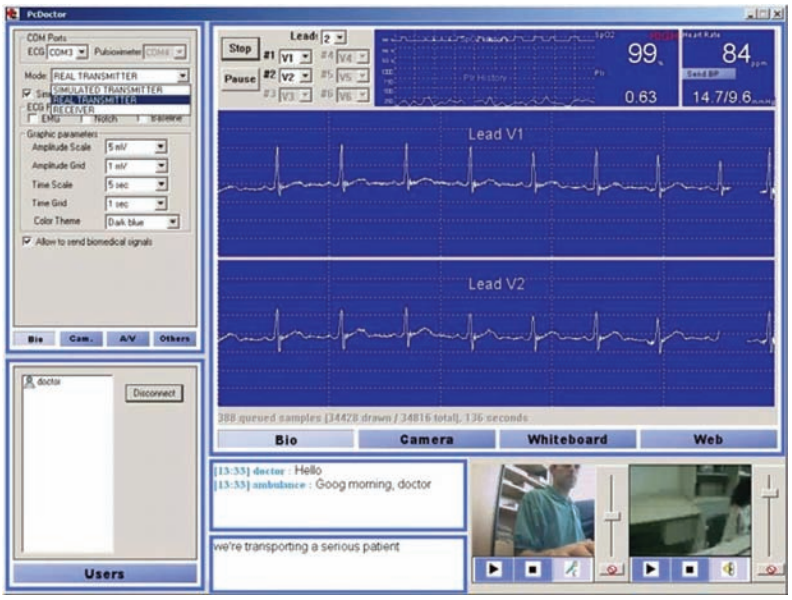
3G M-HEALTH SYSTEM ARCHITECTURE

In this section, the 3G m-health system structure is described in detail. The system (see the system components in Figure 2 and the main application graphical user interface (GUI) in Figure 3) has been built using standard off-the-shelf hardware, instead of developing propriety hardware as in Cullen, Gaasch, Gagliano, Goins, and Gunawardane (2001), uses free software and commercially available 3G wireless UMTS cellular data services. In addition, it provides simultaneous transfer of, among other services, videoconference, high-resolution still medical images and medical data, rather than only one media at a time (Kyriacou, 2003; Pavlopoulos, Kyriacou, Berler, Dembeyiotis, & Koutsouris, 1998).

The 3G m-health system consists of different modules that allow the acquisition, treatment, representation and transmission of multimedia information. The modular design of each medical user service allows great flexibility. In addition, there exist other medical user services like chat and electronic whiteboard that allow data exchange in order to guide the operations performed by remote users.

Figure 2. 3G m-Health system

Figure 3. m-Health application GUI



The medical signals module acquires, compresses, codes, represents, and transmits medical signals in real time. The medical signals acquisition devices included are: a portable electrocardiograph that allows the acquisition of 8 real and 4 interpolated leads of the ECG signal, and that follows the standard communication protocol-ECG (SCP-ECG); a tensiometer that provides systolic and diastolic blood pressure values; and a pulsioximeter that offers the blood oxygen saturation level (SpO_2) and the cardiac pulse.

The details of the 3G system architecture are shown in Figure 4 and Figure 5. As it can be seen, the system comprises of the signalling and session control, medical user services and application control sub-systems, which will be described later.

This architecture allows the 3G system to offer real-time services such as medical data transmission (ECG, blood pressure, heart rate and oxygen saturation), full-duplex videoconference, high-resolution still images, chat, electronic whiteboard, and remote database Web access.

Communication between the remote ambulance personnel and medical specialists is established by means of multipoint multi-collaborative sessions through several network environments capable of supporting the different types of multimedia traffic (Figure 4). The selected conference model (tightly coupled conference model (Rosenberg, 2004)) requires the existence of a MCU (multipoint control unit) to facilitate multipoint operation. The MCU maintains a dialog with each participant in the conference and is responsible for ensuring that the media streams which constitute the conference are available to the appropriate participants. The MCU can belong to mobile's home network (SIP application server) or to a external service platform. Furthermore, the m-health system is developed to support IPv4/IPv6 interworking (Wiljakka & Soinien, 2003) and can be integrated in a 3G network scenario (see Figure 4).

The MCU receives the information generated by each participant in the multipoint conference, processes and forwards it to the appropriate destinations. System users and the MCU exchange information associated with the different services provided (medical user services) and its presentation (application control). Moreover, they

Figure 4. 3G network scenario

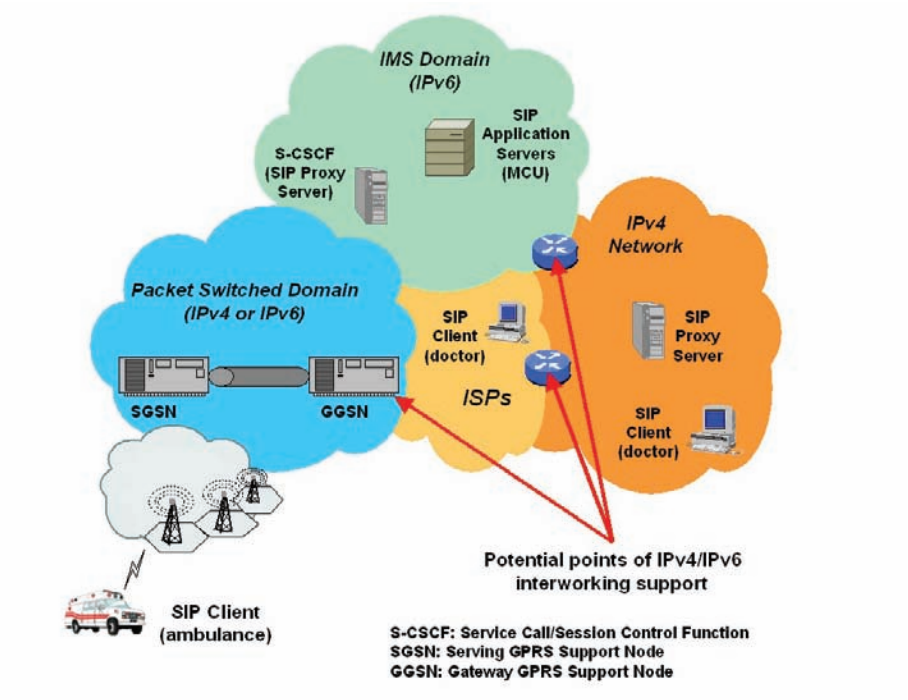
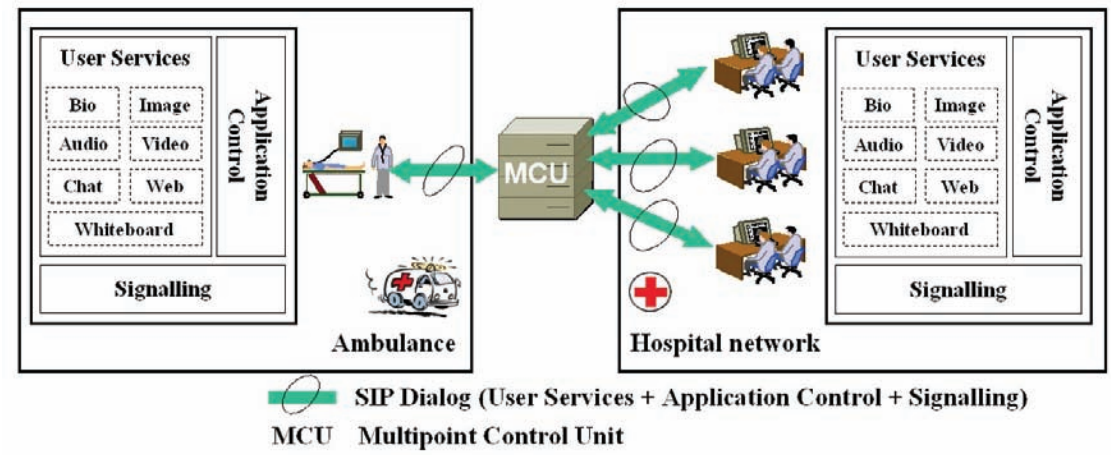


Figure 5. Block diagram of the m-Health system architecture and sub-systems



exchange information related to communication and service quality management (signalling). Next, a detailed description of the sub-systems is presented.

Signalling and Session Control

The developed signalling allows the exchange of the characteristics associated to the different information flows between system elements

and is based on standard protocols that favour interoperability. Signalling tasks, performed by the SIP protocol, begin with the establishment of a SIP dialog with the MCU in which, by means of session description protocol (SDP) messages, the different services are described. In order to do that, each element in the system has a SIP user agent (UA), slightly modified in the MCU to allow the use of multiple simultaneous dialogs.

In addition to session control functions (establishment, management and termination of the multipoint conference), the SIP protocol is also useful for user mobility purposes inside the IMS environment.

Multipoint conference establishment, management and termination is performed by exchanging SIP messages between the different users. When a user connects, he creates a SIP dialog with the MCU, joining the conference. During the conference, SIP messages are exchanged between users and the MCU, varying conference characteristics and therefore allowing its management. In a similar process to that of conference joining, when a user wants to leave it, this fact must be communicated to the MCU with the necessary SIP messages. SIP messages also serve as the mean of transport of SDP messages with the description of medical user services.

The QoS in this system is mainly determined by the characteristics of the UMTS link. Mobile links are very variable, therefore a QoS-monitoring process is required in order to obtain a good system performance. This process is especially important in the MCU because it is there where the QoS-related decisions are taken. When the MCU detects that a particular conference participant needs to modify the characteristics of its multimedia session in order to improve QoS, it renegotiates the corresponding session by sending SIP/SDP messages. Hence, conference participants can modify certain upper-level protocol parameters (codecs used, transmission rates, compression ratios, etc.) in order to adapt the transmitted information to network performance.

The QoS monitoring process is possible thanks to a transport library that provides a uniform interface to send the information generated by medical user services and different QoS estimation tools developed for several types of links. This transport library offers different queuing policies and tools designed to measure the following QoS-related parameters: delay, bandwidth and packet loss rate. Due to the variable nature of wireless links, reception buffers have been properly dimensioned to minimize jitter, delay and packet loss.

Wireless Medical User Services

The medical user services included in the m-health system are associated with information shared in a multi-collaborative environment. Specifically, the system has services to share audio, video, medical data information, high-resolution still images, and graphical and textual information, as well as a Web service that allows remote access to clinical information databases. In addition to these services, there is a service designed to exchange control information (application control), which is discussed later.

Each kind of information is associated with a medical user service and uses a transport protocol and a codec according to its characteristics (see Table 1). Hence, real-time services (audio, video, and medical data information) use the real-time transport protocol (RTP), whereas the rest of the services use the transmission control protocol (TCP). Furthermore, the exchanged information can be very sensitive and requires a secure communication. The 3G m-health system uses an IP security protocol (IPSec) implementation supporting public key certificates in tunnel mode. This protocol ensures private communication of selected services.

The ECG signal is stored both in transmission and reception following the SCP-ECG standard. It is well known that for an efficient transmission an ECG compression technique has to be used. In

Table 1. Codec operation modes for 3G real-time wireless medical user services

	CODEC	CODEC RATE
Audio	AMR*	4.74 5.15 5.9 6.7 7.4 7.95 10.0 12.2 (Kbps)
Video	H.263	5 10 (Frames per second)
Biomedical Signals	WT**	5 10 20 (Kbps)

** Adaptive multi-rate, ** Wavelet transform*

our implementation, a real-time ECG compression technique based on the wavelet transform is used (Alesanco, Olmos, Istepanian, & García, 2003). This is a lossy compression technique, therefore the higher the compression ratio (lower the transmission rate), the higher the distortion at reception. It is clear that there is a trade-off between transmission rate and received ECG signal quality. From the transmission efficiency point of view, a very low transmission rate is desired but from the clinical point of view, a very distorted ECG is useless. Therefore, there exists a minimum transmission rate to be used so as the transmitted ECG is useful for clinical purposes, which was selected in collaboration with cardiologists after different evaluation tests. The minimum transmission rate used in our implementation (625 bits per second and per ECG lead) leads to a clinically acceptable received ECG signal.

Regarding blood pressure, oxygen saturation, and heart rate, these signals have low bandwidth requirements and, therefore, are not compressed.

The videoconference module captures, sends, and plays audio and video information obtained by means of a Web camera and a microphone. In order to reduce the bandwidth, these data are compressed and coded. The video signal is compressed following the H.263 standard, whereas the audio signal uses the adaptive multi-rate (AMR) codec, recommended for UMTS by the 3G Partnership Program (3GPP) (3GPP 2005b). This module provides the basic functionality for starting, pausing and stopping video signals acquisition and representation, as well as volume control for the microphone (capture) and the speakers (reproduction). Due to the fact that each participant in the conference receives a unique video signal, the system allows the user to select the particular video signal among all the users connected.

The high-resolution still image module obtains high quality images with a charge copled device (CCD) colour camera connected to the computer through an image acquisition card. This module includes options to preview the captured images and modify their main characteristics in real time: brightness, contrast, hue, etc. Captured images can be stored and transmitted in different formats, with various qualities and compression levels. These images are sent automatically to the electronic whiteboard module of the remote users, allowing to select and mark fixed areas in a multi-collaborative fashion to facilitate a diagnostic clinical procedure.

Application Control

The MCU forwards the information generated by each medical service according to the presentation spaces defined using the control service. Each medical service has a presentation space associated with it that defines the way in which the information has to be transferred and its destination. The MCU simply forwards the in-

formation it receives, but has a special treatment for the audio, video, medical data, and control services. Regarding the Audio service, the MCU decodes the signal of each user, mixes it with the decoded signal of the other conference participants and codes the result in order to transfer a unique audio signal to each user (Figure 6). On the other hand, the MCU only forwards one video signal to each conference participant. The particular video signal forwarded to each user is selected by using the control service. Finally, the medical data service is similar to the video service: medical data are only generated in the remote location, whereas the other conference participants can only receive them.

The 3G m-health system was adapted to the critical and emergency medical scenarios. In the first stages of its design, user requirements and functional specifications were established in collaboration with medical specialists, in order to create a portable and modular m-health system that could be easily integrated in any environment, using any underlying network technology capable of supporting IP multimedia services

3G M-HEALTH SYSTEM PERFORMANCE

In order to measure the 3G m-health system performance, several tests have been carried out using the system over 64 Kbps (at IP level) UMTS accesses. Table 2 presents the results about average IP-level bandwidth used by real-time network services.

As it can be observed, considering more audio samples per network packet reduces the used bandwidth, since transmission efficiency (information carried by each packet to total packet size ratio) is increased. However, there is a limit in the number of audio samples per packet that can be used because more audio samples per packet yield more audio delay. For example, a more efficient transmission mode including four audio samples per packet every 80 ms causes four times the delay including only one audio sample per packet every 20 ms. Moreover, if an audio packet is lost, all the audio samples carried by it are lost and, therefore, a reduced number of audio samples per packet is more suitable to error-prone environments. Regarding the video user service, it is worth noting that the bandwidth

Figure 6. Application control (audio presentation spaces)

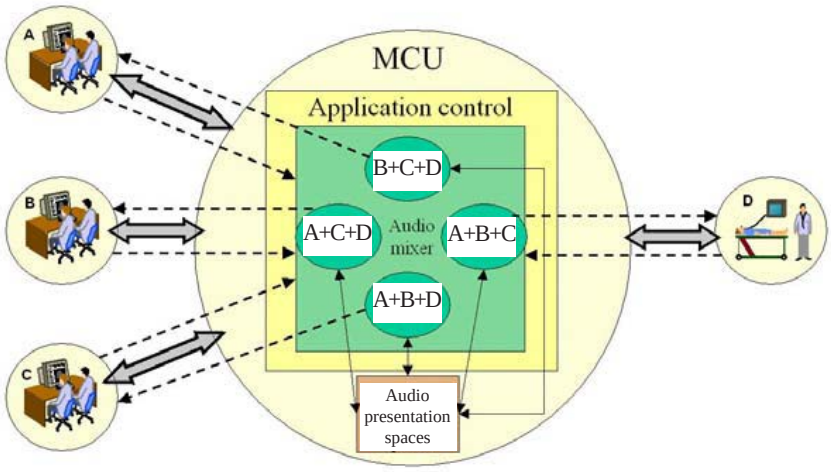


Table 2. Average IP-level bandwidth used by real-time user services

	Operation Mode		IP Bandwidth (Kbps)
	Samples/packet	Codec rate (Kbps)	
Audio	1	4.75	21.2
	1	12.2	28.8
	3	**4.75**	**10.5**
	3	12.2	18.1
	Frames per second		
Video	**5**		**16**
	10		24
	Bit Rate		
Biomedical Signals	**5**		**5.3**
	10		10.3

shown in Table 2 can vary substantially with the movement of the video scene captured. Finally, the medical data service adapts well to the codec rate specified because medical data frame sizes are long enough to obtain a good transmission efficiency.

As it can be checked, the total bandwidth consumed by all real-time medical user services fits in a 64 Kbps UMTS channel, even when the most bandwidth-consuming codec rates and the lowest transmission efficiencies are used. If all medical user services are used (including non real-time services), lower codec rates should be selected. Thus, according to the previous discussions, the codec operation modes selected in this m-health system have been those highlighted in Table 2, achieving a reasonable trade-off between bandwidth, transmission efficiency, delay and loss ratio.

NEXT-GENERATION OF M-HEALTH SYSTEMS

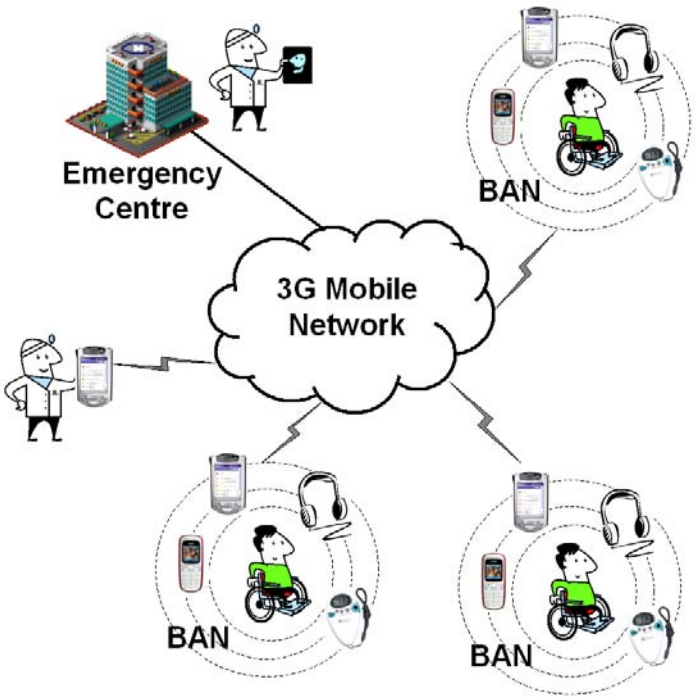
It is evident that organizations and the delivery of health care are being underpinned by the advances in m-health technologies. In the future, home medical care and remote diagnosis will become common, check-up by specialists and prescription of drugs will be enabled at home and virtual hospitals with no resident doctors will be realized.

Hence, the deployment of emerging mobile and wireless technologies will face new challenges: inter-operability between heterogeneous networks (fourth-generation, 4G) and smart medical sensor design integrating sensing, processing, communications, computing, and networking.

With the aid of wireless intelligent medical sensor technologies, m-health can offer healthcare services far beyond what the traditional telemedical systems can possibly provide. The individual sensors can be connected wirelessly to a personal monitoring system using a wireless body area network (WBAN) and can be integrated into the user's clothing, providing wearable and ubiquitous m-health systems. A typical scenario comprises of a WBAN system that communicates with cardiac implantable devices (pacemakers, defibrillators, etc.) and that is linked to the existing 3G infrastructure, achieving "Health Anytime, Anywhere" (Figure 7).

It is expected that 4G will integrate existing wireless technologies including UMTS, wireless LAN, Bluetooth, ZigBee, Ultrawideband, and other newly developed wireless technologies into a seamless system. Some expected key features of 4G networks are: high usability, support for multimedia services at low transmission cost and facilities for integrating services. 4G advances

Figure 7. Typical WBAN-3G scenario



will make easier for mobile patients to interactively get the medical attention and advice they need. When and where is required and how they want it regardless of any geographical barriers or mobility constraints.

The concept of including high-speed data and other services integrated with voice services is emerging as one of the main points of future telecommunication and multimedia priorities with the relevant benefits to citizen-centered healthcare systems. The new wireless technologies will allow both physicians and patients to roam freely, while maintaining access to critical patient data and medical knowledge.

CONCLUSION

This chapter has presented a feasible 3G-based m-health system targeted specifically for critical and emergency medical scenarios. The system architecture is based on 3G networks and advanced signalling protocols (SIP/SDP) that allow the integration of real-time multimedia services over multiple access channels that support IPv4 and IPv6 interworking depending on current commercial UMTS releases.

The system has the following features: simultaneous transmission of real-time clinical data (including ECG signals, blood pressure, and blood oxygen saturation), videoconference, high-resolution still image transmission, and other facilities such as multi-collaborative whiteboard, chat, and Web access to remote databases. The system has been optimized specifically to operate over 3G mobile networks using the most appropriate codecs. Evaluation results show a reliable performance over UMTS accesses (64 Kbps in the uplink).

Home telecare and chronic patient telemonitoring are other application areas in which this m-health system can be used, thus further work is currently undergone to adapt it and to evaluate its performance in each particular scenario.

REFERENCES

3GPP TS 23.228 V6.8.0. (2005). IP Multimedia Subsystem (IMS); Stage 2 (Release 6).

3GPP TS 26.235 V6.3.0. (2005). Packet switched conversational multimedia applications; Default codecs (Release 6).

Alesanco, A., Olmos, S., Istepanian, R. S. H., & García, J. (2003). A novel real-time multilead ECG compression and de-noising method based on the wavelet transform. *Proceedings of IEEE Computers Cardiology* (pp. 593-596). Los Alamitos, CA: IEEE Comput. Soc. Press.

Chu, Y., & Ganz, A. (2004). A mobile teletrauma system using 3G networks. *IEEE Transactional Information Technology in Biomedicine*, *8*(4), 456-462.

Cullen, J., Gaasch, W., Gagliano, D., Goins, J., & Gunawardane, R. (2001). *Wireless mobile telemedicine: En-route transmission with dynamic quality-of-service management.* National Library of Medicine Symposium on Telemedicine and Telecommunications: Options for the New Century.

Curry, G. R., & Harrop, N. (1998). The Lancashire telemedicine ambulance. *Journal Of Telemedicine Telecare*, *4*(4), 231-238.

Istepanian, R. S. H., Kyriacou, E., Pavlopoulos, S., & Koutsouris, D. (2001). Wavelet compression methodologies for efficient medical data transmission in wireless telemedicine system. *Journal of Telemedicine and Telecare*, *7*(1), 14-16.

Istepanian, R. S. H., Laxminarayan, S., & Pattichis, C. S. (2005). *M-Health: Emerging mobile health systems.* New York: Springer. To be published.

Istepanian, R. S. H., & Lacal, J. C. (2003). Emerging mobile communication technologies for health: Some imperative notes on m-health. *Proceedings of the 25th Silver Anniversary International Conference of the IEEE Engineering in Medicine and Biology Society 2* (pp. 1414-1416).

Istepanian, R. S. H., Woodward, B., & Richards, C. I. (2001). Advances in telemedicine using mobile communications. *Proceedings of the IEEE Engineering Medicine and Biology Society 4* (pp. 3556-3558).

Kyriacou, E., Pavlopoulos, S., Berler, A., Neophytou, M., Bourka, A., Georgoulas, A., Anagnostaki, A., Karayiannis, D., Schizas, C., Pattichis, C., Andreou, A., & Koutsuris, D. (2003). Multipurpose healthcare telemedicine systems with mobile communication link support. *BioMedical Engineering OnLine*, *2*(7).

Laiho, J., Wacker, A., & Novosad, T. (2000). *Radio network planning and optimization for UMTS*. New York: Wiley.

Pavlopoulos, S., Kyriacou, E., Berler, A., Dembeyiotis, S., & Koutsouris, D. (1998). A novel emergency telemedicine system based on wireless communication technology–AMBULANCE. *IEEE Trans. Inform. Technol. Biomed*, *2*, 261-267.

Rosenberg, J. (2004). *A framework for conferencing with the session initiation protocol.* Internet draft. Work in progress.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., & Schooler, E. (2002). *SIP: Session initiation protocol* (IETF RFC 3261).

Wiljakka, J., & Soinien, J. (2003). Managing IPv4-to-IPv6 transition process in cellular networks and introducing new peer-to-peer services. *Proceedings of IEEE Workshop on IP Operations and Management* (pp. 31-37).

KEY TERMS

4G: The fourth-generation (4G) is the continuation of the first, second, and third generations of mobile networks. It is a wireless access technology that provides high-speed mobile wireless access with a very high data transmission speed (2-20 Mbps) and enables users to be simultaneously connected to several wireless access technologies and seamlessly move between them in an all-IP environment. These access technologies can be any existing or future access technology. Smart antennas, low power consumption, and software-defined radio terminals will also be used to achieve even more flexibility for the user of 4G systems.

IMS: The IP multimedia subsystem (IMS) is a new open and standardized framework, basically specified for mobile networks, for providing Internet protocol (IP) telecommunication services. It offers a next generation network (NGN) multimedia architecture for mobile and fixed services, based on the session initiation protocol (SIP), and runs over the standard IP. It is used by telecom operators in NGN networks (combining voice and data in a single packet switched network), to offer network controlled multimedia services. The aim of IMS is not only to provide new services but to provide all the services, current and future, that the Internet provides. In addition, users have to be able to execute all their services when roaming as well as from their home networks. To achieve these goals the IMS uses IETF protocols. This is why the IMS merges the Internet with the cellular world; it uses cellular technologies to provide ubiquitous access and Internet technologies to provide new services.

IPv6: Internet protocol version 6 (IPv6) is the next generation protocol designed by the Internet Engineering Task Force (IETF) to replace the current version of the Internet protocol, IP version 4 (IPv4). Today's Internet has been using IPv4 for twenty years, but this protocol is beginning to become outdated. The most important problem of IPv4 is that there is a growing shortage of IPv4 addresses, which are needed by all new machines added to the Internet. IPv6 is the solution to several problems in IPv4, such as the limited number of available IPv4 addresses, and also adds many improvements to IPv4 in other areas. IPv6 is expected to gradually replace IPv4, with the two coexisting for a number of years during a transition period.

QoS: ITU-T recommendation E.800 defines the term quality of service (QoS) as "the collective effect of service performance which determines the degree of satisfaction of a user of the service." Service performance comprises of very different parts (security, operability, etc), so the meaning of this term is very broad. In telecommunications, the term QoS is commonly used in assessing whether a service satisfies the user's expectations. QoS evaluation, however, depends on functional components and is related to network performance via measurable technical parameters. A QoS-enabled network has the ability to provide better service (priority, dedicated bandwidth, controlled jitter, latency, and improved loss characteristics) to selected network traffic over various technologies.

SIP: The session initiation protocol (SIP) is a signalling protocol developed by the IETF intended for setting up multimedia communication sessions between one or multiple clients. It is currently the leading signalling protocol for voice over IP, and is one of the key components of the IMS multimedia architecture. The most important functions of SIP include name mapping and redirection (user location), capabilities negotiation during session setup, management of session participants and capabilities management during the session.

Telemedicine: Telemedicine can be defined as the rapid access to shared and remote medical expertise by means of telecommunications and

information technologies, no matter where the patient or the relevant information is located. Any application of information and communications technologies which removes or mitigates the effect of distance in healthcare is telemedicine. The terms e-health and tele-health are terms often interchanged with telemedicine.

This work was previously published in Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 521-533, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.5
The M-Health Reference Model:
An Organizing Framework for Conceptualizing Mobile Health Systems

Phillip Olla
Madonna University, USA

Joseph Tan
Wayne State University, USA

ABSTRACT

The reference model presented in this article encourages the breakdown of M-Health systems into the following five key dimensions: (1) Communication Infrastructure: a description of mobile telecommunication technologies and networks; (2) Device Type: the type of device being used, such as PDA, sensor, or tablet PC; (3) Data Display: describes how the data will be displayed to the user and transmitted, such as images, e-mail, and textual data; (4) Application Purpose: identification of the objective for the M-Health system; (5) Application Domain: definition of the area in which the system will be implemented. Healthcare stakeholders and system implementer can use the reference model presented in this article to understand the security implications of the proposed system and to identify the technological infrastructure, business requirements, and operational needs of the M-Health systems being implemented. A reference model that encapsulates the emerging M-Health field is needed for cumulative progress in this field. Currently, the M-Health field is disjointed, and it is often unclear what constitutes an M-Health system. In the future, M-Health applications will take advantage of technological advances such as device miniaturizations, device convergence, high-speed mobile networks, and improved medical sensors. This will lead to the increased diffusion of clinical M-Health systems, which will require better understanding of the components that constitute the M-Health system.

INTRODUCTION

M-Health is defined as "mobile computing, medical sensor, and communications technologies for healthcare" (Istepanian, Jovanov, & Zhang, 2004, p. 405). The first occurrence of the term *M-Health* in the literature was in the "Unwired E-Med" special issue on Wireless Telemedicine Systems (Istepanian & Laxminaryan, 2000). Since then, there has been an increased use of the term, encapsulating various types of healthcare systems. The use of the M-Health terminology relates to applications and systems such as telemedicine (Istepanian & Wang, 2003), telehealth (Istepanian & Lacal, 2003), and biomedical sensing system (Budinger, 2003). Until now, there have been considerable confusion and overlap with the use of these terms (Tulu & Chatterjee, 2005).

Rapid advances in Information Communication Technology (ICT) (Godoe, 2000), nanotechnology, biomonitoring (Budinger, 2003), mobile networks (Olla, 2005a), pervasive computing (Akyildiz & Rudin, 2001), wearable systems, and drug delivery approaches (Amy et al., 2004) are transforming the healthcare sector. The insurgence of innovative technology into the healthcare practice not only is blurring the boundaries of the various technologies and fields but also is causing a paradigm shift that is blurring the boundaries among public health, acute care, and preventative health (Hatcher & Heetebry, 2004). These developments not only have had a significant impact on current e-health and telemedical systems (Istepanian, Jovanov, & Zhang, 2004), but they also are leading to the creation of a new generation of M-Health systems with a convergence of devices, technologies, and networks at the forefront of the innovation.

This article proposes the use of a five-dimensional reference model in order to assist system implementers and business stakeholders in understanding the various components of an M-Health system. The approach used by the this article focuses on identifying different dimensions of a Mobile Healthcare Delivery System (MHDS) (Wickramasinghe & Misra, 2005), which then can be used to identify user security requirements for different categories in an organized manner. These dimensions were driven from our literature review (Bashshur, 2002; Bashshur, Reardon, & Shannon, 2000; Raskovic & Jovanov, 2004; Istepanian, Laxminaryan, & Pattichis, 2006; Jovanov, Milenkovic, Otto, & Groen, 2005; Field, 1996; Moore, 2002; Olla & Patel, 2003), and the model reflects a combination of various classification schemes proposed in earlier studies in order to classify telemedicine and telehealth systems.

Based on the previous definition, M-Health is a broad area that transcends multiple disciplines and utilizes a broad range of technologies. There is a variety of applications, devices, and communication technologies that are emerging in the M-Health arena and that can be combined to create the M-Health system. The dimensions consist of the following:

1. **Communication Infrastructure.** Description of the mobile telecommunication technologies that will be used, such as Bluetooth, wireless local area networks, or third-generation technologies (Olla, 2005a).
2. **Device Type.** Relates to the type of device being used to collect the medical data, such as Personal Digital Assistance (PDA), sensor, or tablet PC (Parmanto, Saptono, Ferrydiansyah, & Sugiantara, 2005).
3. **Data Display.** Describes how the data will be displayed and transmitted to the user through images, e-mail, textual data, and other types of data presentation languages (Tulu & Chatterjee, 2005).
4. **Application Purpose.** Identification of the objective for the M-Health system (Field, 1996).
5. **Application Domain.** Definition of the area in which the system will be implemented, such as clinical (e.g., dermatology, radiology, etc.) or non-clinical (e.g., billing, maintenance, etc.) domains (Bashshur et al., 2000).

Selecting different alternatives in the five dimensions will have implications on the functionality of the system; however, the key focus of this discussion will be the emphasis on how the security and integrity of the M-Health system is maintained, based on the dimension choices.

The rest of the article is organized as follows. "Background of M-Health" features a brief background of the evolution of M-Health. "Mobile Healthcare Delivery System Networks" provides an overview of existing and new mobile technologies suitable for the healthcare sector. "The M-Health Reference Model" introduces the five dimensions of the M-Health reference model. The penultimate section uses the reference model to decompose a real-life mobile health delivery system in order to illustrate the security concerns. This is followed by the conclusion.

BACKGROUND OF M-HEALTH

The phenomenon of providing care remotely using ICT can be placed into a number of areas, such as M-Health, telemedicine, and e-health, but, as summed up by R. L. Bashur, President Emeritus of the American Telemedicine Association, the terminology is not the important aspect. "It does not really matter what we call it or where we draw boundaries. … collective and collaborative efforts from various fields of science, including what we call now telemedicine is necessary" (Bashshur, 2002, p. 7). Emphasis on the various components and objectives of the various types of systems should be the priority, as this will increase the chances of implementing efficient and effective systems, of generating viable business models, and of creating secure systems that meet the needs of the stakeholders (Olla & Patel, 2003). Over the evolution of telemedicine, new terminologies have been created as new health applications and delivery options became available and as application areas extended to most healthcare domains. This resulted in confusion and identification of what falls under telemedicine, and what falls under telehealth or e-health became more complicated as the field advanced. New concepts such as Pervasive Health and M-Health also are adding to this confusion. Before understanding the scope and components of M-Health, it is important to mention briefly the history of telemedicine and the advancements of mobile networks, which are collectively the foundation of M-Health. The evolution and growth of telemedicine is correlated highly with ICT advancements and software development. Telemedicine advancements can be categorized into three eras (Tulu & Chatterjee, 2005; Bashshur et al., 2000).

The first era of telemedicine focused solely on medical care as the only function of telemedicine. This era can be named as the telecommunications era of the 1970s. The applications in this era were dependent on broadcast and television technologies, in which telemedicine applications were not integrated with any other clinical data. The second era of telemedicine, a result of digitalization in telecommunications, grew during 1990s. The transmission of data was supported by various communication mediums ranging from telephone lines to Integrated Service Digital Network (ISDN) lines. During this period, there were high costs attached to the communication mediums that provided higher bandwidth. The bandwidth issue became a significant bottleneck for telemedicine in this era. Resolving the bandwidth constraints has been a critical research challenge for the past decade, with new approaches and opportunities created by the Internet revolution. Now, more complex and ubiquitous networks are supporting telemedicine. The third era of telemedicine was supported by a networking technology that was cheaper and more accessible to an increasing user population. The improved speed and quality offered by Internet is providing new opportunities in telemedicine. In this new era of telemedicine, the focus shifted from a technology assessment to a deeper appreciation of the functional relationships between telemedicine technology and the outcomes of cost, quality, and access.

This article proposes a fourth era that is characterized by the use of Internet Protocol (IP) technologies, ubiquitous networks, and mobile/wireless networking capabilities, which can be observed by the proliferation of M-Health applications that perform both clinical and non-clinical functions. Since the proliferation of mobile networks, telemedicine has attracted a lot more interest from both academic researchers and industry (Tachakra, Istepanian, Wang, & Song, 2003). This has resulted in many mobile/wireless telemedicine applications being developed and implemented (Budinger, 2003; Istepanian & Laxminaryan, 2000; Jovanov, Lords, Raskovic, Cox, Adhami, & Andrasik, 2003; Pattichis et al., 2002; Istepanian & Lacal, 2003; Webb, 2004). Critical healthcare information regularly travels with patients and clinicians, and therefore, the need for information to become securely and accurately available over mobile telecommunication networks is key to reliable patient care and reliable medical systems.

Health organizations are required by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) to maintain the privacy of protected health information and to provide individuals with notice of its legal duties and privacy policies with respect to this information (HIPPA, 2005). Specifically, the law requires the following:

- Any medical information that identifies a person must be kept private.
- Notice of legal duties and privacy practices with respect to protected health information must be made available.
- The organization must abide by the terms of its notices currently in effect.

HIPAA also requires:

- Computers and data containing protected health information (PHI) must be protected from compromise or loss.
- Audit trails of access to PHI must be kept.
- Electronic transmissions of PHI must be authenticated and protected from observation or change.

Failure to comply with these requirements (for both HIPAA privacy and security) can result in civil and criminal penalties ranging from fines of $100 to $250,000 and up to 10 years in prison. Medical information always has been considered sensitive and never more so than now due to the security issues created by mobility. Many countries have imposed strict regulations with heavy penalties in order to ensure the confidentiality and authorized distribution of personal medical information. Following the lead of European Union Directive 95/46/EC that protects both medical and financial personal information, the United States and Canada passed important legislation (HIPAA and PIPED, respectively) that imposes substantial penalties, both civil and criminal, for negligent or intentional exposure of personal medical information to unauthorized parties.

The telecommunication industry has progressed significantly over the last decade. There has been significant innovation in digital mobile technologies. The mobile telecommunication industry has advances through three generations of systems and is currently on the verge of designing the fourth generation of systems (Olla, 2005b). The recent developments in digital mobile technologies are reflected in the fast-growing commercial domain of mobile telemedical services (Istepanian et al., 2006). Specific examples include mobile ECG transmissions, video images and teleradiology, wireless ambulance services to predict emergency and stroke morbidity, and other integrated mobile telemedical monitoring systems (Istepanian, Jovanov, & Zhang, 2004; Future trends: Convergence is all, 2005; Warren, 2003). There is no doubt that mobile networks can introduce additional security concerns to the healthcare sector.

Implementing a mobile trust model will ensure that a mobile transaction safely navigates multiple technologies and devices without compromising the data or the healthcare systems. M-Health transactions can be made secure by adopting practices that extend beyond the security of the wireless network used and by implementing a trusted model for secure end-to-end mobile transactions. The mobile trust model proposed by Wickramasinghe and Misra (2005) utilizes both technology and adequate operational practices in order to achieve a secure end-to-end mobile transaction. The first level highlights the application of technologies in order to secure elements of a mobile transaction. The next level of the model shows the operational policies and procedures needed in order to complement technologies used. No additional activity is proposed for the mobile network infrastructure, since this element is not within the control of the provider or the hospital.

The next section will discuss the mobile network technologies and infrastructure, which is a key component of any M-Health system. The network infrastructure acts as a channel for data transmission and is subject to the same vulnerabilities, such as sniffing, as in the case of fixed network transaction. The mobile networks discussed in the next section are creating the growth and increased adoption of M-Health applications in the healthcare sector.

MOBILE HEALTHCARE DELIVERY SYSTEM NETWORKS

The implementation of an M-Health application in the healthcare environment leads to the creation of a Mobile Healthcare Delivery System (MHDS), which can be defined as the carrying out of healthcare-related activities using mobile devices such as a wireless tablet computer, Personal Digital Assistant (PDA), or a wireless-enabled computer. An activity occurs when authorized healthcare personnel access the clinical or administrative systems of a healthcare institution using a mobile device (Wickramasinghe & Misra, 2005). The transaction is said to be complete when medical personnel decide to access medical records (patient or administrative) via a mobile network either to browse or to update the record.

Over the past decade, there has been an increase in the use of new mobile technologies in healthcare, such as Bluetooth and Wireless Local Area Networks (WLAN) that use protocols different than the standard digital mobile technologies such as 2G, 2.5, and 3G. A summary of these technologies is presented next, and a précis of the speeds and range is presented in Table 1.

These mobile networks are being deployed in order to allow physicians and nurses easy access to patient records while on rounds, to add observations to the central databases, to check on medications, as well as to perform a growing number of other functions. The ease of access that wireless networks offer is matched by the security and privacy challenges presented by the networks. This serious issue requires further investigation and research in order to identify the real threats for the various types of networks in the healthcare domain.

Second Generation Systems (2G /2.5G). The second-generation cellular systems (2G) were the first to apply digital transmission technologies such as Time Division Multiple Access (TDMA) for voice and data communication. The data transfer rate was on the order of tens of kbits/s. Other examples of technologies in 2G systems include Frequency Division Multiple Access (FDMA) and Code Division Multiple Access (CDMA).

The 2G networks deliver high quality and secure mobile voice and basic data services, such as fax and text messaging, along with full roaming capabilities around the world. Second-generation technology is in use by more than 10% of the world's population, and it is estimated that 1.3 billion customers in more than 200 countries and territories around the world use this technology (GSM_Home, 2005). The later advanced

Table 1. Comparison of mobile networks based on range and speed

Networks	Speed	Range and Coverage	Main Issues for M-Health
2nd Generation GSM	9.6 kilobits per second (KBPS)	Worldwide coverage, dependent on network operators roaming agreements	Bandwidth limitation, interference
High Speed Circuit Switched Data (HSCSD)	Between 28.8 KBPS and 57.6 KBPS	Not global, only supported by service provider's network	Not widely available, scarcity of devices
General Packet Radio Service (GPRS)	171.2 KBPS	Not global, only supported by service provider's network	Not widely available
EDGE	384 KBPS	Not global, only supported by service provider's network	Not widely available, scarcity of devices
UMTS	144 KBPS - 2 MBPS, depending on mobility	When fully implemented, should offer interoperability between networks, global coverage	Device battery life, operational costs
Wireless Local Area	54 MBPS	30-50 m indoors and 100-500 m outdoors. Must be in the vicinity of hot spot	Privacy, security
Personal Area Networks — Bluetooth	400 KBPS symmetrically 150-700 KBPS asymmetrically	10-100 m	Privacy, security, low bandwidth
Personal Area Networks — Zigbee	20 KBPS-250 KBPS, depending on band	30 m	Security, privacy, low bandwidth
WiMAX	Up to 70 MBPS	Approx. 40 m from base station	Currently no devices and network cards
RFID	100 KBPS	1 m Non-line of sight and contactless transfer of data between a tag and reader	Security, privacy
Satellite Networks	400 to 512 KBPS New satellites have potential of 155 MBPS	Global coverage	Data costs, shortage of devices with roaming capabilities, bandwidth limitations

technological applications are called 2.5G technologies and include networks such as General Packet Radio Service (GPRS) and Enhanced Data rates for GSM Evolution (EDGE). GPRS-enabled networks provide functionality such as always-on, higher capacity, Internet-based content and packet-based data services enabling services such as color Internet browsing, e-mail on the move, visual communications, multimedia messages and location-based services. Another complimentary

2.5G service is EDGE, which offers similar capabilities to the GPRS network.

Third-Generation Systems (3G). The most promising period is the advent of 3G networks, which also are referred to as the Universal Mobile Telecommunications Systems (UMTS). A significant feature of 3G technology is its ability to unify existing cellular standards, such as code-division multiple-access (CDMA), global system for mobile communications (GSM_Home, 2005), and time-division multiple-access (TDMA), under one umbrella (Istepanian, Jovanov, & Zhang, 2004). More than 85% of the world's network operators have chosen 3G as the underlying technology platform to deliver their third-generation services (GSM-Information, 2004). Efforts are underway to integrate the many diverse mobile environments and to blur the distinction between the fixed and mobile networks. The continual rollout of advanced wireless communication and mobile network technologies will be the major driving force for future developments in M-Health systems (Istepanian, Jovanov, & Zhang, 2004). Currently, the GSM version of 3G alone saw the addition of more than 13.5 million users, representing an annual growth rate of more than 500% in 2004. As of December 2004, 60 operators in 30 countries were offering 3GSM services. The global 3GSM customer base is approaching 20 million and already has been launched commercially in Africa, the Americas, Asia Pacific, Europe, and the Middle East (GSM_Home, 2005), thus making this technology ideal for developing affordable global M-Health systems.

Fourth Generation (4G). The benefits of the fourth-generation network technology (Istepanian et al., 2006; Olla, 2005a; Qiu, Zhu, & Zhang, 2002) include voice-data integration, support for mobile and fixed networking, and enhanced services through the use of simple networks with intelligent terminal devices. 4G also incorporates a flexible method of payment for network connectivity that will support a large number of network operators in a highly competitive environment. Over the last decade, the Internet has been dominated by non-real-time, person-to-machine communications (UMTS-Forum-Report14, 2002). The current developments in progress will incorporate real-time, person-to-person communications, including high-quality voice and video telecommunications along with extensive use of machine-to-machine interactions in order to simplify and to enhance the user experience.

Currently, the Internet is used solely to interconnect computer networks. IP compatibility is being added to many types of devices such as set-top boxes to automotive and home electronics. The large-scale deployment of IP-based networks will reduce the acquisition costs of the associated devices. The future vision is to integrate mobile voice communications and Internet technologies, bringing the control and multiplicity of Internet application services to mobile users (Olla, 2005b). 4G advances will provide both mobile patients and citizens with the choices that will fit their lifestyle and make it easier for them to interactively get the medical attention and advice they need when and where it is required and how they want it, regardless of any geographical barriers or mobility constraints.

Worldwide Interoperability for Microwave Access (WiMAX). WiMAX is considered to be the next generation of Wireless Fidelity (WiFi /Wireless networking technology that will connect you to the Internet at faster speeds and from much longer ranges than current wireless technology allows (http://wimaxxed.com/). WiMax has been undergoing testing and is expected to launch commercially by 2007. Research firm Allied Business Research predicts that by 2009, sales of WiMax accessories will top $1 billion (Kendall & Christopher, 2005), and Strategy Analytics predicts a market of more than 20 million WiMAX subscriber terminals and base stations per year by 2009 (ABI Research, 2005).

The technology holds a lot of potential for M-Health applications and has the capabilities to provide data rates up to 70 mbps over distances

up to 50 km. The benefits to both developing and developed nations are immense. There has been a gradual increase in the popularity of this technology. Intel recently announced plans to mass produce and release processors aimed to power WiMax-enabled devices (WiMax, 2005). Other technology organizations that are investing in the further advancement of this technology include Qwest, British Telecom, Siemens, and Texas Instruments. They aim to get the prices of the devices powered by WiMax to affordable levels so that the public can adopt them in large numbers, making it the next global wireless standard. There are already Internet Service Providers in metropolitan areas that are offering pre-WiMAX services to enterprises in a number of cities, including New York, Boston, and Los Angeles (WiMax, 2005).

Wireless Local Area Networks. Wireless Local Area Networks (WLAN) use radio or infrared waves and spread spectrum technology in order to enable communication between devices in a limited area. WLAN allows users to access a data network at high speeds of up to 54 Mb/s, as long as users are located within a relatively short range (typically 30–50m indoors and 100–500m outdoors) of a WLAN base station (or antenna. Devices may roam freely within the coverage areas created by wireless access points, the receivers, and transmitters connected to the enterprise network. WLANs are a good solution for healthcare today; in addition, they are significantly less expensive to operate than wireless WAN solutions such as 3G (Daou-Systems, 2001).

Personal Area Networks. A wireless personal area network (WPAN) (Personal area network, n.d.; Istepanian, Jovanov, & Zhang, 2004) is the interconnection of information technology devices within the range of an individual person, typically within a range of 10 meters. For example, a person traveling with a laptop, a personal digital assistant (PDA), and a portable printer could interconnect wirelessly all the devices using some form of wireless technology. WPANs are defined by IEEE standard 802.15 (IEEE 802.15 Working Group for WPAN, n.d.). The most relevant enabling technologies for M-Health systems are Bluetooth (n.d.) and ZigBee Alliance (n.d.). ZigBee is a set of high-level communication protocols designed to use small, low-power digital radios based on the IEEE 802.15.4 standard for wireless personal area networks. ZigBee is aimed at applications with low data rates and low power consumption. ZigBee's current focus is to define a general-purpose, inexpensive, self-organizing network that can be shared by industrial controls, medical devices, smoke and intruder alarms, building automation, and home automation. The network is designed to use very small amounts of power so that individual devices might run for a year or two with a single alkaline battery, which is ideal for use in small medical devices and sensors. The Bluetooth specification was first developed by Ericsson and later formalized by the Bluetooth Special Interest Group established by Sony, Ericsson, IBM, Intel, Toshiba, and Nokia and later joined by many other companies. A Bluetooth WPAN is also called a *piconet* and is composed of up to eight active devices in a master-slave relationship. A piconet typically has a range of 10 meters, although ranges of up to 100 meters can be reached under ideal circumstances. Implementations with Bluetooth versions 1.1 and 1.2 reach speeds of 723.1 kbit/s. Version 2.0 implementations feature Bluetooth Enhanced Data Rate (EDR) and thus reach 2.1 Mbit/s (Bluetooth, n.d.; Wikipedia, n.d.).

Radio Frequency Identification (RFID). RFID systems consist of two key elements: a tag and a reader/writer unit capable of transferring data to and from the tag. An antenna linked to each element allows power to be transferred between the reader/writer and a remotely sited tag through inductive coupling. Since this is a bi-directional process, modulation of the tag antenna will be reflected back to the reader's /writer's antenna, allowing data to be transferred in both directions.

Some of the advantages of RFID that make this technology appealing to the healthcare sector are as follows:

- No line of sight required between tag and reader.
- Non-contact transfer of data between tag and reader.
- Tags are passive, which means no power source is required for the tag component.
- Data transfer range of up to one meter is possible.
- Rapid data transfer rates of up to 100 Kbits/sec.

The use of RFID in the healthcare environment is set to rise and is currently being used for drug tracking. RFID technology is expected to decrease counterfeit medicines and to make obtaining drugs all the more difficult for addicts (Weil, 2005). There are also applications that allow tagging of patients, beds, and expensive hospital equipment.

Satellite Technologies. Satellite broadband uses a satellite to connect customers to the Internet. Two-way satellite broadband uses a satellite link both to send and to receive data. Typical download speeds are 400 to 512 kbps, while upload speeds on two-way services are typically 64 to 128 kbps. Various organizations (Inmarsat announces availability of the 64 kbit/s mobile office in the sky, 2000) have been investigating the development of an ultra-high-data-rate Internet test satellite for making a high-speed Internet society a reality (JAXA, 2005). Satellite-based telemedicine service will allow real-time transmission of electronic medical records and medical information anywhere on earth. This will make it possible for doctors to diagnose emergency patients even from remote areas and also will increase the chances of saving lives by receiving early information, as ambulance data rates of 155 mbps are expected. One considerable drawback associated with using this technology is cost (Olla, 2004).

This section has summarized the various mobile network technologies that are being used in the healthcare sector. The mobile technologies described have a significant impact on the ability to deploy M-Healthcare application and systems; however, there are combinations of other important factors described in the next section. The next section will present a five-layered model that uses mobile network technologies along with device type, data display, application purpose, and application domain to categorize M-Health systems.

THE M-HEALTH REFERENCE MODEL

The financial cost of delivering quality healthcare is increasing exponentially not only in North America but also around the world. To satisfy increasing healthcare challenges, organizations in the healthcare sector are investing innovative technological solutions in order to meet the higher expectations placed on practice management (Wickramasinghe & Misra, 2005; Wickramasinghe & Silvers, 2002). The evolution of M-Health solutions, such as the use of wireless networks to access patient records and other healthcare services such as billing and prescription, are becoming popular, especially in the U.S. and Canada (Goldberg & Wickramasinghe, 2002) and in many European countries (Sorby, 2002). However, these wireless solutions also bring with them their own challenges, with security being one of the most important issues in M-Health. This is because M-Health systems transmit highly sensitive information such as patient data over cyberspace, and there is a need for high-level, end-to-end security, confidentiality, and privacy.

This section presents a reference model to describe each M-Health system in order to address the security challenges. Specifically, we present a reference model and show how our model can facilitate a higher level of end-to-end security

in a wireless or mobile healthcare environment. Despite the advancements of wireless technologies, the use of wireless security in healthcare is in its infancy, and a robust mobile healthcare model needs to be developed in order to allow further innovation of M-Health applications and services that will fulfill the growing needs of the healthcare sector.

This section presents a reference model that can be utilized to classify M-Health systems by identifying the system components for a specific M-Health system. As previously indicated, the approach used to develop the reference model concentrated on identifying different dimensions of a Mobile Healthcare Delivery System (MHDS), which provides the capability to identify user security requirements for different categories in an organized manner. The reference model was formulated by reviewing the literature, and the model reflects an amalgamation of various classification schemes proposed by previous experts to categorize telemedicine and telehealth systems. The following subsections will provide a description of these five domains for the M-Health reference model illustrated in Figure 1. These domains include communication infrastructure, device type, data display, application purpose, and application domain.

Mobile Communication Infrastructure

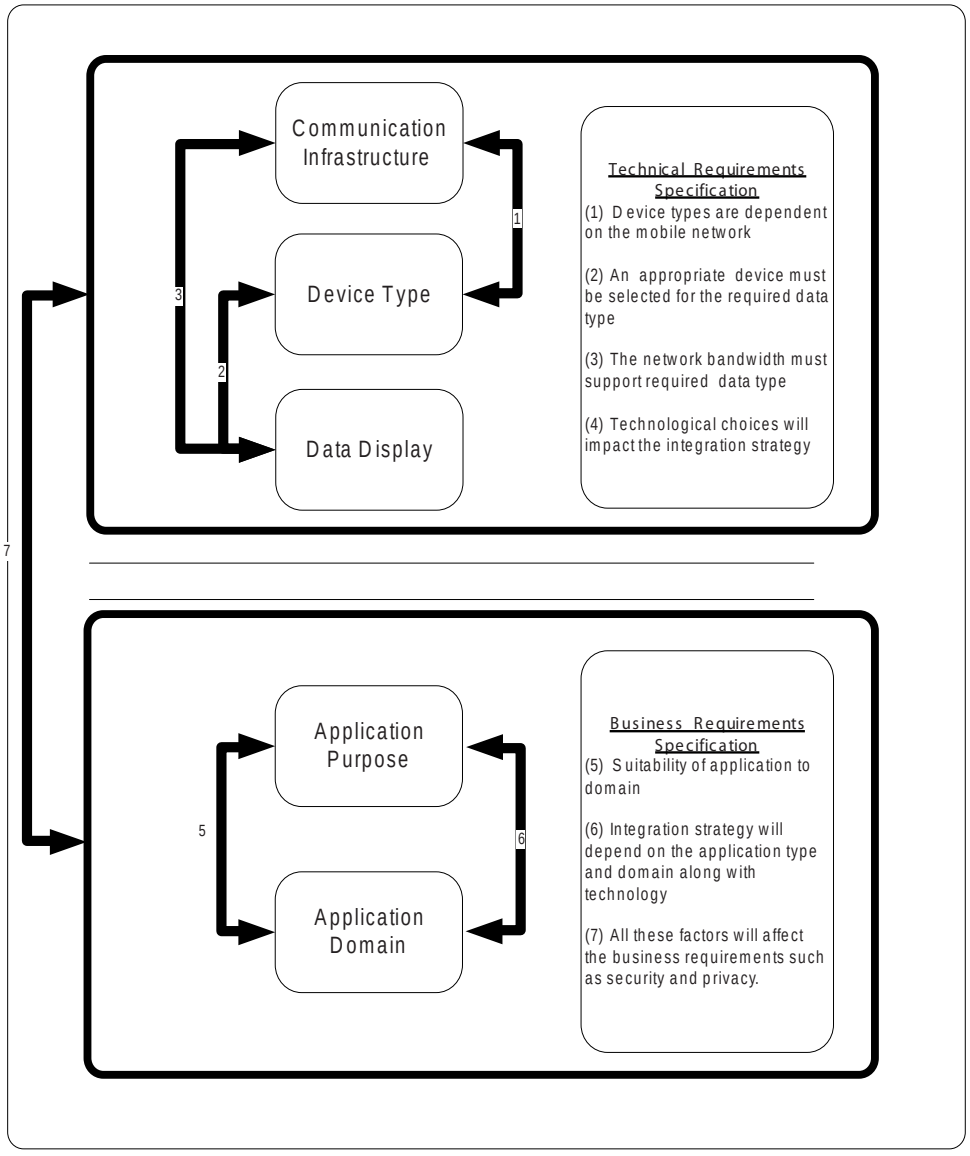
The first dimension defines the mobile infrastructure used to transmit, encode, and receive data. There are numerous wireless infrastructures available from which healthcare providers can choose. Examples of networks that provide personal area networks include Bluetooth and Zigbee. Mobile networks that provide connectivity within buildings include Wireless Local Area Networks (WLAN), which use different protocols from the standard digital mobile technologies such as 2G, 2.5, and 3G, which provide wide area connectivity. A summary of these technologies was provided in "Mobile Healthcare Delivery System Networks."

Device Type

The second section is referred to as Device Type. Due to the advances in medical sensor technologies and handheld computers, M-Health has the potential to deliver services beyond the scope of mobile telemedicine discussed in the previous section. The integration of medical sensors with ICT allows physicians to diagnose, monitor, and treat patients remotely without compromising standards of care. Advances of new materials and signal processing research is allowing the design of smart medical sensors that utilize real-time data recording and processing of multi-physiological signals.

There is a multiplicity of sensors that are available commercially, such as piezo-electrical materials for pressure measurements, infrared sensors for body temperature estimation, and optoelectronic sensors that monitor heart rate and blood pressure. These sensors are being embedded into wearable items and accessories such as sunglasses and rings that can be carried easily. With the continual improvements to the sensors and the miniaturization of computing devices, these wearable devices for monitoring, diagnosing, and treating illnesses are becoming more readily available. Hitachi has developed a wristwatch that can measure the wearer's health conditions such as pulse, moves, and body temperature. The monitoring wristwatch can send data wirelessly via the Internet to remote monitoring services. Hitachi aims to launch the wireless wristwatch commercially within three years (Hitachi's wireless health monitoring wristwatch, 2005), targeting the healthcare service industry for elderly people. The benefit of wearable sensors for physiological data collection is that they can be used to monitor human health constantly without unsettling the patient's day-to-day activities. Examples include rings, vests, watches, wearable intelligent sensors,

Figure 1. M-health reference model for identifying components of healthcare delivery systems



and systems for e-medicine (Winters, Wang, & Winters, 2003). Wireless communication such as WAN is used for transmitting information and accessing healthcare databases wherever appropriate in order to allow free movement of the user (Jovanov et al., 2005).

Personal Digital Assistants (PDAs) have been popular among healthcare practitioners in the last few years. PDA penetration among physicians is at 25%, much higher than the 4% penetration for the general population (Fischer et al., 2003; Holmes, Brown, & Twist, 2001; Harris-Interactive, 2001; Parmanto et al., 2005). Studies have shown that the integration of PDAs into clinical practice has led to decreased medication error rates (Grasso et al., 2002) and the improvement in physicians' adherences to clinical practice guidelines (Shiffman, 1999, 2000).

Although the PDA has great potential to support evidence-based medicine, there are some considerable drawbacks (Parmanto et al., 2005). The small screen is a poor match for information resources designed for full-sized desktop computers. Presenting vast amounts of information in the limited space of the PDA display is a significant technical barrier to the realization of the PDA's potential (Fischer et al., 2003; Larkin, 2001; Peterson, 2002). Other technical PDA drawbacks include low resolution, limited memory, slow processor, and problematic data input. An alternative for the PDA is a tablet PC. These are being adopted in the healthcare sector for information capture at the patient's bedside.

Data Display

This key dimension describes how the information from the M-Health application is to be processed and transmitted. The chosen delivery options can have an important effect on the final quality of the telemedicine event and the outcome. Delivery options in telemedicine can be categorized under two main groups, according to Tulu and Chatterjee (2005): (1) synchronous and (2) asynchronous. Information transactions that occur between two or more participants simultaneously are called synchronous communications. In asynchronous communications, these transactions occur at different points in time. The data displays for both synchronous and asynchronous presentations can be grouped into the following categories: text, data, video, and multimedia (combination).

Application Purpose

The fourth dimension is the Application Purpose. This domain describes the intended use of the application. Field (1996) categorized this field into two main groups: clinical and non-clinical systems.

In 1996, the Committee on Evaluating Clinical Applications of Telemedicine (1996) grouped clinical applications into five categories:

1. Initial urgent evaluation
2. Supervision of primary care
3. Provision of specialty care
4. Consultation
5. Monitoring

Due to rapid advances in the M-Health field, three additional categories were added by Tulu and Chatterjee (2005):

6. Use of remote information and decision analysis resources to support or guide care for specific patients
7. Diagnostic
8. Treatment (surgical and non-surgical)

In addition to these groups, two additional categories have been added by the authors in order to reflect future trends of M-Health systems:

9. Drug delivery
10. Patient identification

The use of M-Health systems for non-clinical purposes includes medical education and administrative duties that do not involve decisions about care for particular patients. Some examples of M-Health non-clinical applications include:

1. Mobile access to the latest drug reference database
2. Bedside access to patient records (increase efficiency by reducing demand for paper records)
3. ePrescribing (mobile prescription writing and verification of drug interactions)
4. Prescription formulary reference (electronically identify most economic pharmaceuticals for a patient)

5. Electronic billing for in-home healthcare workers
6. Patient/drug verification (scan patient and drug bar codes to help to ensure that the appropriate medicine is being administered to the correct patient)
7. Delivery applications (healthcare supply delivery, tracking, and billing)
8. Patient encounter data capture

Application Area

The final dimension is called Application Area. This dimension describes the medical field implementing the M-Health technology. The application area also can be subdivided into clinical and non-clinical use. Clinical use relates to medical departments such as emergency, ophthalmology, pediatrics, and surgery. Non-clinical use relates to maintenance areas such as, charge capture, billing, and administration. The importance of this domain is to highlight the differences in the environments and to identify procedures that are specific to a particular healthcare domain.

In summary, the reference model highlights some of the important elements of an M-Health system. It is important to understand all of the dimensions of the reference model for each mobile healthcare system. Using the previous dimensions to describe an application will allow an M-Health system to be broken down into meaningful components.

The next section presents a case that features the use of an M-health application that uses a PDA to collect private medical data in remote African communities. The case in the next section illustrates how the reference model can be used to derive the components of an M-Health system, which allows issues such as security, privacy, hosting requirements, interference of devices, and integration to be investigated simultaneously.

OVERVIEW OF BLOOD DONOR RECRUITMENT (BDR) PROJECT

As an increased number of healthcare industry professionals adopt mobile-enabled handheld devices to collect, store, and retrieve critical medical information, the need for security has become a top-priority IT challenge. These mobile benefits come with immense corporate and regulatory risk. PDA devices left unsecured while electronic health information is being hotsynced to a PC can become a primary source for the intentional malicious interception of confidential information. Furthermore, the recent adoption of the Patient Privacy and Federal Health Insurance Portability and Accountability Act (HIPPA) is a strong reason not to have protected health information on any unsecured devices, especially a PDA. This section will provide an example of how the reference model described in the previous sections can be used to describe a project initiated by the Red Cross.

The medical informatics data analyst/IT manager of the Uganda Red Cross Society identified a potential weakness in the use of mobile devices in the community health field. One of the authors was consulted to assist in developing a model that addressed data security in order to ensure the main IT security goals: confidentiality, integrity, and availability. Mobile devices were used in Blood Donor Recruitment (BDR) activities to register blood donors' details. The medical history was stored for blood screening purposes. The data are very confidential because some of the blood donors' results are positive for HIV, hepatitis, or syphilis. The system specifications are as follows:

Palm M125 with the following software:

- Palm Desktop software for Windows v. 4.0

- Pendragon form 3.2
- Pendragon Distribution Toolkit 2.0
- Windows XP on the PC
- Microsoft Access 2000
- The volume of data is about 10mb records on the PDA

This information is transferred to the computer server either by wired synchronization (HotSynced) or by wireless network. The hotsync option requires the user to be in the confines of the hospital or office, because it involves the use of the wired cradle. The other alternative that is being considered is the use of a wireless network such as WiFi or GSM to transfer data to the computer. Once the data are stored on the computer, they are manipulated with various software tools.

Using the M-Health reference model described in the previous section, the BDR system was broken down into the various components, as highlighted in Tables 2a and 2b. The following general recommendations were suggested to the client.

1. Upgrade memory on mobile device.
2. Formulate an enforceable end-to-end security policy.
3. Evaluate encryption software that can run both on the PDA and on the PC
4. Implement a solution that ensures that captured data on the PDA can be encrypted before being hotsynced on the PC.
5. Try out software prior to purchase.

There are various software solutions that can be used in order to secure the health system described previously. These solutions vary by encryption method, price, and device compatibility; however, there is no single solution to securing a system. It is important that each scenario be addressed differently, irrespective of how similar they seem.

A software solution that would fulfill the requirements for the scenarios described in Tables 2a and 2b, the PDASecure™by Trust Digital software, provides six different selectable encryption algorithms. Additional protection features include strong password protection to prevent unauthorized synchronizing or beaming of data, unauthorized deletion of files due to viruses or malicious code, and user authentication to devices before data can be decrypted. IT administrators can control who can access data and networks with wireless handheld devices, encrypt 100% of the data, and password-protect devices so that they are useless if stolen or compromised.

Using mobile networks allows data to travel over the open air. When data transmission is wireless, there is a possibility of interception of the radio transmission as well as unauthorized access to the hospital system. Poorly designed WiFi local area networks can be leaky and accessible beyond the intended boundary of use. If an unauthorized access of any kind occurs, it not only could lead to the loss of a patient's privacy but also could lead to other potential consequences.

In Table 2a, the wireless network is not used, and the mobile device must be colocated with the computer system and server in order to transfer and copy data. This option also presents a different set of security problems. Due to the storage capacities of the mobile devices and the growing trend of today's mobile workforce, every port, external disk drive, or JumpDrive can become a security risk. In the corporate environment, IT experts have turned to soldering and gluing USB ports in order to prevent intrusion; they also have installed titanium chastity belts around computers. USB ports and PDA cradles can be used for a variety of functions from input devices as inconspicuous as iPODs, which now are capable of downloading more than 30 gigabits of data.

In summary, using the reference model presented in this article to decompose a Red Cross blood donor system ensured that the security issues were identified and appropriately addressed. The model also was valuable for discovering technical specifications of the system and for understanding

Table 2a. Using the reference model to describe the blood donor system using a cradle solution

System	Purpose	Data	Network	Device	Application Area
Blood Donor Recruitment (BDR)	Blood donor data capture and transfer	Textual data	Fixed cable	Palm PDA	Rural community health
Security Model • Data sent from the PDA and to the PDA must be secured with encryption. • User and server authentication must be in place. • Access to data in the server from a computer or PDA different from where the server runs must be secured with encryption and user/server authentication. • Advances authentication methods such as biometrics, which will provide added security features.					

Table 2b. Using the reference model to describe the blood donor system using a wireless solution

System	Purpose	Data	Network	Device	Application Area
Blood Donor Recruitment (BDR)	Blood donor data capture and transfer	Textual data	WIFI/ GPRS	Palm PDA	Rural community health
Security Model • Data sent to/from the PDA must be secured with encryption. • User and server authentication must be in place. • No data storage in network operator environment (GPRS/UMTS Operator). • WiFi network must be secured.					

what upgrades were required in order to reduce the vulnerabilities of the system. In the future, this model can be improved by the creation of a set of guidelines and standards that are appropriate for a particular type of M-Health system. This will be valuable for governments and private vendors that implement innovative healthcare systems. The model also aids the healthcare stakeholders to identify the technological infrastructure, business requirements, and operational needs for the different types of M-Health systems.

CONCLUSION

In the clinical domain, considerable leaps in the fields of biomedical telemetry, nanotechnology, ICT, and drug delivery techniques will encourage M-Health applications to evolve rapidly over the next decade. These new M-Health applications will take advantage of technological advances, such as device miniaturizations, device convergence, high-speed mobile networks, reduction in power consumption, and improved medical sensors. This will lead to the increased diffusion of clinical M-Health systems, which will impact the reshaping of the healthcare sector.

The non-clinical use of M-Health enterprise systems also will see an increased rate of adoption due to the potential benefits of the mobile solutions. Potential quality improvements may be achieved by implementing platforms that contain functionality that allows organizations to perform clinical results viewing, e-Prescribing, medication administration, specimen collection, charge capture, and physician dictation capture. These systems help to reduce errors by making the patient's medical record and appropriate medical knowledge available at the point of care and at the point of decision.

The reference model presented in this article aids in the breakdown of M-Health systems into the following five key dimensions: (1) Communication Infrastructure — describes mobile telecommunication technologies and networks; (2) Device Type — relates to the type of device being used, such as PDA, sensor, or tablet PC; (3) Data Display — describes how data will be displayed to the user and transmitted, such as images, e-mail, or textual data; (4) Application Purpose — identification of the objective for the M-Health system; and (5) Application Domain — definition of the area in which the system will be implemented.

The reference model described in this article encapsulates the emerging M-Health field and is needed to assist in clarifying the rapid progress in this field. Currently, the M-Health field is disjointed, and it is often unclear what component constitutes an M-Health system. This is addressed by the reference model. Further work is needed in order to identify the appropriate standards and guidelines for implementing the various types of M-Health systems. This would benefit vendors and system implementers by allowing them to understand the various conditions that may apply to a system that uses a specific set of the M-Health reference model variables. The reference model also would benefit from further work that defines how the choice of dimensions can impact the business models and the security policy for the implementation of the system.

REFERENCES

ABI Research. (2005). *WIFI/WIMAX*. Retrieved from http://www.abiresearch.com/category/Wi-Fi_WiMAX

Akyildiz, I., & Rudin, H. (2001). Pervasive computing. *Computer Networks — The International Journal of Computer and Telecommunications Networking, 35*(4), 371-371.

Amy, C., & Richards, G. (2004). A BioMEMS review: MEMS technology for physiologically integrated devices. *Proceedings of the IEEE, 92*(1), 6-21.

Bashshur, R. L. (2002). Telemedicine/telehealth: An international perspective. *Telemedicine and e-Health, 8*(1), 5-12.

Bashshur, R. L., Reardon, T. G., & Shannon, G. W. (2000). Telemedicine: A new health care delivery system. *Annual Review of Public Health, 21*(1), 613-637.

Bluetooth. (n.d.). Retrieved from http://www.bluetooth.org/

Budinger, T. F. (2003). Biomonitoring with wireless communications. *Annual Review of Biomedical Engineering, 5*, 412.

Committee on Evaluating Clinical Applications of Telemedicine. (1996). *Telemedicine: A guide to assessing telecommunications in health care.*

Daou-Systems. (2001). *Going mobile: From eHealth to mHealth*. Retrieved from http://www.daou.com/emerging/pdf/mHealth_White_Paper_April_2001.PDF

Holmes, B. J., Brown, E. G., & Twist, A. E. (2001, June 25). Doctors connect with handhelds. *The Forrester Report*. Retrieved September 2005,

from http://www.forrester.com/ER/Research/Report/Summary/0,1338,11449,00.html

Field, M. J. (1996). *Telemedicine: A guide to assessing telecommunications in health care.* Paper presented at the National Academy Press, Washington, DC.

Fischer, S., Stewart, T. E., Mehta, S., Wax, R., & Lapinsky, S. E. (2003). Handheld computing in medicine. *Journal of the American Medical Informatics Association, 10*(2), 139-149.

Future trends: Convergence is all. (2005, March 12). *NewScientist*, 53. Retrieved January 2006, from http://www.sciencejobs.com/data/pdf/insider/insider129.pdf

Godoe, H. (2000). Innovation regimes, R&D and radical innovations in telecommunications. *Research Policy, 29*(9), 1033-1046.

Goldberg, S., & Wickramasinghe, N. (2002). *Mobilizing healthcare*. Paper presented at the 5th ICECR Conference, Montreal.

Grasso, B. C., Genest, R., Yung, K., & Arnold, C. (2002). Reducing errors in discharge medication lists by using personal digital assistants. *Psychiatric Services, 53*(45), 1325-1326.

Hatcher, M., & Heetebry, I. (2004). Information technology in the future of health care. *Journal of Medical Systems Issue, 28*(6), 673-688.

Health Insurance Portability and Accountability Act of 1996 (HIPPA). (2005). United States Department of Health and Human Services. Retrieved September 30, 2005, from http://www.hhs.gov/ocr/hipaa/

Hitachi's wireless health monitoring wristwatch. (2005). Nikkei-Business-Publications. Retrieved from http://techon.nikkeibp.co.jp/english/NEWS_EN/20050526/105103/?ST=english

IEEE 802.15 Working Group for WPAN. (n.d.). IEEE 802.15. Retrieved from http://www.ieee802.org/15/

Inmarsat announces availability of the 64kbit/s mobile office in the sky. (2000). Inmarsat Swift64. Retrieved from http://www.inmarsat.com/swift64/press_1.htm

Istepanian, R. S. H., & Lacal, J. (2003). *M-Health systems: Future directions.* Paper presented at the 25th Annual International Conference of IEEE Engineering Medicine and Biology, Cancun, Mexico.

Istepanian, R. S. H., & Laxminaryan, S. (2000). UNWIRED, the next generation of wireless and Internetable telemedicine systems [Editorial paper]. *IEEE Transactions on Information Technology in Biomedicine, 4*(3), 189-194.

Istepanian, R. S. H., Laxminarayan, S., & Pattichis, C. (Eds.). (2006). *M-health: Emerging mobile health systems.* London: Springer-Verlag.

Istepanian, R. S. H., & Wang, H. (2003). Telemedicine in UK. In L. Beolchi (Ed.), *European telemedicine glossary of concepts, standards, technologies and users* (5th ed.) (pp. 1159-1165). Brussels, Belgium: European Commission Information Society Directorate-General.

Istepanian, R. S. H., Jovanov, E., & Zhang, Y. T. (2004). M-health: Beyond seamless mobility and global wireless healthcare connectivity [Editorial paper]. *IEEE Transactions on Information Technology in Biomedicine, 8*(4), 405-412.

JAXA. (2005). Aerospace Exploration Agency.

Jovanov, E., Lords, A., Raskovic, D., Cox, P., Adhami, R., & Andrasik, F. (2003). Stress monitoring using a distributed wireless intelligent sensor system. *IEEE Engineering in Medicine and Biology Magazine, 22*(3), 49-55.

Jovanov, E., Milenkovic, A., Otto, C., & Groen, P. C. (2005). A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. *Journal of NeuroEngineering and Rehabilitation, 2*(6). Retrieved Janu-

ary 17, 2006, from http://www.jneuroengrehab.com/content/2/1/6

Kendall, P., Christopher, T. (2005, May 12). *WiMAX — 3G killer or fixed broadband wireless standard?* Retrieved September 2005, from http://www.strategyanalytics.net/default.aspx?mod=ReportAbstractViewer&a0=2393

Larkin, M. (2001). Can handheld computers improve the quality of care? *Lancet, 358*(9291), 1438.

GSM_Home. (2005). Retrieved from http://www.gsmworld.com/index.shtml

GSM-Information. (2004). Retrieved from http://www.gsmworld.com/index.shtml

Moore, S. K. (2002). Extending healthcare's reach: Telemedicine can help spread medical expertise around the globe. *IEEE Spectrum, 39*(1), 66-71.

Olla, P. (2004). A convergent mobile infrastructure: Competition or co-operation. *Journal of Computing and Information Technology, 12*(4), 309-322.

Olla, P. (2005a). Evolution of GSM network technology. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (pp. 290-294). Hershey, PA: Idea Group Reference.

Olla, P. (2005b). Incorporating commercial space technology into mobile services: Developing innovative business models. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities* (pp. 82-113). Hershey, PA: IRM Press.

Olla, P., & Patel, N. (2003). Framework for delivering secure mobile location information. *International Journal of Mobile Communications, 1*(3), 289-300.

Parmanto, B., Saptono, A., Ferrydiansyah, R., & Sugiantara, W. (2005). *Transcoding biomedical information resources for mobile handhelds.* Paper presented at the 38th Hawaii International Conference on System Sciences, Hawaii.

Pattichis, C. S., Kyriacou, E., Voskarides, S., Pattichis, M. S., Istepanian, R. S. H., & Schizas, C. N. (2002). Wireless telemedicine systems: An overview. *IEEE Transaction of Antennas Propagation Magazine, 44*(2), 143-153.

Personal area network. (n.d.). *IBM mobile computing.* Retrieved from http://www.research.ibm.com/topics/popups/smart/mobile/html/phow.html

Peterson, M. F. T. (2002). *The right information at the point of care library delivery via hand-held computers.* Paper presented at the EAHIL Conference of Health and Medical Libraries, Cologne.

Physicians' use of handheld personal computing devices increases from 15% in 1999 to 26% in 2001. (2001). *Harris Interactive Health Care News, 1*(25). Retrieved September 2005, from http://www.harrisinteractive.com/news/allnewsbydate.asp?NewsID=345

Qiu, R. C., Zhu, W., Zhang, Y. Q. (2002, May 26-29). *Third-generation and beyond (3.5G) wireless networks and its applications.* Paper presented at the IEEE International Symposium on Circuits and Systems (ISCS), Scottsdale, Arizona.

Raskovic, D., Martin, T., & Jovanov, E. (2004). Medical monitoring applications for wearable computing. *Computer Journal, 47*(4), 495-504.

Shiffman, R. N. (1999). *User satisfaction and frustration with a handheld, pen-based guideline implementation system for asthma.* Paper presented at the AMIA Symp.

Shiffman, R. N. (2000). A guideline implementation system using handheld computers for office management of asthma: Effects on adherence and patient outcomes. *Pediatrics, 105*(4), 767-773.

Sorby, I. (2002). Characterising cooperation in the ward — A framework for producing require-

ments to mobile electronic healthcare records. In *Proceedings of the 2nd Hospital of the Future Conference*, Chicago, IL. Retrieved January 2005, from http://www.hctm.net/events/papers_2002.html

Tachakra, S., Istepanian, R. S. H., Wang, H., & Song, Y. H. (2003). Mobile e-health: The unwired evolution of telemedicine. *Telemedicine and E-Health Journal, 9*(3), 247-257.

Tulu, B., & Chatterjee, S. (2005). *A taxonomy of telemedicine efforts with respect to applications, infrastructure, delivery tools, type of setting and purpose.* Paper presented at the 38th Hawaii International Conference on System Sciences, Hawaii.

UMTS-Forum-Report14. (2002). *Support of third generation services using UMTS in a converging network environment.* UMTS Forum. Retrieved from http://www.umts-forum.org/servlet/dycon/ztumts/umts/Live/en/umts/Resources_Reports_index

Warren, S. (2003). *Beyond telemedicine: Infrastructures for intelligent home care technology.* Paper presented at the Pre-ICADI Workshop Technology for Aging, Disability, and Independence, London.

Webb, C. E. (2004). Chip shots. *IEEE Spectrum, 41*(10), 48-53.

Weil, N. (2005). Companies announce RFID drug-tracking project: Unisys and SupplyScape plan to track Oxycontin through the supply chain. *Computerworld*. Retrieved January 2006, from http://www.computerworld.com/softwaretopics/erp/story/0,10801,102142,00.html

Wickramasinghe, N., & Misra, S. K. (2005). A wireless trust model for healthcare. *International Journal of Electronic Healthcare, 1*(1), 90-110.

Wickramasinghe, N., & Silvers, J. B. (2003). IS/IT the prescription to enable medical group practices to manage managed care. *Health Care Management Science, 6*(2), 75-86.

Wikipedia. (n.d.). Encyclopedia. Retrieved from http://en.wikipedia.org/wiki

WiMax. (2005). Retrieved September, 2005, from http://wimaxxed.com

Winters, J. M., Wang, Y., & Winters, J. W. (2003). Wearable sensors and telerehabilitation: Integrating intelligent telerehabilitation assistants with a model for optimizing home therapy. *IEEE Engineering in Medicine and Biology Magazine, 1*(22), 56-65.

ZigBee Alliance. (n.d.). Retrieved from http://www.zigbee.org/

This work was previously published in International Journal of Healthcare Information Systems and Informatics, Vol. 1, Issue 2 edited by J. Tan, pp. 1-19, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.6
Design Methodology for Mobile Information Systems

Zakaria Maamar
Zayed University, UAE

Qusay H. Mahmoud
University of Guelph, Canada

INTRODUCTION

Mobile information systems (MISs) are having a major impact on businesses and individuals. No longer confined to the office or home, people can use devices that they carry with them, along with wireless communication networks, to access the systems and data that they need. In many cases MISs do not just replace traditional wired information systems or even provide similar functionality. Instead, they are planned, designed, and implemented with the unique characteristics of wireless communication and mobile client use in mind. These unique characteristics feature the need for specific design and development methodologies for MISs. Design methods allow considering systems independently of the existing information technologies, and thus enable the development of lasting solutions. Among the characteristics that a MIS design method needs to consider, we cite: unrestricted mobility of persons, scarcity of mobile devices' power-source, and frequent disconnections of these devices.

The field of MISs is the result of the convergence of high-speed wireless networks and personal mobile devices. The aim of MISs is to provide the ability to compute, communicate, and collaborate anywhere, anytime. Wireless technologies for communication are the link between mobile clients and other system components. Mobile client devices include various types, for example, mobile phones, personal digital assistants, and laptops. Samples of MIS applications are mobile commerce (Andreou et al., 2002), inventory systems in which stock clerks use special-purpose mobile devices to check inventory, police systems that allow officers to access criminal databases from laptops in their

patrol cars, and tracking information systems with which truck drivers can check information on their loads, destinations, and revenues using mobile phones. MISs can be used in different domains and target different categories of people.

In this article, we report on the rationale of having a method for designing and developing mobile information systems. This method includes a conceptual model, a set of requirements, and different steps for developing the system. The development of a method for MISs is an appropriate response to the need of professionals in the field of MISs. Indeed, this need is motivated by the increased demand that is emerging from multiple bodies: wireless service providers, wireless equipment manufacturers, companies developing applications over wireless systems, and businesses for which MISs are offered. Besides all these bodies, high-speed wireless data services are emerging (e.g., GPRS, UMTS), requiring some sort of new expertise. A design and development method for MISs should support professionals in their work.

MOBILE COMPUTING MODEL

The general mobile computing model in a wireless environment consists of two distinct sets of entities (Figure 1): mobile clients (MCs) and fixed hosts. Some of the fixed hosts, called mobile support stations (MSSs), are enhanced with wireless interfaces. An MSS can communicate with the MCs within its radio coverage area called wireless cell. An MC can communicate with a fixed host/server via an MSS over a wireless channel. The wireless channel is logically separated into two sub-channels: an uplink channel and a downlink channel. The uplink channel is used by MCs to submit queries to the server via an MSS, whereas the downlink channel is used by MSSs to disseminate information or to forward the responses from the server to a target client. Each cell has an identifier (CID) for identification purposes. A CID is periodically broadcasted to all the MCs residing in a corresponding cell.

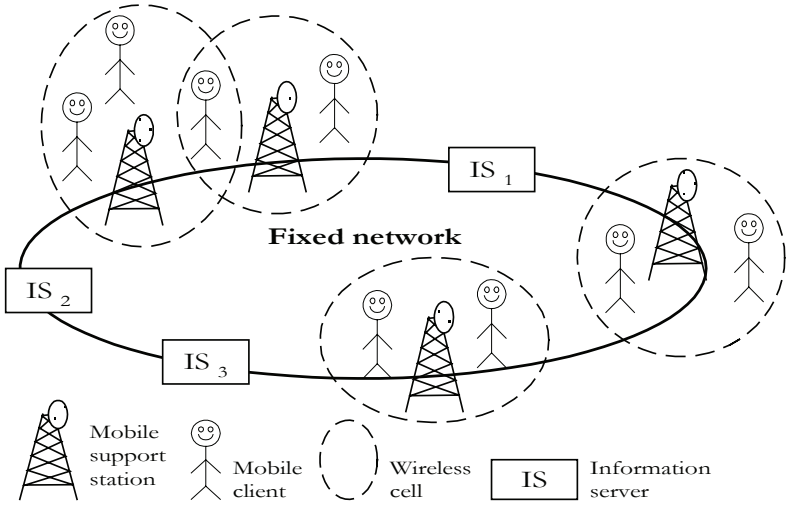
The wireless application protocol (WAP) is a technology that plays a major role in the field deployment of the mobile computing model (Open Mobile Alliance). WAP is an open, global specification that empowers users with mobile devices to easily access and interact with information and services instantly. It describes how to send requests and responses over a wireless connection, using the wireless session protocol (WSP), which is an extended and byte-coded version of HTPP 1.1. A WSP request is sent from a mobile device to a WAP gateway/proxy to establish an HTTP session with the target Web server. Over this session, the WSP request, converted into HTTP, is sent. The content, typically presented in the Wireless Markup Language (WML), is sent back to the WAP gateway, where it is byte-coded and sent to the device over the WSP session.

REQUIREMENTS FOR MISs

The role of an MIS is to provide information to mobile users through wireless communication networks. Two aspects are highlighted here: information and network. Information has to be available, taking into account terrain and propagation techniques. Plus, the information exchange has to be secured. A security problem inherent to all wireless communication networks consists of third parties being able to easily capture the radio signals while in the air. Thus, appropriate data protection and privacy safeguards must be ensured. Regarding the network element, this latter needs to consider failure cases and recover from them.

1. **Information Availability Requirement:** This illustrates the need for a user to have uninterrupted and secure access to information on the network. Aspects to consider are: survivability and fault tolerance, abil-

Figure 1. Representation of the mobile computing model



ity to recover from security breaches and failures, network design for fault tolerance, and design of protocols for automatic reconfiguration of information flow after failure or security breach.

2. **Network Survivability Requirement:** This illustrates the need to maintain the communication network "alive" despite of potential failures. Aspects to consider are: understand system functionality in the case of failures, minimize the impact of failures on users, and provide means to overcome failures.
3. **Information Security Requirement:** This illustrates the importance of providing reliable and unaltered information. Aspects to consider are: confidentiality to protect information from unauthorized disclosure, and integrity to protect information from unauthorized modification and ensure that information is accurate, complete, and can be relied upon.
4. **Network Security Requirement:** This illustrates the information security using network security. Aspects to consider are: confidentiality, sender authentication, access control, and identification.
5. **Additional Requirements of MIS Have Been Put Forward:** Indeed, the increasing reliance and growth in information-based wireless services impose three requirements—availability, scalability, and cost efficiency—on the services to be provided. Availability means that users can count on accessing any wireless service from anywhere, anytime, regardless of the site, network load, or device type. Availability also means that the site provides services meeting some measures of quality such as short, acceptable, and predictable response time. Scalability means that service providers should be able to serve a fast-growing number of customers with minimal performance degradation. Finally, cost effectiveness means that the quality of wireless services (e.g., availability, response time) should come with adequate expenditures in IT infrastructure and personnel.

CHALLENGES AND POSSIBLE SOLUTIONS IN MISs

The requirements discussed above pose several crucial challenges, which must be faced in order for MIS applications to function correctly in the target environment.

- **Transmission Errors:** Messages sent over wireless links are exposed to interference (and varying delays) that can alter the content received by the user, the target device, or the serer. Applications must be prepared to handle these problems. Transmission errors may occur at any point in a wireless transaction and at any point during the sending or receiving of a message. They can occur after a request has been initiated, in the middle of the transaction, or after a reply has been sent.
- **Message Latency:** Message latency, or the time it takes to deliver a message, is primarily affected by the nature of each system that handles the message, and by the processing time needed and delays that may occur at each node from origin to destination. Message latency should be handled, and users of wireless applications should be kept informed of processing delays. It is especially important to remember that a message may be delivered to a user long after the time it is sent. A long delay might be due to coverage problems or transmission errors, or the user's device might be switched off or have a dead battery.
- **Security:** Any information transmitted over wireless links is subject to interception. Some of that information could be sensitive, like credit card numbers and other personal information. The solution needed really depends on the level of sensitivity.

Here are some practical hints useful to consider when developing mobile applications. These hints back the development of the proposed method for designing mobile information systems.

- **Understand the Environment and Do Some Research Up Front:** As with developing any other software application, we must understand the needs of the potential users and the requirements imposed by all networks and systems the service will rely on.
- **Choose an Appropriate Architecture:** The architecture of the mobile application is very important. No optimization techniques will make up for an ill-considered architecture. The two most important design goals should be to minimize the amount of data transmitted over the wireless link, and to anticipate errors and handle them intelligently.
- **Partition the Application:** Think carefully when deciding which operations should be performed on the server and which on the handheld device. Downloadable wireless applications allow locating much of an application's functionality of the device; it can retrieve data from the server efficiently, then perform calculations and display information locally. This approach can dramatically reduce costly interaction over the wireless link, but it is feasible only if the device can handle the processing the application needs to perform.
- **Use Compact Data Representation:** Data can be represented in many forms, some more compact than others. Consider the available representations and select the one that requires fewer bits to be transmitted. For example, numbers will usually be much more compact if transmitted in binary rather than string forms.
- **Manage Message Latency:** In some applications, it may be possible to do other work while a message is being processed. If the delay is appreciable—and especially if the information is likely to go stale—it

is important to keep the user informed of progress. Design the user interface of your applications to handle message latency appropriately.

- **Simplify the Interface:** Keep the application's interface simple enough that the user seldom needs to refer to a user manual to perform a task. To do so, reduce the amount of information displayed on the device, and make input sequences concise so the user can accomplish tasks with the minimum number of button clicks.

PROPOSED METHOD

The first step towards a successful wireless implementation project is a thorough business analysis, which serves as the backbone of any project bearing a fruitful return of investment. The analysis ensures that the project's requirements result in a wireless implementation that will successfully meet users' expectations and needs. Next, determinations about development features, approach, and constraints are made. This ensures that the wireless implementation is a good fit with the planned usage and infrastructure of the company. Finally, a choice needs to be made with regard to the software and hardware systems.

Business Analysis

When considering a wireless implementation, several questions have to be considered:

- What are the overall goals for implementing wireless services?
- What are the new markets to be targeted?
- What are the goals for giving mobile customer/staff wireless remote access?
- What technologies are currently in place towards supporting a wireless enterprise?
- Is interactivity important to the company?
- What current functions are suitable for wireless use?
- How prepared is the infrastructure to develop and host wireless applications?
- Are resources available to develop, implement, and support the wireless project?
- Is it more economical to have a wireless solution compared to a wired one?

Development

There are several factors to take into account when determining the wireless development solution. Indeed, MISs are expanding rapidly and changing from largely voice-oriented to increasingly data and multimedia systems.

- **Online vs. Off-Line:** On one hand, online applications include functions that require continuous connectivity, for example, looking for an inventory status or checking for available flights. Online wireless applications require real-time connectivity to be effective and useful. On the other hand, offline applications do not require real-time connectivity. Instead, they reside locally on a particular wireless device and are always available for use, but not always in real time. In addition, their use is limited to that particular device.
- **Screen Size:** With a much smaller display area than traditional desktops/laptops, it is important to fit in as much user-required functionality as possible, while trying to format the information in such a way that it appears attractive and appealing to end users.
- **Color:** Not all wireless devices support color, and some of them support a broader palette than others. Therefore, it is important to consider each device's color support, especially if multicolor content is to be provided, such as maps or advertisement banners.

- **Ergonomics:** Wireless devices vary widely in their standards and capabilities from one to the other. Which devices should be supported, which ones are best suited for the application's needs, and can all these devices be supported at the same time?

The above-listed questions have to be associated with a development lifecycle of the MIS. We advocate the consideration of four stages to constitute that lifecycle:

1. Requirements Stage
   - Identify key information that users need when mobile.
   - Establish use-case scenarios for such information.
   - Illustrate these scenarios to users for validation purposes.
2. Analysis Stage
   - Analyze and compare similar systems (wired or mobile) to the future mobile system.
   - Identify the elements that are directly linked to wireless aspects.
   - Highlight features of different wireless devices that the future wireless system will support.
   - Identify the needed wireless communication technologies as well as the network topologies on which the future wireless system will be built.
   - Analyze the various technologies to get users' queries and return responses (e-mails, SMS, WML, etc.). Dempsey and Donnelly (2002) listed some of the key features of an m-interface: usability, intelligent and personalized services, security, consultation capabilities, and pervasive and flexible payment mechanisms.
   - Analyze security and scalability problems.
3. Design Stage
   - Analyze security and scalability problems.
   - Use existing information resources or tailor them for mobile use.
   - Develop the architecture of the future application at data and process levels.
   - Discuss the location of data and processes, and who is in charge of maintaining these data and implementing these processes.
   - Provide solutions to potential security and scalability problems.
4. Implementation Stage
   - Develop and test the new application using for example Java 2 Micro Edition (J2ME) platform (http://java.sun.com).
   - Deploy the new application on the field.

CONCLUSION

In this article, we overviewed our vision of the importance of having a dedicated design method for mobile information systems. This importance is motivated by the continuous pressure on the professionals of MISs, who are demanded to put new solutions according to the latest advances in the mobile field. For instance, it is no longer accepted to postpone operations just because there is no connection to a fixed computing desktop. Mobile devices are permitting new opportunities when it comes to banking, messaging, and shopping, just to cite a few.

REFERENCES

Andreou, A. S., Chrysostomou, C., Leonidou, C., Mavromoustakos, S., Pitsillides, A., Samaras, G., Samaras, C., & Schizas, C. (2002). Mobile commerce applications and services: A design

and development approach. *Proceedings of the 1st International Conference on Mobile Business (MBusiness 2002),* Athens, Greece.

Bellavista, P., Corradi, A., & Stefanelli, C. (2002). The ubiquitous provisioning of Internet services to portable devices. *IEEE Pervasive Computing, 1*(3).

Campo, C. (2002). Service discovery in pervasive multi-agent systems. *Proceedings of the 1st International Workshop on Ubiquitous Agents on Embedded, Wearable, and Mobile Devices* (in conjunction with *AAMAS'2002*), Bologna, Italy.

Castano, A., Ferrara, S., Montanelli, S., Pagani, E., & Rossi, G. P. (2003). Ontology-addressable contents in P2P networks. *Proceedings of the 1st Workshop on Semantics in Peer-to-Peer and Grid Computing* (in conjunction with *WWW'2003*), Budapest, Hungary.

Dempsey, S., & Donnelly, W. (2002). Identifying the building blocks of mobile commerce. *Proceedings of the 1st International Conference on Mobile Business (MBusiness'2002),* Athens, Greece.

Elsen, I., Hartung, F., Horn, U., Kampmann, M., & Peters, L. (2001). Streaming technology in 3G mobile communication systems. *IEEE Computer, 34*(9).

Jose, R., Moreira, A., & Rodrigues, H. (2003). The AROUND architecture for dynamic location-based services. *Mobile Networks and Applications, 8*(4).

Karakasidis, A., & Pitoura, E. (2002). DBGlobe: A data-centric approach to global computing. *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops (ICDCSW 2002),* Vienna, Austria.

Konig-Ries, B., & Klein, M. (2002). Information services to support e-learning in ad-hoc networks. *Proceedings of the 1st International Workshop on Wireless Information Systems* (in conjunction with *ICEIS 2002*), Ciudad Real, Spain.

Maamar, Z., Ben-Younes, K., & Al-Khatib, G. (2003). Scenarios of supporting mobile users in wireless networks. *Proceedings of the 2nd International Workshop on Wireless Information Systems* (in conjunction with *ICEIS 2003*), Angers, France.

Open Mobile Alliance. (2005). Retrieved June 2005 from http://www.wapforum.org

Raghu, T. S., Ramesh, R., & Whinston, A. B. (2002). Next steps for mobile entertainment portals. *IEEE Computer, 35*(5).

Ratsimor, O., Chakraborty, D., Tolia, S., Kushraj, D., Kunjithapatham, A., Gupta, G., Joshi, A., & Finin, T. (2002). Allia: Alliance-based service discovery for ad-hoc environments. *Proceedings of the 2nd ACM Mobile Commerce Workshop* (in conjunction with *MOBICOM 2002*), Atlanta, GA.

KEY TERMS

Cell: A geographic area defining the range in which a mobile support station supports a mobile client. Each cell has a cell identifier (CID) that uniquely describes it.

General Packet Radio Service (GPRS): A mobile data service available to users of global system for mobile communications (GSM) users.

Java 2 Micro Edition (J2ME): An edition of the Java platform for developing applications that can run on consumer wireless devices such as mobile phones.

Mobile Client (MC): A user with a handheld wireless device that is able to move while maintaining its connection to the network.

Mobile Information System (MIS): A computing information system designed to support users of handheld wireless devices.

Mobile Support Station (MSS): A static host that facilitates the communication with mobile clients. An MSS supports mobile clients within a geographic area known as a cell.

Short Message Service (SMS): A service for sending text messages, up to 160 characters each, to mobile phones.

Universal Mobile Telecommunications System (UMTS): A third-generation (3G) mobile phone technology.

Wireless Application Protocol (WAP): A standard specification for enabling mobile users to access information through their handheld wireless devices.

Wireless Markup Language (WML): A scripting language that is part of the WAP specification.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 190-194, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.7
Distribution Patterns for Mobile Internet Applications

Roland Wagner
Johannes Kepler University Linz, Austria

Franz Gruber
RISC Software GmbH, Austria

Werner Hartmann
FAW Software Engineering GmbH, Austria

ABSTRACT

After the enormous success of the internet and mobile networks, the next upcoming boost for information technology will be the combination of both. But developing applications for this domain is challenging, because first, most mobile devices provide only small memory and processor footprints, prohibiting resource intensive code at client side and second, mobile networks offer only limited bandwidth, and the probability to connection losses is relatively high compared to wired networks. Selecting the appropriate software architecture in terms of distributing the functionality of the system between server and client device is crucial. Application distribution patterns, known from conventional system development, are analysed for their applicability for the mobile environment. After the more abstract analysis of the patterns, the IP multimedia subsystem (IMS) which is part of the current specification of 3G mobile networks is introduced and its support for different application distribution patterns is examined.

OVERVIEW

The success of mobile applications strongly depends on optimal utilization of client, server, and network resources. The distribution of the application functionality between client and server has strong impact on the grade of the resource utilization. Therefore, we present a schema for application distribution patterns and analyze architectural locations where an application can

be distributed and in this chapter we move our focus to the inherent problem of mobile applications to keep the data on the device and an the backend consistent.

With these distribution patterns, we will analyze several approaches for mobile Web access. In the last part of the chapter, we introduce an advanced architecture for representing mobile multimedia Web content: the IMS (IP multimedia system) with prerequisites and features as an example of a modern approach.

APPLICATION DISTRIBUTION PATTERNS

For the following we consider an application, whether it is mobile or not, consisting of three parts; The Presentation Layer, responsible for representing the visual parts of the application and doing the consummation of user input events. For distribution purposes, we divide the Presentation Layer in two sup-parts: The Dialog Representation, which is the visual painting and the reaction to events of the user and the Dialog Control which defines the sequence of the dialogs through the application.

The Business Logic Layer or the Application Kernel, responsible for the implementation of the business process, which means the origin and flow of data, which derives from the Backend Layer (Persistence Layer, Database Layer), which is responsible for retrieving and storing the data according to the requirements of the Business Logic Layer. Also for distribution analysis we divide the Backend Layer into the two subparts: The Database Access, which encapsulates the interface of the application programming language to the database (e.g., JDBC or ADO.NET) and the Database, which represents the database management system (DBMS) itself with tables, data, and stored procedures, etc.

Starting from this general architecture of an application, we want to derive a schema for making applications mobile, which means the separation of the whole application of parts of the application to a specified mobile device.

To be able to identify different software techniques to realize a device independent representation of an application, different architectural approaches have to be analyzed. First different levels where presentation and business logic can be separated are described. Finally, software techniques implementing the described design patterns are analyzed.

DISTRIBUTION LEVELS

This section describes possible distribution levels for every kind of client/server applications. The design of the client/server applications must provide functionality on the server. The client's functionality is mainly to display data ("thin client architecture"). This architectural design should ease distribution of new client versions. According to these needs, the design patterns "distributed presentation," "remote user interface" and "distributed application kernel" (taken from Renzel & Keller, 1997) are studied in more detail than "remote database" and "distributed database" pattern.

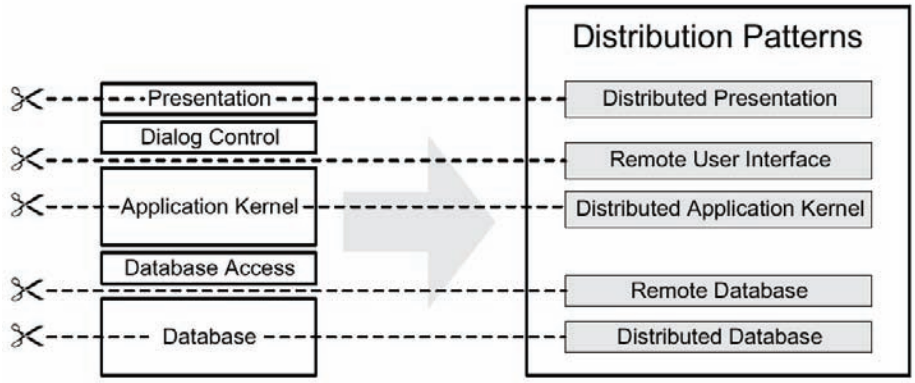
These patterns (see Figure 1) are interesting separating the representation from the business logic.

The remaining two design patterns "remote database" and "distributed database" in Figure 1 are applicable for realizing Web-based clients. A realization of these design patterns results in a "fat client" architecture, respectively the whole functionality of an application is located in the client.

Distributed Presentation

An application design according to the "distributed presentation," design pattern means to partition the application at the presentation layer.

Figure 1. Distribution levels (Renzel & Keller, 1997)



The implementation of one part of the presentation layer acts as a client, the other part together with the rest of the layers represents server. Client/server applications designed following this design pattern result in very small clients. The whole functionality of the application as well as the user interaction handling is located at the server side. On the client side, only user interface presentation is done.

Realizing this architectural design pattern, following points have to be considered:

- This design reduces complexity of the application development, because all functionality is located in the server part of the application.
- Testing of the application is eased, because functionality of the application is not spread over various nodes in the network.
- If the client owns a Windows-based user interface (UI), network traffic can get high due to intense presentation data (the user often modifies the appearance of the UI and/or many different windows are needed) and event handling.
- Batch and transaction processing is eased, because data processing and transaction control is located at the same node in the network.
- Security mechanisms are easy to implement, although authentication, authorization, and secure communication must be realized.
- As clients do not store any data, data consistency need not be considered at client side.
- Distribution cost is limited to the system software required on the terminal nodes. The presentation data is sent to the client via network connection.

Remote User Interface

An application design according to the "remote user interface" design pattern means to partition the application between the dialog control layer and the application kernel layer (Figure 1). The client part processes all user interactions and sends, according to these actions, requests to the server part. The server part executes application functionality and sends its results back to the client part. The client program does not have any functionality; it is just responsible for user interaction logic and UI presentation and is therefore named "thin client."

Realizing this architectural design pattern, the following points have to be considered:

- Communication between client and server has to be implemented by the application programmer
- Communication effort between client and server is less than with the "Distributed Presentation" pattern as the whole user interaction and UI presentation is calculated on client side
- Security mechanisms are easy to realize because no application code is situated at client side. Nevertheless authentication, authorization and secure communication must be realized
- The client system needs an operating system, a graphical presentation system and communication software to interact with the server application
- Software distribution is more expensive than with the "distributed presentation" pattern, because clients are larger and have more functionality. Therefore; it is more difficult to distribute them at runtime

Distributed Application Kernel

An application design according to the "distributed application kernel" design pattern means to partition the application kernel level (Figure 1). In the client part, user interaction, UI presentation and parts of application functionality are implemented; the rest of the application functionality and data access is implemented in the server part. Depending on how much functionality is implemented on the client side, the client will develop rather large ("fat client").

Realizing this architectural design pattern, following points have to be considered:

This design pattern is suitable for highly interactive applications and provides good utilization of the underlying hardware.

Performance of the application is dependent on where the application kernel is cut. If done right, excellent performance can be reached.

As application functionality is located at the client and the server side, the use of batch jobs adds further complexity to the application. On the one hand you could run a batch job from the client, resulting in high network traffic; on the other hand you can run it on the server, replicating the necessary functionality for the batch job from the client (replication vs. network traffic).

Security mechanisms are more complex to realize as application data is processed at the client side.

As data is needed at the client side, consistency of the data has to be considered.

Software distribution is more complicated, as there may also be configuration issues at client side.

Remote Database

With the Remote Database application distribution pattern, one can realize applications where the full presentation logic and the business logic and part of the persistence layer but not the data itself reside on one logical tier and the storage is done on a remote location.

Note that for encapsulating the database functionality we divide the persistence layer (backend layer) into two logical partitions.

First the Database Access Layer, which consists of the connecting layer from a programming language or technology (e.g., Java or .NET) and the DBMS (database management system) itself. Well known examples of Database Access Layer software are the Java Database Connectivity (JDBC) and the Action Data Objects for .NET (ADO.NET), and second the DBMS consisting of the actual interface to the physical data stored and eventually in the DBMS located persistence logic (e.g., realized by stored procedures).

The characteristic of the remote database application distribution pattern is that the DBMS is separated from the other partitions of the application but the application programming

interface (API) to the DBMS resides where the application is.

Regarding a Java application the remote database can be achieved by packaging the application and the JDBC libraries on the client and the DBMS on a server.

Realizing this architectural design pattern, following points have to be considered:

- This design pattern is suitable for highly interactive applications and provides good utilization of the underlying hardware
- Performance of the application solely depends on the client computer as mostly all the computation is done on the client tier
- Network traffic only occurs when synchronizing the persistence layer with the data displayed or manipulated on the client side
- Security mechanisms are more complex to realize as application data is processed at the client side
- As data is needed at the client side, consistency of the data has to be considered
- Software distribution is more complicated, as there may also be configuration issues at client side

Distributed Database

Additionally to the remote database application distribution pattern (see the previous section for details) the client tier gets partitions of the DBMS functionality. This pattern can be realized if the DBMS supports the division of its functionality to several clients.

This pattern is required if one needs DBMS functionality on the client tier without having the possibility to run the whole application locally.

One typical scenario for this application distribution pattern is an application realized with the lightweight additions from IBM DB2 Everyplace, which runs on a mobile device and keeps itself synchronized with a DB2 server can be located anywhere. Also Oracle, Sybase, and other major DBMS vendors have lightweight versions of their DBMSs.

Realizing this architectural design pattern, following points have to be considered:

- This design pattern is suitable for highly interactive applications and provides good utilization of the underlying hardware
- Performance of the application solely depends on the client computer as mostly all the computation is done on the client tier
- Network traffic only occurs when synchronizing the persistence layer with the data displayed or manipulated on the client side. This is fully handled by the DBMS, the application itself does no network processing (if data is not fetched from other 3rd party locations). Note that the synchronization logic can not be influenced as it is fully handled by the DBMS
- Security mechanisms are more complex to realize as application data is processed at the client side
- Software distribution is more complicated, as there may also be configuration issues at client side

Conclusion

Understanding mobile applications, one can use any of the application distribution patterns. Which one the application developer chooses, depends on the application scenario itself. Hereby following questions can help leading the adequate application distribution patterns:

- Is there a network connection always available?
- Do I have to keep data on the client device
- Do I have enough computational resources on the client mobile device?
- Is persistent data on the client a requirement?

In answering these and other similar questions one can choose an application scenario for realization.

Up to now, we have only talked about the application distribution pattern (exactly speaking on how can one distribute and divide an application to several parts). Now we are going on to the practical implications and real life application architecture scenarios for mobile Web access.

This is done via representation of common architectures for mobile Web applications and an analysis to which application distribution pattern it belongs.

ARCHITECTURES FOR MOBILE WEB ACCESS

In this chapter we want to present several well-known architectures for mobile access to the World Wide Web. As we are talking about mobile multimedia issues, we want to put our focus especially on representing multimedia content on the Web via the introduced technologies.

The architectures we are looking at are:

- HTML/WML/ASP/JSP Web access
- Local applications and socket access
- MIDP-Java-Applications

HTML/WML/ASP/JSP Web Access

Access to the Web via Internet browser is regarding the efforts on mobile devices the simplest way to applications and data in the Internet.

Nearly the whole work (with the exception of rendering the user interface) is done on the Web server. Therefore these technologies are completely independent from the mobile device. User interfaces generated via HTML, see (Hypertext Markup Language, 2002), and (Extensible Hypertext Markup Language, 2005) or WML (Wireless Markup Language, 2002) are not as comfortable as one is used to have on a personal computer, although Active Server Pages (Keyton, 2000) and JavaServer Pages (JavaServer Pages Technology, 2005), etc. have released several advanced controls and technologies, like JavaServer Faces (JavaServer Faces, 2005) for the construction of advanced user interfaces.

However, from the basic design these technologies have no built-in mechanisms for representing multimedia content. Several extensions and plug-ins for browsers have been established; one of the most used is the Macromedia Flash Player, which is available for many platforms.

Another main disadvantage of server-based approaches is that supporting various mobile devices requires in the worst case a complete new generation of the user interface. Specialties of a mobile device cannot be regarded as in the server side technology the client is not regarded automatically. The required always-on connection to the network can be a crucial issue (e.g., on bypassing temporary network failures or availability).

The main advantage of this approach is that one does not have to redistribute software on new releases and updates. On the client side, only the browser and the required plug-ins have to be installed. Data is only stored on the server, therefore no synchronization logic is needed. Data storage is mostly realized with the XForms standard (W3C, 2005).

As application distribution pattern (see previous sections) the only realizable one is the distributed presentation pattern. Business logic and dialog control as well as backend are not accessible directly from the client device.

Applications Accessing Remote Services

Applications realized for mobile devices can access remote services provided by Web services, sockets, remote method invocation (RMI), remote procedure calls (RPC), etc, physical network access via wireless networking technology (UMTS, GSM, GPRS, etc.) is needed.

In this approach, the client runs the presentation of the application and the main dialog control. Business logic and backend need not be on the client machine, but can be located on the client. Therefore, from the architectural point of view all application distribution patterns from the previous section can be realized; only the backend storage in the DBMS has to be separated from the device, which is in fact not a restriction.

User interfaces are developed for one special device, therefore one can develop rich and full-featured user interfaces respecting the client's specialties. Multimedia content has to be integrated by the application developer and is not standardized. However, a new release requires software distribution on each client device.

Support for new versions of mobile devices with another software platform (operating system etc.) mostly causes a full new development of the application.

The architecture itself depends on the application scenario. For communication the following data transport mechanisms can be used:

- remote method invocation,
- Web services (SOAP over HTTP(S)),
- sockets,
- HTTP, and
- proprietary protocols.

A special more portable creation of user interfaces can be achieved via user interface description languages like XCC (Schmidt & Weinstein, 2002), XUL (Introduction to a XUL Document, 1999), (XWT) and UIML (User Interface Markup Language (UIML) Specification. 2002).

In this approach, a runtime is established which parses the user interface description from a server or from local file system und then renders the interface on the mobile device. Here we combine the declarative description of the user interface as we do it in HTML, etc. and the integration into a special device. Having realized such an application architecture, applications from purely local to an application with remote presentation application distribution pattern is possible, depending on the usage scenario.

MIDP Applications Accessing Remote Services

As MIDP (mobile information device profile) defines a portable way to write client managed network aware applications with respect of special concerns of mobile devices, MIDP seems to be the perfect solution for mobile multimedia applications. Technically MIDP is nothing else than a local application with a standardized runtime environment (Mobile Information Device Profile, 2002), therefore all the arguments and issues supplied in the previous section also apply here.

From an architectural point of view, MIDP supports every kind of application distribution pattern with the big advantage of being portable over many platforms.

As a more advanced and sophisticated example of a possibility for accessing multimedia data over the Web we will introduce the IP multimedia system (IMS) in the next section.

IP MULTIMEDIA SUBSYSTEM

After the introduction of wireless, mobile data connections as enabling technology for media rich, mobile applications, demand for an appropriate operation and deployment environment that eases roll out and operation of such applications arose. The IP multimedia subsystem (IMS), defined and specified by the 3GPP consortium as part of the UMTS standard, aims to provide such an infrastructure, (3GPP.TS 23.228 IP Multimedia Subsystem (IMS); Stage 2 (Release 6) 2005). 3GPP standardizes the interfaces between components, and third-party developers can build components without having to sell the entire service network. For instance, a company might specialize in mobile positioning servers and sell that service separately

to operators. (Andersson, 2001) Although IMS is standardized as part of UMTS, it supports internetworking with WLAN (since release 6). Further, the CDMA2000 Multimedia Domain is based on 3GPP IMS.

IMS Architecture Overview

The IMS architecture can be subdivided into a three layered architecture, as depicted in Figure 2. The connectivity layer builds the interface to the underlying network infrastructure, while the service layer provides APIs and infrastructure to build and deploy services and applications. The control layer is responsible for session handling and user subscription management.

Connectivity Layer

The connectivity layer consist in routers and switches and is responsible for the connection to transport bearers, like 2,5G, 3G mobile networks, or WLAN. It provides an abstraction layer to enable uniform transport capacity and quality negotiation with underlying networks.

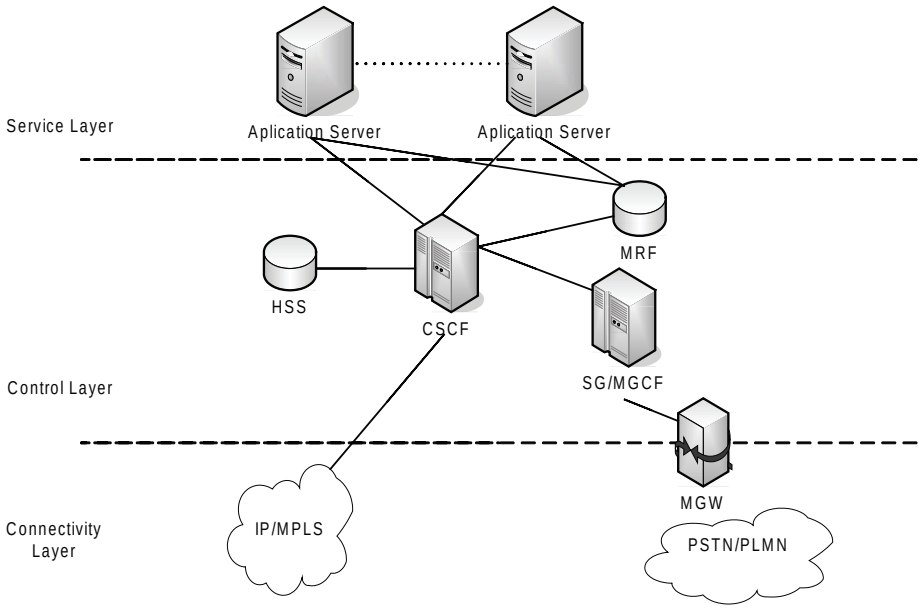
Control Layer

The control layer of IMS is responsible for session setup, modification, and release, and provides gateways to existing IP and circuit switched networks. The key elements of the control layer are the call session control functions (CSCF) and the home subscriber server (HSS). The role of CSCF can be split up into proxy CSCF (P-CSCF), interrogating CSCF (I-CSCF) and serving CSCF (S-CFCS).

The P-CSCF provides roaming functionality. If a mobile user wants to access IMS services while roaming in a foreign network, the P-CSCF is contacted. The P-CSCF detects the home network of the user and forwards the request to the home network of the user.

The I-CSCF is contacted if a user requests IMS services in his home network. The main functionality of the I-CSCF is to find out an S-

Figure 2. IMS architecture overview



CSCF which can handle the request. By routing service request to different equivalent S-CFCS servers according to their current load, it also provides the functionality of a load balancer.

The S-CSCF performs the session management of the IMS network. It handles SIP (session initialization protocol) messages, establishes sessions and negotiates with underlying transport layers to guarantee requested service qualities. SIP is an application-layer control (signalling) protocol for creating, modifying, and terminating sessions with one or more participants. (Rosenberg et al., 2002)

The home subscriber server (HSS) maintains a database with the unique service profile of all users. The user service profile stores all user service information and preferences in a central location. This information includes the end users current registration information, roaming information, instant messaging service information etc.

The media resource function (MRF) provides an interface to application servers and the S-CSCF to control media streams. It consists in a MRF controller which interprets incoming information, and a MRF processor which provides functionality like audio transcoding, media analysis etc. The MRFP is controlled by the MRFC.

The gateway functionality of the IMS is implemented by the media gateway control function (MGCF), the media gateway (MGW) and the signaling gateway (SG). The task of the MGCF is to control one or more MGWs, which enables scalability of the gateway system. The MGCF translates SIP messages from the CSCF, into a format that can be processed by the connected network. The MGW processes the ingoing and outgoing media streams between end users. Its primary function is to convert media from one format to another. The signalling gateway (SG) forms the signalling interface to legacy public switched telephone networks (PSTN) by transforming SS7 signalling information, which is the only format supported by PSTN, into IP and vice versa.

Service Layer

The service layer of IMS provides foundations to implement applications and services for the end user. It comprises in one or more SIP application servers that communicate with the underlying IMS functions provided by the S-CSCF. The S-CSCF is the anchor point for delivering new services since it manages the SIP sessions. (Parameshwar & Reece, 2004) application servers and S-CSCF interact via the IMS service control (ISC) interface which is based on SIP and its extension. ISC provides an event notification service, which allows Application Servers to subscribe for user specific events.

SIP application servers are programmable through scripting languages, SIP-CGI or APIs like SIP-servlets to provide the logic of value added services. The Java API for integrated networks (JAIN) and the JAIN server logic execution environment (JSLEE), which are APIs available for SIP application servers, are activities led by Sun Microsystems to create a standardized, Java-based service development API that abstracts from the underlying network architecture. See (Ferry & Lim Boon, 2004).

The architecture of the IMS service layer allows deployment of services by the operator and by 3rd party service providers through the Opens Service Access (OSA) API. The OSA API enables the secure integration of 3rd party services.

IMS Features

IMS provides a communication and service infrastructure which eases implementation and deployment of mobile multi media applications. Further, it introduces new features to the mobile domain like flexible session handling, and quality of service, which are introduced here.

Flexible Session Handling

The session initiation protocol (SIP) is the core communication protocol of the IMS. SIP is a text-based client-server protocol, similar to HTTP or SMTP that initiates session setup, routing, authentication within an IP domain. SIP enables the creation of sessions that hold different services, like voice and video. Those services can be synchronized if desired (e.g., for video telephony). Sessions can be dynamically modified, which allows adding a video component to an existing voice session. Further, IMS supports establishing multiple sessions of unrelated services, like an asynchronous textual chat session running concurrently and independent of a video conference session at one device.

Quality of Service

Real-time mobile IP communication is difficult due to fluctuating bandwidths, which severely affect the transmission of IP packets through the network. In traditional IP networks, transport quality is defined as "best effort," meaning that the network will do its best to ensure the required bandwidths, but there is no guarantee. IMS introduces a quality of service (QoS) feature to ensure that critical elements of IP transmission, such as transmission rate, gateway delay and error rates can be measured, improved and guaranteed in advance. By these means, users are able to specify the level of quality they require depending on the type of service and the user's circumstances. The policy decision function (PDF) node of the IMS negotiates service quality issues like resource authorisation and reservation, approval and removal of QoS commit with underlying network resources.

Standardized IMS Service Enablers

IMS facilitates the creation and delivery of multimedia services based on common enablers in a "write once, use many" way. These key elements in the IMS architecture are so-called service enablers. They represent generic and reusable building blocks for service creation. The service enablers developed for successful applications can become "global enablers" that are automatically included in new applications and services. Examples for these service enablers are group, list, and presence management and multi-party conferencing.

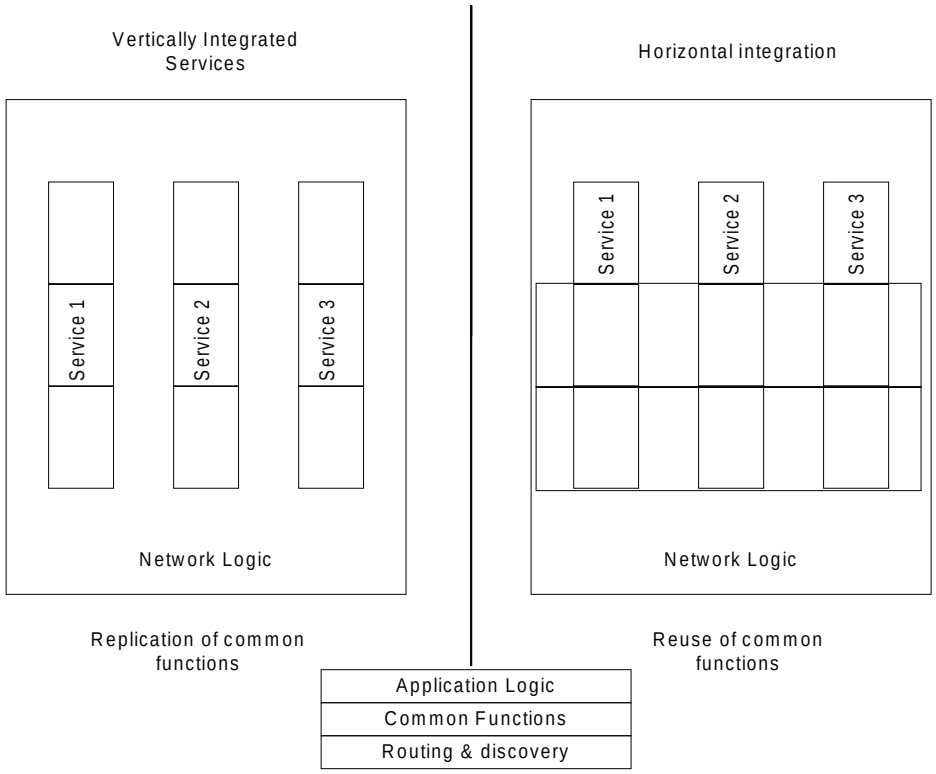
Application Integration

The architecture of the IMS server layer together with the standardized IMS service enablers fosters horizontal service/application integration. Figure 3 compares traditional vertical integrated services with the horizontal integration approach. The horizontal architecture in IMS also specifies interoperability and roaming, provides bearer control, charging and security. Horizontal integration enables the reuse of common infrastructure and service enablers, which speeds up and simplifies service creation and delivery.

Application Distribution with IMS

The mobile and wireless domain is a challenging environment for application developers. On the one hand, bandwidth and reliability of mobile networks is still not as high as desired, which could partly be antagonized by putting more application logic to the client, and facilitating caching mechanisms to enhance performance. On the other hand, the limited memory and processing resources provided by currently available mobile devices aggravates the implementation of such application architectures. To mitigate this situation, a service deployment infrastructure should show great flexibility in terms of application distribution capabilities. The SIP-based client communication of IMS enables multiple application distribution architectures. IMS even allows deployment of peer-to-peer applications, but the services pro-

Figure 3. Service reuse instead of replication



vided by the IMS infrastructure, like presence, group management or billing are centralized. Thus, applications that take full advantage of the services offered by IMS are inherently client-server applications. The following sections discuss how different application distribution patterns can be implemented within IMS.

Distributed Presentation

The distributed presentation approach leaves the entire work load at the server. The client only needs to interpret and render the presentation according to the received description. No or only marginal additional software is required at the client, which eases deployment and maintenance of such applications. The centralized application logic also avoids data synchronisation and consistency problems, because data access control is located at a (logically) single server node.

The distributed presentation architecture enables applications to take full advantage of the IMS infrastructure, like service enablers, billing etc. The application server centric architecture of IMS, with the ability of horizontal application integration, fosters this distribution pattern.

The high server dependency of this application architecture makes it more vulnerable by network failures, because a lost server connection is immediately recognized by the user.

Remote User Interface

The remote user interface architecture leaves the business logic relevant tasks at the server, while moving user interface related functionality like consistency and plausibility checks etc. to the client. This includes applications which use WMLscript to validate user input, but also PoC (push to talk over cellular) applications, which perform

basic consistency checks like preventing double adding of users to groups. As the business logic is located at the server, this architecture allows utilisation of the IMS infrastructure.

The effort for deployment and maintenance is increased, if the application requires a dedicated client.

Distributed Application Kernel

The distributed kernel architecture splits the application logic between client and server. Applications that utilize this distribution pattern can take full advantage of IMS features like flexible session management and QoS, horizontal service integration can only be utilized by the server side part of the application. The interface between client and server part of the application need to communicate by IMS protocols like SIP or SDP (session description protocol) which may add some protocol translation overhead to the application and increase the complexity of the application.

This architecture enables to decrease the server dependency by employing a more intelligent, application specific connection management at the client. This can be used to "hide" short time disconnections from the user, and the centralized business logic avoids data synchronisation.

Remote/Distributed Database

With remote or distributed database architectures, the whole application functionality and in case of the distributed database architecture, also data is located at the client. The network infrastructure is used to synchronize distributed data bases. Although these architectures can be implemented within IMS enabled networks, taking advantage of the centralized IMS features like service enablers or billing requires considerable more effort than with server-based applications.

CONCLUSION

The IMS is the key element in the 3G architecture that makes it possible to provide ubiquitous cellular access to all the services that the internet provides (Camarillo & Garcia-Martin, 2004). The architecture of IMS clearly emphasizes server centric application architectures, which reside on application servers and take advantage of the various service enablers provided by IMS. The feature rich infrastructure of IMS fosters rapid application and service development, reuse of existing functionality and integration of different applications.

REFERENCES

3GPP.TS 23.228 IP Multimedia Subsystem (IMS); Stage 2 (Release 6). (2005). Retrieved May 16, from http://www.3gpp.org/ftp/Specs/archive/23_series/23.228/

Andersson, C. (2001). *GPRS and 3G wireless applications: Professional developer's guide.* Mississauga, Ontario, Canada: John Wiley & Sons Ltd.

Camarillo, G., & Garcia-Martin, M. (2004). *The 3G IP Multimedia Subsystem (IMS): Merging the internet and cellular worlds.* Mississauga, Ontario, Canada: John Wiley & Sons Ltd.

Extensible Hypertext Markup Language. (2005). Retrieved May 16, 2005, from http://www.w3.org/MarkUp/

Ferry, D., & Lim Boon, S. L. (2004). *JAIN SLEE 1.0 Specification, Final Release.* Sun Microsystems Inc.

Hypertext Markup Language. (2002). Retrieved May 16, 2005, from http://www.w3.org/MarkUp/

Introduction to a XUL Document. (1999). Retrieved May 16, 2005, from http://www.mozilla.org/xpfe/xptoolkit/xulintro.html

JavaServer Faces. (2005). Retrieved May 16, 2005, from http://java.sun.com/j2ee/java serverfaces/index.jsp

JavaServer Pages Technology. (2005). Retrieved May 16, 2005, from http ://java.sun.com/products/jsp/index.html

Keyton, W. A. (2000). *ASP in a nutshell*. Sebastopol, CA: O'Reilly Associates.

Megacz, A. (2003).The XWT reference windowing toolkit. Retrieved May 16, 2005, from http://www.xwt.org/reference.html

Mobile Information Device Profile. (2002). Retrieved May 16, 2005, from http://java.sun.com/products/midp/

Parameshwar, N., & Reece, C. (2004). *Advanced SIP Series: SIP and 3GPP*. Award Solutions. Retrieved from http://www.awardsolutions.com/downloads

Renzel, K, & Keller, W. (1997). Client server distribution—A pattern language. *Proceedings of the Pattern Languages of Programming Conference (PLoP)* (Tech. Rep. No. #wucs-97-34). Monticello, WA: Washington University.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R. et al. (2002). *RFC 3261—SIP: Session Initiation Protocol*. Network Working Group. Retrieved May 16, 2005, from http://rfc.net/rfc3261.html

Schmidt, A., & Weinstein, T. (2002). *Design und Implementierung ultraleichter Java Clients.* NetObject Days. Retrieved May 16, 2005, from http://www.old.netobjectdays.org/pdf/01/papers/node/weinstein_schmidt.pdf

User Interface Markup Language (UIML) Specification. (2002). Retrieved May 16, 2005, from http://www.uiml.org/specs/docs/uiml30-revised-02-12-02.pdf

Wireless Markup Language. (2002). Retrieved May 16, 2005, from http://www1.wapforum.org/tech/documents/WAP-191-WML-20000219-a.pdf

W3C. (2005). *XForms—The next generation of Web forms.* Retrieved May 16, 2005, from http://www.w3.org/MarkUp/Forms/

KEY TERMS

3GPP: The 3rd Generation Partnership Project (3GPP) is a collaboration agreement that was established between European, Japanese, and North American telecommunication standardization organizations to create a globally applicable third generation (3G) mobile phone system specification.

IMS: The IP multimedia subsystem (IMS) is an open, standardized multimedia architecture for mobile and fixed services. It is based on a 3GPP variant of SIP and runs over the standard Internet protocol (IP). It enables telecom operators to offer network controlled multimedia services.

MIDP: Mobile information device profile (MIDP) is part of the J2ME framework and stands for mobile information device profile. It is specified by Sun Microsystems for the use of Java on embedded devices like cell phones or PDAs.

OMA: The Open Mobile Alliance (OMA) is an initiative of major manufacturers of end user equipment and infrastructure for mobile telecommunication networks. It aims to create interoperable services enablers to work across countries, operators, and mobile terminals. The OMA is driven by market requirements.

PoC: Push to talk over cellular (PoC) is a service that enables half duplex one to many communication over cellular networks.

SIP: Session initiation protocol (SIP) is a protocol developed by the IETF MMUSIC Working Group and proposed standard for setting up sessions between one or more clients. SIP is similar to HTTP and shares some of its design principles: it is human readable, very simple, and request-response.

Three Tier Architecture: Three-tier is a client-server architecture in which the user interface, functional process logic ("business rules"), data storage, and data access are developed and maintained as independent modules, most often on separate platforms.

This work was previously published in Handbook of Research on Mobile Multimedia, edited by Ibrahim, pp. 507-520, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.8
Web-Based Seamless Migration for Task-Oriented Mobile Distance Learning

Degan Zhang
University of Science and Technology of Beijing, China

Yuan-chao Li
China University of Petroleum, P.R. China

Huaiyu Zhang
Northwest University, China

Xinshang Zhang
Jidong Oilfield, P.R. China

Guangping Zeng
University of Science and Technology of Beijing, China

ABSTRACT

As a new kind of computing paradigm, pervasive computing will meet the requirements of human being that anybody maybe obtain services in anywhere and at anytime, task-oriented seamless migration is one of its applications. Apparently, the function of seamless mobility is suitable for mobile services, such as mobile Web-based learning. In this article, under the banner of seamless mobility, we propose a kind of approach supporting task-oriented mobile distance learning paradigm. Web-based seamless migration, which has the capability that task for mobile distance learning (MDL) dynamically follows the learner from place to place and machine to machine without learner's awareness or intervention by active service. Our key idea is this capability can be achieved by architecture of component smart platform and agent-based migrating mechanism. In order to

clarify the approach, firstly, a description of the task for mobile distance learning and migrating granularity of task has been suggested. Then, the mechanism of seamless migration has been described, including solving several important sub-problems, such as transferring delay, transferring failure, residual computation dependency. Finally, our implemented platform for Web-based seamless migration has been explained, the validity comparison and evaluation of this kind of mobile distance learning paradigm is shown by an experimental demo. Suggested Web-based learning paradigm by seamless migration is convenient to distance learn during mobility and is useful for the busy or mobile distance learner.

INTRODUCTION

It is known to all that pervasive/ubiquitous computing (Weiser, 1991) is a new computing paradigm fusing the technologies of computing, communication, and digital multimedia, which integrates information space and physical space of human being's life, so it makes the computing and communication just like the life necessity, such as water, electricity, and air. This paradigm meets the requirements of human being that anybody maybe obtain services in anywhere and at anytime, so it is full of future. Nowadays, many ambitious projects have been proposed and carried on to welcome the advent of pervasive computing. There are a bunch of branch research fields under the banner of it, such as Seamless Mobility (Satyanarayanan, 2001).

For seamless mobility, the history and context of computing task will be migrated with person's mobility, and the computing device and software resource around this task will make adaptive change. The chief function requirement of seamless mobility is on the continuity and adaptability of computing task. The continuity is that the application can pause and continue the work without the loss of the current state and the running history. The adaptability is that the application is not restricted by computing device and context of service but adaptable to its environment.

Apparently, this function of seamless mobility is suitable for mobile learning paradigm (Takasugi, 2001, 2003). For learner, it is necessary and accessible when he or she can NOT complete his or her learning task/courseware, such as video, audio, text, picture, etc., in one specified scene, he or she can go on learning the uncompleted task/courseware in other spots by seamless mobility based on the Web. In our opinion, this is a kind of mobile working paradigm — learning by seamless migration with computing task. But when seamless migration for computing task of learning is realized on PC, laptop, or PDA, there are several difficult problems to be solved: (1) Meet different networked Web environment, such as different OS platform. (2) Manage the seamless-service among multiple machine devices. (3) Describe computing task of learning and only migrate the relative parts of task interested by learner in order to reduce the delay produced by migrated data.

In this article, we propose a test bed of learning by seamless migration for mobile learning, which can be suitable for the required dynamic changes to the network and environment without learner awareness or intervention, and the condition of only sitting in front of the desktop PC for mobile learning is unnecessary. The structure, mechanism, result of experimental evaluation of the test bed is reported. It makes the ultimate mobile system possible by dynamically implementing the changes required to follow the learner from place to place and machine to machine.

The rest of this article will be organized as follows. Firstly, we give formal description of task of mobile distance learning and migrating granularity of task of learning. After that, we design and discuss efficient approach of Seamless Mobility based on agent for task-oriented mobile distance learning, along with the description of our implemented platform for Seamless Mobility.

Finally, we evaluate the validity of the approach and platform for mobile distance learning and draw a conclusion.

DESCRIPTION FOR TASK OF LEARNING

In order to clarify and realize how to transfer tasks of learning among different distance computing environments, firstly, a formal description and classification of task is required, which is independent of the realization mechanism. To adapt the environment of pervasive computing, a universal description language for task of learning should be used. Nowadays, the description languages for workflow or task of learning are mainly based on stationary computing environment (Simmons & Apfelbaum, 2001). However, the computing environment of seamless mobility is dynamic and mobile, so the description language should be abstract and self-adapted. Based on our knowledge, XML (Extended Markup Language) and SMIL (Synchronized Multimedia Integration Language) released by W3C can be used (Shi & Xie, 2003).

The task or transaction of learning cared by learner is our alleged *Task* (in brief, T), which consists of subtask or sub-transaction T_i, each T_i is an independent unit of function. Because of the diversity of task, its subtask or atomic task may be different from each other. In order to keep the compatibility, the description of subtask should be abstract, mainly, the key and necessary parameters, such as Qos of subtask, environments, etc.

During mobile distance learning, the task can be classified into three kinds based on DATA TYPE of its contents:

1. **Event-Type:** It is strict with the delay, the transferred bytes of subtask is few, but timely, semantic and no loss during transferring. Once the command of operation is done, the result should be shown.
2. **Stream-Type:** It is not strict with delay and semantics, permitting a certain loss during transferring, but strict with jitteriness of transferring.
3. **Bulk-Type:** It is different from event type because the transferred byte of subtask is much larger (maybe several Mbytes), and it is also different from stream type because when executed, it requires the integrality of data.

Described formally task of learning by SMIL is as follows:

```
 <smil xmlns="http://www.w3.org/2001/SMIL20/
Language">
<head>
     <layout>
          <root-layout width="" height=""/> <region
id=" " ……/>
      </layout>
     </head>
     <body>
       <seq>
         <par>
                    …….
         </par>
       </seq>
     </body>
</smil>
```

MOBILE GRANULARITY OF THE TASK

How to deal with the problem of task-oriented mobile distance learning under the banner of seamless mobility? Currently, the usable technology is based on Mobile IP used as network-level protocol and stationary or mobile agent used in application-level. Active and intelligent mobile agent controlled by running container can deploy or adjust dynamically its services according to application requirement or running status of network. As a kind of special computing resource, agent supports deployment of computing resource and mobility freely, which makes the system

manage and adjust easily, so it is suitable for application of seamless mobility.

During the task-oriented mobile distance learning, mobile granularity of the task should be traded off reliability, the communication volume of network, and so forth. According to integrity of transferred contents of the task, the mobile granularity of the task may be divided into "Strong Transfer" and "Weak Transfer," and the mode of transferring may be controlled by the Travel Schedule/Plan.

Strong Transferring means that total information involved in the current task must be transferred, after reaching the target terminal, the task can execute continuously from snapshot point. But in mobile WWW, it is difficult to collect total information of current task, to describe and record the executed status and necessity of task under the high bandwidth network, so the burden of this mode is very heavy and complex. Nowadays, the JVM is not supported this mode. Weak Transferring is only done for partial executing status and data, its speed is much faster then that of Strong Transferring, and its delay is much shorter then that of Strong Transferring. Of course, Weak Transferring has its shortcomings, for instance, the total historical executing status of task is difficult to be restored. So it is decided only by detailed scenario that which mode should be adopted in the application.

AGENT-BASED APPROACH OF TASK-ORIENTED SEAMLESS MOBILITY

Because the proliferation of mobile devices and the appearance of runtime environment give new challenges to support user mobility, we give a mechanism of integrating mobile devices into runtime environments to provide more computational, communication and storage capabilities to mobile distance learner. In this mechanism, we think that the data migrating is needed between different computing devices by different network, such as wireless infrastructure-based communication, multi-hop ad-hoc networks, dynamic topology without any infrastructure-based communication, Internet-based networks and different computing devices interconnected using IEEE 802.11x and Bluetooth technology. So a seamless and transparent migrating mechanism between different networking interfaces is needed. Seamless migrating between different networks for different computing devices by mobile agents is a basic feature for improving the quality of a perceived service under the pervasive computing mode. However, the heterogeneity also implies that the services are also distributed over the accessible networks. Based on context information, our mechanism spontaneously interoperates with available resources discovered in runtime environment so that it can improve the performance of mobile interactive distance learning.

Classification of Agent

On the attribute of agent, it can be classified into two types. One is Stationary Agent, which is not mobile and kept in the agent environment (AE). The other is Mobile Agent (MA), which may be transferred in the AE. On the function of agent, it can be classified into three types, Terminal Agent (TA) (including User Agent (UA)), Navigation Agent (VA) (including Network Capability Agent (NCA) and Location management Agent (LMA)), Task Agent (KA) (including Execution/code Agent (EA) and Data Agent (DA, such as Service Agent (SA), User Document Database Agent (UDDA))). These agents may be Stationary or Mobile Agent.

Stationary or Mobile Agent (Danny & Mitsuru, 2001) is a kind of program with its name and can interact with other agents or resources when transferring from one network to another in different heterogeneous network (Karnik, 1991). This program can dynamically decide when and where to transfer. It can suspend at any running

point or transfer to another computer and execute continuously. Mobile Agent also can clone itself or produce its sub-agent and transfer to other computer to do complex task in cooperation mode. Besides common attributes (such as autonomous, initiative, intelligent), mobility is its main attribute, which makes it roam among different networks. Mobile Agent is suitable for computing of large heterogeneous network like Internet. Mobile Agent System (MAS) (Ciancarini, 2002) is a kind of system for creating, explaining, executing, scheduling, transferring, and terminating agent. Each system can run many agents. TA is one interactive interface for human and terminal devices. Users may search, browse, subscribe, and publish new service by TA. UA is on half of the user, which may transfer among different environments with the user. UA can partly store attribute data of the user and be used as buffer of current terminal. VA is used as addressing in the MAS and carry KA in its "MessageBox (**MB**)." KA is used as restoring runtime environment, executing task in the new environment, and recording each snapshot as current and history execution status.

New Approach of Seamless Mobility Based on Agent

The basic strategy of transferring based on agents has two modes:

1. Strong Transfer from source node to target node.
2. Weak Transfer from source node to target node.

The first mode is adapted for seamless mobile scenarios with smaller amount of transferred data, such as task of "Event-Type." If the task is "Stream-Type" or "Bulk-Type," the delay of transferring and unnecessary amount of data are larger, which is not adapted for seamless transferring on Mobile WWW.

The second one may be implemented through two methods: partial information has been loaded on the target node/terminal station, downloaded the relative partial information timely during the runtime of task. This mode is adapted for the task with "Stream-Type" or "Bulk-Type," which can reduce the transferred amount of data, the delay of transferring and improve the running efficiency, but occupied a certain storage space of target node.

The basic transferring step of agents is as follows:

1. Determine the agents running on the source node.
2. Suspend the agents.
3. Snapshot or record the information of running agents and transfer them to target node.
4. Reconstruct or restore the information of running agents on the target node.

In our opinion, the design of transferring method is considered from two aspects:

1. The transferring amount of data is total or partial.
2. The occasion of suspending agent on the source node and restoring agent on the target node.

The existing transferring mechanisms (Takasugi, 2003) have NOT analyzed and discussed both of them deeply, especially, how to adapt for the application requirement of seamlessness of pervasive computing. In our opinion, it must deal with the problems of seamless transferring method, transferring delay, transferring failure and residual dependency, and so forth.

Based on the basic strategy of transferring mentioned previously, now we discuss the new efficient mechanism of seamless mobility suggested by us.

If the transferred amount of data is partial, and this part of information must be transferred firstly so that the task can restore the runtime environment and run continuously on the target node, this part of information is named "Key Set," such as executing code, running status, and so on.

According to the classification of agent discussed earlier, we make the following rule:

1. **Navigation Agent** (VA) need NOT do direct relative works with the task, which is familiar with the topological structure of subnet of target node and addressing in the network. The Data Structure of VA may be divided into two parts: one is itself "function body," another is MessageBox (MB, mark as □) used as loading moved object and transferring in the subnet.
2. **Task Agent** (KA) does detail jobs, which includes executing the code, managing the data and environmental status, and so on. It can transfer with the Navigation Agent (VA) in the network and need NOT know the subnet's structure.
3. When transferred, KA looks up relative VA and joins in its MB firstly, and then sent to target node by VA.

For the sake of convenience, we give a kind of general case:

TA wants KA on the logic node PA_2 (Persona Avatar 2) to be transferred to another logic node PA_3 (Persona Avatar 3), according to the time-topological relation of transferred object, the "TRAVEL SCHEDULE/PLAN" which is a kind of DATA STRUCTURE independent of agent has been made. The current scenario is that the TA is connected with logic node PA_1 which is connected to logic node PA_2 through double direction link, The arrow shows the connected direction and solid line with arrow shows KA can transit the logic link.

The designed algorithm of seamless mobility includes eight main steps:

1. According to the subscribed TRAVEL SCHEDULE/PLAN for transferring, logic node PA_2 lets VA begin addressing in the network according to the address supplied by logic node PA_3, when the connection is successful, VA sends instruction "TransferNode" to Logic node PA_3 as target node, VA+ transfer to PA_3 after packing, the packet consists of the recorded structure of KA, the association relationship between KA, the space occupied by KA, the type of KA, the information of VA for task transferring and "Messenger" information for scheduling all agent (including VA, EA and DA), the state of arrived Messenger is not "Executing" but "Waiting" and storing in the queue of PA_3.
2. Logic node PA_3 sends instruction "UpdateLinking" to all logic nodes connected to PA_2, such as logic node PA_1 (Persona Avatar 1). The instruction includes the information modifying the link address, such as link ID, IP and Port of two ends. During the transferring, the Messengers to PA_3 store in the relative queue and wait for executing unless the Key Set or the total task is finished to be transferred.
3. When PA_1 has received the instruction "UpdateLinking," it creates the association to new link, and sends instruction "LinksUpdated" to logic node PA_2.
4. When PA_2 has received all expected message "LinkUpdated," and then sends instruction "ActiveNode" to logic node PA_3. The message includes the list of all arrived Messenger. At the same time, PA_2 delete the transferred VA+.
5. When PA_3 has received the message "ActiveNode," received Messenger from PA_2 listed in the tail of the queue. According to the topological relationship, under the rule

of FIFO, PA_3 activates the Messenger. Up to now, the transferring work is finished.

6. When all Messengers are activated, each KA will restore running environment and do instruction "ExecuteTask."
7. During the executing of each KA, on the one hand, the historical snapshots will be recorded and saved (including the structure of KA, the association relationship between KA, the space occupied by KA, the type of KA, the information of VA for task transferring and Messenger information for scheduling all agent (including VA, EA and DA), the state of Messenger for scheduling), on the other hand, VA do the instruction "ListenTask" continuously and get the next transferring instruction "TransferSignal."
8. During the executing of KA, if no instruction "TransferSignal" is got by VA, KA will execute its task continuously until the task is completed, otherwise, it will stop executing and go to Step 1 for preparing the new turn of transferring. The new turn process will be subject to the subscribed TRAVEL SCHEDULE/PLAN. The whole process of transferring is seamless.

The previous approach is for a kind of general case. The other special cases is similar to this one, such as if TA wants to interact with PA2/PA3, PA1/PA2 should be involved according to the subscribed TRAVEL SCHEDULE/PLAN.

Transferring Failure Problem

In the pervasive computing environment, because the position of agent is often variable, the cases may be occurred. When the agent1 is being transferred to agent2 and wanted to be embedded in agent2 on node C to deal with the task together, but the agent2 has moved from node C to node D during the transferring of the agent1. That is to say, when agent1 arrives at node C, the agent2 can't be found by the agent1. This case is so-called *transferring failure problem*. This kind of problem can't keep the continuity of transferring of task.

In order to solve this problem, there are three factors should be considered:

1. When the position of agent has moved, how to know this change by other relative agents.
2. When the transferring of agent, how to deal with the message sent to it.
3. During the transferring of agent, whether the receiving agent can be transferred freely or not.

Our solution is as follows:

1. When the agent moves to new node, it should send "Notification" Message to all other relative agents.
2. Before the agent prepares to be transferred, it should look up the current position of receiving agent, at the same time, the transferring relationship and event should be sent to it by message
3. The transferring topological relationship of agent should be determined. We select the rule "FIFO (First In First Out)" for it. When some agents begin to be transferred, the other receiving agent should be locked. After the transferring process is over, the receiving agent should be unlocked at once. Based on this rule, a signal semaphore may be set up.

The "Notification" Message may adopt three kinds: Unicast, Multicast, Broadcast.

For different applications, in detail, it can be divided as follows:

1. Unicast, Multicast, Broadcast in the GROUP
2. Unicast, Multicast, Broadcast among the GROUPS
3. Broadcast from a agent to MAS

4. Unicast, Multicast, Broadcast among different MAS

The formal expression of "Notification" Message of agent is that:

Agent_Message_Notify (Sender, Receiver , Message)

Correspondingly, the formal expression of Unicast, Multicast, Broadcast is that:

Agent_Message_Notify (agent1, agent2, Message)
Agent_Message_Notify (agent, Multicast (agentX) , Message) where "agentX" shows other agents
Agent_Message_Notify (agent, Broadcast (Any), Message) where "Any" shows any agent of Groups

Based on the "time-topological" relationship, we design a kind of synchronal mechanism "addressing first, then locking and transmitting," which can realize the synchronization between transferring agent and receiving agent. So the transferring failure problem can be solved radically and adapted for all kinds of application pattern. The "time-topological" relationship can be used in the "Travel Schedule for transferring." The schedule may consist of certain travel sequences, each travel sequence includes the following DATA STRUCTURE:

Schedule ID TP_ID, Schedule Name TP_Name, Schedule Made Date TP_Time, Travel Number No, Transferring Object TP_Object, Transferring granularity TP_Granule, Source Address of Transferring Object TO_IPC, Target Address of Transferring Object TO_IPD, Transferring Condition TP_Condi, Transferring Mark TP_ SnapshotPoint, Entry Address for Re-running TP_RunEntry.

Transferring Condition TP_Condi, Transferring Mark TP_SnapshotPoint, Entry Address for Re-running TP_RunEntry are important for basic transferring operation, that is to say, Only the TP_Condi is OK, the "Travel Schedule for transferring" may be run, meanwhile, record and save "TP_ Snapshot Point" and "TP_RunEntry," both for restoring the running environment. Whether TP_Condi is OK or NOT, the following aspects should be set and checked:

1. Whether or not the current status (may be divided into five kinds: Ready 1, Waiting 2, Transferring 3, Running 4, Destroyed 5) of agent is "Waiting 2."
2. Whether the target address TO_IPD may be reached or not.
3. Whether the threshold of transferring delay is OK or not
4. Whether the residual dependency cases may occur or not.

Transferring Delay

It is named "Transferring Delay" that the time interval from the suspended snapshot point of a running agent to the re-run snapshot point on the target node. The delay is one of main parameters to access the seamless mobility. The information for transferring an agent includes that instruction sets, address sets, runtime state when suspending, executing code, data, Messenger schedule information, and so on. Messenger schedule information and runtime state after being suspended must be transferred totally, but other information may not be transferred totally. Once the necessary information has been restored on the target node, especially the snapshot point of runtime state before being transferred, the agent may be re-run.

Three factors is mainly involved in the transferring delay of agent:

1. Which kind of transferring granularity, Strong Transfer or Weak Transfer.
2. When will the agent be transferred, which is the suspending time of the agent. The occasion of suspending agent is divided

into three kinds: Suspend immediately after determining to transfer, Suspend after the total information is transferred completely, Suspend after the Key Set is transferred successfully.

3. When will the agent be restored and run, which is the restoring time of agent. The occasion of restoring agent on the target node is divided into two cases: restore after the total information is transferred completely, restore after the Key Set is transferred successfully.

In fact, there are three valid mapping forms based on the previous three factors:

1. **Strong Transfer:** Suspend after transferring, restore and run after finishing the whole information.
2. **Strong Transfer:** Suspend after transferring, restore and run timely after finishing the Key Set information (at the same time, the whole residual information will continue transmitting).
3. **Weak Transfer:** Suspend after transferring, restore and run after finishing the Key Set information (at the same time, the selected partially information will continue transmitting to the target node).

In the first mode, the transferring delay of agent includes that packing the whole information and transmitting, restoring the agent and the whole information on the target node. In the second mode, includes that packing the Key Set information and transmitting, restoring the agent and the Key Set. In the third mode, includes that packing the Key Set information and transmitting, restoring the agent and the Key Set. In the same condition, the delay of the first one is longest, the third one in shortest.

Residual Computation Dependency Problem

In the second and third transferring mode, because of selecting a part information as the "Key Set" and being transferred firstly, but different applications, it is NOT known which part of information is necessary. So a certain "Key Set" may NOT transferred to the target node timely, the running agent must wait for it. That is to say, running agent is still dependent on part of information on the source node. This case is named "residual computation dependency." This problem may lengthen the transferring delay of agent, when serious, it will influence the seamlessness. So the cases must be avoided.

In our opinion, the "residual dependency" problem will be solved from two aspects:

1. Tuning reasonably the transferring granularity of Execution/code Agent (EA), Data Agent (DA) and other agents (such as environment-state agent). If too larger, it is restricted by the bandwidth. If too smaller, transferring time is much more. Both may lengthen the delay. Based on analyzing from theory and application tests, we suggests a partition method named "subsection" or "pagination," which determines the size or number of "section" or "page" by bandwidth, buffer, volume of MessageBox. When the cases are occurred, the necessary information may be transmitted through "section interruption" or "page interruption," but the frequency should be adjusted automatically according to historical record information.
2. Optimize the Key Set. The Key Set will be determined automatically according to nearest principle and used frequently principle and cut off the redundancy information. The relative adapted strategy may be referred the bibliography (Milojicic, 2002; Takasugi, 2001).

Communication Primary for Migration

In order to support Web-based seamless migration, we have designed several communication primaries for transferring agent, a part of them are as follows:

1. **BeginToListen Primary:** VA will invoke the BeginToListen Primary to listen the port nPort, meanwhile, register the callback function OnAccept to receive connection message from other agents, when VA have received the connection request, the OnAccept will call back the connection_ID Connection_ ID. The Primary is:
   BeginToListen(UINT nPort,ACCEPT_CALLBACK callback);
   Void (CAgent::* ACCEPT_CALLBACK) (UINT& Connection_ID).
2. **BeginToRequest Primary:** When VA wants to set up connection to target agent, it will invoke BeginToRequest Primary to send request to the agent with IP:nPort, if it is successful, the connection ID nConnection_ID will be called back. The Primary is:
   BeginToRequest(UINT &nConnection_ID, CString IP,UINT nPort).
3. **Tranfer Primary:** When agent wants to send messages or transfer task to other objects, it will invoke Tranfer Primary to do it. The Primary is:

   Transfer (UINT nConnection_ID, CString strMsg, CDate time_stamp).

Similarly:

PrepareForRecv Primary for receive message or data stream:

PrepareForRecv(UINT nConnection_ID, RECEIVE_CALLBACK callback);

Void (CAgent::* RECEIVE_CALLBACK) (CString &strMsg, UINT& ConnectionID).

PrepareForClose Primary for close or end the connection:

PrepareForClose(UINT nConnection_ID, CLOSE_CALLBACK callback);

Void (CAgent::* CLOSE_CALLBACK) (UINT &Connection_ID).

THE IMPLEMENTED PLATFORM

The function and service of software platform (Simon, 2002) supporting task-oriented mobile distance learning paradigm — Web-based seamless migration should include:

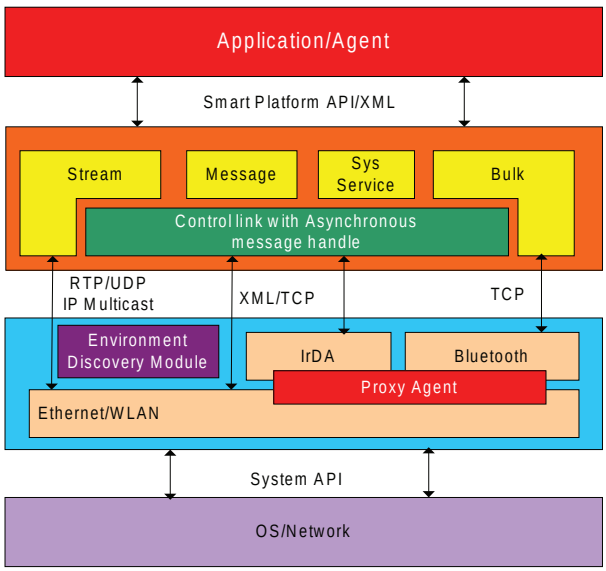
1. **Management method of resource and services:** When a new mobile device is brought into a space or new module or component used in the old device, the software manager can know how to spontaneously discovery them and what is wanted to be interactive. Because the resources of device are not same in a system, they may be embedded device, wearable computing device, mobile computing device, etc. their computing capability, memory capability, interactive mode are different. When the device is mobile or nomadic in the different environment, the interconnection problem is existed. The infrastructure can transform or translate the contents.
2. **Support for message-oriented, stream-oriented or bulk-oriented communication:** Here we argue that actually there are three catalogs of communications needs in runtime environment of mobile distance learning with different QoS requirements. Message-oriented communication is a kind of communication that occasionally happens and usually has high-level semantics for distance learning, e.g. a command asking a module or component to play the specified information. These communications are sensitive to the loss of messages; whereas their requirements on the delivery

latency are moderate, as long as it is within a reasonable boundary, say, 50 ms, according to the cognitive character of human. Stream-oriented communication is a kind of communication that constantly occurs. Their semantic level usually is relatively low and the drop of several data units is usually tolerable. However, they are sensitive to the variation of the delivery latency, while in most cases their requirement on the delivery latency is also moderate. Bulk-oriented communication is a kind of communication with much larger Bytes amount (maybe several K/Mbytes) to be delivered, which is not much sensitive to the variation of the delivery latency. The work paradigm may be Client/Server, Browse/Server and Peer-to-Peer, so the protocol stack of communication is a certain kind of link of IP — TCP/UDP/RTP — HTTP/FTP.

3. **Coordination mechanism of continuity and self-adaptability:** As a distributed mode, the infrastructure can coordinate the relationship of association, communication, collaboration of modules, so coordination mechanism of continuity and self-adaptability among modules or component is more important to the whole function and services. Of course it is important that supporting one-to-many communication, heterogeneous platforms and implementation languages. It is common in runtime environment of mobile distance learning that a message should be delivered to many modules or components simultaneously. Even in the case of stream-oriented communication, there will be multiple learners of a single stream. Therefore, it is necessary for software platform to have the one-to-many communication capability. Modules or components in a runtime environment usually impose different requirements on the underlying hardware and OS. Moreover, they are often implemented in different languages. A software platform should have the adequate capability to deal with all these diversities.

Based on analyzing to the function and service of software platform, we have designed and developed it. Figure 1 is the structure description of our implemented platform supporting task-oriented

Figure 1. Structure description of platform



mobile distance learning paradigm — Web-based seamless migration. Figure 2 is the structure of seamless migration embedded in this implemented platform, which includes four layers: SM-link layer, SM-path layer, SM-connection layer and SM-session layer. In Figure 2, "T" stands for "Task," "A" stands for "Agent," "MA" stands for "Mobile Agent" and "C" stands for "Container," which is a daemon threads component installed in each relative mobile devices. Their working principle has been previously mentioned.

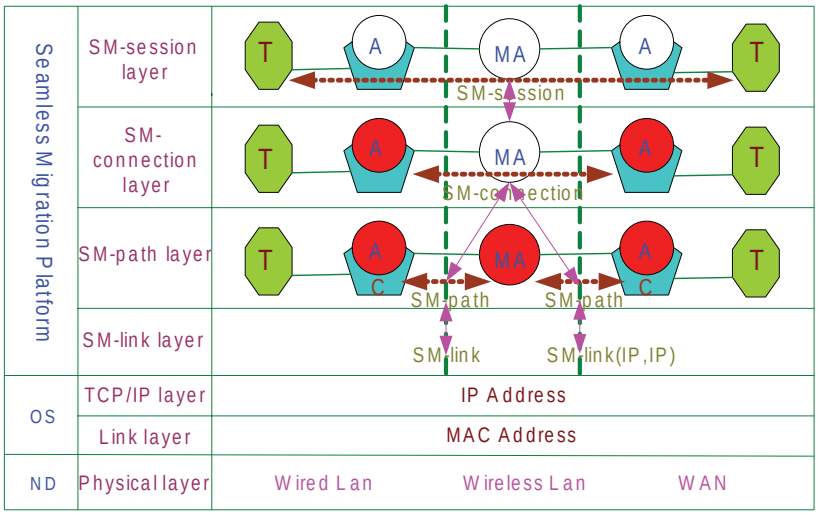
Our implemented platform can work in Client/Server, Browse/Server and Peer-to-Peer paradigm. The platform is a multi-agent system. The structure can be divided into multiple levels. Multiple agents are collaborated for seamless mobility. Each one has its special function. The agent class and container class can be partially defined as follows:

```
class CAgent  {
public:
        CAgent();
        virtual ~CAgent();
        BOOL Register();
        BOOL Quit();
BOOL Subscribe(CString strGrpName, NOTIFY_
CALLBACK callback, CString strTemplate="");
        UINT GetSharedFile(LPCTSTR
url,LPCTSTR lpszTagInfo=NULL);
        virtual void OnConnect();
        virtual void OnDisconnect();        ...
};

class Ccontainer {
public:
    CContainer();
    virtual ~CContainer();
    BOOL LaunchAgentByName(CString strAgt-
Name);
    BOOL LaunchAgentByPath(CString strPath);
    void ProcessDSCmd(CDSMsg & msg); …;
typedef struct _MINIHTTP_REQUEST {
    SOCKET          socket;
    char*           http_data;
    unsigned long   http_data_size;
    MINIHTTP_FIRST_LINE*  first_line;
} MINIHTTP_REQUEST;
typedef struct _MINIHTTP_RESPONSE {
    unsigned int    range_begin;
    unsigned int    range_end;
    unsigned long   http_data_size;
    int             http_response_code;
} MINIHTTP_RESPONSE; ... }
```

There are four kinds of components in our platform: Management Interface Component, Task Manager Component, Continuity Manager Component, and Service Manager Component.

Figure 2. Structure of seamless migration embedded in the platform



The interface is often used as defining the attributes of agent, such as ID of Agent, Name of Agent, Type of Agent (such as TA, SA, UA, VA, EA, DA, and so on), Current Status of Agent (five status are "Ready," "Waiting," "Transferring," "Running," "Dead" or "Destroyed"), Association Relationship of agent (including relationship between agent and task, relationship between two agents). The task manager is for application service, which manages the application/task array, including task description, task analyzing, mapping, or binding between task and service, loading, executing, scheduling of task, etc. The continuity manager is for agent management, context-awareness computing, and history/status recording, such as making of "Transferring Travel Schedule" of agent, addressing of VA, determining of transferring granularity which is for avoiding the transferring failure, reducing the residual dependency and contracting the transferring delay, etc. The service manager conducts the registration of service, discovery of service, service association, and mapping or binding between task and service of mobile distance learning. Service discovery is the base of seamless mobility of task of learning. Currently, several discovery ideas have been designed or used, such as Service Location Protocol, Jini, Salutation, Universal Plug and Play, Bluetooth Service Discovery Protocol, and others (Garlan & Siewiorek, 2002).

These components can communication each other, and may be controlled by application interactive interface including agents and global control of task, which is interface of human computers, such as PC, laptop, PDA, Mobile phone, embedded devices. The stationary or mobile agent is the basic encapsulation of the software modules in the system for management of service and mobility. Each computer in the runtime environment of mobile distance learning will host a dedicated process called Container, which provides system-level services for the agents that run on the computer and manages them as well. It makes the details of other parts of the system transparent to agent and provides a simple communication interface for stationary or mobile agent. There is one global dedicated process in the environment, which mediates the "delegated communication" between stationary or mobile agent and provides services such as directory service, dependency resolution.

TEST OF MOBILE DISTANCE LEARNING BASED ON OUR PLATFORM

Our implemented platform can show many scenarios, such as Web-based seamless migration with "Event-Type" task, "Bulk-Type" task, "Stream-Type" task for task-oriented mobile distance learning.

Here is an example that includes Web-based seamless migration for task of learning on PC, laptop, or PDA under dynamic changes of the network and environment without user awareness or intervention. Just like Figure 3, the task of learning can follow me form one device to another device or from my house to other places, such as my office, stadium, coffee house, park, airport, etc., and vice versa.

Now supposing the task of learning consists of three sub-tasks of mobile distance learning: playing video, playing mp3, and reading documents. As a demo of many scenarios, this kind of task of learning is described partially by SIML as follows:

```
<smil>
<head>
<layout>
<root-layout background-color="#D3DD86"
  width="640" height="480" />
<region id="videoregion" top="0" left="0"
  width="320" height="240" />
<region id="textregion" top="0" left="321"
  width="320" height="159" ->
</layout>
</head>
<body>
```

```
<seq><par>
<video src=”Pervasiv.avi” region= ”videoregion”
   begin=”0.6s” />
<audio src=”IloveChina.mp3” begin=”2.6s” />
<textstream src=”Pervasive.txt” region=
   ”textregion” begin=”5s”
end=”9000s” />
</par></seq>
</body>
</smil>
```

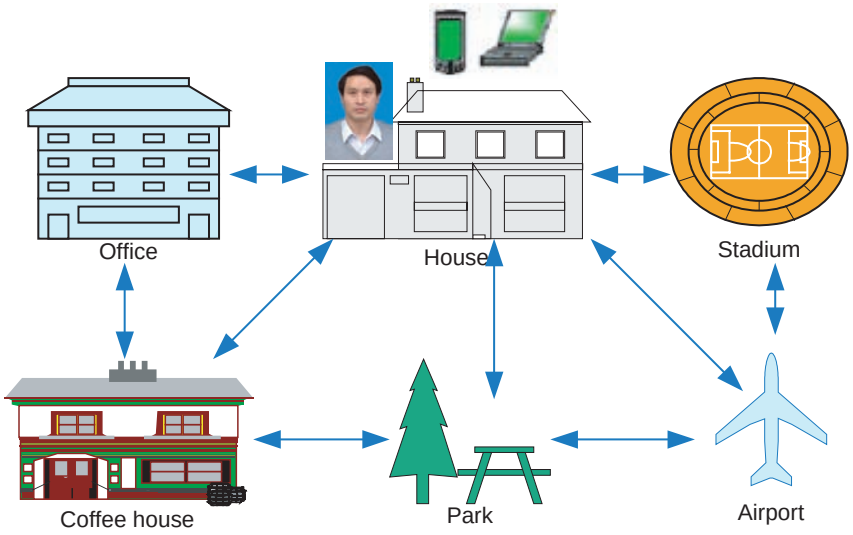
The previous task of learning has three sub-tasks of learning: Playing *Pervasiv.avi*, playing *IloveChina.mp3*, reading *Pervasive.txt*. They will be done according to the time-sequence in parallel mode based on the algorithm mentioned previously. With the learner's movement from one station (such as House) to another station (such as Airport), these uncompleted sub-tasks of distance learning can seamlessly migrate from PC of his or her house to laptop, or PDA with him or her in Airport and go on learning (watching, listening, and reading) continuously by mobile agent on our platform supporting Web-based seamless migration.

In our experiments of mobile distance learning, the deployment of device is the CPU frequency, RAM of PC and laptop (a kind of mobile device) are 1.2 GHz, 512 MBytes,respectively, and 450 MHz, 64 MBytes RAM are for PDA (another kind of mobile device), the speed of wired network and wireless network is 10 M/100 MHZ, 1-3 MHZ, respectively. The nodes were connected by wireless and wired Web network — Internet.

CONCLUSION

In order to meet the application requirements of mobile distance learning, we have proposed a kind of novel distance learning paradigm — task-oriented Web-based seamless migration, which supplies the function that the task of distance learning dynamically follows the learner from place to place and machine to machine, so it is convenient to learn during mobility, and is useful or helpful for the mobile learner or mobile attendee. Our key idea is that this capability can be achieved by layering architecture of component platform and agent-based migrating mechanism. In this article, we have given the formal description of the task of mobile distance learning, discussed the migrating granularity of the task of learning. The

Figure 3. Task can be migrated with me from one place to another place



innovative significance is that we have designed a kind of approach of Web-based seamless migration, including solving these problems, such as shortening migration delay, avoiding migration failure and residual computation dependency. The validity of this approach and its corresponding software platform for mobile distance learning has been tested by many demos.

REFERENCES

Ciancarini, P. (2002). Coordinating multi-agent applications on the WWW: A reference architecture. *IEEE Trans. on Software Engineering, 24*(5), 363-375.

Danny, B. L., & Mitsuru, O. (2001). Seven good reasons for mobile agents. *Communications of the ACM, 42*(3), 86-89.

David, K., & Robert, S. G. (2002). Mobile agents and the future of the Internet. *ACM Operating Systems Review, 33*(3), 7-13.

Garlan, D., & Siewiorek, D. P. (2002). Project aura: Toward distraction-free pervasive computing. *IEEE Pervasive Computing, 1*(2), 22-31.

Karnik, N. M. (1991). Design Issues in mobile agent programming systems. *IEEE Concurrency, 6*(3), 125-132.

Milojicic, D. (2002). Mobile agent applications. *IEEE Concurrency, 7*(3), 80-90.

Satyanarayanan, M. (2001). Pervasive computing: Vision and challenges. *IEEE Personal Communications, 8*(8), 10-17.

Shi, Y. C., & Xie, W. K. (2003). The smart classroom: Merging technologies for seamless tele-education. *IEEE Pervasive Computing Magazine, 2*(2), 25-33.

Simmons, R., & Apfelbaum, D. (2001). A task description language for robot control. In *Proceedings Conference on Intelligent Robotics and Systems*, New York (Vol. 1, No. 10, pp. 138-147).

Simon, S. (2002). A model for software configuration in ubiquitous computing environments. In *Proceedings of Pervasive. LNCS 2414*, Zürich (Vol. 1, No. 7, pp. 181-194).

Takasugi, K. (2001). Adaptive system for service continuity in a mobile environment. In Proceedings of *IEEE APCC*, Tokyo, Japan (Vol. 1, No. 9, pp. 75-83).

Takasugi, K. (2003). Seamless service platform for following a user's movement in a dynamic network environment. In *Proceedings of PerCom'03* (Vol. 1, No. 8, pp. 125-132).

Weiser, M. (1991). The computer for the twenty-first century. *Scientific American, 265*(3), 94-104.

This work was previously published in International Journal of Distance Education Technologies, Vol. 4, Issue 3, edited by S. Chang and T. Shih, pp. 62-76, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.9
TCP Enhancements for Mobile Internet

Bhaskar Sardar
Jadavpur University, India

Debashis Saha
Indian Institute of Management (IIM) Calcutta, India

INTRODUCTION

Transmission Control Protocol (TCP), the most popular transport layer communication protocol for the Internet, was originally designed for wired networks, where bit error rate (BER) is low and congestion is the primary cause of packet loss. Since mobile access networks are prone to substantial noncongestive losses due to high BER, host motion and handoff mechanisms, they often disturb the traffic control mechanisms in TCP. So the research literature abounds in various TCP enhancements to make it survive in the mobile Internet environment, where mobile devices face temporary and unannounced loss of network connectivity when they move. Mobility of devices causes varying, increased delays and packet losses. TCP incorrectly interprets these delays and losses as sign of network congestion and invokes unnecessary control mechanisms, causing degradation in the end-to-end goodput rate. This chapter provides an in-depth survey of various TCP enhancements which aim to redress the above issues and hence are specifically targeted for the mobile Internet applications.

BACKGROUND

As wireless devices are becoming the fastest growing segments of the computer industry, the networking picture has changed radically in the last decade. Millions of people now want to access the Internet at any time from wherever in the world they may be. To allow this, mobile IP [PER96] has been developed to route packets to these mobile users. As a best effort type of protocol, mobile IP has fulfilled its task fairly well; but TCP [POS81] has to glue well to the mobile IP in order to provide the applications with an end-to-end and connection-oriented packet transport mechanism that ensures reliable and the ordered delivery of data. However, in the absence of wireless enhancements for TCP to work over

mobile Internet, several known problems affect its performance [CAC95]. Nevertheless, most of the wireless data applications (e.g., FTP, Web, telnet, multicasting, etc.) use TCP as the default ***transport layer protocol,*** as they want to achieve reliable and guaranteed delivery of data. But TCP, having faced several problems specific to this network, poses a huge bottleneck to reaching a high goodput rate.

As a result, over the years TCP has been modified several times to improve its performance, and, hence, several important TCP versions have emerged, such as TCP-Tahoe [FAL96], TCP-Reno [FAL96], TCP-Vegas [BRA95], TCP-New Reno [FLO99], and TCP-SACK [FAL96],[MAT96]. However, all these mechanisms and various versions do not work the same, when called to work in diverse environments such as satellite networks, last hop wireless networks, and mobile ad-hoc networks. In [TSA02], [TIA05] authors have compared several TCP enhancing schemes for mobile/wireless networks. In [TSA02], Tsaoussidis and Matta have considered the effect of high BER, unexpected disconnection, and battery power for comparing various TCP enhancing schemes. They are of the opinion that the error detection mechanism must be able to classify different types of errors (e.g., congestion, transient wireless error, persistent wireless error, and handoff), and, based on the error classification, an appropriate recovery strategy must be employed that differs from congestion-oriented mechanisms employed by TCP. They argued for the importance of defining a new performance metric (e.g., energy efficiency) to measure protocol stability and fairness in last hop wireless networks. In [TIA05], Tian et al., have considered different application areas (e.g., cellular, satellite, ad-hoc, and heterogeneous networks) for TCP. But they concentrated on the effect of high BER and channel asymmetry on the performance of TCP in all four application areas.

TCP IN MOBILE INTERNET

Problem of Running TCP in Mobile Internet

The following characteristics have major impact on the performance of TCP in Mobile Internet [SAR06]:

- **High BER:** The bit error rate in wireless networks is much higher than those experienced in traditional wired networks. High BER results in a large number of packet drops. TCP treats these drops as congestion loss and starts congestion control procedures resulting in a degraded performance.
- **Limited spectrum:** Bandwidth is a scarcer resource in wireless networks than its wireline counterparts (e.g., the bandwidth of fast Ethernet is 100 Mbps, whereas GPRS has a bandwidth of 384 Kbps). So, sharing wireless bandwidth efficiently between mission critical and non-critical traffic is a very important task.
- **Handoff:** When a user leaves a cell and enters a new one, handoff takes place. During handoff, a mobile host may lose connection to the base station, and any data transmitted for the mobile host are lost. TCP treats this packet loss as congestion and slows down transmission rate resulting.
- **Unpredictable delay:** As a mobile user moves randomly, distance from a BS varies, resulting in temporally varied delay. This unpredictable delay is difficult for TCP to handle gracefully.
- **Frequent disconnection:** Mobile hosts often get disconnected (when in motion and/or discharged battery) without any warning. Transmission during this period causes huge packet drops leading to pseudo-congestion and hence degraded performance.
- **Limited energy:** Mobile devices are battery powered, and, hence, cannot afford too many

retransmissions, unlike electrically powered devices. In other words, TCP is not designed as an energy-efficient protocol.

Classification of TCP Enhancements

At first approximation, TCP enhancements for mobile Internet can be divided into two main categories as shown in the Figure 1.

In this section we will describe notable proposals which have been made in the literature to improve the performance of TCP in both ***3G*** cellular networks and ***WLAN***. We will provide a comprehensive comparison of these proposals and show which problems are solved by these proposals and which have not yet been solved.

TCP for 3G Cellular Networks

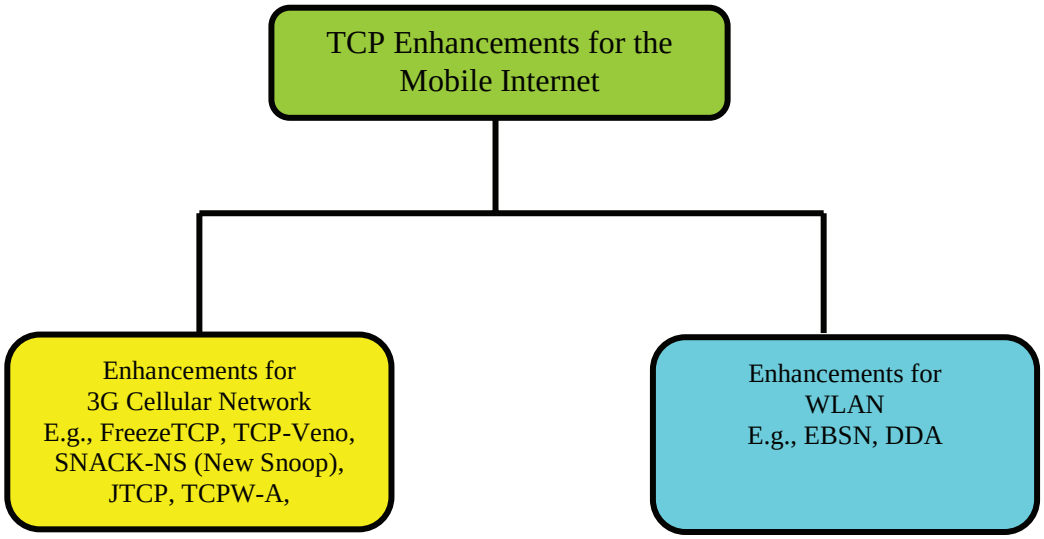
Freeze-TCP [GOF00]

Freeze-TCP was designed mainly to deal with frequent disconnection and handoff. The main idea is to move the onus of signaling an impending disconnection to the Mobile Host (MH). It avoids timeouts at the sender during periodic disconnection and handoff, since a timeout shrinks the sending window and reduces the performance. It exploits the ability of the MH to advertise a window of zero. Based on the signal strength, the MH detects the handoff and advertises a zero window adjustment (ZWA). The sender then freezes transmission and timeout value and enters the persist mode. Once the MH is reconnected, it advertises a non-zero window, so that the sender resumes with their window and timeout values unaffected due to handoff. To implement Freeze-TCP, the network stack needs to be aware of mobility (at least to some extent). In essence, some cross-layer (layers of the protocol stack) efforts and information exchange are needed. It is also necessary for the MH to predict impending disconnection within the round-trip time. If a disconnection cannot be predicted, the behavior and performance will be exactly that of standard TCP.

TCP-Veno [FU03]

TCP-Veno is a novel end-to-end congestion control mechanism and effective for dealing with random

Figure 1. Classification of TCP enhancements



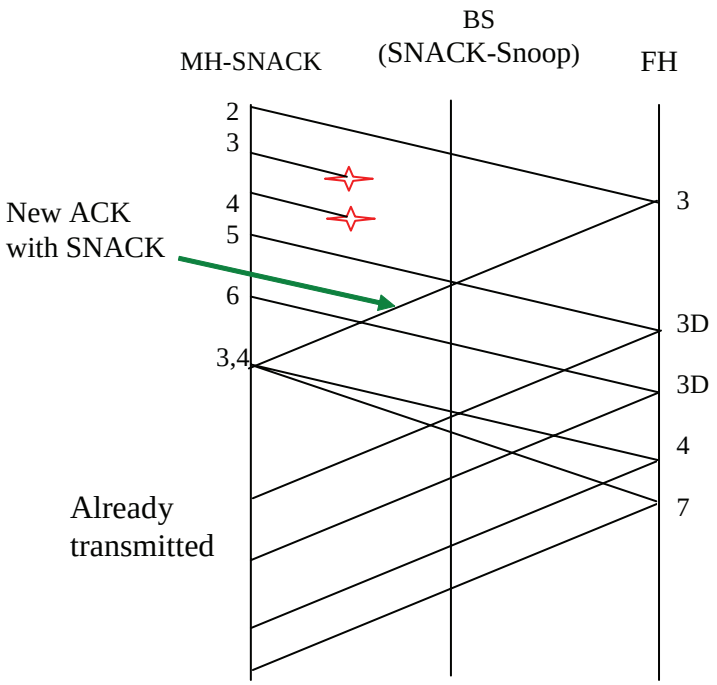
packet loss. TCP-Veno exploits a mechanism similar to Vegas [BRA95] to estimate the state of the connection and applies the AIMD schemes of TCP-Reno. Specifically: (1) it refines the multiplicative decrease algorithm of TCP-Reno by adjusting the slow start threshold according to the perceived network congestion level rather than a fixed drop factor, and (2) it refines the linear increase algorithm so that the connection can stay longer in an operating region in which the network bandwidth is fully utilized. It calculates the number of backlog packets as in Vegas but uses this value as an indication whether the connection is in a congestive state. If packet loss is detected when the connection is in the congestive state, it will be considered as congestion loss; otherwise, it is assumed as random loss. TCP-Veno can reduce the amount of window size degradation but might not have good behavior when the random loss is high. It does not deal with disconnection and handoff.

SNACK-NS (New snoop) [SUN04]

SNACK-NS is a link layer retransmission protocol with the ability to overcome the limitations of batched ACK used by TCP. SNACK-NS consists of two protocol components: SNACK-Snoop and SNACK-TCP. The SNACK-Snoop is deployed at the base station (BS) and SNACK-TCP is deployed at the MH. SNACK mechanism is designed to provide explicit information on multiple packet losses over wireless link. In SNACK, explicit loss information is conveyed by several loss blocks, and each block stores the sequence number of the lost packets.

Figure 2 shows the recovery steps after the drops of packets three and four. During transmission from MH to fixed host (FH), SNACK-Snoop does not require the storage of arriving packets. It only stores the sequence number of the received packets to determine the losses. When an ACK

Figure 2. Recovery from two losses in MH to FH direction

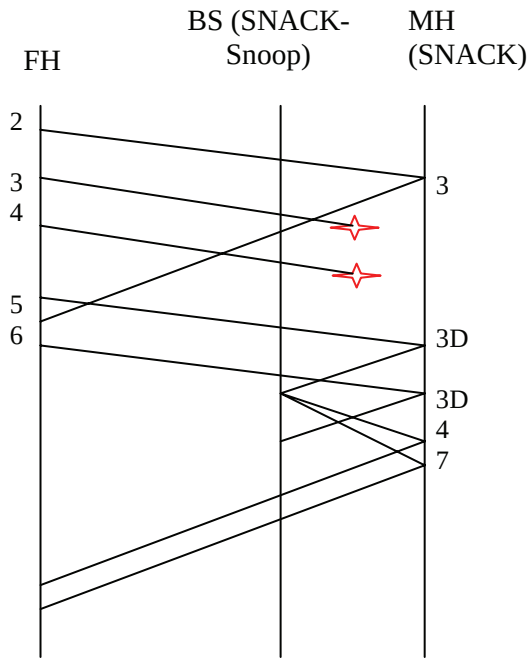


arrives at the BS, whether new or duplicate, SNACK-Snoop can detect the wireless losses, if any, and piggyback all their sequence numbers to the ACK as primary multiple wireless loss information. For data transmission from fixed host to mobile host (see Figure 3), SNACK-Snoop stores all the packets received by the base station. These packets are used to judge whether a loss is due to congestion or wireless error, and supports fast local retransmission over wireless links. The protocol fails when IP payload is encrypted.

JTCP [WU04]

JTCP is a Jitter-based, end-to-end, robust, and fair protocol for the heterogeneous wireless networks. It was designed to identify the causes of packet lose and react accordingly. To differentiate between congestion and wireless losses, the protocol uses jitter ratio, which is the loss predictor formulated by the inter-arrival jitter. A congestion event is defined as when the timer expires or three DUPACKs are received and the jitter ratio is significant. As three DUPACKs are received, JTCP determines if these DUPACKs arrive within a RTT. If the period for DUPACKs has been extended to the next RTT and jitter ratio is significant, the protocol treats it as one congestion event and enters into fast recovery phase as TCP-Reno. Otherwise it enters the immediate recovery phase as it assumes that losses are due to wireless error. After a timeout, if the jitter ratio is significant, it enters slow start phase, otherwise it assumes burst losses caused the timer expiration and enters fast recovery phase because of the non-congestion event. As multi-losses occur in the same RTT, it considers them as one congestion event and reduces the sending rate once. It performs poorly in the presence of frequent disconnection and handoff. It reduces the sending rate even if there were no congestion or wireless errors.

Figure 3. Recovery from two losses in FH to MH direction



TCP-Westwood with Agile Probing (TCPW-A) [WAN05]

TCPW-A is a sender-side only enhancement of TCPW [GER01] that deals with highly dynamic bandwidth, large propagation time/bandwidth, and random loss. Along with eligible rate estimate (ERE) mechanism of TCPW, TCPW-A uses two more mechanisms: agile probing and persistent non-congestion detection (PNCD). The PNCD mechanism is concerned with how to detect extra unused bandwidth. PNCD identifies the availability of persistent extra bandwidth in congestion avoidance, and invokes agile probing accordingly. Agile probing adaptively and repeatedly resets slow start threshold-based ERE. Each time the slow start threshold is reset to a value higher than the current one, the congestion window climbs exponentially to the new value. The result is fast convergence of the congestion window to a more appropriate slow start threshold value. Even if the PNCD algorithm can accurately detect non-congestion, there is always the possibility that the network becomes congested immediately after the connection switches to agile probing phase. One such scenario is after a buffer overflow at the bottleneck router. Many of the TCP connections may decrease their congestion window after a buffer overflow, and congestion is relieved in a short time period.

TCP for WLAN

The protocols described in this section are mainly designed to work in 3G cellular networks. But it has been seen that these protocols (e.g., [WAN98]) perform better when called to work in a WLAN environment.

Explicit Bad State Notification (EBSN) [BAK97]

[BAK97] proposed the EBSN scheme. When the wireless link is inbad state, BS sends an EBSN to the source, which causes the previous time out to be cancelled and a new time out put in place based on an existing estimate of round trip time. The major downside of this scheme is that end-to-end semantic is violated.

Delayed Duplicate Acknowledgement (DDA) [VAI99]

DDA scheme attempts to mimic the behavior of Snoop protocol [BAL95]. In this scheme, the BS does not need to look into the TCP header. This scheme may be preferred when encryption is used. The BS implements a link level retransmission scheme for packets those are lost on wireless link. This scheme uses link level ACK to trigger retransmissions. The TCP receiver attempts to reduce interference between TCP and link level retransmission by delaying the third and subsequent DUPACKs for some interval d. If d is chosen large enough to allow time for link level retransmission of the lost packet, then the retransmitted packet would reach the receiver before the third and subsequent DUPACKs could be sent. Since the TCP sender does not receive more than two DUPACKs, it will not fast retransmit. In the presence of real congestion, the performance of this scheme is degraded compared to standard TCP. Standard TCP will send the third DUPACK without any delay, thus initiating fast retransmission sooner than this scheme.

Overall Comparison of the Protocols

It is extremely difficult to do a comprehensive comparison of the proposals because each aims to solve a different problem (e.g., high BER, handoff, frequent disconnection, battery constraint, etc.) of the wireless link. Nevertheless, we make a bold attempt in Table 1 to compare the protocols against the issues discussed earlier. We add two more issues here: end-to-end TCP semantic and encrypted TCP payload. Although these two are not typical issues for Mobile Internet only, we

Table 1. General comparison of TCP enhancement schemes for mobile Internet

	Intermediate node TCP mod?	*End-to-End TCP semantic*	*Handing High BER*	*Frequent disconnection*	*Handoff*	*Real time handoff*	*Wireless bandwidth sharing*	*Encrypted TCP payload*	*Energy efficiency*
Freeze-TCP	No	Yes	No	Yes	Yes	No	No	Yes	Medium
TCP-Veno	No	Yes	No	No	No	No	No	Yes	Medium
SNACK-NS	Yes	Yes	Yes	No	No	No	No	No	Medium
JTCP	No	Yes	Yes	No	No	No	No	Yes	Medium
TCPWA	No	Yes	Yes	No	No	No	No	Yes	Medium
EBSN	Yes	No	Yes	No	No	No	No	No	Medium
DDA	No	Yes	Yes	No	No	No	No	Yes	Low

believe that these two features must not be violated in any TCP enhancement scheme.

It is clear from Table 1 [SAR06] that most the proposals try to maintain end-to-end semantics and are effective in dealing with high BER, but none of the proposals provides real-time handoff and roaming facility. However, only Freeze-TCP effectively deals with frequent disconnection and handoff. It is interesting to note that none of the proposals satisfies the criteria of efficient sharing of wireless bandwidth, which is very important in real life.

FUTURE TRENDS

Although the existing protocols provide some possible solutions to alleviate the problems of TCP in mobile networks, a careful scrutiny of the protocols indicates that, none of the protocols solves all the wireless specific issues of TCP. To design an efficient TCP for mobile Internet, the desirable properties of the protocol must include the solutions for all wireless specific problems of TCP. But we believe that it is difficult to create a "one size fits all" TCP for Mobile Internet. Also 3G cellular networks have substantially higher uplink rates. The channel interference characteristic on the uplink is very different from downlink. Uploading is becoming more and more important. So, TCP enhancement schemes must perform equally well in the case of transmission from MH to FH direction.

CONCLUSION

When TCP encounters packet drop, it usually invokes congestion control and avoidance pro-

cedures. Due to the characteristics specific to wireless networks, such as signal fading and mobility, packets may be lost due to congestive and noncongestive losses. So it might mistake a channel loss or losses due to temporal disconnection and handoff as congestion event, and reduce the window size immediately. This makes TCP a bad choice for Mobile Internet (cellular or WLAN). In this chapter, we have provided a comprehensive and in-depth survey on recent research in TCP for mobile Internet. The taxonomy and characteristics of TCP enhancements for mobile wireless access networks are introduced, and a categorized analysis of different existing solutions show that researchers are yet to find an elegant technique to detect the exact cause of packet loss.

REFERENCES

Bakshi, B. S., Krishna, P., Vaidya, N. & Pradhan, D. K. (1997). Improving performance of TCP over wireless networks. *ICDCS*, 365-373.

Balakrishnan, H., Seshan, S., & Katz, R. H. (1995). Improving reliable transport and handoff performance in cellular wireless networks *ACM Wireless Networks, 1*(4), 469-481.

Brakmo, L., & Peterson, L. (1995). TCP Vegas: End to end congestion avoidance on a global Internet. *IEEE JSAC, 13*(8), 1465-1480.

Caceres, R., & Iftode, L. (1995). Improving the performance of reliable transport protocol in mobile computing environment. *IEEE Journal of Selected Areas in Communications, 13*(5), 850-857.

Fall, K., & Floyd, S. (1996). Simulation-based comparisons of Tahoe, Reno, and SACK TCP. *Computer communication review, 26*(3), 5-21.

Floyd, S., & Henderson, T. (1999). *The new-Reno modification to TCP's fast recovery algorithm*. RFC 2582.

Fu, C. P., & Liew, S. C. (2003). TCP Reno: TCP enhancement for transmission over wireless access networks. *IEEE JSAC, 21*(2), 216-228.

Gerla, M., Sanadidi, M., Wang, R., Zanella, A., Casetti, C., & Masco, S. (2001). TCP Westwood: Congestion window control using bandwidth estimation. *IEEE GLOBECOM*, 3, 1698-1702.

Goff, T., Moronski, J., Phatak, D.S., & Gupta, V. (2000). Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments. *IEEE INFOCOM*, 3, 1537-545.

Ka-Cheong, L., & Li, V. O. K., (2006). Transmission Control Protocol (TCP) in Wireless Networks: Issues, Approaches and Challenges. *IEEE Communications Surveys & Tutorials*, *8*(4), pp. 64-79.

Mathis, M., Mahdavi, J., Floyd, S., & Romanow, A. (1996). *TCP selective acknowledgement options.* RFC 2018.

Perkins, C. (1996). *IP mobility support.* RFC 2002.

Postel, J. (1981). *Transmission control protocol.* RFC 793.

Sardar, B., & Saha, D., (2006). A Survey of TCP Enhancements for Last-Hop Wireless Networks. *IEEE Communications Surveys & Tutorials, 8*(3), pp. 20-34.

Sun, F., Soung, V. L., & Liew, C. (2004). Design of SNACK mechanism for wireless TCP with New Snoop. *IEEE WCNC, 5*(1), 1046-1051.

Tian, Y., Xu, K., & Ansari, N. (2005). TCP in wireless environments: Problems and solutions. *IEEE Communication Magazine, 43*(3), S27-S32.

Tsaoussidis, V., & Matta, I. (2002). Open issues on TCP for mobile computing. *Journal on Wireless Communication and Mobile Computing, 2*(1), 3-20.

Vaidya, N., & Mehta, M. (1999). *Delayed duplicate acknowledgements: A TCP-unaware approach to improve performance of TCP over wireless.* Technical Report 99-003.

Wang, R., Yamada, K., Sanadidi, M. Y., & Gerla, M. (2005). TCP with sender-side intelligence to handle dynamic, large, leaky pipes. *IEEE JSAC, 23*(2), 235-248.

Wang, T. S. (1998). Mobile-end transport protocol: An alternative to TCP/IP over wireless links. *IEEE INFOCOM*, 3, 1046-1053.

Wu, E. H. K., & Chen, E. H. K. (2004). JTCP: Jitter-based TCP for heterogeneous wireless networks. *IEEE JSAC, 22*(4), 757-766.

Xu, K., Tian, Y., & Ansari, N. (2004). TCP-Jersey for wireless IP communications. *IEEE JSAC, 22*(4), 747-756.

KEY TERMS

Congestion Window: The maximum number of TCP packets that a sender is allowed to send at a time to the network, or in other words, it is the maximum carrying capacity of the network.

DUPACK: An acknowledgement packet that contains the sequence number of last acknowledged packets.

Energy Efficiency: The ratio of minimum energy consumption required to transmit a certain amount of data to that of actual energy consumption.

Goodput: The ratio of number of packets successfully transmitted to number of packets actually transmitted.

Mobile IP: A modified IP protocol to allow users on the move to access the backbone network like the Internet.

TCP: A true end-to-end, connection-oriented protocol for providing reliable and ordered delivery of packets to the application, bypassing the unreliable nature of the Internet.

Throughput: The ratio of number of packets successfully transmitted to amount of time required to complete the communication.

ENDNOTES

[1] Portions reprinted, with permission, from (B. Sardar, and D. Saha, "A Survey of TCP Enhancements for Last-Hop Wireless Networks", IEEE Communications Surveys & Tutorials, Vol. 8, No. 3, pp. 20-34, 2006) ©2006 IEEE".

This work was previously published in the Encyclopedia of Internet Technologies and Applications, edited by M. Freire and M. Pereira, pp. 619-625, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.10
A Cooperative Framework for Information Browsing in Mobile Environment

Zhigang Hua
Chinese Academy of Sciences, China

Xing Xie
Microsoft Research Asia, China

Hanqing Lu
Chinese Academy of Sciences, China

Wei-Ying Ma
Microsoft Research Asia, China

INTRODUCTION

Through pervasive computing, users can access information and applications anytime, anywhere, using any device. But as mobile devices such as Personal Digital Assistant (PDA), SmartPhone, and consumer appliance continue to flourish, it becomes a significant challenge to provide more tailored and adaptable services for this diverse group. To make it easier for people to use mobile devices effectively, there exist many hurdles to be crossed. Among them is small display size, which is always a challenge.

Usually, applications and documents are mainly designed with desktop computers in mind. When browsing through mobile devices with small display areas, users' experiences will be greatly degraded (e.g., users have to continually scroll through a document to browse). However, as users acquire or gain access to an increasingly diverse range of portable devices (Coles, Deliot, & Melamed, 2003), the changes of the display area should not be limited to a single device any more, but extended to the display areas on all available devices.

As can be readily seen from practice, the simplest multi-device scenario is when a user

begins an interaction on a first access device, then ceases to use the first device and completes the interaction using another access device. This simple scenario illustrates a general concern about a multi-device browsing framework: the second device should be able to work cooperatively to help users finish browsing tasks.

In this article, we propose a cooperative framework to facilitate information browsing among devices in *mobile environment*. We set out to overcome the display constraint in a single device by utilizing the cooperation of multiple displays. Such a novel scheme is characterized as: (1) establishing a communication mechanism to maintain *cooperative browsing* across devices; and (2) designing a *distributed user interface* across devices to cooperatively present information and overcome the small display area limited by a single device.

BACKGROUND

To allow easy browsing of information on small devices, there is a need to develop efficient methods to support users. The problems that occur in information browsing on the small-form-factor devices include two aspects: (1) how to facilitate information browsing on small display areas; and (2) how to help user's access similar information on various devices.

For the first case, many methods have been proposed for adapting various media on small display areas. In Liu, Xie, Ma, and Zhang (2003), the author proposed to decompose an image into a set of spatial-temporal information elements and generate an automatic image browsing path to display every image element serially for a brief period of time. In Chen, Ma, and Zhang (2003), a novel approach is devised to adapt large Web pages for tailored display on mobile device, where a page is organized into a two-level hierarchy with a thumbnail representation at the top level for providing a global view and index to a set of sub-pages at the bottom level for detail information. However, these methods have not considered utilizing multiple display areas in various devices to help information browsing.

For the second case, there exist a number of studies to search relevant information for various media. The traditional image retrieval techniques are mainly based on content analysis, such as those content-based image retrieval (CBIR) systems. In Dumais, Cutrell, Cadiz, Jancke, Sarin, and Robbins (2003), a desktop search tool called Stuff I've Seen (SIS) was developed to search desktop information including email, Web page, and documents (e.g., PDF, PS, MSDOC, etc.). However, these approaches have not yet taken into account the phase of information distribution in various devices. What's more, user interface needs further consideration such as to facilitate user's access to the information that distributes in various devices.

In this article, we propose a cooperative framework to facilitate user's information browsing in mobile environment. The details are to be discussed in the following sections.

OUR FRAMEWORK

Uniting Multiple Displays Together

Traditionally, the design of user interface for applications or documents mainly focus on desktop computers, which are commonly too large to display on small display areas of mobile devices. As a result, readability is greatly reduced, and users' interactions are heavily augmented such as continual scrolling and zooming.

However, as users acquire or gain access to an increasingly diverse range of the portable devices, the thing changes; the display area will not be limited to a single device any more, but extended to display areas on all available devices. According to existing studies, the user interface of future applications will exploit multiple coordinated

modalities in contrast to today's uncoordinated interfaces (Coles et al., 2003). The exact combination of modalities will seamlessly and continually adapt to the user's context and preferences. This will enable greater mobility, a richer user experience of the Web application, and a more flexible user interface. In this article, we focus on overcoming display constraints rather than other *small form factors* (Ma, Bedner, Chang, Kuchinsky, & Zhang, 2000) on mobile devices.

The Ambient Intelligence technologies provide a vision for creating electronic environments sensitive and responsive to people. Brad (2001) proposed to unite desktop PCs and PDAs together, in which a PDA acts as a remote controller or an assistant input device for the desktop PC. They focused on the shift usage of mobile devices mainly like PDAs as extended controllers or peripheries according to their mobility and portability. However, it cannot work for many cases such as people on the move without access to desktop computers.

Though multiple displays are available for users, there still exist many tangles to make multiple devices work cooperatively to improve the user's experience of information browsing in mobile devices. In our framework, we design a distributed interface that crosses devices to cooperatively present information to mobile users. We believe our work will benefit users' browsing and accessing of the available information on mobile devices with small display areas.

Communication Protocol

The rapid growth of wireless connection technologies, such as 802.11b or Bluetooth, has enabled mobile devices to stay connected online easily. We propose a communication protocol to maintain the cooperative browsing with multiple devices. When a user manipulates information in one device, our task is to let other devices work cooperatively. To better illustrate the communication, we introduce two notations as follows: (1) *Master device* is defined as the device that is currently operated on or manipulated by a user; and (2) *Slave device* refers to the device that displays cooperatively according to user's interactions with a master device.

We define a whole set of devices available for users as a cooperative group. A rule is regulated that there is only one master device in a cooperative group at a time, and other devices in the group act as slave devices. We call a course of cooperative browsing with multiple devices as a cooperative session. In such a session, we formulate the communication protocol as follows:

- A user selects a device to manipulate or access information. The device is automatically set as the master device in the group. A cooperative request is then multicast to the slave devices.
- The other devices receive the cooperative request and begin to act as slave devices.
- When the users manipulate the information on the master device, the features are automatically extracted according to the analysis of interactions, and are then transferred to slave devices.
- According to the received features, the corresponding cooperative display updates are automatically applied on the slave devices.
- When a user quits the manipulation of information in the master device, a cooperative termination request is multicast to the slave devices to end the current cooperative session.

Two-Level Browsing Scheme

We set out to construct distributed user interfaces by uniting the multiple display areas on various devices to overcome the display constraint in a single device. In our framework, we propose a two-level cooperative browsing scheme, namely within-document and between-document. If a

document itself needs to be cooperatively browsed across devices, we define this case as within-document browsing. Otherwise, if a relevant document needs to be cooperatively browsed across devices, we consider this case as between-document browsing.

1: Within-Document Cooperative Browsing

There exist many studies to improve the browsing experiences on small screens. Some studies proposed to render a thumbnail representation (Su, Sakane, Tsukamoto, & Nishio, 2002) on mobile devices. Though users can browse an overview through such a display style, they still have to use panning/zooming operations for a further view. However, users' experiences have not been improved yet since these operations are difficult to be finished in a thumbnail representation.

We propose a within-document cooperative strategy to solve this problem, where we develop a so-called two-level representation for a large document: (1) presenting an index view on the top level with each index pointing to detailed content portion of a document in the master device; and (2) a click in each index leads to automatic display updates of the corresponding detailed content in the slave devices.

We believe such an approach can help users browse documents on small devices. For example, users can easily access the interesting content portions without scrolling operations but a click on the index view.

2: Between-Document Cooperative Browsing

As shown in previous studies (Hua, Xie, Lu, & Ma, 2004, 2005; Nadamoto & Tanaka, 2003), users tend to view similar documents (e.g., image and Web page) concurrently for a comparative view of their contents. User's experience will be especially degraded in such scenarios due to two reasons. Firstly, it's difficult for users to seek out relevant information on a mobile device, and the task becomes more tedious with the increase of the number of devices for finding. Secondly, it's not feasible to present multiple documents simultaneously on a small display area, and it's also tedious for users to switch through documents for a comparative view.

In our system, we propose a between-document cooperative mechanism to address this problem. Our approach comprises of two steps: (1) relevant documents are automatically searched based on the information a user is currently focusing on the master device; and (2) the searched documents are presented on the slave devices. Therefore, this method can facilitate users' comparative view without manual efforts. Thus, users can easily achieve a comparative view with a simple glimpse through devices.

APPLICATION OF OUR COOPERATIVE FRAMEWORK

To assess the effectiveness of our framework, we apply it to several types of documents that are ubiquitous in mobile devices, including images, text documents (such as e-mail, PDF file, MS documents like DOC or PPT files) and Web pages. In the following sections, we illustrate each in detail.

Cooperative Browsing of Documents

1: Within-Document Cooperative Browsing

Document readability is greatly degraded due to the small display areas on current mobile devices; users have to continually scroll through the content to browse each portion in detail. In this case, we believe our between-document solution is capable of solving this problem: (1) partitioning a document into a series of content sections according to

paragraph or passage information; (2) extracting a summary description from each portion using a title or sub-title (summary instead if no titles); and (3) generating an index view for the document where each index points to the relevant detailed content portion in a large document.

Figure 1 shows an example of our solution, where an MSWord document is represented through its outline index, and a click leads to the display of detailed content in its slave devices. The design for slides is really useful in practice. For instance, for a speaker who often moves around to keep close contact with his/her audiences, it's necessary to develop a mechanism to facilitate the speaker's interaction with the slides when he or she moves around. We present an indexed view on small devices like SmartPhone, which can be taken by users, and the interaction with this phone generates the display updates on the screen.

2: Between-Document Cooperative Browsing

In our multi-device solution, we search relevant documents automatically and then deliver them to the slave devices to help browsing. We automatically identify the passages that are currently displayed on the center screen in the master device as the indicative text to find out relevant information. As has been pointed out by many existing studies, it is sometimes better to apply retrieval algorithms to portions of a document text than to all of the text (Stanfill & Waltz, 1992). The solution adopted by our system was to create new passages of every appropriate fixed length words (e.g., 200 words). The system searches a similar document from each device by using the passage-level feature vectors with keywords. The similarities of passages are computed using the Cosine distance between the keyword feature vectors (e.g., TFIDF vector model). In this way, our system searches for similar passages in the available document set, and the document with the greatest number of similar paragraphs becomes the similar page.

Figure 2 shows an example of this case, where (b) is the search results by our approach according to the content information that is extracted from (a). Furthermore, our system automatically scrolls the relevant document to the similar passages that hold the maximal similarity.

Figure 1. An example for within-document cooperative browsing

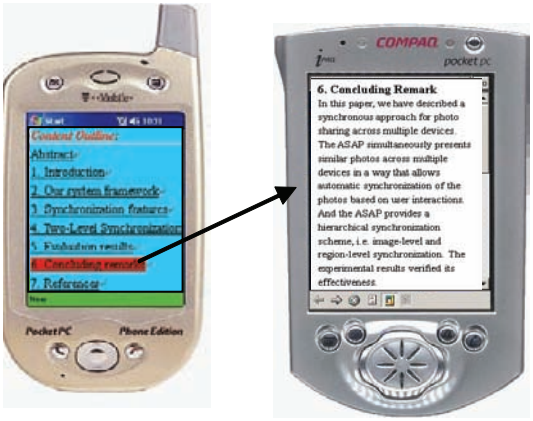
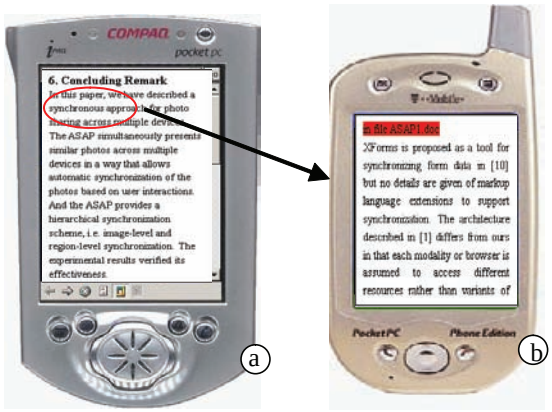


Figure 2. An example for between-document cooperative browsing



Cooperative Browsing of Web Pages

With the pervasive wireless connection in mobile devices, users can easily access the Web. However, Web pages are mostly designed for desktop computers, and the small display areas in mobile devices are consequently too small to display them. Here, we apply our framework to employ a cooperative way to generate a tailored view of large Web pages on mobile devices.

1: Within-Page Cooperative Browsing

Different from documents (e.g., MSWord), Web pages include more structured contents. There exist a lot of studies on page segmentation to partition Web pages into a set of tailored blocks. Here, we adopt the methods by Chen, Xie, Fan, Ma, Zhang, and Zhou (2003), in which each page is represented with an indexed thumbnail with multiple segments and each of them points to a detailed content unit. Figure 3 shows an example of this case. In our system, we deliver the detailed content blocks to various devices. Additionally, each detailed content block is displayed on a most suitable display area.

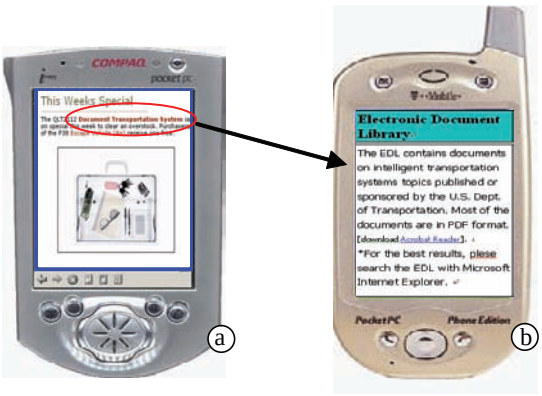
2: Between-Page Cooperative Browsing

Besides improving page readability in small display areas of mobile devices, users also need to browse relevant pages that contain similar information. In common scenarios, users need to manually search these relevant pages through a search engine or check-through related sites. Here, we develop an automatic approach to present relevant Web pages through a cross-device representation.

Our method to find out similar pages comprises three steps: (1) extracting all links from the page, which are assumed to be potential candidates that contain relevant information; (2) automatically downloading content information for each extracted links, and representing each document as a term-frequency feature vector; and (3) comparing the similarity of extracted pages and current page based on the Cosine distance through the feature vector. Thus, the page with the maximal similarity is selected as the relevant one, and its URL is sent to other devices for an automatic display update. An example is shown in Figure 4.

Figure 3. An example for within-page cooperative browsing

Figure 4. An example for between-page cooperative browsing



Cooperative Browsing of Images

Lpictures are not fitful for the display on mobile devices with small display areas. Here we apply our framework to facilitate users' image browsing through a cooperative cross-device representation.

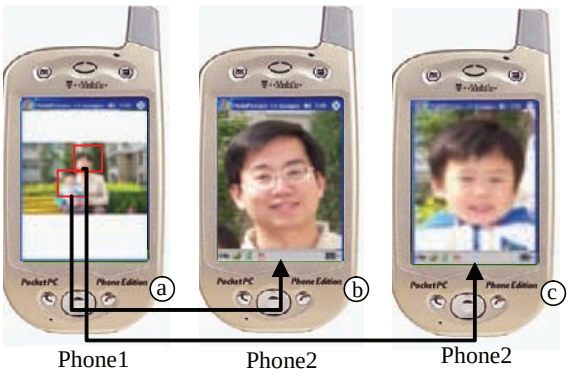
1: Within-Image Cooperative Browsing

In addition to all previous automatic image browsing approaches in a single small device (Chen, Ma, & Zhang, 2003; Liu, Xie, Ma, & Zhang, 2003), we provide in our approach a so-called smart navigation mode (Hua, Xie, Lu, & Ma, 2004). In our approach, an image is decomposed into a set of attention objects according to Ma and Zhang (2003) and Chen, Ma, and Zhang (2003), and each is assumed to contain attentive information in an image. Switching objects in a master device will result in a detailed rendering of the corresponding attention object on the slave device (e.g., Figure 5).

2: Between-Image Cooperative Browsing

In our previous work (Hua, Xie, Lu, & Ma, 2004), we proposed a synchronized approach called ASAP to facilitate photo viewing across multiple devices, which can simultaneously present similar photos on various devices. A user can interact with either of the available devices, and the user's interaction can automatically generate the synchronized updates in other devices. In Figure 6, there are two PDAs denoted PDA1 and PDA2, and each stores a number of pictures. When a user clicks a photo in PDA1 (Figure 6a), there are two steps to be done simultaneously: (1) PDA1 searches out similar images and displays them (b); (2) PDA1 sends image feature to PDA2, which then search out the similar photos (c).

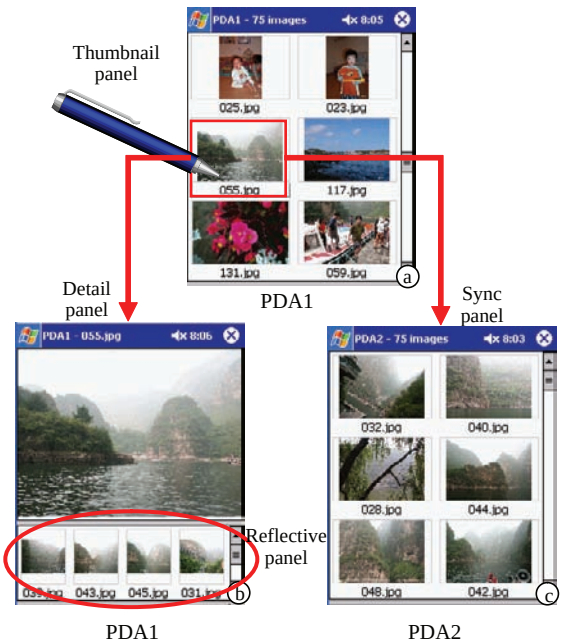
Figure 5. An example for within-image cooperative browsing



FUTURE TRENDS

Now, we are planning to improve our work in three aspects. First, we will develop more accurate algorithms to search out relevant information of various media types including image, text and Web page. Second, we plan to devise more advanced

Figure 6. The synchronization between PDA1 and PDA2



distributed interfaces to facilitate users' information browsing tasks across various devices. Third, we plan to apply our work to other applications such as more intricate formats of documents and execution application GUIs. We also plan to conduct a usability evaluation among a wide number of users to collect their feedbacks, which will help us find the points they appreciate and the points that need further improvements.

CONCLUSION

In this article, we developed a cooperative framework that utilizes multiple displays to facilitate information browsing in mobile environment. A two-level browsing scheme is employed in our approach, namely within- and between- document browsing. We apply our framework to a wide variety of applications including documents, Web pages and images.

REFERENCES

Brad, A. (2001). Using handhelds and PCs together. *Communications of the ACM, 44*(11), 34-41.

Chen, L. Q., Xie, X., Fan, X., Ma, W. Y., Zhang, H. J., & Zhou, H. Q. (2003). A visual attention model for adapting images on small displays. *ACM Multimedia Systems Journal, 9*(4), 353-364.

Chen, Y., Ma, W.Y., & Zhang, H. J. (2003, May). Detecting Web page structure for adaptive viewing on small form factor devices. In the *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary (pp. 225-233).

Coles, A., Deliot, E., & Melamed, T. (2003, May). A framework for coordinated multi-modal browsing with multiple clients. In the *Proceedings of the 13th International Conference on World Wide Web*, Budapest, Hungary (pp. 718-726).

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D. C. (2003, July). Stuff I've seen: A system for personal information retrieval and re-use. In the *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada (pp. 72-79).

Hua, Z., Xie, X., Lu, H. Q., & Ma, W. Y. (2004, November). Automatic synchronized browsing of images across multiple devices. In the *Proceedings of the 2004 Pacific-Rim Conference on Multimedia*, Tokyo, Japan. Lecture Notes in Computer Science (pp. 704-711). Springer.

Hua, Z., Xie, X., Lu, H. Q., & Ma, W. Y. (2005, January). ASAP: A synchronous approach for photo sharing across devices. In the *11th International Multi-Media Modeling Conference*, Melbourne, Australia.

Liu, H., Xie, X., Ma, W. Y., & Zhang, H. J. (2003, November). Automatic browsing of large pictures on mobile devices. In the *Proceedings of ACM Multimedia 2003 Conference*, Berkeley, California (pp. 148-155).

Ma, W. Y., Bedner, I., Chang G., Kuchinsky, A., & Zhang H. J. (2000, January). A framework for adaptive content delivery in heterogeneous network environments. *Multimedia Computing and Networking 2000*, San Jose.

Ma, Y. F., & Zhang, H. J. (2003, November). Contrast-based image attention analysis by using fuzzy growing. In the *Proceedings of ACM Multimedia 2003 Conference*, Berkeley, California (pp. 374-381).

Nadamoto, A., & Tanaka, K. (2003, May). A comparative Web browser (CWB) for browsing and comparing Web pages. In the *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary (pp. 727-735).

Stanfill, C., & Waltz D. L. (1992). Statistical methods, artificial intelligence, and information

retrieval. In P. S. Jacobs (Eds.), *Text-based intelligent systems* (pp. 215-225). Lawrence Erlbaum.

Su, N. M., Sakane, Y., Tsukamoto, M., & Nishio, S. (2002, September). Rajicon: Remote PC GUI operations via constricted mobile interfaces. *The 8th Annual International Conference on Mobile Computing and Networking,* Atlanta (pp. 251-262).

KEY TERMS

Ambient Intelligence: Represents a vision of the future where people will be surrounded by electronic environments that are sensitive and responsive to people.

ASAP System: The abbreviation of a synchronous approach for photo sharing across devices to facilitate photo viewing across multiple devices, which can simultaneously present similar photos across multiple devices at the same time for comparative viewing or searching.

Attention Object: An information carrier that delivers the author's intention and catches part of the user's attention as a whole. An attention object often represents a semantic object, such as a human face, a flower, a mobile car, a text sentence, and so forth.

Desktop Search: The functionality to index and retrieve personal information that is stored in desktop computers, including files, e-mails, Web pages and so on.

Small Form Factors: Mobile devices are designed for portability and mobility, so the physical size is limited actually. This phase is called small form factors.

TFIDF Vector Model: TF is the raw frequency of a given term inside a document, which provides one measure of how well that term describes the document contents. DF is the number of documents in which a term appears. The motivation for using an inverse document frequency is that terms that appear in many documents are not very useful for distinguishing a relevant document from a non-relevant one.

Visual Attention: Attention is a neurobiological conception. It implies the concentration of mental powers upon an object by close or careful observing or listening, which is the ability or power to concentrate mentally.

This work was previously published in the Encyclopedia of Human Computer Interaction, edited by C. Ghaoui, pp. 120-127, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.11
Describing the Critical Factors for Creating Successful Mobile Data Services

Anne Tseng
Helsinki School of Economics, Finland

Jukka Kallio
Helsinki School of Economics, Finland

Markku Tinnilä
Helsinki School of Economics, Finland

ABSTRACT

Mobile operators play a central role in the development of the mobile data services market. They have primary access to the customer relationship, a key source of revenue, and are responsible for how revenue is distributed to other participants in the value chain. As a result, a successful operator-driven business model is essential to the survival of the mobile data industry. The purpose of this chapter is to describe the critical factors that have influenced the results of operators based on countries that have been at the forefront of mobile data services innovation. Then, by comparing the key characteristics of operator-driven business models in these four cases around the world, we will describe the critical factors used in designing successful mobile data services.

INTRODUCTION

Given the high penetration of mobile phones and the PC Internet, it has long been predicted that mobile data usage would increase substantially as a result of the intersection of the two channels. Unfortunately, however, mobile data usage has been slow to materialize. Despite the business potential, entrants and incumbents alike have been confounded by a host of unexpected challenges such as insufficient demand, competition from substitutes, and, most important, lack of profitable business models. In this chapter, we will attempt to address the last issue by conducting an exploratory study of the development of business models in Japan, Western Europe, South Korea, and China in order to describe the critical factors

used in successful operator business models for mobile data services.

For the purposes of this analysis, we define *mobile data services* as any mobile non-voice service. This includes wireless data transfer technologies such as instant messaging and SMS as well as e-mail and the mobile Internet. We define *mobile Internet* as Internet access using mobile devices including but not limited to cell phones, personal digital assistants, and so forth. Moreover, since creating and capturing value is one of the most challenging issues in the mobile data services business, we shall use increased Average Revenue Per User (ARPU) as a key objective for mobile operators.[1]

ANALYSIS BY COUNTRY

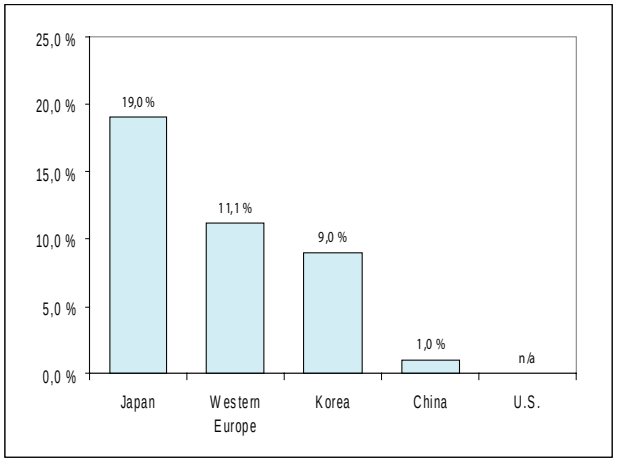
Mobile operators play a central role in the development of the mobile data services market. They have primary access to the customer relationship, a key source of revenue, and are responsible for how revenue is distributed to other participants in the value chain. As a result, a successful operator-driven business model is essential to the survival of the mobile data industry. In the following four empirical cases, we will study the experiences of four operator groups that constitute some of the largest and most active mobile data service markets in the world; namely, the countries/regions of Japan, South Korea, China, and Europe.

We begin the discussion by comparing the success stories of selected operators in countries that have the highest mobile data services revenues. Mobile data revenues are the highest in Japan at 19% of total operator revenues, followed by Western Europe, where data revenues represent 11% of total revenues, most of which is SMS.[2] Korea comes in third with mobile data representing 9% of South Korean operators' revenues. China comes in fourth with 1% (see Figure 1).

Japan

With NTT DoCoMo's launch of i-mode in February 1999, Japan became the first country to successfully introduce a mobile data service and, with the exception of countries like South Korea, has been one of the few countries able to grow

Figure 1. Data as percentage of operator revenues



Source: Merrill Lynch Telecom Services Research, Wireless Matrix 3Q02, 10 Dec 2002 issue

and sustain a significant subscriber base in mobile data services. As of March 2003, there were 61.8 million subscribers of mobile data services in Japan (Henten et al., 2003).

The operator made several departures from industry-accepted norms in its introduction of i-mode, which ultimately proved instrumental to its success. First, in contrast to the European operators who targeted mobile Internet products and service offerings toward the high-end business user, NTT DoCoMo wanted to create a product that would attract the average consumer. To do so, the design team insisted that the product be easy to use and affordable (Matsunaga, 2000). This affected the carrier's marketing approach, which did not focus on the technology, as was done with WAP in Europe, but rather on the user benefits of the i-mode service (Natsuno, 2000). Second, as early adopters, the operator targeted younger segments of the market to help initiate the positive network cycle necessary to spur mass consumer demand (Ratliff, 2002). In keeping with this segment, subscription prices were kept at impulse-buy levels in order to compete with magazines (Ratliff, 2002).

Most importantly, NTT DoCoMo viewed i-mode as a complete product. In order to implement its vision, DoCoMo played a key role as a coordinator of the value chain. By deciding every detail from the design of the handsets to what constituted official content on its mobile portal, the company ensured that its phones, applications, and Web sites were fully functional at launch (Matsunaga, 2000). Japanese manufacturers tailor-made i-mode cell phone models to DoCoMo's exact specifications, including screen size, weight, battery life, and functionality (Matsunaga, 2000). DoCoMo also chose to partner with content providers instead of purchasing or creating the information, deciding that rather than provide content, it would serve as a gateway for quality content (Natsuno, 2000). This emphasis would later contribute to its successful branding as a quality service offering.

The decision to focus on quality content also dictated the company's choice of technologies. Therefore, instead of using WAP, which was at the time being touted by the world's largest handset manufacturers (Nokia, Ericsson, and Motorola) as the global standard for mobile data services, DoCoMo chose cHTML, a subset of the HTML language used by the fixed Internet (Matsunaga, 2000). This made the transition from the fixed Internet to mobile data services easier for content providers (Natsuno, 2000). The company also decided to use packet-switched vs. circuit-switched technology, a move that also allowed users to access the Internet without charging them for the time spent online (Henten et al., 2003). Furthermore, DoCoMo did not charge content providers to participate and offered instead a revenue-sharing agreement in which it kept 9% of the content fees generated for its billing service, and content providers received 91% of the content fees (Ratliff, 2002). The end result was a positive network effect that led to a substantial increase in subscribers and wireless content. Interestingly, NTT DoCoMo's competitors, KDDI and J-Phone, have been able to successfully copy the i-mode business model, despite the use of different technological platforms and target segments (Bohlin et al., 2003).

South Korea

South Korean mobile operators have followed business models that are similar to Japan's i-mode model (i.e., coordinate the value chain through strong relationships with handset manufacturers and content providers with a generous revenue sharing agreement between operators and service providers) (Kelly et al., 2003). What makes their success more striking is that they have succeeded, in spite of using different technologies and platforms (i.e., CDMA and WAP). Moreover, operators have proven to be innovative in dealing with government regulations. In 2000, when the Korean government banned handset subsidies, Korean mobile operators responded by segmenting their

voice offerings into multiple branded products (Strand, 2002). This meant that children, teenagers, single women, business men, and the elderly all had their own branded products and distinct portals as well as separate promotions and pricing plans. The novel approach allowed operators to successfully differentiate their services on brand and price, and increased demand for mobile voice and data services.

Europe

Despite China's market size potential, as a region, Western Europe still represents the largest market with 300 million cellular subscribers (Merrill Lynch, 3Q, 2002).[3] However, aside from SMS, Europe has been unsuccessful in introducing the mobile Internet to the mass consumer. Few users adopted WAP, and GPRS adoption rates have thus far been relatively low.[4]

Mobile operators made several critical mistakes in the introduction of WAP services. The service was mismarketed as fixed Internet on the mobile phone, suffered from poor service and inadequate technical support, lacked content, and was expensive (Baldi et al., 2002). Moreover, the service was modular, which meant that users had to custom download applications as they needed them, making it frustratingly difficult and time-consuming for customers to use. While WAP has improved significantly since its initial launch three years ago, it continues to suffer from the stigma of being over hyped. The launch of GPRS networks should improve users' reactions to the new service as the network is packet-switched, which is more conducive to transmitting data services. However, the success of GPRS is far from assured, and Europe's choice of 3G, W-CDMA, remains unproven.

The relative success of the fixed Internet in Europe also has affected operators' approach to the mobile Internet. Fear of becoming a bit-pipe plus the pressure to recoup the high prices paid for 3G spectrum have resulted in higher consumer prices and revenue-sharing models that have been less generous for content providers in Europe than in Asian countries (Quigley, 2001). Another consequence has been the decision by some operators to develop content in-house, a practice that has proven costly and time-consuming (Maitland et al., 2003). Gradually, however, operators are learning to partner, which means allowing content and service providers to co-exist with them in the mobile data services value chain, offering more generous revenue sharing agreements and focusing more on creating a positive user experience than on selling faster and more advanced technologies. Operators like Vodafone have begun to successfully copy elements of the i-mode business model, such as partnering with content providers, sharing revenues, and working more closely with handset manufacturers (Salz, 2003). After being shut out of 2G, Japanese handset manufacturers are seeing a resurgence of business from European operators, and even Nokia, the largest handset manufacturer, has started to conform its handsets to operators' demands (Longino, 2003). Whether or not Europe will be able to grow its position in the mobile data services market will depend on its ability to develop sustainable, profitable business models. Vodafone has been among the leaders, releasing its own version of I-mode—Vodafone Live!—and has successfully attracted a million subscribers after only five months in service. Meanwhile, Virgin Mobile's continued expansion into the U.S. and other countries provides a useful lesson in branding and targeting.

China

As a latecomer into the mobile data services market, China was able to benefit from the experiences of both Europe and Japan. Like Europe, SMS is extremely popular in China (Albright, 2001; China Unicom, 2003, Merrill Lynch, 2002). In the second half of 2002, some 23.7 billion messages were transmitted over the country's mobile networks (China Unicom, 2003; Merrill Lynch

Research, 2002). Like Europe, China Mobile, the country's largest operator, was unable to attract subscribers to its WAP services; only 8,000 users had signed up within five months of launch (Yan, 2001). Then, in November 2000, the operator introduced Monternet, a mobile portal initiative, patterned after NTT DoCoMo's i-mode business model. Moreover, China Mobile made substantial investments to facilitate its use, developing a common platform to facilitate mobile data roaming capabilities and a standardized billing system, requiring an upgrade of mobile operations in every province and metropolitan area to support the new billing requirements (Sigurdson, 2002). Mobile service providers collect the full retail price for their services and then compensate the operator for using their network and billing services (Zhang, 2001). Following the Japanese and Korean model, the operator also introduced a generous revenue share agreement of 91% to the content provider and 9% for the operator, with an additional 6% collected by the operator in the case of credit issues (Yan, 2001). The subsequent success of the mobile portal has since helped the three largest independent portals to become profitable for the first time (Kurtenbach, 2003).

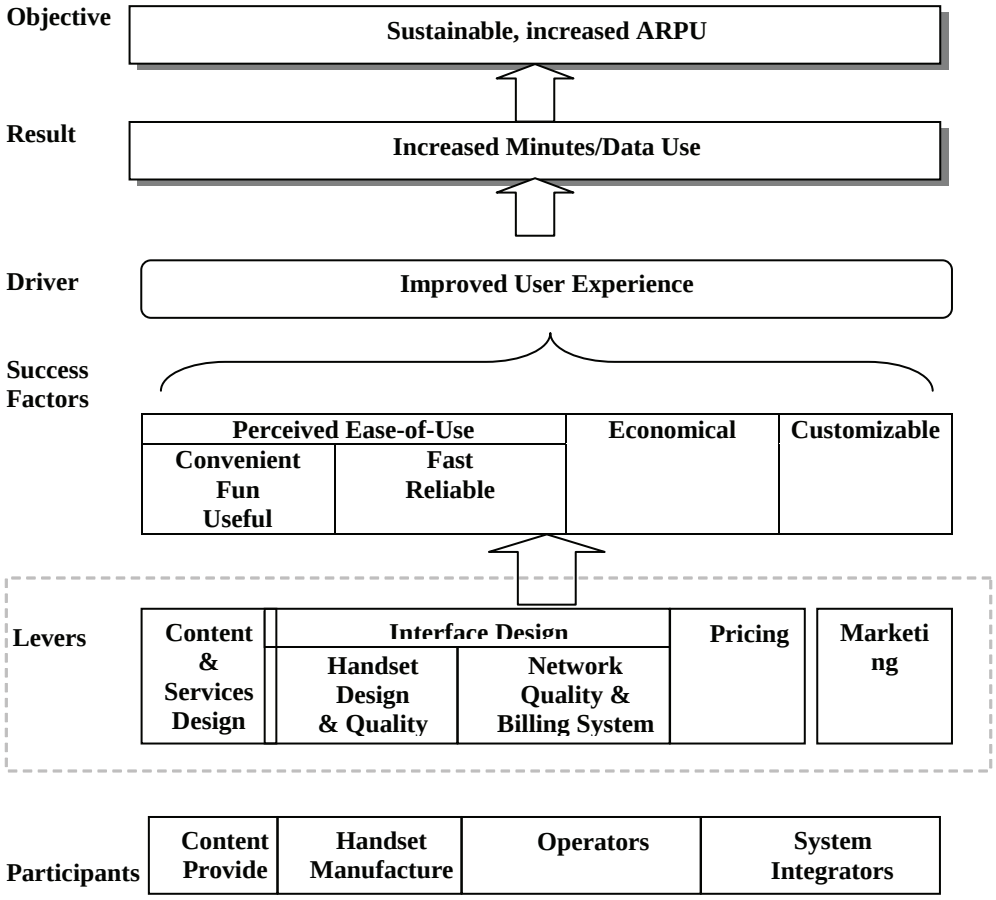
DESCRIBING CRITICAL FACTORS FOR INTRODUCTION OF SUCCESSFUL MOBILE DATA SERVICES

As the experiences of selected mobile operators show, successful business models are essential to the development of the mobile data services industry. Services need to be developed closely with technologies in order to create a positive user experience that will attract and retain customers. In order to create a positive user experience, time, effort, and cost must be invested to shepherd products from development to implementation. This requires the joint efforts of manufacturers, network operators, content and services providers, and systems integrators. Based on our comparisons of successful operator business models, we will describe these critical success factors for mobile operators.

To increase ARPU, we begin by focusing on the universal levers that an operator can influence and use to its advantage. As the experience of NTT DoCoMo demonstrates, these levers include content/handset/interface design, network quality, pricing, billing, marketing, and customer support (see Figure 2). By changing these levers, the operator can affect the user experience so that he or she will increase the amount of minutes used and data downloaded, which, in turn, will lead to increased ARPU.

To encourage increased usage, users must have a positive experience. A Nielsen Norman report on WAP usability, published in December 2000, concluded that usability drives mobile data service adoption. The average consumer may be unforgiving, if services are found difficult to use. Operators have to focus on creating a positive user experience through continual innovation and compelling content and services. In particular, subscribers have been attracted to services that are easy to use, reliable, fast, entertaining/informative, and reasonably priced. These elements can be achieved best through the cooperation of the handset manufacturer, operator, third party systems integrators, and content and application providers. This implies that if there is any discord between the three parties, then demand and the market for new content, and, hence, mobile data services will not develop. Thus, as NTT DoCoMo's experience suggests, the operator should focus on coordinating and controlling the quality of the final product produced as a result of the interaction between the different participants in the value chain. By doing so, the other components, such as technology, pricing, billing, and marketing, will fall into place (see Figure 2). Next, we describe several factors that may influence the successful deployment of mobile data services.

Figure 2. Achieving operator's objective



Content and Services Design

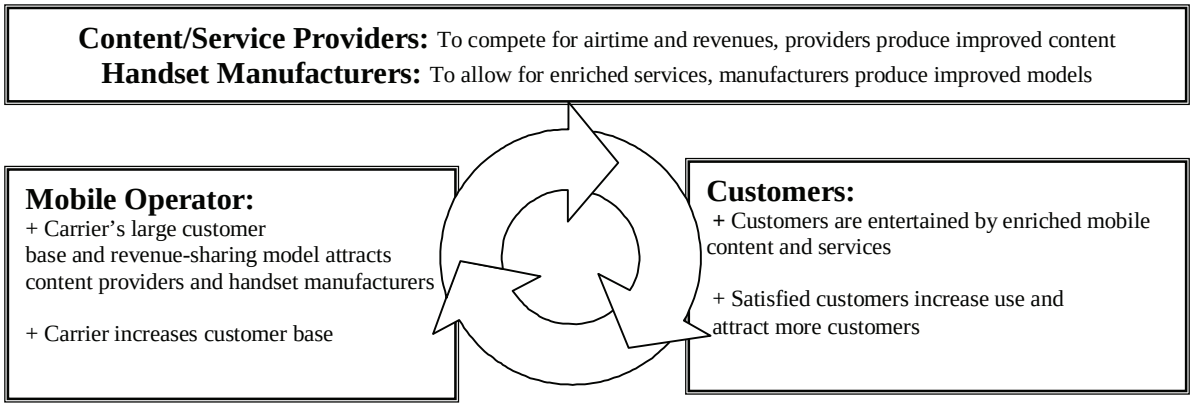
Operators who have been able to launch successful mobile data service offerings, like Vodafone Live! and i-mode, have learned that attracting quality content is imperative to attracting customers and increasing usage. NTT DoCoMo held its content providers to strict quality standards and acted as a certifier for its official sites. However, DoCoMo also tried to make the experience easier for the content provider. Its choice of cHTML as a platform instead of WAP had more to do with making things easier on the content provider than on how advanced the technology was.

The resulting content selection and quality has been astounding. There are 1,800 official sites accessible through i-mode, and services include photo exchanges, GPS location services, video-conferencing, and music/video/application downloads. The situation is similar in Korea, where even Korea's smallest carrier has some 300 providers and 5,000 services. Services include photo exchanges, GPS location services, video-conferencing, and music/video/application downloads.

Revenue Sharing

To attract a sufficient number of customers to their mobile data services business, successful mobile operators also need to work with a critical mass of content/service providers who can design

Figure 3. Partners in the virtuous cycle



innovative services, which, in turn, will attract customers. A revenue-sharing model that allows content providers to thrive and promotes healthy competition so that only the best services and content are delivered to the customer is crucial.

European mobile operators have been lambasted by content providers for exerting their market power to demand excess revenues from content providers in exchange for a good location on their mobile portals. Operators have been reluctant to share much of their own revenues because of their desire to improve ARPU. Worried they would become little more than a bit-pipe through which providers send their data, mobile operators attempted to control wireless Internet access by partnering with or developing their own services. This may be short-sighted, as the real revenue driver is in communication services (i.e., voice, e-mail, and data traffic) as opposed to content, so the operators should be creating an unattractive market in the long term for the sake of a relatively small short-term return. At the same time, however, operators may be more willing to share revenues, once wireless content services start to take off and content creators are perceived as value-adding.

Unlike their Western counterparts, Japanese operators do not want to be content providers and have not used their positions to demand an exess share of the revenue from data services. This has allowed all the parties involved to benefit from the revenue generated and aligns their strategies to a common goal of revenue expansion. As a result, content providers have been encouraged to invest in value-added services. DoCoMo calls this three-way relationship the virtuous cycle, with all four parties—subscriber, content provider, handset manufacturer, and operator driving the market (see Figure 3).

Marketing

With voice, operators have typically relied on two methods to segment the market: pricing and technology. However, with wireless data services, the rules appear to have changed. The operator needs to educate the consumer about the user benefits of new technologies (e.g., GPRS). The more successful operators have paid significant attention to marketing in order to educate the public about new services and to reach out to specific segments of the population who would be especially attracted to the service through targeted marketing approaches.

Teens are commonly and effectively targeted as a segment of the population most likely to be interested in mobile data services. This group includes teens that are supported by their parents

as well as young adults in their 20s and 30s with high disposable incomes. They played a significant role in the eventual mass adoption of mobile data services in Japan and were credited with discovering and initiating the widespread adoption of SMS in Europe. Not surprisingly, Vodafone has chosen to target this segment with its new Vodafone Live service offering. In mobile voice services, Korean and Japanese operators also have used multiple brands to personalize users' experiences, thereby reducing the likelihood of churn and increasing revenue. Virgin Mobile has built its MVNO business on this segment, as well.

In South Korea, operators went even further, splitting their standard voice services into sub-brands, specifically targeted at different age and gender segments of the mobile users. These new sub-brands have their own portals, marketing, content and services offerings, and pricing, and have been instrumental in getting the mobile users to discard their 2G mobile phones and buy the new non-subsidised 2.5G mobile phones. The sub-brands have proven to be such a hit that only 24% of Korea's largest mobile operator subscribers remain on a standard price plan.

The way mobile data services are marketed also can impact perception and, therefore, its uptake. Operators have differed in the marketing messages they have used to reach out to potential users. Japan, Korea, and China have focused primarily on the entertainment aspect of mobile data services, stressing user benefits, whereas European operators have tended to focus on the technology, lower prices, prepaid plans, and switching operators in order to save money. U.S. operators have followed a similar path, highlighting the technical aspects of high-speed and always-on connectivity in addition to picture e-mail.

For example, Vodafone Live, in a departure from the customary marketing habits of European operators, is using trendy advertisements to sell its new service as fun and entertaining in attempt to make users aware of the fact that there is more one can do with his or her cell phone than text messaging or calling. This approach may have contributed to its exceptional customer uptake within the first five months of launch.

Equally important is making sure products are not hyped and that they perform as they are billed. For example, WAP service was billed as the fixed Internet over the phone, and users were led to believe that the mobile Internet would be as exciting as the fixed-line Internet and were disappointed when reality did not match expectations.

Pricing

While operators in the U.S. and Europe struggle to find a pricing structure that works for mobile data, Japanese and Korean operators provide some possible solutions. Japanese operators use a hybrid of subscription and per-data usage fees. I-mode and EZWeb uses low-cost monthly subscription fees in addition to a set data traffic fee per packet of data sent/received. J-Phone has no subscription but charges per data sent/received, a practice that Vodafone Live now is trying to copy.

Like J-Phone, Korean operators do not charge subscription prices; however, they go one step further—they distinguish between different types of data traffic. Users pay a lower rate for multimedia downloads, which tend to be data intensive but relatively low in value, than for text-based downloads, which are less data intensive, but have relatively high user value.

The advantages of these pricing plans are that they are relatively simple for users to understand and allow the carrier to further segment the market by price. However, they may be difficult to implement and may require expensive conversions from existing billing systems that have been based on billing for voice minutes as opposed to data transactions.

Billing Systems

Operators are finding that flexible billing systems are essential for the development of sophisticated

services and business models and for maintaining customers. Once users are willing to pay for content, the method of payment becomes critical in ensuring the customer goes through with the sale. At this point, the success of content or service depends on the billing system. While European operators have found some success with reverse billing, many still have a long way to go before GPRS billing systems accurately capture the needs of all its customers (e.g., prepaid customers' real-time needs). U.S. operators are facing similar difficulties with legacy billing platforms that were built to track voice usage and can't easily or accurately account for the number of data transactions. Again, Vodafone has made advances with its M-Pay Service, which allows users to pay for low-cost digital and hard items valued at less than £5.00 and a pay-as-you-go system that allows it to offer mobile data services to its prepaid customers. In the meantime, Asian operators continue to make strides in the billing front. In 2002, Korean SK Telecom agreed to team up with LG Telecom to cooperate on offering mobile settlement and payment services in a preliminary step to open m-commerce opportunities. NTT DoCoMo upgraded its billing system so costs for premium services could be added to a user's phone bill, as did China Mobile as part of the launch of its mobile Internet portal.

Handset Design

The importance of handset design in enhancing the user experience should not be underestimated. Both Japanese and Korean operators have worked closely with handset manufacturers to create handsets that optimize user experience. As a result, Japanese phones are laden with features, function according to operator and manufacturer's specifications, and fit with the usage patterns of their target customer—the Japanese consumer. The latest handset models in Korea are among the most advanced in the world and can access the Internet at a rate up to 2.4 megabits per second, four times as fast as GPRS phones. This creates a virtuous cycle in itself as handset manufacturers continually improve handset technologies so that more technically advanced services can be offered. This, in turn, attracts customers to buy new handsets, replacing their older models and increasing traffic, thereby creating a win-win situation for both the operator and the handset manufacturer. In contrast, Western operators have a more contentious relationship with their dominant manufacturers, which include Nokia, Motorola, and Siemens, and they have had less influence over handset design than their Asian counterparts, which may be influencing the less-than-seamless quality of services that have been provided in Europe. This however, may soon be changing with Vodafone Live's precedent-setting move.

Network Quality

While increased bandwidth alone will not necessarily increase adoption, poor network quality can have a negative effect on user experience. Initial users of WAP experienced dropped data sessions, poor data throughput, and session instability as a result of poor network quality. In fact, U.S. networks suffer from poor voice quality due to the practice of compressing signals as a result of limited spectrum. Although GSM networks still dominate, operators have made inroads using CDMA 1x technology as an emerging 2.5G network. With the advantages of a lower cost to upgrade networks, better capacity and speed, and affordable and attractive handsets, CDMA networks offer distinct advantages over other types of networks at this time. W-CDMA deployment unfortunately is still very limited, handsets are rare, and technology is expensive. (GPRS has scale, but suffers from a limited number of color handsets, slower speeds, and the risk of cannibalizing voice capacity.) Table 1 compares the major levers/drivers that determine the different business models of operators around the world.

Table 1. Comparison of business models

Lever	Japan	South Korea	Europe	China
Content/ Service	1,800 sites from I-mode (800-900 each from J-Phone and KDDI). Services include photo exchanges, GPS location services, video-conferencing, and music/video/ application downloads	Korea's smallest carrier has 300 providers and 5,000 services. Services include photo exchanges, GPS location services, video-conferencing, and music/video/ application downloads	Limited - Vodafone Live! claims it has been able to attract 250+providers. Services include photo exchange, location-based services, application and ringtone downloads, SMS	Monternet reportedly has 130+ content providers
Revenue Sharing	Yes. Japanese operators pass 88-91% of subscription fees to content providers	Yes.	For Premium SMS mostly with European operators getting a larger cut than in Japan	Yes. Chinese operators relinquish 85-91% of content fee to providers
Marketing	Subbranding by age, gender, and corporate type	Subbranding by age, gender, and corporate type	Vodafone Live! and Virgin Mobile targeting youths or young adults	N/a
Marketing Message	Fun and entertaining services tailored to customer segments	Fun and entertaining services tailored to customer segments	Technology, low price, prepaid plans, and switching operators to save $$$	Free Inter-connections, Limitless Possibilities tailored services to customer segments
Pricing	Subscription fees and per-packet data traffic fees	No subscription; only per-packet data traffic fees	Per kilobyte under GPRS and per message for SMS	Subscriptions and data traffic fees
Billing Systems	Billed directly to mobile phone bill	Billed directly to mobile phone bill	Premium SMS bills directly to phone bill. Different macro- and micro-payment systems co-exist	Charged directly to mobile phone bill (China Mobile)
Handset Design	Domestic suppliers offer 65,000 color screen, 12-line displays, polyphonic speakers, digital cameras, and Internet browsers	Domestic suppliers offer 65,000 color screen, 12-line displays, polyphonic speakers, digital cameras, and Internet browsers	Nokia, Motorola, and entries from Panasonic and Sharp increase the number of camera phones	Nokia, Motorola, Samsung, Domestics gaining ground
Network Quality	Good – GPRS & PDC in Japan	Excellent – CDMA & WAP	Excellent-GSM/GPRS/WAP	Good-GSM/ GPRS/CDMA

CONCLUSION

Our findings suggest that a successful transition from mobile voice services to mobile data services represents a significant shift in focus and approach for all industry participants and, in particular, for operators. Instead of focusing purely on increasing bandwidth and complex technologies, successful operators are increasingly focusing on improving the user experience through coordinated handset and service design. They also are focusing on creating effective billing systems, offering services at reasonable prices, and targeting marketing strategies. We have described several factors that can influence the strategies of mobile operators that wish to implement successful mobile data services.

The achievement of successful operators, especially in Japan and Korea, can be distilled to this key phrase: Applications drive traffic and traffic drives revenue. In order to manage the process, successful mobile operators coordinate

the relationship between different members of the mobile services value chain. By coordinating, the successful operator ensures that every participant has a clearly defined role in the value chain and is compensated for it. This includes fostering an open platform in which research and development can be shared, ensuring that participants' interests are aligned with the customer's (in a model that is essentially customer-driven), thus allowing all participants to maximize their value in the value chain.

These findings represent opportunities for further research in other issues surrounding mobile operator-driven business models. One possibility is to develop a new theory into which factors are more difficult to transfer and why. Continued research in this area could address the difficulties of copying successful operator-driven business models.

REFERENCES

Albright, P. (2001). China may join SMS craze [electronic version]. *Wireless Week*. Retrieved May 20, 2003, from *http://www.wirelessweek.com/index.asp?layout=article&articleid=CA82361*

Baldi, S., & Thaung, H.P. (2002). The entertaining way to m-commerce: Japan's approach to the mobile Internet—A model for Europe? *Electronic Markets, 12*(1), 6-13.

Bohlin, E., Bjorkdahl, J., Lindmark, S., & Burgelman, J. (2003). Strategies for making mobile communications work for Europe: Implications from a comparative study. *Proceedings of the European Policy Research Conference (EuroCPR)*, Stockholm, Sweden.

China Unicom. (2003). Announcement in respect of statistics data for the month of April 2003. Retrieved May 19, 2003, from *http://www.chinaunicom.com.hk/main/eng*

Henten, A., Olesen, H., Saugstrup, D., & Tan, S. (2003). New mobile systems and services in Europe, Japan and South Korea. *Proceedings of the Stockholm Mobility Roundtable 2003*, Stockholm, Sweden.

Kelly, T., Gray V., & Minges, M. (2003). Broadband Korea: Internet case study. *ITU*. Retrieved May 20, 2003, from *http://www.itu.int/ITU-D/ict/cs/korea/material/CS_KOR.pdf*

Kurtenbach, E. (2003). China's Internet companies roar back to life—Thanks to mobile phones. *Wireless Business & Technology Magazine, 3*(2).

Longino, C. (2003). Live!—An agent of change: Will Vodafone Live! shake up the mobile world beyond carrier portals? *The Feature: It's All About the Internet*. Retrieved May 20, 2003, from *http://www.thefeature.com/*

Maitland, C.F., van de Kar, E.A.M., & Wehn de Montalvo, U. (2003). Network formation for provision of mobile information and entertainment services. *Proceedings of the 16th Bled Electronic Commerce Conference e-Transformation*, Bled, Slovenia.

Matsunaga, M. (2000). *The birth of i-mode: An analogue account of the mobile Internet*. Singapore: Chuang-Yi Publishing Pte Ltd.

Merrill Lynch Telecom Services Research. (2002). *Quarterly update on global wireless industry metrics*. Merrill Lynch Global Securities Research & Economics Group. Global Fundamental Equity Research Department.

Natsuno, T. (2000). *I-mode strategy*. London: John Wiley & Sons, Inc.

Poulbere, V. (2003). Mobile data services: A market of 15.4 billion EUR in 2002 in Western Europe thanks to the rapid growth of SMS services (2/3). *IDATE Newsletter No. 223*. Retrieved May 20, 2003 from *http://www.idate.fr*

Quigley, P. (2001). Warning to operators: Overhaul or say goodbye. *Wireless Week, 7*(38), 15. Retrieved September 17, 2001, from *http://www.wire lessweek.com/index.asp?layout=article&articleid= CA159353*

Ratliff, J. (2002). NTT DoCoMo and its i-mode success: Origins and implications. *California Management Review, 44*(3), 55.

Salz, P. (2003). The right stuff the feature: It's all about the Internet. Retrieved May 20, 2003, from *http://www.thefeature.com*

Sigurdson, J. (2002). China as number one — in mobiles. *International Institute for Asian Studies Newsletter, 29,* 1.

Strand, J. (2002). Sub-brands can help European mobile operators become profitable. *Wireless Watch Japan*. Retrieved May 20, 2003, from *http://technologyreports.net/wirelessreport/?articleID=817*

Yan, X. (2001). Mobile data communications in China. *Proceedings of the Sixth Asia-Pacific Regional Conference of the International Telecommunications Society,* Hong Kong.

Zhang, N. (2001). The wireless virtuous circle emerges Linktone. *Proceedings of the I and I Asia,* Hong Kong, China.

ENDNOTES

1. Although we will not focus on cost, it should be taken into consideration along with the business models, as many of these aspects will require significant investments in not only time and effort but also money.
2. Western Europe in this context includes the countries of Austria, Belgium, Denmark, Finland, France, Germany, Greece, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the UK.
3. This is evidenced by the three largest mobile operators, which are China Mobile followed by Vodafone (UK-based) and Deutsche Telekom (Germany-based).
4. Even SMS, which generated an estimated 11.7 billion of European operators' data revenues in 2002, was due more to chance discovery than through skilled marketing on the part of the operators (Poulbere, 2003).

This work was previously published in Managing Business in a Multi-Channel World: Success Factors for E-Business, edited by T. Saarinen, M. Tinnila and A. Tseng, pp. 204-219, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.12
A Design Framework for Mobile Collaboration

Pedro Antunes
University of Lisboa, Portugal

INTRODUCTION

Mobile collaboration involves people working together and moving in space. Research in mobile collaboration has primarily focused on technical issues like connectivity support or remote information access. We argue there is a lack of research on many nontechnical issues vital to design mobile collaboration systems, disentangling the relationships between collaboration, work context, and mobility.

Our fundamental concern is to go beyond the technical issues towards the assimilation of the mobility dimension in all processes shaping collaborative work, including information sharing, context awareness, decision making, conflict management, learning, etc. This article aims to codify into a design framework:

- Some fundamental human factors involved in mobile collaboration.
- Several guidelines for developing mobile collaboration systems.

The design framework provides general constructs identifying phenomena of interest necessary to inquire about the work context, human activities, and system functionality. The framework identifies *what* information may interest designers, bounding their relationships with the other stakeholders. The framework also guides the design process, identifying *how* user requirements may be applied during the implementation phase.

The framework has been validated in several real-world design cases. Two cases will be briefly described. This research contributes to the design of mobile collaborative systems. The most significant contributions are related to artifacts and emphasize that designers shall explore the potential of artifacts to support concerted work and sensemaking activities.

BACKGROUND

Several conceptual frameworks have been proposed in the group support systems (GSS) field (DeSanctis & Gallupe, 1987; Nunamaker, Dennis,

Valacich, Vogel, & George, 1991; Pinsonneault & Caya, 2005). However, these frameworks capture the notion of place in a very restrictive way, more tied to group proximity than mobility, where geographical references play a central role in tying information together (Mackay, 1999).

The above limitation is being tackled in two closely related research areas: collaborative spatial decision-making (CSDM) and spatial decision support systems (SDSS) (Nyerges, Montejano, Oshiro, & Dadswell, 1997). SDSS address the combination of DSS with geographical information systems (GIS), while CSDM studies the integrated support to collaboration, decision, mobility, and geographical information.

We find several studies on the infrastructural basis of SDSS. Zhao, Nusser, and Miller (2002) identify the infrastructural requirements for SDSS. Gardels (1997) and Touriño et al. (2001) contribute with the integration of multimedia with geo-referenced data. Hope, Chrisp, and Linge (2000) tackle the access to remote databases by fieldworkers, while Pundt (2002) addresses data visualization in the same context. All of these research projects do not directly address mobile collaboration but explore basic features necessary to support this functionality.

Regarding the human factors of SDSS, we account for studies of user interaction with multimodal and tangible GIS interfaces (Coors, Jung, & Jasnoch, 1999; Rauschert, Agrawal, Sharma, Fuhrmann, Brewer, & MacEachren, 2002). In the same line, we also cite developments in synthetic collaborative environments for geo-visualization (Grønbæk, Vestergaard, & Ørbæk, 2002; Manoharan, Taylor, & Gardiner, 2002). However, these research studies address fixed work settings.

More in line with collaboration studies, we find several research emphasizing the need to support group modeling in CSDM (Armstrong, 1994, 1997). Some propose very specific solutions, such as the integration of workflow management with SDSS (Coleman & Li, 1999).

Finally, addressing the broad-spectrum CSDM design, we find the work from Tamminen, Oulasvirta, Toiskallio, and Kankainen (2004), who propose an integrated framework with guidelines for eliciting innovative ideas for mobile technology based on context-awareness (although not collaboration). Nyerges et al. (1997) also propose an integrated framework for CSDM, but the framework is specific for the transportation context.

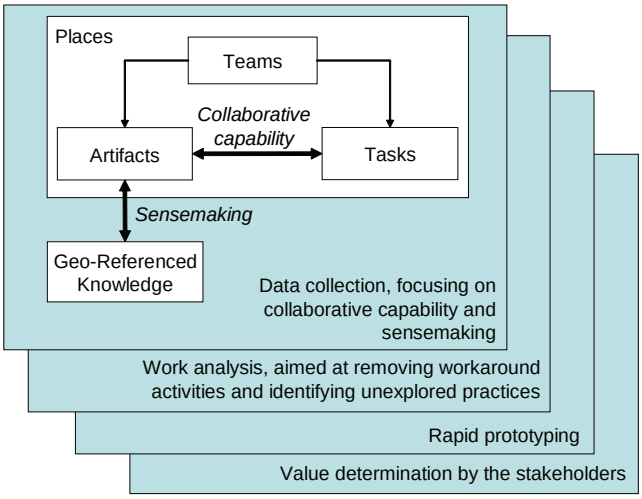
As demonstrated by the research previously cited, there is a whole new perspective over GSS brought by the mobility dimension, making CSDM quite distinct from GSS. However, the most important distinctions are not captured by current GSS and CSDM frameworks: (1) the central role of geo-references in the information architecture; (2) the interaction support to obtain, manage and share geo-referenced data while in the field; (3) the role of geo-references in modeling group work; and (4) the added impact of context awareness in the system design, regarding in particular work place mobility. Our perspective is that we need to integrate these various phenomena into a meaningful and purposeful framework.

THE FRAMEWORK

The framework is bounded by two major requirements: It has to be open for exploring and interpreting mobile collaboration in various settings, thus requiring relatively abstract elements and constructs, and it has to link them in a purposeful way. Our major goal is to set the initial boundaries for inquiring about mobile collaboration, setting at the same time a design roadmap.

The framework, shown in Figure 1, is structured around five basic elements and four design phases. The basic elements are teams, tasks, artifacts, and places, while the design phases consider data collection, work analysis, prototyping, and value determination. As described below in more detail, the basic elements have an important role throughout the design phases,

Figure 1. Design framework for mobile collaboration



structuring the various design activities taking place in each phase.

The relationships between the five basic elements are defined as follows. Teams manipulate artifacts to accomplish tasks in certain places. This combination of elements affords the most common spatial arrangements that we find in collaborative settings. The same argument applies to artifacts and tasks, were we may consider having artifacts/tasks fixed in a single place, distributed, or moving through several places. We assume these elements are consensual in the CSDM field, so that no further considerations are necessary.

In contrast, the relationship between artifacts and tasks, noted as *collaborative capability*, deserves further consideration. The notion of collaborative capability (Nunamaker, Romano, & Briggs, 2002) identifies several categories of increasing ability for successful creation of meaning, ranging from the individual, collective, and coordinated to the concerted creation of meaning. The theory is that organizations will increase their potential to create value by increasing their collaborative capability. Further details and validity tests of this theory can be found in Bach, Belardo, and Faerman (2004) and Qureshi and Briggs (2003). We realize this theory has an immediate impact in CSDM design, because work processes are affected by geographical constraints and thus there may be an opportunity for increasing the organizational effectiveness. From this theory we draw an implication for design: The development of shared artifacts, supporting concerted tasks, should be preferred to the development of individualized artifacts so that work processes become independent of geographical constraints.

The final framework basic element is geo-referenced knowledge. We regard the manipulation of artifacts, in mobile collaboration, not an end in itself but a mean to construct and augment shared knowledge about the work space and the objects found on it. This shared knowledge is necessarily tied to geographical references and mediated through artifacts. We may characterize the relationship between artifacts and geo-referenced knowledge as *sensemaking*: an ongoing process aiming to create order and make retrospective sense of what occurs (Weick, 1993). We argue sensemaking precisely captures the fundamental nature of mobile collaboration: people handling together information in fluid contexts. As the sensemaking theory posits, the outcomes from

mobile collaboration result from "thinking by doing" (Weick, 1993), since problems and solutions are highly context dependent. The presence of this element in the framework introduces one more implication for design: artifacts must enrich sensemaking by integrating mechanisms for searching, browsing, visualizing or summarizing geo-referenced information.

We now turn our attention to the design phases. The first phase concerns data collection aimed at understanding the work context. In this phase we adopt the contextual inquiry method (Beyer & Holtzblatt, 1998), which utilizes a mix of ethnography and interviews to understand the work. While contextual inquiry is context independent, this phase is structured around the framework basic elements, and specifically collects data about collaborative capability and sensemaking (how users organize themselves and make sense of geo-referenced data).

The second phase is dedicated to analyze work from the field data. Again, the framework plays an important role centering the analysis around places, artifacts and geo-referenced knowledge, focusing the modeling activity on the phenomena of most interest to mobile collaboration. We also suggest that attention to collaborative capability and sensemaking will raise new opportunities for removing workaround activities and identifying unexplored work practices, which are characteristic of innovative design solutions (Vicente, 1999).

The third phase is rapid prototyping. Here, low- or mid-fidelity prototypes serve to communicate with the stakeholders and evaluate the feasibility of the design ideas. The prototypes are fundamentally built around artifacts, task support and geo-referenced knowledge management.

Finally, the last step concerns the value determination by the stakeholders. We have been using context interviews (Beyer & Holtzblatt, 1998) to gather feedback from the stakeholders about the design solutions. Next, we describe two cases where this framework has been applied.

CASE STUDY ONE

This case addressed work redesign at a national agency responsible for inventorying geological resources. One major problem with this organization was that an inventory process took a long time to complete, mostly because experts had to go repeatedly to the field to retrieve information and resolve conflicts.

The framework helped organizing the field observations and interviews with experts involved in the process. This way we came to understand how work moved between the office and the field, what artifacts were used, and how geological information was gathered, analyzed, organized and consolidated. The inventory process required a combination of individual and collaborative activities, since expertise from different fields had to be combined.

Then, we began to analyze the work process, focusing on the five basic framework elements: teams, tasks, artifacts, places, and geo-referenced knowledge. At this stage we realized that a typical geological inventory took about 2 years to complete, as a consequence of several visits to the field, multiple activities in the office and many gap periods. Several critical incidents concurred to this situation: (1) bad initial data; (2) the occurrence of doubts when in the field or in the office; (3) the occurrence of conflicts between experts, which could only be resolved by sending someone to the field for confirmation; and (4) the concurrent execution of multiple inventory processes, causing management and planning difficulties. The framework had also a crucial role in the identification of the major design requirements:

- Fieldwork evolved around two artifacts: the field book and the combination of a map with a transparent overlay. The map/overlay allowed drawing inventory data, while the field book was used to annotate supplementary information, including doubts and concerns arising in fieldwork. All relevant knowledge

was geo-referenced, both in the map/overlay and field book.

- The field book was personal, signifying a reduced collaborative capability. This indicated that sharing the field book could increase the collaborative capability.
- Sensemaking was problematic because of the many unresolved doubts arising during fieldwork and difficulties reconstructing the field context in the office. Also, geo-referenced knowledge was distributed between the field book and map/overlay, which were difficult to co-relate. These observations indicated there was ample opportunity to develop information management mechanisms aiming to increase sensemaking.
- The inventory process was delayed by the need to swap work between the office and the field, a situation which could be resolved by increasing the team's collaborative capability: Bringing all relevant stakeholders together to resolve problems as they were appearing in the field or in the office.

These requirements lead us to prototype a digital artifact integrating the field book and map/overlay, and supporting cross-referencing and searching. We also allowed the fieldworker to contact the office workers using GPRS and an instant messaging mechanism. The redesigned work process allowed the fieldworker to get in contact with the office workers and immediately exchange comments on any occurring problem. The elements in the field book were synchronized to keep the conversation in context and facilitate sensemaking. Also, the fieldworker had an easier task when moving back to the office. Because doubts were resolved in the field, there was less time spent in the office. Addressing our observation that all knowledge were geo-referenced, the instant messages exchanged between the field and office workers were preserved in the field book with automatic associations to the geographical position of the fieldworker, thus keeping the doubts, comments or opinions in their context.

The prototype was evaluated with a field test and contextual interviews with several experts from the national agency. The obtained results indicate that the system increased sensemaking and collaborative capability. Related to sensemaking, the participants regarded very positively the expeditious way to locate points and associate them in the field book. Related with collaborative capability, the participants were extremely favorable to the communication between field and office workers, effectively resolving problems occurring in the field and thus simplifying the whole inventory process. More details about this case study can be found in (Antunes & André, 2006).

CASE STUDY TWO

This case involved work optimization in a small accountancy company, where meetings were the primary coordination mechanism. The company was not satisfied with the meetings productivity and regarded technology as a silver bullet. Different alternatives were experimented, which included the use of GSS and workflow tools, but cultural factors contributed to an unenthusiastic view of these technologies since they imposed too much structure to meetings. We proposed an alternative approach, which would not conflict with their informal work organization. The proposal considered the use of personal digital assistants (PDA) in meetings.

The framework allowed organizing the several data collected from interviews and meeting observations. We observed that the company had three types of meetings: (1) briefings, aimed to discuss ongoing projects; (2) planning meetings, where tasks and personnel were allocated to new projects; and (3) process definition meetings, where the whole collection of projects was taken in perspective to ensure an adequate allocation of resources. Different teams participated in these

meetings, accomplishing different tasks and using different artifacts and knowledge.

One issue raised by the framework during data collection was to identify people and information mobility related with meetings. During work analysis we characterized the specific nature of the artifacts moved by the accountants, such as meeting agendas, "to do" lists, and calendaring information. We came to understand two fundamental problems related with collaborative capability and sensemaking:

- It was difficult to move artifacts out of meetings. Sensemaking was affected by the lack of context (e.g., when a meeting outcome was delivered to someone that did not participate in the meeting).
- Meetings were affected by reduced collaborative capability, in particular the absence of a shared whiteboard capable to integrate the data brought by the participants.

These problems lead to the development of a prototype with the following characteristics: use PDA to bring information into and out of meetings; integrate the meeting information in a shared whiteboard; and supply a sensemaking mechanism capable to display the information flows across several meetings in an integrated way.

This case study was evaluated in two dimensions: framework and prototype. Selected accountants participated in evaluation tasks carried out at each design stage, evaluating the quality of data collected, work analysis, design ideas and prototype. The obtained feedback indicated that the framework was useful to elicit the organizational context of the problem. The evaluators also considered the data collection phase very useful and efficient. The work analysis phase was also considered very useful to help them understand the possibilities and limitations of the proposed solution.

Concerning the prototype, the evaluators considered the sensemaking functionality very useful and adjusted to their needs and thought that the simplicity of the PDA role bringing information in and out of meetings was adequate to their expectations, provided that not much text editing was required. More details about this case study can be found in Antunes and Costa (2002) and Costa, Antunes, and Dias (2002).

CONCLUSION

One important advantage of design frameworks is codifying current knowledge and best practices into design guidelines directly pointing towards where innovation may emerge. Our framework leads designers to identify meaningful ways to articulate places, users, tasks, artifacts and geo-referenced knowledge. The framework also guides the design process, keeping the designer focused on the issues most relevant to mobile collaboration.

The presented case studies highlight two different contexts where the framework pointed directly towards these concerns and definitely was useful informing the adopted designs. The evaluations conducted within the case studies confirmed the relevance of the framework as well as the relevance of the adopted design solutions. Artifacts emerged as the most important area of concern in mobile collaboration, mostly because they have potential to increase the collaborative capability and sensemaking.

REFERENCES

Antunes, P., & André, P. (2006). A conceptual framework for the design of geo-collaborative systems. *Group Decision and Negotiation, 15*, 273-295.

Antunes, P., & Costa, C. (2002). Handheld CSCW in the meeting environment. *LNCS, 2440*, 47-60.

Armstrong, M. (1994). Requirements for the development of GIS-based group decision support systems. *Journal of the American Society for Information Science, 45*(9), 669-677.

Armstrong, M. (1997). *Emerging technologies and the changing nature of work in GIS*. Bethesda, MD: American Congress on Surveying and Mapping.

Bach, C., Belardo, S., & Faerman, S. (2004). Employing the intellectual bandwidth model to measure value creation in collaborative environments. In *Proceedings of 37th Hawaii Int. Conference on System Sciences.*

Beyer, H., & Holtzblatt, K. (1998). *Contextual design: Defining customer-centered systems*. San Francisco: Morgan Kaufmann.

Coleman, D., & Li, S. (1999). Developing a groupware-based prototype to support geomatics production management. *Computers, Environment and Urban Systems, 23*, 1-17.

Coors, V., Jung, V., & Jasnoch, U. (1999). Using the virtual table as an interaction platform for collaborative urban planning. *Computers & Graphics, 23*(4), 487-496.

Costa, C., Antunes, P., & Dias, J. (2002). Integrating two organisational systems through communication genres. *LNCS, 2315,* 125-132.

DeSanctis, G., & Gallupe, R. (1987). A foundation for the study of group decision support systems. *Management Science, 33*(5), 589-609.

Gardels, K. (1997). Open GIS and on-line environmental libraries. *SIGMOD Record, 26*(1), 32-28.

Grønbæk, K., Vestergaard, P., & Ørbæk, P. (2002). Towards geo-spatial hypermedia: Concepts and prototype implementation. In *Proceedings of the 30th ACM Conference on Hypertext and Hypermedia.*

Hope, M., Chrisp, T., & Linge, N. (2000). Improving co-operative working in the utility industry through mobile context aware geographic information systems. In *Proceedings of the 8th ACM International Symposium on Advances in Geographic Information Systems.*

Mackay, S. (1999). Semantic integration of environmental models for application to global information systems and decision-making. *ACM SIGMOD Record, 28*(1), 13-19.

Manoharan, T., Taylor, H., & Gardiner, P. (2002). A collaborative analysis tool for visualisation and interaction with spatial data. In *Proceedings of the 7th International Conference on 3D Web Technology.*

Nunamaker, J., Dennis, A., Valacich, J., Vogel, D., & George, J. (1991). Electronic meeting systems to support group work: Theory and practice at Arizona. *Communications of the ACM, 34*(7), 40-61.

Nunamaker, J., Romano, N., & Briggs, R. (2002). Increasing intellectual bandwidth: Generating value from intellectual capital with information technology. *Group Decision and Negotiation, 11*(2), 69-86.

Nyerges, T., Montejano, R., Oshiro, C., & Dadswell, M. (1997). Group-based geographic information systems for transportation site selection. *Transportation Research, 5*(6), 349-369.

Pinsonneault, A., & Caya, O. (2005). Virtual teams: What we know, what we don't know. *International Journal of e-Collaboration, 1*(3), 1-16.

Pundt, H. (2002). Field data collection with mobile GIS: Dependencies between semantics and data quality. *GeoInformatica, 6*(4), 363-380.

Qureshi, S., & Briggs, R. (2003). Revision the intellectual bandwidth model and exploring its use by a corporate management team. In *Proceedings of 36th Hawaii International Conference on System Sciences.*

Rauschert, I., Agrawal, P., Sharma, R., Fuhrmann, S., Brewer, I., & MacEachren, A. (2002). Designing a human-centered, multimodal GIS interface to support emergency management. In *Proceedings of 10th ACM International Symposium on Advances In Geographic Information Systems.*

Tamminen, S., Oulasvirta, A., Toiskallio, K., & Kankainen, A. (2004). Understanding mobile contexts. *Personal and Ubiquitous Computing, 8*(2), 135-143.

Touriño, J., Rivera, F., Alvarez, C., Dans, C., Parapar, J., Doallo, R., et al. (2001). COPA: A GE-based tool for land consolidation projects. In *Proceedings of the 9th ACM International Symposium on Advances in Geographic Information Systems.*

Vicente, K. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Erlbaum.

Weick, K. (1993). The collapse of sensemaking in organizations: The Mann Gulch disaster. *Administrative Science Quarterly, 38*, 628-652.

Zhao, P., Nusser, S., & Miller, L. (2002). Design of field wrappers for mobile field data collection. In *Proceedings of 10th ACM international symposium on Advances in geographic information systems.*

KEY TERMS

Mobile Collaboration: People collaborating and moving through space.

Design Framework: A collection of general constructs identifying phenomena of interest and guiding the design process.

Collaborative Spatial Decision Making: The integrated study of collaboration, decision making, and mobility support.

Collaborative Capability: Defines four levels in increasing ability to create meaning: individual, collective, coordinated, and concerted.

Sensemaking: An ongoing process aiming to create order and make sense of what occurs.

Geo-Referenced Knowledge: Knowledge that is tied to a geographical reference.

Group Support System: A technological system supporting and mediating group work.

This work was previously published in the Encyclopedia of E-Collaboration, edited by N. Kock, pp. 133-138, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.13
Interface Design Issues for Mobile Commerce

Susy S. Chan
DePaul University, USA

Xiaowen Fang
DePaul University, USA

INTRODUCTION

Effective interface design for mobile handheld devices facilitates user adoption of mobile commerce (m-commerce). Current wireless technology poses many constraints for effective interface design. These constraints include limited connectivity and bandwidth, diverse yet simplistic devices, the dominance of proprietary tools and languages, and the absence of common standards for application development.

The convergence of mobile Internet and wireless communications has not yet resulted in major growth in mobile commerce. Consumer adoption of m-commerce has been slow even in countries such as Finland, which have broadly adopted wireless technology (Anckar & D'Incau, 2002). An international study of mobile handheld devices and services suggests that mobile commerce is at a crossroads (Jarvenpaa, Lang, Takeda & Tuunainen, 2003). The enterprise and business use of wireless technology holds greater promise, but it demands the transformation of business processes and infrastructure. Poor usability of mobile Internet sites and wireless applications for commerce activities stands out as a major obstacle for the adoption of mobile solutions. For example, even with the latest 3G phones in Japan, consumers still find the small screen display and small buttons on these devices difficult to use (Belson, 2002).

BACKGROUND

Mobile Commerce

Mobile commerce broadly refers to the use of wireless technology, particularly handheld mobile devices and mobile Internet, to facilitate transaction, information search, and user task performance in business-to-consumer, business-to-business, and intra-enterprise communications (Chan & Fang, 2003). Researchers have proposed

several frameworks for the study of m-commerce. Varshney and Vetter's framework (2001) presents 12 classes of m-commerce applications, ranging from retail and online shopping, auction, mobile office, and entertainment to mobile inventory emphasizing the potential of mobile B2B and intra-enterprise applications. The framework by Kannan, Chang, and Whinston (2001) groups mobile services into goods, services, content for consumer e-commerce, and activities among trading partners.

Waters (2000) proposes two visions for the potential and opportunities of m-commerce. One perspective argues that the mobile, wireless channel should be viewed as an extension of the current e-commerce channel or as part of a company's multi-channel strategies for reaching customers, employees, and partners. The second, more radical view suggests that m-commerce can create markets and business models.

Recent development in m-commerce has substantiated the first perspective. Major e-commerce sites have implemented their mobile Internet sites as an extension of wired e-commerce to support existing customers (Chan & Lam, 2004; Chan et al., 2002). Consumers have shown relatively low willingness to use m-commerce, but adopters of e-commerce are more likely to embrace this new technology (Anckar & D'Incau, 2002). Furthermore, perceived difficulty of use can affect consumers' choice of m-commerce as a distribution channel (Shim, Bekkering & Hall, 2002). These findings suggest that in a multi-channel environment, m-commerce *supplements* e-commerce instead of becoming a *substitute* for e-commerce.

Enterprise and business applications of m-commerce technologies seem to hold greater promise, because it is easier for companies to standardize and customize applications and devices to enhance current work processes. An Ernst & Young study (2001) of the largest companies in Sweden shows that, except for the retail industry sector, most industries have viewed m-commerce as being vital for growth and efficiency strategies, but not necessarily for generating new revenue. However, integrating the wireless platform in an enterprise requires significant structural transformation and process redesign.

Research on Wireless Interface Design

Several recent studies have examined interface design for mobile applications using handheld devices. Researchers have found that direct access methods were more effective for retrieval tasks with small displays (Jones, Marsden, Mohd-Nasir, Boone & Buchanan, 1999). Novice WAP phone users perform better when using links instead of action screens for navigation among cards, and when using lists of links instead of selection screens for single-choice lists (Chittaro & Cin, 2001). Ramsay and Nielsen (2000) note that many WAP usability problems echo issues identified during the early stage of Web site development for desktop computers, and could be alleviated by applying good user interface design. Such design guidelines for WAP applications include: (1) short links and direct access to content, (2) backward navigation on every card, (3) minimal level of menu hierarchy, (4) reduced vertical scrolling, (5) reduced keystrokes, and (6) headlines for each card (Colafigi, Inverard & Martriccian, 2001; Buchanan et al., 2001). Buyukkokten, Garcia-Molina, and Paepcke (2001) have found that a combination of keyword and summary was the best method for Web browsing on PDA-like handheld devices.

Diverse form factors have different interface requirements. The study by Chan et al. (2002) of 10 wireless Web sites across multiple form factors reveals that user tasks for the wireless sites were designed with steps similar to the wired e-commerce sites, and were primarily geared towards experienced users. Many usability problems, such as long download and broken connections, information overload, and excessive horizontal and vertical scrolling, are common to three

form factors—WAP phone, wireless PDA, and Pocket PC. Interface design flaws are platform independent, but the more limitations imposed on the form factors, the more acute the design problems become.

Mobile users access information from different sources and often experience a wide range of network connectivity. Context factors have a particular impact on the usability of mobile applications. Based on a usability study conducted in Korea, three use context factors—hand (one or two hands), leg (walking or stopping), and co-location (alone or with others)—may result in different usability problems (Kim, Kim, Lee, Chae & Choi, 2002). Therefore, the user interface design has to consider various use contexts. Researchers also suggest a systems-level usability approach to incorporating hardware, software, "netware," and "bizware" in the design of user-friendly wireless applications (Palen & Salzman, 2002). Perry, O'Hara, Sellen, Brown, and Harper (2001) have identified four factors in "anytime anywhere" information access for mobile work: the role of planning, working in "dead time," accessing remote technological and informational resources, and monitoring the activities of remote colleagues.

Multimodal interfaces are gaining importance. The MobileGuiding project developed in Spain is aimed at building a European interactive guide network on a common, multimodal, and multilingual platform in which contributors will provide leisure information and cultural events in their locations (Aliprandi et al., 2003). Furthermore, there has been a study conducted in Finland that addresses the design and evaluation of a speech-operated calendar application in a mobile usage context (Ronkainen, Kela & Marila, 2003).

MAIN THRUST OF THIS ARTICLE

Five issues are essential to the interface design for mobile commerce applications, including: (a) technology issues, (b) user goals and tasks, (c) content preparation, (d) application development, and (e) the relationship between m- and e-commerce.

Technology Issues

Limitation of Bandwidth

Most mobile communication standards only support data rates that are less than 28.8 kbps. Connections to the wireless service base stations are unstable because signal strength changes from place to place, especially on the move. These constraints limit the amount of information exchanged between device and base station. Indication of the download progress and friendly recovery from broken connections are necessary to help users gain a better sense of control.

Form Factor

Mobile commerce services are accessible through four common platforms: wireless PDA devices using Palm OS, Pocket PCs running Microsoft Windows CE/Pocket PC OS, WAP phone, and two-way pagers. Within the same platform, different form factors may offer different functionalities. A developer should consider the form factor's unique characteristics when developing m-commerce applications.

User Goals and Tasks

Mobile users can spare only limited time and cognitive resources in performing a task. Services that emphasize mobile values, and time-critical and spontaneous needs, add more value for m-commerce users. These mobile services may include the ability to check flight schedules, check stock prices, and submit bids for auction (Anckar & D'Incau, 2002). In addition, mobile tasks that demonstrate a high level of perceived usefulness, playfulness, and security are the ones most likely

to be adopted by users (Fang, Chan, Brzezinski & Xu, 2003).

Content Preparation

Constraints in bandwidth and small screen size demand different design guidelines. Most design guidelines for e-commerce (e.g., Nielsen, Farrell, Snyder & Molich, 2000) support the development of rich product information sets and a complete shopping process. In contrast, wireless Web sites have to simplify their content presentation.

Amount of Information

Content adaptation is necessary to convert information for the mobile Web (Zhou & Chan, 2003). However, users should have sufficient, if not rich, information to accomplish the goals for the application.

Navigation

Navigation systems vary from one form factor to another because the design of handheld devices differs. Currently, there is no consensus on which functions or features should be provided by the application, or built into the device itself.

Depth of Site Structure

Since mobile users have limited time for browsing wireless applications, the organization of information is critical. A flatter structure with fewer steps for wireless applications would allow users to review more options in the same step, and to locate the desired information more quickly.

Graphics or Text

Text is a better choice for displaying information on small screen browsers. However, better technology may improve the screen quality of handheld devices to display more complicated graphics. When determining the format of information to present, it is important to consider the form factor, because it may pose additional constraints on the format.

Development Environment

Mobile computing alters the assumption of "fixed" context of use for interface design and usability testing (Johnson, 1998). Traditional means of user interviews or usability testing in a laboratory environment cannot reveal insights into users' activities and mobility in real life. Contextual consideration is critical for gathering information about user requirements. For example, when developing and testing a mobile application for grocery shoppers, user requirement gathering and prototype evaluation should be conducted in a grocery store (Newcomb, Pashley & Stasko, 2003). The method of contextual inquiry can augment user interface design by exploring the versatility of usage patterns and usage context (Väänänen-Vainio-Mattila & Ruuska, 1998). While contextual inquiry may help developers gain a realistic understanding of contextual factors affecting user behaviors in motion, it is difficult to conduct non-obtrusive observations and inquiries. Developers for mobile applications need to consider the application context surrounding the relationship between the mobile device and user goals and tasks.

Relationship Between M-Commerce and E-Commerce

The wireless channel for e-commerce has raised many new questions regarding coordination between interactions with users across multiple channels. Some researchers suggest that because of the "transaction aware" and "location aware" characteristics of the wireless technology, mobile consumers may increase impulse purchases, especially in low-value, low-involvement product categories, such as books and CDs (Kannan et al.,

2001). At present, many Web sites have extended the wireless channel to leverage relationships with exiting customers (Chan et al., 2002). The current state of technology and poor usability of mobile Web sites makes it difficult to expand m-commerce as an independent channel. Many analysts believe that the wireless channel is promising for customer relationship management (CRM) because of its ability to: (1) personalize content and services; (2) track consumers or users across media and over time; (3) provide content and service at the point of need; and (4) provide content with highly engaging characteristics (Kannan et al., 2001). The challenge is how to coordinate interface and content across multiple channels so that experienced users and repeat customers can handle multiple media and platforms with satisfaction.

FUTURE TRENDS

Technology Trends

User interface design for mobile commerce will likely be influenced by four trends. First, multiple standards for wireless communication will not be resolved quickly, especially in North America. Second, the high cost of third-generation (3G) technology may delay the availability of broadband technology for complex functionality and content distribution for mobile applications. Third, instead of the convergence of functionalities into a universal mobile handheld device, there may be a variety of communication devices operating in harmony to support users in their everyday lives. Fourth, input and output format may expand to incorporate voice and other formats, as well as expandable keyboards. The introduction of the voice-based interfaces may complement the text-based interface and remedy some of the information input/display problems of the handheld devices. These trends suggest opportunities to conceptualize wireless user interface beyond text-based interaction. The new challenges are to design better multimodal interfaces for inter-device communication in order to simplify tasks for mobile users.

Development Trends

Alternative methods for interface design and evaluation will be necessary to support m-commerce applications development. First, requirement analysis should focus on the context of mobile users' behaviors and tasks. Contextual inquiry and other methods may be developed to facilitate the understanding of interactions between mobility and usability. Second, usability testing should be conducted with an understanding of contextual variables beyond user behavior. Third, mapping form factors, user tasks, data needs, and content across multiple channels and platforms is necessary to synchronize content and coordinate functionality in a distributed system. Fourth, user-centered design guidelines for mobile applications will be important. These trends require a fresh look at current methodology and will help determine new ways of incorporating user interface design and usability testing for distributed wireless application development. The reference framework proposed by Lee and Benbasat (2003) may be useful in this regard. Their framework incorporates seven design elements for m-commerce interface: context, content, community, customization, communication, connection, and commerce.

M-Commerce Business Models

Wireless technology for m-commerce is likely to evolve in two areas. For intra-enterprise and business-to-business uses, wireless technology provides location-aware and mobility-aware solutions for mobile workers. There is a broad range of possibilities for B2B applications because such deployment can be controlled more easily. Content distribution may be integrated with the enterprise systems. Context-based applications,

interfaces, functionality, and even devices can be customized according to the mobile tasks and user groups in the B2B context. This approach makes application development, deployment, and integration easier to manage. In contrast, it is far more challenging to manage the design, development, and deployment of wireless applications for customers. Wireless technology's capability for personalization seems to be the strongest argument for m-CRM services to enhance customer retention (Chan & Lam, 2004). A careful mapping of tasks, data, form factors, and the CRM process is essential for user interface design.

CONCLUSION

Wireless technology and the mobile Internet continues to evolve. Until the technology matures and bandwidth improves, wireless applications will be geared toward users requiring limited bandwidth, short exchange of data and text, and simple functionality. Two areas of wireless applications, CRM and enterprise efficiency, may reap greater success. Consumer e-commerce Web sites should focus on the selection of tasks that are most suitable for the wireless channel and demonstrate mobile values, especially for experienced users. Such mapping process requires a solid understanding of the CRM strategy, user preferences, and the constraints imposed by a mobile environment. For enterprise adoption, consolidating the wireless platforms and form factors will facilitate interface design. In either case, additional research to improve usability for mobile commerce is essential.

REFERENCES

Aliprandi, C., Athenour, M., Martinez, S.C., & Patsis, N. (2003). MobileGuiding: A European multimodal and multilingual system for ubiquitous access to leisure and cultural contents. In C. Stephanidis & J. Jacko (Eds.), *Proceedings of the 10th International Conference on Human-Computer Interaction* (vol. 2, pp. 3-7). Mahwah, NJ: Lawrence Erlbaum.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. *Journal of Information Technology Theory & Application*, *4*(1), 43-64.

Belson, K. (2002, April 22). Japan is slow to accept the latest phones. *The New York Times*, C4.

Buchanan, G., Farrant, S., Jones, M., Thimbleby, H., Marsden, G., & Pazzani, M. (2001). Improving mobile Internet usability. *Proceedings of the 10th International World Wide Web Conference* (pp. 673-680). New York: ACM Press.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the whole in parts: Text summarization for Web browsing on handheld devices. *Proceedings of the 10th International World Wide Web Conference*. New York: ACM Press.

Chan, S., & Fang, X. (2003). Mobile commerce and usability. In K. Siau & E. Lim (Eds.), *Advances in mobile commerce technologies* (pp. 235-257). Hershey, PA: Idea Group Publishing.

Chan, S., Fang, X., Brzezinski, J., Zhou, Y., Xu, S., & Lam, J. (2002). Usability for mobile commerce across multiple form factors. *Journal of Electronic Commerce Research*, *3*(3), 187-199.

Chan, S., & Lam, J. (2004). Customer relationship management on Internet and mobile channels: A framework and research direction. In C. Deans (Ed.), *E-commerce and m-commerce technologies*. Hershey, PA: Idea Group Publishing.

Chittaro, L., & Cin, P.D. (2001). Evaluating interface design choices on WAP phones: Single-choice list selection and navigation among cards. In M.D. Dunlop & S.A. Brewster (Eds.), *Proceedings of Mobile HCI 2001: Third International Workshop on Human Computer Interaction with Mobile Devices*.

Colafigli, C., Inverard, P., & Martriccian, R. (2001). Infoparco: An experience in designing an information system accessible through WEB and WAP interfaces. *Proceedings of the 34th Hawaii International Conference on System Science.* Los Alamitos, CA: IEEE Computer Society Press.

Ernst & Young. (2001). *Global online retailing: An Ernst & Young special report.* Gemini Ernst & Young.

Fang, X., Chan, S., Brzezinski, J., & Xu, S. (2003). A study of task characteristics and user intention to use handheld devices for mobile commerce. *Proceedings of the 2nd Annual Workshop on HCI Research in MIS* (pp. 90-94).

Jarvenpaa, S., Lang, K., Takeda, Y., & Tuunainen, V. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41-44.

Johnson, P. (1998). Usability and mobility: Interactions on the move. In C. Johnson (Ed.), *Proceedings of the 1st Workshop on Human Computer Interaction with Mobile Devices.*

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K., & Buchanan, G. (1999). Improving Web interaction on small displays. *Computer Networks: The International Journal of Distributed Informatique, 31,* 1129-1137.

Kannan, P., Chang, A., & Whinston, A. (2001). Wireless commerce: Marketing issues and possibilities. *Proceedings of the 34th Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press.

Kim, K., Kim, J., Lee, Y., Chae, M., & Choi, Y. (2002). An empirical study of the use contexts and usability problems in mobile Internet. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press.

Lee, Y. ,& Benbasat, I. (2003). Interface design for mobile commerce. *Communications of the ACM, 46*(12), 49-52.

Newcomb, E., Pashley, T., & Stasko, J. (2003). Mobile computing in the retail arena. *Proceedings of the Conference on Human Factors in Computing Systems, 5*(1), 337-344.

Nielsen, J., Farrell, S., Snyder, C., & Molich, R. (2000). *E-commerce user experience: Category pages.* Nielsen Norman Group.

Palen, L. & Salzman, M. (2002). Beyond the handset: Designing for wireless communications usability. *ACM Transactions on Computer-Human Interaction, 9*(2), 125-151.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: Understanding access anytime, anywhere. *ACM Transactions on Computer-Human Interaction, 8*(4), 323-347.

Ramsey, M., & Nielsen, J. (2000). *WAP usability: Déjà vu: 1994 all over again.* Nielsen Norman Group.

Ronkainen, S., Kela, J., & Marila, J. (2003). Designing a speech operated calendar application for mobile users. In C. Stephanidis & J. Jacko (Eds.), *Proceedings of the 10th International Conference on Human-Computer Interaction* (vol. 2, pp. 258-262). Mahwah, NJ: Lawrence Erlbaum.

Shim, J.P., Bekkering, E., & Hall, L. (2002). Empirical findings on perceived value of mobile commerce as a distributed channel. *Proceedings of the 8th Americas Conference on Information Systems* (pp. 1835-1837).

Varshney, U., & Vetter, R. (2001). A framework for the emerging mobile commerce applications. *Proceedings of the 34th Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press.

Väänänen-Vainio-Mattila, K., & Ruuska, S. (1998). User needs for mobile communication devices: Requirements gathering and analysis through contextual inquiry. In C. Johnson (Ed.),

Proceedings of the 1st Workshop on Human Computer Interaction with Mobile Devices.

Waters, R. (2000, March 1). Rival views emerge of wireless Internet. *Financial Times FT-IT Review,* 1.

Zhou, Y., & Chan, S. (2003). Adaptive content delivery over the mobile Web. *Proceedings of the 9th Americas Conference on Information Systems* (pp. 2009-2019).

KEY TERMS

Contextual Inquiry: This interface design method employs an ethnographic approach such as observing user activities in a realistic context.

Fixed Context of Use: Traditional user interface design and testing assumes a single domain, with the users always using the same computer to undertake tasks alone or in collaboration with others.

Form Factor: This platform or operating system runs on a handheld device. Major form factors include Palm, Pocket PC, and WAP.

Interface Design: Design of the interactions between humans and computers.

Location-Aware Service: Mobile services that provide information based on a user's location through the support of a global positioning system. Such services include mobile maps, weather, restaurants, and movie directories.

M-CRM: Interactions between a company and its customers for marketing, sales, and support services through the mobile Web and wireless channel.

Multimodal Interface: An interface that communicates with users through multiple modes.

Usability: Usability refers to how well an application is designed for users to perform desired tasks easily and effectively.

This work was previously published in the Encyclopedia of Information Science and Technology, Vol. 3, edited by M. Khosrow-Pour, pp. 1612-1617, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.14
Handheld Computing and Palm OS Programming for Mobile Commerce

Wen-Chen Hu
University of North Dakota, USA

Lixin Fu
The University of North Carolina at Greensboro, USA

Hung-Jen Yang
National Kaohsiung Normal University, Taiwan

Sheng-Chien Lee
University of Florida, USA

INTRODUCTION

It is widely acknowledged that mobile commerce is a field of enormous potential. However, it is also commonly admitted that the development in this field is constrained. There are still considerable barriers waiting to be overcome. One of the barriers is most software engineers are not familiar with handheld programming, which is the programming for handheld devices such as smart cellular phones and PDAs (personal digital assistants). This article gives a study of handheld computing to help software engineers better understand this subject. It includes three major topics:

- **Mobile commerce systems:** The system structure includes six components: (1) mobile commerce applications, (2) mobile handheld devices, (3) mobile middleware, (4) wireless networks, (5) wired networks, and (6) host computers.
- **Handheld computing:** It includes two kinds of computing: client- and server-side handheld computing.
- **Palm OS programming:** The Palm OS Developer Suite is used to develop applications for palm devices by handheld programmers.

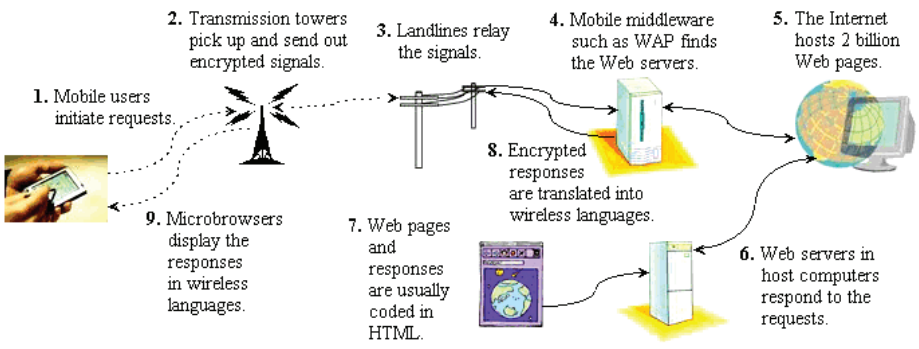
This article focuses on Palm OS programming by giving a step-by-step procedure of a palm application development. Other client-side handheld computing is also discussed.

BACKGROUND

With the introduction of the World Wide Web, electronic commerce has revolutionized traditional commerce and boosted sales and exchanges of merchandise and information. Recently, the emergence of wireless and mobile networks has made possible the extension of electronic commerce to a new application and research area: *mobile commerce*, which is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of mobile handheld devices. In just a few years, mobile commerce has emerged from nowhere to become the hottest new trend in business transactions. To explain how the mobile commerce components work together, Figure 1 shows a flowchart of how a user request is processed by the components in a mobile commerce system, along with brief descriptions of how each component processes the request (Hu, Lee, & Yeh, 2004).

1. **Mobile commerce applications:** Electronic commerce applications are numerous, including auctions, banking, marketplaces and exchanges, news, recruiting, and retailing to name but a few. Mobile commerce applications not only cover the electronic commerce applications, but also include new applications, which can be performed at any time and from anywhere by using mobile computing technology, for example, mobile inventory tracking.
2. **Mobile handheld devices:** An Internet-enabled mobile handheld device is a small general-purpose, programmable, battery-powered computer that is capable of handling the front end of mobile commerce applications and can be operated comfortably while being held in one hand. It is the device with which mobile users interact directly with mobile commerce applications (Hu, Yeh, Chu, Chu, Lee, & Lee, 2005).
3. **Mobile middleware:** The term middleware refers to the software layer between the operating system and the distributed applications that interact via the networks. The primary mission of a middleware layer is to hide the underlying networked environment's complexity by insulating applications from explicit protocols that handle disjoint

Figure 1. A flowchart of a user request processed in a mobile commerce system



memories, data replication, network faults, and parallelism (Geihs, 2001). The major task of mobile middleware is to seamlessly and transparently map Internet contents to mobile stations that support a wide variety of operating systems, markup languages, microbrowsers, and protocols. WAP (Open Mobile Alliance Ltd., n.d.) and i-mode (NTT DoCoMo Inc., 2006) are the two major kinds of mobile middleware.

4. **Wireless and wired networks:** Wireless communication capability supports mobility for end users in mobile commerce systems. Wireless LAN, MAN, and WAN are the major components used to provide radio communication channels so that mobile service is possible. In the WLAN category, the Wi-Fi standard with 11 Mbps throughput dominates the current market. However, it is expected that standards with much higher transmission speeds, such as IEEE 802.11a and 802.11g, will replace Wi-Fi in the near future. Compared to WLANs, cellular systems can provide longer transmission distances and greater radio coverage, but suffer from the drawback of much lower bandwidth (less than 1 Mbps). In the latest trend for cellular systems, 3G standards supporting wireless multimedia and high-bandwidth services are beginning to be deployed.
5. **Host computers:** A user request such as database access or updating is actually processed at a host computer, which contains three major kinds of software: (1) Web servers, (2) database servers, and (3) application programs and support software.

MAIN FOCUS OF THE CHAPTER

Handheld computing is a fairly new computing area and a formal definition of it is not found yet. Nevertheless, the authors define it as follows:

Handheld computing is to use handheld devices such as smart cellular phones and PDAs (personal digital assistants) to perform wireless, mobile, handheld operations such as personal data management and making phone calls.

Again, handheld computing includes two kinds of computing: client- and server- side handheld computing, which are defined as follows:

- **Client-side handheld computing:** It is to use handheld devices to perform mobile, handheld operations, which do not need the supports from server-side computing. Some of its applications are (a) address books, (b) video games, (c) note pads, and (d) to-do-list.
- **Server-side handheld computing:** It is to use handheld devices to perform wireless, mobile, handheld operations, which require the supports from server-side computing. Some of its applications are (a) instant messages, (b) mobile Web contents, (c) online video games, and (d) wireless telephony.

Client- and Server-Side Handheld Computing

Some popular mobile environments/languages for client-side handheld computing are listed below:

- **BREW (Binary Runtime Environment for Wireless):** It is an application development platform created by Qualcomm Inc. for CDMA-based mobile phones (Qualcomm Inc., 2003).
- **J2ME (Java 2 Platform, Micro Edition):** J2ME, developed by Sun Microsystems Inc., provides an environment for applications running on consumer devices, such as mobile phones, PDAs, and TV set-top boxes, as well

as a broad range of embedded devices (Sun Microsystem Inc., 2002).

- **Palm OS:** Palm OS, developed by Palm Source Inc., is a fully ARM-native, 32-bit operating system running on handheld devices. Using Palm OS to build handheld applications will be introduced later.
- **Symbian OS:** Symbian Ltd. is a software licensing company that develops and supplies the advanced, open, standard operating system—Symbian OS—for data-enabled mobile phones (Symbian Ltd., 2005).
- **Windows Mobile:** Windows Mobile is a compact operating system for mobile devices based on the Microsoft Win32 API. It is designed to be similar to desktop versions of Windows (Microsoft Corp., 2005).

They apply different approaches to accomplishing handheld computing. Table 1 shows the comparison among these five handheld-computing languages/environments.

Most applications of server-side handheld computing such as instant messaging require network programming such as TCP/IP programming will not be covered in this article. The most popular application of server-side handheld computing is database-driven mobile Web sites, whose structure is shown in Figure 2. A database-driven mobile Web site is often implemented by using a three-tiered client/server architecture consisting of three layers.

A database-driven mobile Web site is often implemented by using a three-tiered client/server architecture consisting of three layers:

1. **User interface:** It runs on a handheld device (the client) and uses a standard graphical user interface (GUI).
2. **Functional module:** This level actually processes data. It may consist of one or more separate modules running on a workstation or application server. This tier may be multi-tiered itself.
3. **Database management system (DBMS):** A DBMS on a host computer stores the data required by the middle tier.

The three-tier design has many advantages over traditional two-tier or single-tier designs, the chief one being: The added modularity makes it easier to modify or replace one tier without affecting the other tiers.

Table 1. A comparison among five handheld-computing languages/environments

	BREW	**J2ME**	**Palm OS**	**Symbian OS**	**Windows Mobile**
Creator	Qualcomm Inc.	Sun Microsystems Inc.	PalmSource Inc.	Symbian Ltd.	Microsoft Corp.
Language/ Environment	Environment	Language	Environment	Environment	Environment
Market Share (PDA) as of 2005	N/A	N/A	3rd	4th	1st
Market Share (Smartphone) as of 2006	?	N/A	4th	1st	5th
Primary Host Language	C/C++	Java	C/C++	C++	C/C++
Target Devices	Phones	PDAs & phones	PDAs	Phones	PDAs & phones

Palm OS

Palm OS is a fully ARM-native, 32-bit operating system designed to be used on palm handhelds and other third-party devices. Its popularity can be attributed to its many advantages such as its long battery life, support for a wide variety of wireless standards, and the abundant software available. The plain design of the Palm OS has resulted in a long battery life, approximately twice that of its rivals. It supports many important wireless standards including Bluetooth and 802.11b local wireless and GSM, Mo-bitex, and CDMA wide-area wireless networks (PalmSource Inc., 2002). Two major versions of Palm OS are currently under development:

- **Palm OS Garnet:** It is an enhanced version of Palm OS 5 and provides features such as dynamic input area, improved network communication, and support for a broad range of screen resolutions including QVGA.
- **Palm OS Cobalt:** It is Palm OS 6, which focuses on enabling faster and more efficient development of smartphones and integrated wireless (WiFi/Bluetooth) handhelds.

Palm OS Programming

The *Palm OS Developer Suite*, which is the official development environment and tool chain from PalmSource, is intended for software developers at all levels. It is a complete IDE (integrated development environment) for:

- Protein applications (all ARM-native code) for Palm OS Cobalt and
- 68K applications for all shipping versions of the Palm OS.

The following steps show how to develop a Palm OS application, a simple "Hello, Mobile world!" program, under Microsoft Windows XP:

1. Download and install the Palm OS Developer Suite at http://www.palmos.com/dev/tools/dev_suite.html.
2. Activate the Eclipse Workbench IDE as shown in Figure 3 under the Windows environment by selecting the following options:

   Start ▶ All Programs ▶ PalmSource ▶ Palm OS Developer Suite

Figure 2. A generalized system structure of a database-driven mobile Web site

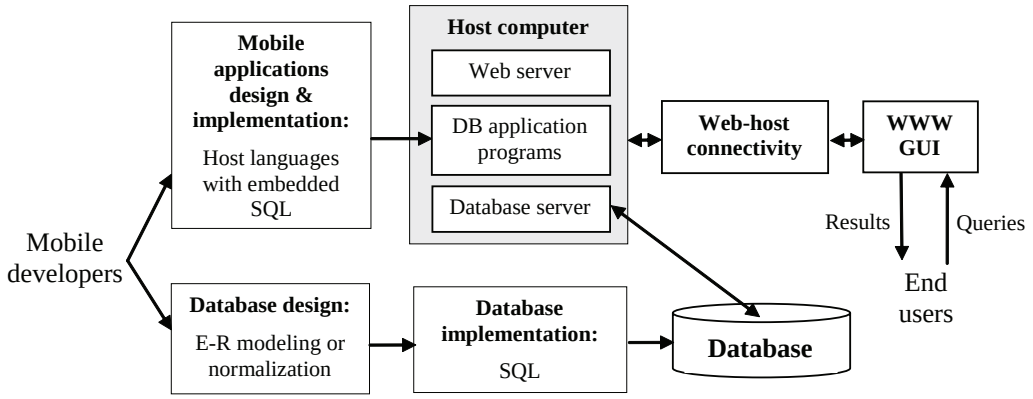
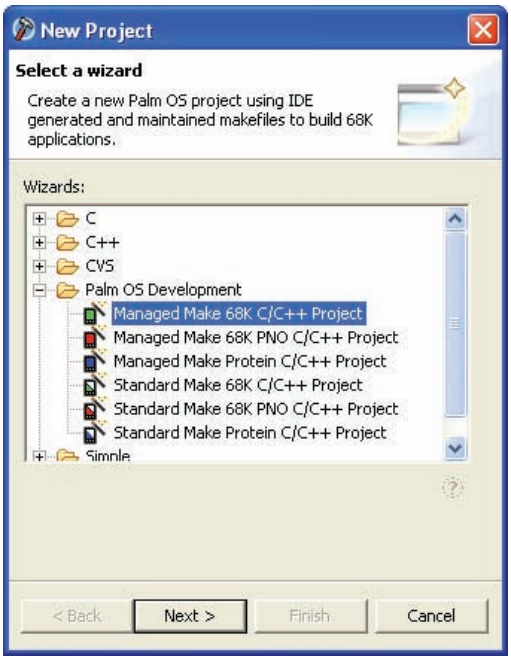


Figure 3. A screenshot of the Palm OS Developer Suite

Figure 4. A screenshot showing Palm OS application and make types



May select a default workspace at "C:\Program Files\PalmSource\Palm OS Developer Suite\workspace."

3. Create a new project by selecting a wizard. There are three Palm OS application types as shown in Figure 4:
   - Palm OS 68K Application
   - Palm OS 68K Application with PACE Native Objects
   - Palm OS Protein Application

   There are also two kinds of *make files*:
   - **Standard make:** It provides a generic set of makefiles that you can modify and tailor for your specific application build.
   - **Managed make:** It dynamically generates your makefile based on the contents of your project folders.
4. Create a Palm OS C/C++ program and put it in the directory "C:\Program Files\PalmSource\Palm OS Developer Suite\workspace\HelloWorld\." Figure 5 gives a Palm example, which displays the text "Hello, Mobile world!" an image, and a button "OK" on a Palm device.

   For how to create Palm OS applications, check Palm OS Developer Documentation at http://www.palmos.com/dev/support/docs/. In order to display the current status on the

Figure 5. An example of a Palm OS HelloWorld program

C:\Program Files\PalmSource\Palm OS Developer Suite\workspace\Hello\HelloWorld.c

```
// This header is from the Palm SDK and contains the needed reference
// materials for the use of Palm API and its defined constants.
#include <PalmOS.h>

// The following IDs are from using Palm Resource Editor.
#define Form1 1000
#define OK  1003

// ----------------------------------------------------------------------
// PilotMain is called by the startup code and implements a simple
// event handling loop.
// ----------------------------------------------------------------------
UInt32 PilotMain( UInt16 cmd, void *cmdPBP, UInt16 launchFlags ) {
 short   err;
 EventType e;
 FormType *pfrm;

 if ( cmd == sysAppLaunchCmdNormalLaunch ) {
  // Displays the Form with an ID 1000.
  FrmGotoForm( Form1 );

  // Main event loop
  while( 1 ) {
   // Doze until an event arrives or 100 ticks are reached.
   EvtGetEvent( &e, 100 );
   // System gets first chance to handle the event.
   if ( SysHandleEvent( &e ) ) continue;
   if ( MenuHandleEvent( (void *) 0, &e, &err ) ) continue;

   switch ( e.eType ) {
    case ctlSelectEvent:
     if ( e.data.ctlSelect.controlID == OK )
                    goto _quit;
     break;
    case frmLoadEvent:
     FrmSetActiveForm( FrmInitForm( e.data.frmLoad.formID ) );
     break;
    case frmOpenEvent:

 pfrm = FrmGetActiveForm( );

 FrmDrawForm( pfrm );

 break;
                    case menuEvent:

 break;
                    case appStopEvent:

 goto _quit;
     break;
                    default:
     if ( FrmGetActiveForm( ) )
      FrmHandleEvent( FrmGetActiveForm( ), &e );
     break;
   }
  }
_quit:
  FrmCloseAllForms( );
 }
```

Eclipse, may need to constantly refresh the project *HelloWorld* by right clicking on the mouse on the project name as shown in Figure 6.

If the project includes resources (with an .xrd filename extension) such as buttons and images, the *Palm OS Resource Editor* at

Start ► All Programs ► PalmSource ► Tools ► Palm OS Resource Editor

could be used to create the resources as shown in Figure 7.

5. Build the project *HelloWorld*.
6. Activate a Palm OS emulator by selecting

   Start ► All Programs ► PalmSource ► Tools ► Palm OS Emulator

7. Drag the icon of Hello.prc (Palm Application file) at "C:\Program Files\PalmSource\Palm OS Developer Suite \workspace\Hello5\Debug\Hello.prc" to the Palm OS emulator. Figure 8 shows the execution result of the project HelloWorld.
8. If the application is finalized, synchronize the application to handheld devices by selecting

   Start ► All Programs ► Palm Desktop ►
   Palm Desktop

   after downloading and installing the Palm Desktop at http://www.palmos.com/dev/tools/desktop/.

Palm References

Since this article is not intended to be a comprehensive Palm programming guide, this section provides more palm information for further references for interested readers. Table 2 shows four documents for Palm OS SDK (software

Figure 6. A screenshot of the Palm OS Developer Suite after the HelloWorld project is created

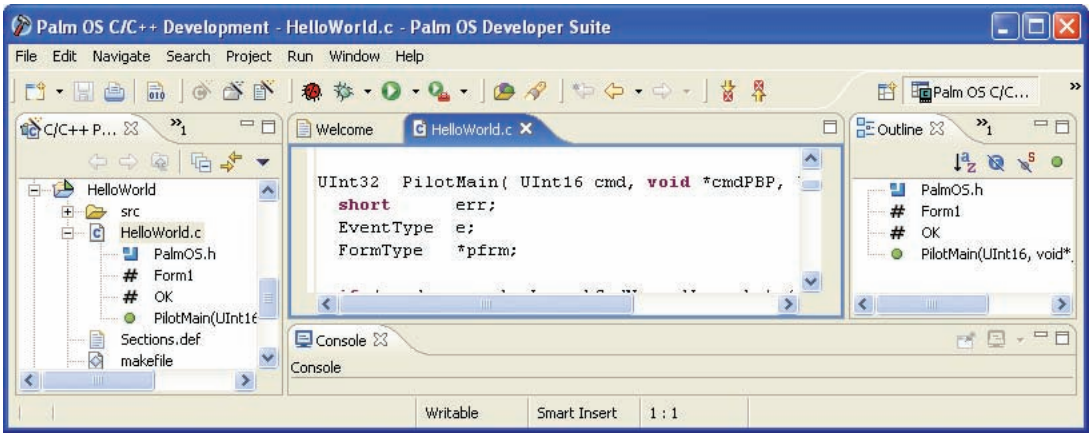


Figure 7. A screenshot of the Palm OS Resource Editor

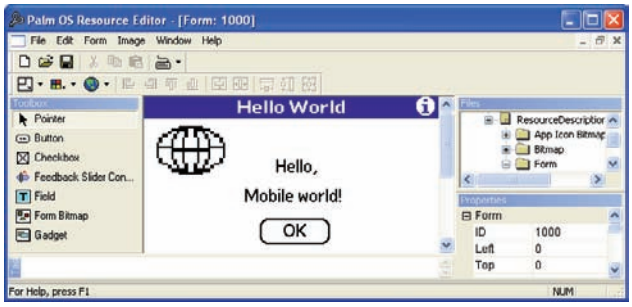


Figure 8. A screenshot of the execution results of the HelloWorld project

Table 2. Palm OS SDK documentation

Document	Description	URLs
Palm OS Programmer's API Reference	An API reference document that contains descriptions of all Palm OS function calls and important data structures.	http://www.palmos.com/dev/support/docs/palmos/PalmOSReference/ReferenceTOC.html
Palm OS Programmer's Companion, Vol. I & II	A multi-volume guide to application programming for the Palm OS. This guide contains conceptual and "how-to" information that complements the reference.	http://www.palmos.com/dev/support/docs/palmos/PalmOSCompanion/CompanionTOC.html and http://www.palmos.com/dev/support/docs/palmos/PalmOSCompanion2/Companion2TOC.html
Constructor for Palm OS	A guide to using constructor to create Palm OS resource files.	http://www.palmos.com/dev/support/docs/constructor/CGRTOC.html
Palm OS Programming Development Tools Guide	A guide to writing and debugging Palm OS applications with the various tools available.	http://www.palmos.com/dev/support/docs/devguide/ToolsTOC.html

Table 3. An overview of the Palm OS Programmer's Companion

Volume	Description
I	Gives fundamental knowledge of Palm OS programming such as event loop and user interface.
II	Describes the handheld's communications capabilities such as Bluetooth and network communication.

Table 4. An overview of the Palm OS programmer's API reference

Function	Description
User Interface	User interface APIs include events, notifications, attention, control, dialogs, forms, lists, menus, scroll bars, and so forth.
System Management	Provides largest number of functions such as alarm, debug, file streaming, graffiti, I/O, memory, pen, sound, time, windows, etc. for system management.
Communications	Provide various communication functions such as IR, modem, network, telephony, and so forth.
Libraries	Include miscellaneous libraries such as Internet, Bluetooth, cryptography, and so forth.

development kit). Details of the two documents *Palm OS Programmer's API Reference* and *Palm OS Programmer's Companion* are given next.

The *Palm OS Programmer's Companion* (PalmSource Inc., 2004b, 2004c) provides extensive conceptual and "how-to" development information, and official reference information of Palm OS 68K functions and data structures. Table 3 gives an overview of the Palm OS Programmer's Companion.

Table 4 gives an overview of the *Palm OS programmer's API* (Application Programming Interface) reference of Palm OS 68K SDK (PalmSource Inc., 2004a). It includes four major sections (1) user interface, (2) system management, (3) communications, and (4) libraries.

FUTURE TRENDS

A number of mobile operating systems with small footprints and reduced storage capacity have emerged to support the computing-related functions of mobile devices. For example, Research In Motion Ltd's BlackBerry 8700 smartphone uses RIM OS and provides Web access, as well as wireless voice, address book, and appointment applications (Research In Motion Ltd., 2005). Because the handheld device is small and has limited power and memory, the mobile OSes' requirements are significantly less than those of desk- or lap- top OSes. Although a wide range of mobile handheld devices are available in the market, the operating systems, the hub of the devices, are dominated by just few major organizations. The following two lists show the operating systems used in the top brands of smart cellular phones and PDAs in descending order of market share:

- **Smart cellular phones:** Symbian OS, Linux, RIM OS, Palm OS, Windows Mobile-based Smartphone, and others (Symbian Ltd., 2006).
- **PDAs:** Microsoft Pocket PC, RIM OS, Palm OS, Symbian OS, Linux, and others (Gartner Inc., 2005).

The market share is changing frequently and claims concerning the share vary enormously. It is almost impossible to predict which will be the ultimate winner in the battle of mobile operating systems.

CONCLUSION

Using Internet-enabled mobile handheld devices to access the World Wide Web is a promising addition to the Web and traditional e-commerce. Mobile handheld devices provide convenience and portable access to the huge information on the Internet for mobile users from anywhere and at anytime. However, most software engineers are not familiar with programming for handheld devices. Handheld computing is the programming for handheld devices such as smart cellular phones and PDAs. This article gives a study of handheld computing by including three major topics:

- **Mobile commerce systems:** Mobile commerce is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of mobile handheld devices. The system structure includes six components: (1) mobile commerce applications, (2) mobile handheld devices, (3) mobile middleware, (4) wireless networks, (5) wired networks, and (6) host computers.
- **Handheld computing:** It includes two kinds of computing:
  - **Client-side handheld computing:** It is to use handheld devices to perform mobile, handheld operations, which do not need the supports from server-side computing.
  - **Server-side handheld computing:** It is to use handheld devices to perform wireless, mobile, handheld operations, which need the supports from server-side computing.
- **Palm OS programming:** Two major versions of Palm OS are currently under development:
  - **Palm OS Garnet:** It is an enhanced version of Palm OS.
  - **Palm OS Cobalt:** It is the Palm OS 6.

This article focuses on Palm OS programming by giving a step-by-step procedure of a palm application development. The Palm OS Developer Suite is used to develop applications for palm devices by handheld programmers.

REFERENCES

Gartner Inc. (2005). Gartner says Worldwide PDA shipments increased 32 percent in the second quarter of 2005. Retrieved January 13, 2006, from http://www.gartner.com/press_releases/asset_133230_11.html

Geihs, K. (2001). Middleware challenges ahead. IEEE Computer, 34(6), 24-31.

Hu, W. C., Lee, C. W., & Yeh, J. H. (2004). Mobile commerce systems. In S. Nansi (Ed.), *Mobile commerce applications (pp.* 1-23). Hershey, PA: Idea Group Publishing.

Hu, W.-C., Yeh, J.-h., Chu, H.-J., & Lee, C.-w. (2005). Internet-enabled mobile handheld devices for mobile commerce. *Contemporary Management Research, 1*(1), 13-34.

Microsoft Corp. (2005). What's new for developers in Windows Mobile 5.0? Retrieved December 21, 2005, from http://msdn.microsoft.com/mobility/windowsmobile/howto/documentation/default.aspx?pull=/library/en-us/dnppcgen/html/whatsnew_wm5.asp

NTT DoCoMo Inc. (2006). i-mode technology. Retrieved October 2, 2005, from http://www.nttdocomo.com/technologies/present/imodetechnology/index.html

Open Mobile Alliance Ltd. (n.d.). WAP (wireless application protocol). Retrieved July 21, 2005, from http://www.wapforum.org/

PalmSource Inc. (2004a). Palm OS programmer's API reference. Retrieved December 15, 2005, from

http://www.palmos.com/dev/support/docs/palmos/PalmOSReference/ReferenceTOC.html

PalmSource Inc. (2004b). Palm OS programmer's companion, Vol. I. Retrieved February 21, 2006, from http://www.palmos.com/dev/support/docs/palmos/PalmOSCompanion/CompanionTOC.html

PalmSource Inc. (2004c). Palm OS programmer's companion, Vol. II. Retrieved February 21, 2006, from http://www.palmos.com/dev/support/docs/palmos/PalmOSCompanion2/Companion2TOC.html

PalmSource Inc. (2002). Why PalmOS? Retrieved June 23, 2005, from http://www.palmsource.com/palmos/Advantage/index_files/v3_document.htm

Qualcomm Inc. (2003). BREW and J2ME—A complete wireless solution for operators committed to Java. Retrieved February 12, 2005, from http://brew.qualcomm.com/brew/en/img/about/pdf/brew_j2me.pdf

Research In Motion Ltd. (2005). BlackBerry application control—An overview for application developers. Retrieved January 05, 2006, from http://www.blackberry.com/knowledgecenterpublic/livelink.exe/fetch/2000/7979/1181821/832210/BlackBerry_Application_Control_Overview_for_Developers.pdf?nodeid=1106734&vernum=0

Sun Microsystem Inc. (2002). Java 2 Platform, Micro Edition. Retrieved January 12, 2006, from http://java.sun.com/j2me/docs/j2me-ds.pdf

Symbian Ltd. (2006). Fast facts. Retrieved June 12, 2006, from http://www.symbian.com/about/fastfacts/fastfacts.html

Symbian Ltd. (2005). Symbain OS Version 9.2. Retrieved December 20, 2005, from http://www.symbian.com/technology/symbianOSv9.2_ds_0905.pdf

KEY TERMS

Client-Side Handheld Computing: It is used by handheld devices to perform mobile, handheld operations, which do not need the supports from server-side computing. Some of its applications are (a) address books, (b) video games, (c) note pads, and (d) to-do-list.

Handheld Computing: It is used by handheld devices such as smart cellular phones and PDAs (personal digital assistants) to perform wireless, mobile, handheld operations such as personal data management and making phone calls.

Mobile Commerce: It is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of mobile handheld devices.

Mobile Handheld Devices: They are small general-purpose, programmable, battery-powered computers, but they are different from desk- or lap- top computers mainly due to the following special features: (1) limited network bandwidth, (2) small screen/body size, and (3) mobility.

Palm OS: Palm OS, developed by PalmSource Inc., is a fully ARM-native, 32-bit operating system running on handheld devices. Two major versions of Palm OS are currently under development: Palm OS Garnet and Palm OS Cobalt.

Palm OS Developer Suite: It is the official development environment and tool chain from PalmSource Inc. and is intended for software developers at all levels. It is a complete IDE (Integrated Development Environment) for (1) Protein applications (all ARM-native code) for

Palm OS Cobalt and (2) 68K applications for all shipping versions of the Palm OS.

Server-Side Handheld Computing: It is used by handheld devices to perform wireless, mobile, handheld operations, which need the supports from server-side computing. Some of its applications are (a) instant messages, (b) mobile Web contents, (c) online video games, and (d) wireless telephony.

This work was previously published in the Encyclopedia of Internet Technologies and Applications, edited by M. Freire and M. Pereira, pp. 205-214, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.15
Privacy-Preserving Transactions Protocol Using Mobile Agents with Mutual Authentication

Song Han
Curtin University of Technology, Australia

Vidyasagar Potdar
Curtin University of Technology, Australia

Elizabeth Chang
Curtin University of Technology, Australia

Tharam Dillon
University of Technology, Australia

ABSTRACT

This article introduces a new transaction protocol using mobile agents in electronic commerce. The authors first propose a new model for transactions in electronic commerce, mutual authenticated transactions using mobile agents. They then design a new protocol by this model. Furthermore, the authors analyse the new protocol in terms of authentication, construction, and privacy. The aim of the protocol is to guarantee that the customer is committed to the server, and the server is committed to the customer. At the same time, the privacy of the customer is protected.

INTRODUCTION

Security and Privacy are the paramount concerns in Electronic Commerce (Eklund, 2006). Mobile agent systems are becoming involved in e-commerce (Claessens, Preneel, & Vandewalle, 2003). However, security and privacy within mobile agents must be addressed before mobile agents can be used in a wide range of electronic commerce applications.

The security in the electronic transactions with mobile agents can be classified into two different aspects:

- One is the security of the hosts, to which the mobile agents will travel
- The other is the security of the mobile agents, by which some sensitive information may be transported to the hosts.

The above first security is used to protect the hosts, since the mobile agents may be malicious. For example, the mobile agent may be in the disguise of a legal mobile agent. Therefore, the host will interact with the mobile agent on an electronic transaction. However, the mobile agent tries to obtain some sensitive information (e.g., the secret development plan, the financial report, etc.) about the host. This case will damage the benefit of the host. Therefore, it is very important to maintain the security of the host if some malicious mobile agents travel to the hosts.

The above second security is used to protect the mobile agents, since the hosts may be hostile. For example, when the mobile agents with some sensitive information arrive at the host, those pieces of sensitive information (e.g., the private key, bank account password, home address, etc.) are of paramount importance to the mobile agents' owner. Therefore, the host may try to attain the information through interacting with the mobile agents. As a result, the customer (the owner of the mobile agents) may be blackmailed by the host, since the host holds some sensitive information obtained from the underlying mobile agents. Therefore, it is imperative to design some security mechanism to maintain the security of the mobile agents.

Hosts' security mechanisms include: (1) authentication; (2) verification; (3) authorisation; and (4) payment for services (Claessens et al., 2003). In this article, we will utilise the method of authentication to preserve the security of the host. Authentication is one of the cryptographic techniques. The implication of authentication is to *assure that the entity (customer, host, mobile agents, etc.) requesting access or interaction is the entity that it claims to be.*

Mobile agents' security mechanisms (Kotzanikolaou, Burmester, & Chrissikopoulos, 2000) include: (1) authentication; (2) encryption algorithms; and (3) digital signatures. In this article, we will utilise the digital signature technique. Digital signature is another cryptographic technique. A digital signature scheme is a method of signing a message stored in electronic form. As such, a signed message can be transmitted over a computer network. Also, a signature can be verified using a publicly known verification algorithm. Therefore, anyone who knows the verification algorithm can verify a digital signature. The essence of digital signatures is to convince the recipient that a message (attached to its valid digital signature) is really sent from the signer.

In a virtual community, delegation of signing rights is an important issue since security and privacy are concerned. Consider the following scenario: An international logistics company, AuHouse's President is scheduled to sign a major contract with an Automobile Company in Europe on February 28. However, because of a management emergency, the President is required to attend a meeting held in the General Building of AuHouse in Australia on the same day. This meeting is vital to the future of the AuHouse. However, the contract in Europe is also very important to the organisation. How then can the President be in two places at once and sign the contract, even though he cannot be physically in Europe? Undetachable signature protocol will help the President to solve this issue since the undetachable signature protocol can provide the delegation of signing power while preserving the privacy of the President.

Undetachable signatures are one of the digital signatures which could provide secure delegation of signing rights while preserving privacy. So far only a few undetachable signatures have been created (Coppersmith, Stern, & Vaudenay, 1993; Kotzanikolaou et al., 2000; Sander & Tschudin, 1998). Sander and Tschudin (1998) first proposed the undetachable signatures. The construction is

based on the birational functions (Shamir, 1993). However, Stern proved that undetachable signatures based on birational functions are insecure and vulnerable to the attacks (Coppersmith et al., 1993). Another construction on RSA cryptosystem was proposed (Kotzanikolaou et al., 2000). This undetachable signature scheme is secure since its security is based on the security of RSA signatures. However, it is known that RSA signatures usually need to be about 1,024 bit-length or much more in order to maintain an optimal security level (Lauter, 2004). At the same time, mobile agents are working in an environment of mobile communications. Therefore, low bandwidth and efficient communications are much more satisfactory for mobile agents, since the mobile agents often migrate from its owner to a server and from this server to other servers.

Two secure transaction protocols having short signatures have been proposed (Han, Chang, & Dillon, 2005a, 2005b). However, their schemes do not have the mutual authentication mechanism. Therefore, their protocol cannot guarantee that the server is committed to the customer, and the customer is committed to the server.

In this article, we will design a mutual authenticated transaction protocol with mobile agents. We will provide the mutual authentication mechanism to assure that the mobile agent is really the one it claims to be, sent by its owner, the customer. Simultaneously, we will provide a new undetachable signature to protect the privacy of the customer.

The organisation of the rest of this article is as follows: We first provide the model of mutual-authenticated electronic transactions with mobile agents, and the definition of undetachable signatures. Secondly, some mathematical preliminaries are presented that will be used in the new protocol. Thirdly, a new transaction protocol with mutual authentication using mobile agents is proposed. Then, the analysis and proofs are provided, mainly including authentication analysis and construction analysis, as well as privacy analysis—a very important property for a practical virtual community. The conclusions appear in the last section.

MODEL OF MUTUAL AUTHENTICATED TRANSACTIONS WITH MA AND DEFINITION OF UNDETACHABLE SIGNATURES

In this section, we will provide a new model of the transactions using mobile agents (MA) with mutual authentication between the server and the mobile agents (de facto, the customer) and the definition of undetachable signatures. Note that it is the first definition for undetachable signatures, to the best of our knowledge. An undetachable signature scheme consists of four algorithms, namely setup, key, sign, and verify.

Model of Mutual Authenticated Transactions with MA

There are at least four participants involving in the model. The participants are: a customer C (which plays the role of the identifier of the customer), a number of servers (i.e., electronic shops)$S_1 S_2,...,S_n$ (which play the roles of all the servers, respectively), a key certificate authority KCA (which plays the role of the identifier of the key certificate authority), and a number of mobile agents MA_1, MA_2, ..., MA_n (which play the roles of these mobile agents, respectively). Besides these participants, there are six procedures for the proposed model. These procedures deliver the specifications for the mutual authenticated electronic transactions protocol using mobile agents (MAs). The details of this model are as follows:

Setup Algorithm

It generates public key and private key for the customer and also some public parameters for the corresponding servers. In this algorithm, the

customer will construct her purchase requirements Req$_c$ according to her purchase plan. The server will construct Bid_S, that defines the bid of the server for a selling activity.

Key Algorithm

In this algorithm, a key certificate authority will also be involved. The customer, a key certificate authority, and some servers will collaborate to assign some public and private parameters. A suitable public key encryption algorithm $E_{pub \otimes prv}$ will be known to the customer and those servers. The private keys and public keys will be certificated by the key certificate authority, respectively. In addition, there is a shared key between the key certificate authority. That will be used for the authentication before the mobile agent migrates to the underlying server.

Mobile Agents Preparing

It involves the interactions between the customer and its mobile agents. The customer will construct some mobile codes for each mobile agent $MA(1 \leq j \leq n)$. These mobile codes include: *TBI*, Req$_c$, and a pair of undetachable signature functions; where *TBI* is the temporary identifier of the customer. The undetachable signature function pair is used to generate the bid tuple on the purchase requirement. Therefore, these mobile agents will travel with the mobile codes to these servers.

Mutual Authentications

This algorithm is used to take authentications between the customer and the servers $S_1 S_2,...,S_n$. Authentication is mutual means that the customer and the underlying server authenticate in a symmetrical way. It will assure that: *the underlying server is committed to the customer; and the customer is committed to the server.*

Mobile Agent Execution

This algorithm will make the server attend the bidding for the purchase brought with the mobile agent. Each mobile agent will take the mobile codes to the server. Then, the server will design its bid and sign on the bid. In the end, the server arranges the mobile agent to travel back to the customer.

Transactions Verifying

The customer first checks whether the timestamp is still valid. If it is valid, the customer will verify the signature on the bid. If it is legal, and the bid is an optimal one, the customer will accept this bid.

Definition of Undetachable Signatures

Undetachable signature is, in fact, a kind of encrypted function. Some private parameters will be embedded in the function, and the recipient of this function can execute the computation of the function with some inputs chosen by her self. We utilise the definition proposed by Han and Chang (2006).

Setup is a probabilistic polynomial time algorithm which takes as input a security parameter k and outputs a family of system parameters.

Key is a probabilistic polynomial time algorithm which is executed by a trusted centre and the signers. The input contains system parameters, as well as random parameters which are chosen by the trusted centre and the signers. The output includes a public key $pk \in \underline{K}$ and a corresponding secret key sk.

Sign is a probabilistic polynomial time algorithm, which takes as input a secret key sk and a message $m \in \underline{M}$ and outputs a signature $Sig_{sk} \in \underline{S}$.

In general, there are many valid signatures for any pair $(m, pk) \in \underline{M} \times \underline{K}$.

Verify is a deterministic polynomial time algorithm. The input includes a message and its allayed signature $Sig_{sk} \in \underline{S}$, as well as system parameters. The output is "Accept" or "Otherwise".

PRELIMINARIES

In this section, we will provide some mathematical knowledge that is used in the design and analysis in the proposed protocol. There are two multiplicative cyclic groups, G_1 and G_2 of prime order q. g_1 is a generator of G_1 and g_2 is a generator of G_2.

A bilinear map is a map $e: G_1 \times G_2 \rightarrow G_T$ with these three properties:

Bi-linearity: for any $P \in G_1$, $Q \in G_2$ and x, $y \in Z$, $e(P^x, q^y)=(P, Q)^{xy}$.

Non-degenerate: if g_1 is a generator of G_1 and g_1 is a generator of G_2, then $e(g_1,g_2)=1$.

Efficient Computability: There is an efficient algorithm to compute $e(P, Q)$ for any P and Q.

We will use the general case $G_1 \neq G_2$ so that we can take advantage of certain families of elliptic curves to obtain short signatures. Specifically, elements of G_1 have a short representation, whereas elements of G_2 might not.

We say that (G_1, G_2) are bilinear groups if there exists a group G_T, an isomorphism $\psi: G_2 \rightarrow G_1$, and a bilinear map $e: G_1 \times G_2 \rightarrow G_T$, and e, ψ, and the group action in G_1, G_2, and G_T can be computed efficiently. Generally, the isomorphism ψ is constructed by the trace map over elliptic curves.

Each customer selects two generators $g_1 \in G_1$, $g_2 \in G_2$, and e(. , .) as above. He will choose $x \in Z^*_p$ and compute $v=g^x_2 \in G_2$. There are four cryptographic hash functions will be used:

H_1, H_2, H_3, and H_4,

where

$H_1 : \{0,1\}^* \times \{0,1\}^* \mapsto Z_p$
$H_2 : Z_p \mapsto Z_p$
$H_3 : \{0,1\}^* \times Z^*_p \mapsto Z_p$
$H_4 : \{0,1\}^* \times \{0,1\}^* \times Z_p \mapsto Z_p$.

TRANSACTIONS PROTOCOL WITH MUTUAL AUTHENTICATION

A new undetachable signature scheme will be proposed for the protocol of secure transactions. This new undetachable scheme belongs to the domain of short signatures (Boneh & Boyen, 2004; Courtois, 2004; Lauter, 2004). As described in the previous section, short signatures have the characteristics of shorter bit-length of signatures, fast signature generation, as well as fast signature verification. These characteristics are imperative for mobile agents, which take part in the secure transactions between a customer and any server.

Setup Algorithm

Setup algorithm is mainly to set up the compulsory parameters assigned to each participant. We will use the mathematical settings of bilinear mapping groups introduced as above. Each customer will do the following steps:

1. Customer selects $g_1 \in G_1$, $g_2 \in G_2$, two generators;
2. Customer selects bilinear mapping $e(.,.)$ as above;
3. Customer randomly selects $x \in Z^*_p$ and computes $v = g^x_2 \in G_2$; and
4. Customer selects two securely cryptographic hash functions H_1 and H_2 : $H_1 : \{0,1\}^* \times \{0,1\}^* \mapsto Z_p$, $H_2 : Z_p \mapsto Z_p$.

In addition, there are another two secure cryptographic hash functions H_3 and H_4, such as SHA-1 (Stinson, 1995, p. 248, pp. 251-253); where $H_3 : \{0,1\}^* \times Z_p^* \mapsto Z_p$ and $H_4 : \{0,1\}^* \times \{0,1\}^* \times Z_p \mapsto Z_p$.

Therefore, the private key of the customer is x; the public key is g_1, g_2, $e(.,.)$, H_1, H_2, H_3, and H_4. All these public parameters are also known to the servers. Note that H_3 and H_4 will be used for the authentication between the underlying server and the mobile agent (de facto, the customer) in subsection 4. D. Another point is that all the public keys are certificated by the key certificate authority, in order to maintain the integration and non-repudiation.

Since we are constructing a transactions protocol, we should specify some corresponding information about the customer and the server. For example, who is the buyer? Who is the bidder (de facto seller). That is, what is the corresponding information of the customer and the server? Here, the server represents the host computer that the mobile agents will visit in the transactions.

For permanent usability, we let C be a permanent identifier for the customer, and let S be a permanent identifier of the server. For a specific purchase (i.e., e-transactions), we define *TSI* as the temporary bidder identifier. In fact, *TSI* is derived from S and the corresponding purchase/selling information (for example, valid period for this bid). We also define *TBI* as the temporary buyer identifier (this may represent the mobile agent). At the same time, t is a time-stamp generated by the underlying server. R is a random element generated by the key certificate authority. R_1 is a random element generated by the customer. These items will be used for the authentication before the mobile agent takes the transactions with the underlying server.

In addition, we denote the constraints of the customer by Req_c, and the bid of the server by Bid_S. The two items are defined as follows:

Req_c defines the requirements of the customer for a specific purchase. It includes: (1)the description of a desired product; (2) an expiration date and time stamp; (3) the maximum price that is acceptable to the customer; and (4) a deadline for the delivery of the product.

Bid_S defines the bid of the server for a selling activity. It includes: (1) the description of the server's product; (2) the minimum price that will be acceptable to the server; (3) a deadline for the delivery of the product; (4) a deadline for paying money into the bank account of the server; and (5) an expiration date and time stamp.

Key Algorithm

The *key algorithm* is a probabilistic polynomial time algorithm. The key certificate authority, the customer, and each server will collaborate to assign some keys. All the keys specified here have two fundamental functions: Some of them are used to maintain the privacy of the underlying participants; the others are used to maintain the authentication between the customer and the underlying server.

1. The key certificate authority determines a practical public key encryption algorithm $E_{pub \otimes prv}$ for the customer and the underlying server. Note that the customer and the underlying server cannot agree on $E_{pub \otimes prv}$ by themselves, since they need mutual authentication in the underlying transactions protocol.
2. The key certificate authority generates a random secret element $k \in Z_p^*$ for the underlying server. Therefore, the key certificate authority and the underlying server will share this element. This key will be used for the authentication before the mobile agent migrates to the server.

3. The key certificate authority sends a pair of public key pub_c and private key prv_C to the customer securely.
4. The key certificate authority sends a pair of public key pub_S and private key prv_S to the underlying server securely.

All these public keys and private keys will be involved when the customer initiates the e-transaction with the server. The public key encryption algorithm can maintain the private communications between the customer and the server.

Mobile Agents Preparing

This algorithm is used to equip the mobile agent with executable codes. The customer constructs the executable codes by using his private key. However, the private key will be presented as a blinded version, since the mobile agent will migrate with the executable does to the underlying server. This will not leak any useful information about the private key. The customer equips the Mobile Agent with executable codes. The executable codes are, in fact, an undetachable signature function pair:

$y() = ()\text{-}x_1 \pmod q$
and $y_{signed}()=x_2 \times g^{H_2(()\text{-}x_1)}$.

where $x_1 = H_1(TBI, \mathrm{Re}q_C)$ is bounded by q;

$x_2 = g_1^{\frac{x_1}{x}} \in G_1$, where the exponentiation is computed modular q. x_2 can be seen as a variant version of the short signature:

$x_1=H_1(TBI, \mathrm{Re}q_C) \pmod q$
$x_2 = g_1^{\frac{x}{x}} \in G_1$

where C is a message, $\mathrm{Re}q_c$ is a random element. Therefore, x_1 and x_2 could be treated as the signature (Han & Chang, 2006): $\sigma=h(m, r)^{\frac{1}{x}}$ on the message m; where $h(m, r)=g_1^{x_1}$. The security is based on an assumption of q-SDH (Han & Chang, 2006).

Equipped with the executable codes, the mobile agent will migrate from the customer to the server. This agent will carry *TBI* and $\mathrm{Re}q_c$ as part of its data. Also, the mobile agent can sign any purchase (restricted by the purchase requirement) on behalf of the customer. Therefore, this algorithm realises the delegation of signing rights from the customer to the mobile agent.

Before the mobile agent migrates to the underlying server, the customer and the server need to authenticate with each other. Note that this process is actually executed between the mobile agent and the server. Here, the mobile agent, in fact, represents its owner, that is, the customer. However, for the simplicity of the deployment, we arrange the customer and the server to take the process of authentication. The process will assure that: (1) the customer is the purchaser (de facto, the buyer); and (2) the underlying server is truly the bidder (de facto, the seller).

Mutual Authentication

If the mutual authentication is successful, this algorithm can verify that: (a) *the underlying server is committed to the customer for the coming transaction;* and (b) *the customer is committed to the server for the coming transaction*. The customer, the underlying server, and the key certificate authority will attend this algorithm. The following are the details:

1. In order to win the indent (transaction order), the server promulgates her selling information and sends an authentication request to the customer. This request includes *TSI* and *t*.
2. The customer sends *TSI* along with *t* and his permanent identifier *C* to the key certificate authority through a secure channel.
3. After the key certificate authority receives the information, it first checks whether the

permanent identifier C is legal and t is valid. If one of them is not valid, the process stops here; otherwise, the key certificate authority computes $AU_C=H_3(t,k)$ and $K_t=H_3(R,k)$. Then, the key certificate authority sends the tuple$\{AU_C, R, K_t\}$ to the customer through a secure channel.

4. Once the customer receives the tuple from the key certificate authority, he computes $AU_S=H_4(t, R, K_t)$ and stores it in his database. Then the customer sends the tuple $\{AU_C, R, t\}$ to the underlying server.
5. Once the server gets the tuple, she first checks whether t is valid. If t is valid, the process stops here; otherwise, the server calculates $AU^*_C=H_3(t,k)$ and compares it with the received AU_C. If they are not equal, the transaction is terminated here; otherwise, the customer is authenticated. The server then computes $K_t^*=H_3(R,k)$ and $AU^*_S=H_4(R, t, K_t^*)$, and sends both of them to the customer.
6. After receiving AU^*_S and K_t^*, the customer compares AU^*_S with the one AU_S stored in this database. If the following two equations hold:

$$K_t^* = K_t$$
$$AU_S^* = AU_S.$$

Then, the server is authenticated successfully. After the mutual authentication is completed successfully, the customer will arrange the mobile agent to take the transaction with the underlying server.

Mobile Agent Execution

After the mobile agent arrives at the server, the agent will give all its data and the executable code to the server. The server will execute the executable code provided by the mobile agent, that is, $y(\ )$ and $y_{signed}(\ )$. The details are as follows:

1. The server computes $y_1= H_1(TBI, TSI, Bid_S)$ with a bid.
2. The server computes $r=y(x)=y_1-x_1 \pmod q$. If $r\equiv 0$, the server will stop.
3. The server computes:

$$\begin{aligned} y_2 &= y_{signed}(y_1) \\ &= x_2 \times g^{H_2(y_1-x_1)} \\ &= g_1^{\frac{x_1}{x}} \times (g_1^x)^{H_2(y_1-x_1)} \\ &= g_1^{\left(\frac{x_1}{x}+xH_2(y_1-x_1)\right) \pmod q} \in G_1 \end{aligned}$$

where $g=g_1^x \in G_1$.

4. The server outputs the x-coordinate x_3 of y_2, where x_3 is an element in Z_q.
5. The server hands the mobile agent a tuple $TBI, TSI, Bid_S, y_1, m, x_3$; This tuple will represent part of the transaction.
6. The mobile agent with the tuple migrates to its owner, that is, the customer.

Remark: In the above algorithm, we let TBI and TSI involve the computation of the transaction. This will help to protect the permanent identity of the customer as well as the underlying server. This principle is reasonable, since the temporary identifier is privately linked to the permanent identifier.

Transaction Verifying

This algorithm is used to verify that the fulfilled transaction is an optimal one. If it is, the customer will accept the transaction. The details are as follows: When the mobile agent returns from the server, the customer will check the returned data provided by the mobile agent. The customer will need to follow these steps:

1. The customer will check the undetachable signature (r, x_3) for this transaction by utilising the following formula.

2. The customer will find whether there is a point in G_1: $g_3 = (x_3, y_3)$ (where t is an element in Z_p) such that the following equation holds in G_1:

$$e\left(g_3, v^{H_2(r)}\right) = e\left(g_1, g_2\right)^{(x_1 + x^2 H_2(r)) H_2(r)}$$

If there is no such point, then the customer will not accept this transaction. Otherwise, she will accept this transaction.

That is to say: If the above equality holds, that certifies that the transaction is valid. Then the customer will accept the transaction. Otherwise, the customer will arrange the current mobile agent or another mobile agent to migrate to another server to seek a desirable bid and accomplish the transaction.

ANALYSIS OF THE TRANSACTIONS PROTOCOL

In this section, we will analyse the proposed protocol of transactions with mobile agents and provide authentication analysis, security proof, and privacy analysis. We first provide the authentication analysis - mutual authentication analysis. We will show how the customer is committed to the server, and how the server is committed to the customer. Construction analysis tells how the protocol works, what the principal of the protocol is, and how the mobile agents help the transactions. Security proof shows how to extract the signature scheme from transactions. Subsequently, we will analyse how the privacy is preserved for both the customer and the server.

Authentication Analysis

In this subsection, we will analyse the authentication mechanism. As previously described, we know that the temporary identifier TSI is derived from the permanent identifier S, and the temporary identifier TBI is derived from the permanent identifier C. Therefore, TSI is linked to S, and TBI is linked to C.

1. Mutual authentication fulfils the mutual commitment between the buyer and the seller, that is, the customer and the underlying server. If the seller is committed to the buyer, and the buyer is not committed to the seller, then it will be probable that the buyer would not accept or confirm the transactions. This will result in the buyer presenting no responsibility for the underlying purchase. On the other hand, only the temporary identifiers of the customer and the server are involved in the mobile agent preparing algorithm and the mobile agent execution algorithm. It is known that the temporary identifier has a specific and short-term valid period. Therefore, it is necessary to accomplish the mutual authentication.
2. The mechanism of authentication is as follows: On the one hand, the underlying server is authenticated through: (1) the customer checks whether $K_t^* = K_t$; and (2) the customer checks whether $AU_S^* = AU_S$. In fact, $K_t^* = K_t$ reflects that the underlying server has a shared private element with the key certificate authority. $AU_S^* = AU_S$ implies that the server constructs AU_S^* using the elements from the customer. On the other hand, the customer is authenticated through: (1) the server checks whether the timestamp t is valid; and (2) the server checks whether $AU_C^* = AU_C$. In fact, that t is valid implies that the customer really receives t and replies correctly. $AU_C^* = AU_C$ implies that the customer attains the element AU_C from the key certificate authority, since only the server and the key certificate authority can compute the value of AU_C^* and AU_C.
3. The proposed transaction protocol indicates that the authentication process cannot be forged by the server as well as the customer. Consequently, anyone else (excluding the

customer, server, and the key certificate authority) cannot forge the authentication. The security of the mutual authentication is based on the property of cryptographic hash functions (Stinson, 1995).

Construction Analysis

We will deploy the proposed transaction protocol from the construction point of view. This will help us to further understand the transaction protocol.

Note that a key certificate authority is involved in the following three algorithms: setup algorithm, key algorithm, and authentication algorithm. Therefore, the function of the key certificate authority can be deployed according to the three aspects:

a. The key certificate authority certifies the public keys (signing algorithm) for the customer in the setup algorithm.
b. The key certificate authority determines the public key algorithm for the customer and the underlying server in the key algorithm. This assures the private communications between the customer and the server.
c. The key certificate authority helps to accomplish the mutual authentication between the customer and the server. This is realised through: (1) a shared private key between the key certificate authority and the underlying server; (2) computing $AU_C = H_3(t,k)$ and $K_t = H_3(R,k)$; and (3) confirming the legality of the permanent identifier of the customer. All these three are presented in the authentication algorithm.

Next, we deploy the proposed transaction protocol from the delegation-of-signing-right point of view.

In the transaction protocol, the mobile agent is awarded a pair of functions ($y(\ )$ and $y_{signed}(\ )$)and migrates with them to the server. This pair of functions maintains the un-leakage of the signing algorithm (actually the signing private key) of the customer. The input x of the server is linked to the server's bid. At the same time, the mobile agent is also given the certified requirements of the customer (a, b), satisfying

$y(\ ) = (\ ) - x^1 \pmod q$,
and $y_{signed}(\ ) = x^2 \times g^{H_2((\ )-x_1)}$ *in* G_1.

The parameters of function $y(\ )$ are such that the output of this function includes the customer's constraints. The server modifies these by including the bid, Bid_S in the input y_1, in such a way as to satisfy:

- The message m links the constraints of the customer to the bid of the server; and
- It gets an undetachable signature (r, x_3) for the transaction, where $r = (y_1 - x_1) \pmod{}$ and x_3 is the x-coordinate of the point *beta*. This serves as a certificate which is authenticated by the customer as follows

$$e\left(g_3, v^{H_2(r)}\right) = e\left(g_1, g_2\right)^{(x_1 + x^2 H_2(r)) H_2(r)}.$$

The certified constraints of the customer Req_C, and the bid of the server, Bid_S restrict the scope of *the context* of *the transaction*, that is, the certificate (r, x_3) to "optimal bid" transactions with the appropriate time-limits (or more generally, to whatever requirements the customer and the server stipulate).

Note that even if a server ignores the customer's constraints Req_C and executes the mobile agent associated with the executable code ($y(\ )$ and $y_{signed}(\ )$) in order to produce an undetachable signature of the customer for a bogus bid, the signature will be invalid. If a server is not willing to bid for a purchase, then the mobile agent will travel to another server to obtain an optimal bid for the transaction.

Privacy Analysis

Privacy is the most concern in respect to financial issues of the participants in the transactions (Eklund, 2006). Therefore, besides the security analysis, it is also necessary to analyse the privacy of the proposed protocol. We will analyse the privacy of the transaction protocol from the following four aspects:

1. Privacy of the signing key of the customer: This privacy is maintained by the mobile agent's executable code, that is, the pair of functions ($y(\,)$ and $y_{signed}(\,)$) since the signing key is implied and embedded in the content of $y_{signed}(\,)$.
2. Privacy of the identity of the customer: This privacy is maintained through the encrypted communication. In fact, when the customer sends the mobile agent to some servers to seek "optimal purchase", she will encrypt the whole or part of the tuple ($y(\,)$, $y_{signed}(\,)$, *TBI*, Req_c) (if necessary for the whole content), by utilising her private key prv_C of the underlying public key encryption.
3. Privacy of the context of the transaction initiated between the customer and a server: This privacy is maintained through the mutual encrypted communications between the customer and the server, who will utilise the public key encryption algorithm established in the setup algorithm of the e-transaction protocol.
4. Privacy of the identity of the underlying server: This privacy is maintained through the fact that when the server hands the tuple *TBI, TSI,* Bid_S, Y_1, r, x_3 to the mobile agent to migrate to the customer, the server will encrypt the part of the tuple in which is related to its identity information, by utilising her private key prv_S of the underlying public key encryption.

CONCLUSION

In this article, we have defined a model of mutual authenticated transactions with mobile agents. Then, a new electronic transaction protocol is presented according to the proposed model. We have provided the corresponding analysis and proofs for the proposed transaction protocol. In detail, there are authentication analysis, construction analysis, and privacy proof. For authentication analysis, we deploy it from the mechanism of mutual authentication, as well as the unforgeable authentication elements. The construction analysis helps to better understand the principles of the proposed protocol. For privacy analysis, it is shown that the privacy is maintained through the involvement of the public key algorithm, an undetachable signature function, and the temporary identifiers.

REFERENCES

Boneh, D. & Boyen, X. (2004). Short signatures without random oracles.In P*roceedings of Eurocrypt 2004, Lecture Notes in Computer Science 3027*, pp. 56-73, Springer Press.

Claessens, J., Preneel, B. & Vandewalle. J. (2003). How can mobile agents do secure electronic transactions on untrusted hosts? *ACM Transactions on Internet Technology 3(1)*, 28-48.

Coppersmith, D., Stern, J. & Vaudenay, S. (1993). Attacks on the birational permutation signature schemes.In P*roceedings of CRYPTO 1993 Lecture Notes in Computer Science 773*, 435-443.

Courtois, N. (2004) *Short signatures, provable security, generic attacks and computational security of multivariate polynomial schemes such as HFE*, Quartz and Sflash. Cryptology Eprint 2004/143.

Digital Signature Algorithm (DSA), RSA (as specified in ANSI X9.31), and Elliptic Curve DSA (ECDSA; as specified in ANSI X9.62), *Federal Information Processing Standard,* 186-2.

Edjlali, G., Acharya, A. & Chaudhary, V. (1998). History-based access control for mobile code. In *Proceedings of ACM Conference in Computer and Communication Security 1998,* 38-48.

Eklund, E. (2006). *Controlling and securing personal privacy and anonymity in the information society.* http://www.niksula.cs.hut.fi/~eklund/Opinnot/netsec.html

Han, S., Chang, E. & Dillon, T., (2005a). Secure e-transactions using mobile agents with agent broker. In *Proceedings of the Second IEEE Conference on Service Systems and Service Management, 2,* 849-855, Jun.13-15, Chongqing University, China.

Han, S., Chang, E. & Dillon, T. (2005b). Secure transactions using mobile agents with TTP. In *Proceedings of the Second IEEE Conference on Service Systems and Service Management, 2,* 856-862, Jun. 13-15, Chongqing University.

Han, S. & Chang, E. (2006). *New efficient undetachable sigantures.* Technical Report, IS-CBS2006, School of Information Systems, Curtin Business School, Curtin University of Technology.

Kolaczek, G. (2003). Specification and verification of constraints in role based access control for enterprise security system. In *Proceedings of the 12th IEEE International Workshops on Enabling Technologies,* Infrastructure for Collaborative Enterprises, 9-11 June 2003, Linz, Austria, IEEE Computer Society 2003: 190-195.

Kotzanikolaou, P., Burmester, M. & Chrissikopoulos, V. (2000). Secure transactions with mobile agents in hostile environments. In *Proceedings of the Fifth Australasian Conference on Information Security and Privacy, 10(12) July 2000,* pp. 289-297, Lecture Notes in Computer Science 1841.

Lauter, K. (2004). The advantages of elliptic curve cryptography for wireless security. *IEEE Wireless Communications Magazine*.

Rivest, R.L., Shamir, A. & Adleman, L.M. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM 21(2),* 120-126.

Sander, T. & Tschudin, C.F. (1998). Protecting mobile agents against malicious hosts. *Mobile Agents and Security 1998*, Lecture Notes in Computer Science 1419, 44-60.

Shamir, A. (1984). Efficient signature schemes based on birational permutations. *Advances in Cryptology - CRYPTO 93*, Lecture Notes in Computer Science 773, 1-12.

Stinson, D.R. (1995) *Cryptography: practice and theory.* CRC Press, Boca Raton.

Patarin, J., Courtois, N. & Goubin, L. (2001). QUARTZ, 128-bit long digital signatures. In *Proceedings of Topics in Cryptology* - The Cryptographer's Track at RSA Conference 2001, San Francisco, CA, USA, April 8-12, 2001, Lecture Notes in Computer Science 2020, 282-297.

This work was previously published in International Journal of Information Security and Privacy, Vol. 1, Issue 1 , edited by H. Nemati, pp. 35-46, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.16
Robust Algorithms for DOA Estimation and Adaptive Beamforming in Wireless Mobile Communications

R.M. Shubair
Etisalat University College, UAE

K.O. AlMidfa
Etisalat University College, UAE

A. Al-Marri
Etisalat University College, UAE

M. Al-Nuaimi
Etisalat University College, UAE

ABSTRACT

This paper presents a tool for the modelling, analysis and simulation of direction-of-arrival (DOA) estimation and adaptive beamforming needed in the design of smart antenna arrays for wireless mobile communications. The developed tool implements the Minimum Variance Distortionless Response (MVDR) algorithm for DOA estimation and the Least Mean Squares (LMS) algorithm for adaptive beamforming. Performance of each algorithm is investigated with respect to the variation of a number of parameters that related to the signal environment and sensor array. Results of numerical simulation are useful for the design of smart antennas systems with optimal performance. Hence, the developed simulation tool can be used to improve and accelerate the design of wireless networks. It can also be used for computer-aided learning of modern communication systems utilizing smart antenna arrays.

INTRODUCTION

The demand for mobile communication resources has increased phenomenally over the past few years. Adaptive, or smart, antenna techniques have emerged as a key way to achieve the ambitious requirements introduced for current and future mobile systems. Examples of these requirements are the ability to support broadband data, maximize spectral efficiency and support new location-based services (AlMidfa, 2000,2003).

A smart antenna is commonly defined as a multi-element antenna where the signals received at each element are intelligently and adaptively combined to improve the overall performance of the wireless system, with the reverse performed on transmit. The benefit of smart antennas is that they can increase range and capacity of systems while helping to eliminate both interference and fading.

Most wireless systems such as Wi-Fi, WiMAX, UWB and GPS can benefit from the addition of an adaptive array antenna. Since Wi-Fi systems are time-division-duplex (TDD) systems, the received weights can be used for transmission to obtain the same gains in both directions with the use of smart antennas on one side only. As an example, Winters (1998) showed that a four-antenna array can provide up to a 13 dB signal-to-noise ratio gain (6 dB array gain plus a 7 dB diversity gain), or the possibility of data rates as high as 500 Mbps (as considered for IEEE 802.11n). Similar gains can be achieved in WiMAX systems (particularly those using TDD) (Winters, 1998).

One way in which adaptive antennas can be exploited is by Direction Finding (DF), where algorithms are used to estimate the direction-of-arrival (DOA) of the received signals at the Base Station (BS). These are used to improve the performance of an antenna array by controlling its directivity to reduce effects such as interference, delay spread and multipath fading (Godara, 2003).

In addition to estimating the directions of the signals from the desired mobile users, adaptive antennas are used to estimate directions of interference signals. The results of DOA estimation are then used to adjust the weights of the adaptive beamformer so that the radiated power is maximized towards the desired users, and radiation nulls are placed in the directions of interference signals (Shubair, 2004; 2005). Hence, a successful design of an adaptive array depends highly on the choice of the DOA estimation algorithm, which should be highly accurate and robust.

This paper investigates the Minimum Variance Distortionless Response (MVDR) algorithm for DOA estimation and the Least Mean Squares (LMS) algorithm for adaptive beamforming. The theory of each algorithm is developed using a realistic model for the signal environment surrounding the sensor array. The performance of each algorithm is then analyzed using a simulation tool developed along with a graphical user interface (GUI). This includes investigating the effect of parameters related to the signal environment such as the number of incident signals and their angular separation. It also investigates effect of parameters related to the design of the sensor array itself including number of array elements and their spacing.

DOA ESTIMATION USING MVDR ALGORITHM

The MVDR algorithm involves estimating the noise subspace from the covariance matrix on which the M array steering vectors are projected. These steering vectors are also known as direction vectors and they represent the response of an ideal array to the signal sources. The signal sources can be derived from the direction vectors which are orthogonal to the noise subspace (Al-Ardi, 2003; 2004; 2005).

The algorithm starts by constructing a real-life signal model. Consider a number of plane waves from M narrow-band sources impinging from different angles θ_i, $i = 1, 2, \ldots, M$, into a uniform

linear array (ULA) of *N* equi-spaced sensors, as shown in Figure 1.

At a particular instant of time *t*, *t*=1,2, …, *K*, where *K* is the total number of snapshots taken, the array output will consist of the signal plus noise components. The signal vector **x**(*t*) can be defined as (Liberti, 1999):

$$\mathbf{x}(t) = \sum_{m=1}^{M} \mathbf{a}(\theta_m) \cdot \mathbf{s}_m(t) \qquad (1)$$

where **s**(*t*) is an *M*×1 vector of source waveforms, and for a particular source at direction θ from the array boresight; **a**(θ) is an *N*×1 vector referred to as the array response to that source or array steering vector for that direction. It is given by:

$$\mathbf{a}(\theta) = [1 \quad e^{-j\phi} \quad ..... \quad e^{-j(N-1)\phi}]^T \qquad (2)$$

where *T* is the transposition operator, and ϕ represents the electrical phase shift from element to element along the array. This can be defined by:

$$\phi = (2\pi/\lambda) d \cos\theta \qquad (3)$$

where *d* is the element spacing and λ is the wavelength of the received signal.

The signal vector **x**(*t*) of size *N*×1 can be written as:

$$\mathbf{x}(t) = \mathbf{A} \cdot \mathbf{s}(t) \qquad (4)$$

where $\mathbf{A}=[\mathbf{a}(\theta_1) \ \ldots \ \mathbf{a}(\theta_M)]$ is an *N*×*M* matrix of steering vectors.

The array output consists of the signal plus noise components, and it can be defined as:

$$\mathbf{u}(t) = \mathbf{x}(t) + \mathbf{w}(t) \qquad (5)$$

where **x**(*t*) and **w**(*t*) are assumed to be uncorrelated and **w**(*t*) is modelled as temporally white and zero-mean complex Gaussian process. Equation 5 can be written in matrix form of size *N*×*K* as:

$$\mathbf{U} = \mathbf{A} \cdot \mathbf{S} + \mathbf{W} \qquad (6)$$

where $\mathbf{S}=[\mathbf{s}(1) \ldots \mathbf{s}(K)]$ is an *M*×*K* matrix of source waveforms and $\mathbf{W}=[\mathbf{w}(1) \ \ldots \ \mathbf{w}(K)]$ is an *N*×*K* matrix of sensor noise. The spatial covariance matrix **R** of the observed signal vector **u**(*t*) can be defined as:

$$\mathbf{R} = E[\mathbf{u}(t) \cdot \mathbf{u}(t)^H] \qquad (7)$$

where *E*[ ] and *H* are the expectation and conjugate transpose operators, respectively. Substituting (5) into (7), the spatial covariance matrix **R** can now be expressed as:

$$\mathbf{R} = E[\mathbf{A} \cdot \mathbf{s}(t) \cdot \mathbf{s}(t)^H \cdot \mathbf{A}^H] + E[\mathbf{w}(t) \cdot \mathbf{w}(t)^H] \qquad (8)$$

For this signal model, the covariance matrix **R** will have *M* signal eigenvalues, and *N*-*M* noise eigenvalues. Let $\mathbf{E}_s$ be the matrix constructed of the corresponding *M* signal eigenvectors $\mathbf{E}_s=[e_1 \ e_2 \ \ldots \ e_M]$, and $\mathbf{E}_n$ be the matrix containing the remaining *N*-*M* noise eigenvectors $\mathbf{E}_n=[e_{M+1} \ e_{M+2} \ \ldots \ e_N]$. The peaks in the MVDR angular spectrum occur whenever the steering vector *E*(ϕ) is orthogonal to the noise subspace.This technique minimizes the contribution of the undesired interferences by minimizing the output power while maintaining the gain along the look direction to be constant, usually unity. That is,

$$\min E[|y(\theta)|^2] = \min \mathbf{w}^H R_{uu} \mathbf{w}, \ \mathbf{w}^H A(\theta_0) = 1 \qquad (9)$$

Using Lagrange multiplier, the weight vector that solves equation (1) can be shown to be:

$$w = \frac{\mathbf{R}_{uu}^{-1} A(\theta)}{A^H(\theta) \mathbf{R}_{uu}^{-1} A(\theta)} \qquad (10)$$

Now the output power of the array as a function

of the DOA estimation, using MVDR beamforming method (Haykin, 2002), is given by MVDR spatial spectrum as:

$$P_{MVDR}(\theta) = \frac{1}{A^H(\theta)\mathbf{R}_{uu}^{-1}A(\theta)} \quad (11)$$

The angles of arrival are estimated by detecting the peaks in this angular spectrum.

GRAPHICAL USER INTERFACE FOR DOA ESTIMATION

The MVDR algorithm has been implemented using MATLAB version 6.5. A graphical user interface (GUI) shown in Figure 2 has been developed to ease the simulation. It was originally developed to utilize MUSIC algorithm (Belhoul, 2003) and is adopted here using MVDR. The user can input the signal parameters including the number of snapshots *K*, the number of mobile users *M* and their angle(s) of arrival θ_i. This information is used to generate a realistic signal model. As for the sensor array, the user may input the number of array elements *N* and their spacing *d*. Default values for these parameters can be retrieved at any time by pressing the "Default" button. The simulation can be started by clicking on the "Run" button. This will produce a plot of the MVDR angular spectrum. This plot can be saved to a file by pressing the "Save" button. Alternatively, the plot may be sent to a printer by pressing the "Print" button.

DOA ESTIMATION PERFORMANCE STUDY

To demonstrate the versatility and accuracy of the developed tool, it is used to study the effect of changing a number of parameters related to the signal environment as well as the sensor array.

Figure 3 shows the MVDR angular spectrum generated due to signals arriving from M=1 user and M=7 users. When there is only one mobile user (M=1) present in the vicinity of the base station, the MVDR algorithm performs better since it produces an angular spectrum with a sharp peak and a lower noise floor. The performance of the algorithm degrades when there are many mobile users because the spatial correlation between the incoming signals makes it difficult for the algorithm to resolve them successfully.

When the mobile users are close to each other, there will be a strong correlation between the signals. In this case the MVDR algorithm finds it difficult to resolve the users. This is illustrated in Figure 4(a) for three adjacent users. However, the performance improves significantly as the users move away from each other, as shown in Figure 4(b) for which the MVDR angular spectrum has sharper peaks and lower noise floor.

Figures 5(a) and 5(b) show the MVDR angular spectrum using an eight-element array (N=8) and three-element array (N=3), respectively. It is evident that using more elements improves the resolution of the algorithm in detecting the incoming signals. This is achieved, however, at the expense of computational efficiency and hardware complexity of the sensor array.

Figures 6(a) and 6(b) show the MVDR spectrum for an element spacing of d=0.25λ and d=0.5λ, respectively. When the elements of the sensor array are placed too close to each other, mutual coupling effects dominate, resulting in inaccuracies in the estimated angles of arrival, as shown in Figure 6(a) for which d=0.25λ. Mutual coupling effects for closely spaced elements must, therefore, be taken into account when designing the sensor array. To overcome this problem, the spacing between the elements of the sensor array must be increased resulting in a better resolution of the estimated peaks, as shown in Figure 6(b) for which d=0.5λ.

ADAPTIVE BEAMFORMING USING LMS ALGORITHM

The signal at the output of the sensor array in Figure 1 at time *n*, *y*(*n*), is given by a linear combination of the data at the *N* sensors at time *n* (Haykin, 2002):

$$y(n) = \sum_{k=1}^{N} w_k^* x_k(n), \quad (12)$$

where * represents the complex conjugate, $x_k(n)$ is the complex envelope representation of the received signal from the *k*-th antenna element at time *n* and is the complex weight applied to $x_k(n)$. Equation (12) can be represented in vector form as:

$$y(n) = \mathbf{w}^H \mathbf{x}(n), \quad (13)$$

where ^{H} denotes the Hermitian operator and:

$$\mathbf{w} = [w_1, w_2, ..., w_N]^T \quad (14)$$

is the complex weight vector (Rong, 1997).

The LMS algorithm is based on the steepest-descent method, which recursively computes and updates the weight. It is intuitively reasonable that successive corrections to the weight vector in the direction of the negative of the gradient vector should eventually lead to the MMSE, at which point the weight vector assumes its optimum value (Tran, 2003). In the LMS algorithm, the weights are updated by the following equation:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu \mathbf{x}(n) e^*(n), \quad (15)$$

where $\mathbf{w}(n+1)$ denotes the weights to be computed at iteration $n+1$ and μ is a positive constant that controls how fast and how close the estimated weights approach the optimal solution that minimizes the error $e(n)$ which is given by:

$$e(n) = d(n) - y(n) \quad (16)$$

where $d(n)$ is the reference signal and $y(n)$ is the output signal. The value of μ that shows stability and convergence of the algorithm should not exceed the following limit:

$$0 < \mu < \lambda_{\max}^{-1} \quad (17)$$

where $\lambda_{\max}$ is the maximum eigenvalue of the covariance matrix **R** given in Equation (8). The block diagram of LMS adaptive beamforming algorithm is shown in Figure 7.

GRAPHICAL USER INTERFACE FOR ADAPTIVE BEAMFORMING

An interactive simulation tool with a graphical user interface (GUI) has also been developed to implement the LMS adaptive beamfoming algorithm. The developed GUI is shown in Figure 8. The user may input parameters related to the signal environment as well as the sensor array. The developed simulation tool has been used to study the performance of the LMS adaptive beamforming algorithm as discussed below.

ADAPTIVE BEAMFORMING PERFORMANCE STUDY

Figure 9 shows the SMI beampattern of only one incident signal and another beampattern of three incident signals. With only one incident signal $M=1$ (solid line), the sensor array can successfully form a beampattern with high performance. By increasing the number of incident signals to $M=3$ (dashed line), the performance is degraded and the correlation between signals is increased. The increased number of incident signals also results in an increase in the computational overhead. This is because more sensor array elements are needed to detect the increased number of incidents signals.

Figure 10 shows that the LMS beamformer can form a sharper beam toward the desired signal

incident at an angle of 20^o if the number of elements N of the beamforming array is increased from N=4 (dashed line) to N=6 (solid line).

Figure 11 shows the effect of increasing the angular separation on the LMS beamformer. The LMS algorithm gives sharper main beams, reduces the correlation between wanted signals and avoids deviation from the exact value of the desired signals as the angular separation between two incident signals is increased from 20^o to 60^o.

Figure 12 shows the LMS beampattern for an element spacing d=0.2λ and also when the elements are placed farther away from each other with a spacing of d=0.5λ between adjacent elements. LMS beamformer can form a better beam pattern if the number of element spacings is increased from d=0.2λ to d=0.5λ. The best performance can be achieved when element spacing d=0.5λ. The element spacing cannot be greater than 0.5λ to avoid spatial aliasing. However, the element spacing cannot be made arbitrarily small since two closely spaced antenna elements will exhibit mutual coupling effects.

Figure 13 shows the LMS beampattern for the step size μ=0.001 and also when the step size μ=0.01. Increasing the step size from 0.001 to 0.01 increases misadjustment noise and causes the algorithm to converge faster. The step size is selected such that $0 < \mu < \lambda_{\max}^{-1}$ as given in Equation (17).

CONCLUSION

A versatile simulation tool that implements both the MVDR DOA estimation algorithm and LMS adaptive beamforming algorithm was developed together with a user-friendly GUI. A number of numerical experiments were conducted to investigate the effect of various parameters on the performance of the MVDR DOA estimation algorithm and its ability to resolve incoming signals accurately and efficiently. The performance of the LMS adaptive beamforming algorithm has also been investigated to verify its ability to produce the desired beam pattern. The developed simulation tool can be used to improve and accelerate the design of wireless networks. It can also be used for computer-aided learning of modern communication systems utilizing smart antenna arrays.

REFERENCES

Al-Ardi, E.M., Shubair, R.M., & Al-Mualla, M.E. (2003). Investigation of high-resolution DOA estimation algorithms for optimal performance of smart antenna systems. *Proceedings of IEE Fourth International Conference on Third-Generation Mobile Communication Technologies* London, UK (pp. 310-314).

Al-Ardi, E.M., Shubair, R.M., & Al-Mualla, M.E. (2004). Computationally efficient DOA estimation in a multipath environment. *IEE Electronics Letters, July, 40*(14), 908-909.

Al-Ardi, E.M., Shubair, R.M., & Al-Mualla, M.E. (2005). Computationally efficient DOA estimation in a multipath environment using covariance differencing, & iterative spatial smoothing. *Proceedings of IEEE International Symposium on Circuits, & Systems* (ISCAS), Kobe, Japan, 3805-3808).

AlMidfa, K. (2003). *Direction finding methods: A theoretical, & practical performance analysis including the eEffect of pPolarisation diversity.* Ph.D. Thesis, Centre for Communications Research, University of Bristol, UK.

AlMidfa, K., Tsoulos, G.V., & Nix, A. (2000, May). Performance evaluation of direction-of-arrival (DOA) estimation algorithms for mobile communication systems. In *Proceedings of IEEE Vehicular Technology Conference* (VTC). Tokyo, Japan (pp. 200-203).

Belhoul, F.A., Shubair, R.M., & Al-Mualla, M.E. (2003). Modeling and performance analysis of DOA estimation in adaptive signal processing arrays**.** In *Proceedings of IEEE International Symposium on Electronics, Circuits, and Systems* (ICECS), Sharjah, UAE, (pp. 340-343).

Godara, L.C. (2003). Application of antenna arrays to mobile communications, part II: Beamforming, & direction-of-arrival considerations. In *Proceedings of IEEE*, 85 (8),1195-1245.

Haykin, S. (2002). *Adaptive filter theory* (4th ed.). NJ: Prentice Hall.

Liberti, J., & Rappaport, T. (1999). *Smart antennas for wireless communications.* NJ: Prentice Hall.

Rong, Z. (1996). *Simulation of adaptive array algorithms for CDMA systems.* M.Sc. Thesis, Virginia Polytechnic Institute and State University, USA.

Shubair, R.M., & Al-Merri, A. (2004) Robust algorithms for direction finding, & adaptive beamforming: performance and optimization. In *Proceedings of IEEE International Midwest Symposium on Circuits, & Systems* (MWSCAS), Hiroshima, Japan (pp. 589-592).

Shubair, R.M., & Al-Merri, A. (2005, June). Performance of adaptive beamforming arrays for spatial interference rejection. In *Proceedings of International Symposium on Antenna Technology, & Applied Electromagnetics* (ANTEM), Saint-Malo, France (pp. 186-189).

Tran, X.N. (2003). *Subband adaptive array for mobilecommunications with applications to CDMA systems.* Ph.D. Thesis, Electro-Communications University, Japan.

Winters, J.H. (1998). Smart antennas for wireless systems. *IEEE Proceedings on Personal Communications, 5*(1), 23-27.

This work was previously published inInternational Journal of Business Data Communications and Networking, Vol. 2, Issue 4, edited by J. Gutierrez, pp. 34-45, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.17
Mobile Information Filtering

Witold Abramowicz
The Pozañ University of Economics, Poland

Krzysztof Banaśkiewicz
The Pozañ University of Economics, Poland

Karol Wieloch
*The Pozañ University of Eco*nomics, Poland

Paweł Żebrowski
The Pozañ University of Economics, Poland

INTRODUCTION

Information filtering techniques have been continuously developed to meet challenges arisen from new requirements of Information Society. These techniques gain even much more on importance in the facet of greater mobility of people. One of the most dynamic and compelling areas is the environment of wireless and mobile devices. Just recently, information filtering and retrieval have begun to take into consideration circumstances in which they are being used. As information needs of mobile users are highly dynamic, this points out the necessity of considering additional set of attributes describing user situation—context. This article presents an information filtering system for mobile users (mobileIF) being developed in the Department of Management Information Systems at The Pozna University of Economics. Architecture of the mobileIF is a result of research done in the field of contexts, their taxonomies and influence on information relevance in dynamic user's environment. The paper shows our approach to contexts, discusses time perspective on filtering systems and finally, describes mobileIF architecture and basic data flow within it. At last, we present our current research in fields related to the mobileIF system.

BACKGROUND

The process of providing a user with relevant information can be viewed in two different ways. On the one hand, it can be described as the process where single query is performed on a set of

documents (information retrieval—IR). On the other hand, it can be understood as applying a set of queries to a single document (information filtering—IF). Although, the aim of both methods is serving users with relevant documents, the way of processing content in information retrieval and information filtering systems significantly differs from each other (Belkin & Croft, 1992). What is more, queries performed in IR represent short-term information needs, whereas profiles, representing information needs in filtering, stand for relatively constant interests in a particular subject (Baeza-Yates & Ribeiro-Neto, 1999). There are many different applications of IR and IF in various areas, however, majority of them utilizes similar techniques such as Boolean model, vector space, and probabilistic models, as well as some brand new ones, like neural network or Bayesian models. Baeza-Yates et al. (1999) provide exhaustive comparison of those techniques.

In the central point of our interests is information filtering domain, that could be divided itself into several additional subdomains according to methods used. The most important ones are content-based (cognitive) filtering and social (collaborative) filtering. The idea that stands for the content-based filtering is to select the right information (relevant to user) by comparing representations of information being searched to representations of user profiles' contents (Oard & Marchionini, 1996). This method of IF has turned out in many systems to be very effective, especially in dealing with textual objects. The latter one overcomes some limitations of content-based filtering (such as problems with filtering multimedia objects, difficult to use for novices, etc.). The collaborative filtering improves results of IF by taking advantage of judgements of multiple users who have similar interests on the read documents (Shardanand & Maes, 1995). Basis for this technique is the assumption that users who judged the same documents in the similar way to others in the group, will most probably proceed like that in future, while judging new documents.

Both of those methods have many specific advantages as well as some drawbacks. The natural way of evolution is combining these techniques in order to achieve better results of filtering. Claypool, Gokhale, Miranda, Murnikov, Netes, and Sartin (1999) and Li and Kim (2003) proposed some hybrid methods.

There are many definitions of context provided in literature. Among them, several deserve special attention as stimulus to our further considerations. In one of the earliest definitions Schilit (1995) distinguishes the following types of context: computing context (network capacity, connectivity, communication costs, and available devices), user context (user's profile, location, people nearby, and social situation), and physical context (lightning, noise level, temperature, and traffic conditions). According to Schmidt, context is divided into two categories, namely: human factors (information of the user, social environment, and user's tasks) and physical environment (location, infrastructure, and conditions) (Schmidt, Beigl & Gellersen, 1999). Both presented definitions try to identify context by simple division of some characteristics into several groups of potentially distinct attributes. However, neither of them is suitable for inferring more aggregated and complex information. This inconvenience is reduced in the definition by Chen and Kotz (2000) who distinguish low-level and high-level context. The former group contains raw contextual information such as location, temperature etc. (mainly acquired from physical sensors), whereas the latter one is specified on the basis of supplied low-level contexts. More formal definition is provided by Dey and Abowd (1999) who argue "context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."

Systems that take into account context changes and adapt to them (to some degree) are defined as context-aware (Pascoe, 1999). Such adapta-

tion may involve adjustment to user's device capabilities (e.g., screen resolution, memory, software attributes, network bandwidth, and user preferences). Context-aware information delivery system takes into account not only semantic information relevance, but also context of the user. Changes in context may suggest changes in user information needs. Information delivery systems that are based on these assumptions are often defined as context-aware retrieval (CAR) systems (Brown & Jones, 2001). Korkea-aho (2000) provides a wider spectrum of CAR application examples.

MOBILE INFORMATION FILTERING

Contextualization in MobileIF

A citizen of Information Society wants to be provided only with such information he or she requires. In case of filtering domain, user information needs are depicted by user profile that expresses rather long-term goals. On the contrary, active goals and current tasks can be supported by contextual information. It is obvious that the latter ones are more significant for mobile users as they can better adjust to their daily rapidly changing activities.

In order to fulfil this requirement, the system has to be able to process several types of user contexts. The notion of contextual information in mobileIF has to be examined from two different points of view.

In the first one, we distinguish two new groups of contexts that influence:

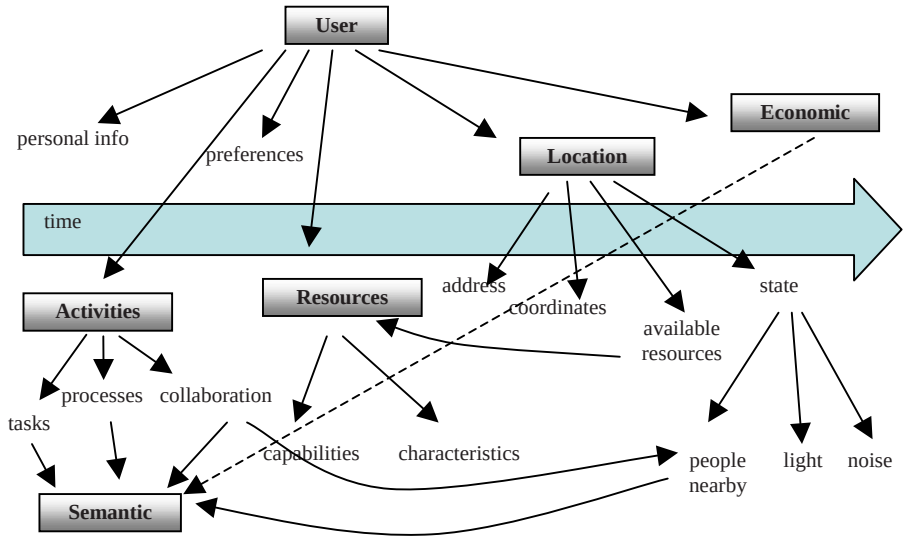
- User information needs (named as semantic contexts). They are used to more precisely define queries to the system obtained out of the user's profile. The context can extend the query with some additional concepts as well as narrow it down (Wen, Lao, & Ma, 2004).
- The way the information is delivered to mobile users (named as distributive context). Such contexts allow an adaptation of the filtering results in order to provide the optimal presentation and delivery. This may be done according to the capabilities of devices or user preferences (Costa, 2003; W3C, 2005).

From the second point of view, both described groups may address several dimensions that are crucial for mobileIF system.

- **Time:** The time context is connected with the occurrence of other contexts.
- **Location:** The location context is intrinsically linked to the geographical context, given by the street-network and other infrastructure, points of interest, environmental and topological features etc.
- **Resources:** They correspond to characteristics and capabilities of utilized resources (e.g., user devices).
- **Social Context:** This interpersonal context gives information about relationships between user and other persons or organizations
- **User Activities:** Describing the user's current tasks, and in a broader sense, the existence of specific conditions or steps in a process.
- **Economic Context:** That handles with relations among various dynamic economic metrics and constraints connected with acquisition of information.
- **Semantic (Cognitive) Context:** Represents current (or future) information needs acquired from user's current tasks, contacts, roles, and preferences. It may be directly derived from other contexts

Most of presented contexts influence each other and should not be treated separately (as shown in Figure 1). It is better to consider user context

Figure 1. Context dependencies



as a point in a multidimensional space with an unbounded number of dimensions. However, the most significant context in our model is time. The special treatment of this context is driven by the nature of mobileIF. As a dynamic system its main characteristic is changeability and continual adaptability to user's future situations.

Attention should be put to duality of time interpretation. On the one hand, it is linked to information and determines its actuality. On the other hand, it is the point of reference for other contextual dimensions. Another crucial assumption of our context model is ability of aggregation of basic contextual information to form more complex contexts. Therefore, it is essential to develop context representation that is interoperable and usable in many cases, independent from the context sources, and also flexible enough to accommodate future needs.

Time Perspective of Information Filtering

An information filtering process consists of several steps (Belkin & Croft, 1992). Although it is not explicitly stated, some of them start in distinct time moments. The most interesting are:

- Publishing texts
- Indexing texts
- Expressing user's information needs
- Comparison of profiles and indexed documents
- Dissemination of retrieved texts to users

Two of them are of a special importance: the time of specifying user needs and the time of delivery. They are related to the semantic context and to the distributive context respectively. A consequence of an existing time gap is that contextual information that influenced retrieval of a particular document is not the same when the document is delivered to the user (and should comply with his information needs). If we want to fully utilize contexts, both moments should be considered and each time moment ought to be treated differently.

Time management systems (TMS) allow the user to plan his future activities, record them in a computer manageable form, and finally trace their execution.

Scheduled jobs enable distinguishing the time of filtering and the time of delivery. A planned task gives two important pieces of information: due time of the task and its description. That information is a source of distributive and semantic context, respectively. On one hand, a task description is an implicit expression of user needs, and on the other one, it instantiates a particular semantic context state in some point in future. When a task reaches its due time the semantic context becomes the current one (the semantic context changes along with time flow). Distributive contextual information can be obtained by analyzing user's resources and time-related properties of scheduled tasks. At each time moment we always have up-to-date state of both components of the context (semantic and distributive).

Architecture of the MobileIF System

In this article we describe prototype architecture of mobileIF system. One of the mobileIF's objectives is to provide the most relevant, quality-oriented information available from well-defined, specific web sources. Developing our system architecture, we assumed possible cooperation with a few predefined web portals. We concentrate on relevance and quality of information that will be presented on different mobile devices with diverse technical capabilities.

To improve information filtering we need to define user profile as precisely as possible. The system extracts the required information from the user's time or project management applications and sends it to proper mobileIF's modules. Raw data is processed and converted into specific parts of user profile. Semantics of an entry and related resources are stored in a form of semantic subprofile and are used later in the filtering process. Context tags determine a distributive profile that helps presenting filtered documents. All filtered documents are stored within an internal repository and are easily accessible within mobileIF system. With help of a GUI, documents are available anytime, anywhere. The mobileIF architecture is presented in Figure 2.

The whole system is divided into four parts:

- User-end module
- User manager module
- Broker module
- Data sources

The User-end module is an interface to mobileIF. We distinguished two main functions of this module: to present results and to extract data defining all parts of a user's profile. When a new entry within a user's TMS application appears for the first time, mobileIF creates a new profile. In this profile we put semantic content extracted from the entry (subject, description, related resources, etc.) as well as preferred time of presenting filtered documents, entry deadline, date of creation, mobile device currently in use.

The User Manager is a central part of the mobileIF system. That is the place, where information is being processed, organized into semantic and distributive profiles, and dispersed among different parts of the system. User data (user's profile) is stored within the Profile Repository. The Profile Repository is a dynamic module in the sense that all changes made by the user within an associated entry are always updated in the module. Furthermore, the Profile Repository is responsible for managing user's profiles linked to tasks marked by the user as done or expired. The next part of the User Manager is the Document Repository, a place where all downloaded documents are stored. The third part of the User Manager is the Profiler. This module is responsible for binding all the filtered documents with appropriate TMS entries and determining, which documents should be presented at a given time, according to distributive profiles.

The Broker module manages filtering processes and collaboration with external sources. The Semantic Profile Repository is a part where copies of all the semantic profiles are stored. The

second part of the Broker is the Document Index that manages creation and updating of full-text documents' indices. The actual filtering process (comparisons) takes place within the Filtering Engine. The results are sent back to the User Manager.

An Exemplary Scenario

A new user starts his cooperation with the mobileIF system by creating a new task (entry) in his TMS application. He fills in a subject, an exact date of meeting, all necessary resources (e.g., people involved), a task category, its priority and marks the whole entry as information need. The user can also provide more details in the task's body.

MobileIF automatically extracts data from this particular task and sends it to the User Manager. There, a new profile connected to the initial task is created. This is the moment of determining all the crucial word entities that will be utilized during the process of information filtering.

Afterwards, the User Manager creates the semantic and distributive subprofiles. The first one is based on entities gathered from textual fields of task's description like subject, body and category. The latter one consists of data extracted from resource fields, mobile device currently in use, and task deadline.

Both the subprofiles are stored in the Profile Repository. The semantic part of the profile is transferred to the Semantic Profile Repository (within the Broker). The Broker keeps the semantic profile as long as an associated entry is valid within user's TMS application. When the entry is removed or its deadline expires, the associated profile is deleted from the Semantic Profile Repository.

In the meantime, the Broker downloads documents from an external web source on a regular basis and stores them in the Document Index. Each new semantic profile that appears in the Semantic Profile Repository is compared against the set of documents indices.

Relevant documents are sent back to the User Manager and are stored in the Document Repository. The User Manager binds them with an appropriate task in the user's TMS. When the user wants to see some relevant documents, Profiler determines the best set of relevant documents with respect to user's distributive subprofile.

FUTURE TRENDS

MobileIF is implemented only partially. There are many several areas that still require further research and testing. One of them is adjusting of graphical form of documents with respect to device capabilities. Many documents are considered relevant because only few parts of them really are interesting. We will try to extract only those relevant excerpts. We expect the task to have more in common with query answering.

Lack of precision in matching document content and user needs is another issue. The use of ontologies, shallow text processing, and automatic building of knowledge bases are approaches that may increase effectiveness of mobileIF system. At first they allow us to smartly expand our queries (from TMS entries) and to index documents.

Another issue we would like to investigate are index structures. Vector based information representation seems to be inappropriate as far as graph-based ontologies will be concerned.

CONCLUSION

Literature brings a great number of definitions of context, however, none of them have been adopted as a standard yet. They were usually developed for the purpose of specific use and for specific systems. They deal mainly with basic and easy detectable contexts like location, temperature, or devices. We present a novel approach to utilization of contextual data for information filtering purposes. There are seven basic contexts we

take into account: time, location, resources, social, activities, economics and semantics. Those contexts can be used twofold. On one hand, they raise content's relevance (according to a semantic profile) and on the other hand they influence the process of a document delivery (according to a distributive profile). Such a partition allows us to use a planning or calendar application as a source of user's context. The architecture we have described addresses contextual issues related to the whole process of information filtering.

REFERENCES

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: Adison-Wesley ACM Press.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and retrieval: Two sides of the same coin? *Communications of the ACM, 35*(12), 29-37.

Brown, P. J., & Jones, G. J. F. (2001). Context-aware retrieval: Exploring a new environment for information retrieval and information filtering. *Personal and Ubiquitous Computing, 5*(4), 253-263.

Chen, G., & Kotz, D. (2000). *A survey of context-aware mobile computing research*. Technical Report: TR2000-381, Hanover, NH: Dartmouth College.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper. *Proceedings of ACM SIGIR Workshop on Recommender Systems*. Berkeley, CA: ACM Press.

Costa, P. D. (2003), *Towards a services platform or context-aware applications*. Dissertation, University of Twente.

Dey, A. K., & Abowd, G. D. (1999). *Toward a better understanding of context and context-awareness.* GVU Technical Report GIT-GVU-99-22, Georgia Institute of Technology.

Korkea-aho, M. (2000). *Context-aware applications survey*. Helsinki University of Technology. Retrieved February 25, 2003, from http://www.hut.fi/mkorkeaa/doc/context-aware.html

Li, Q., & Kim, B. M. (2003). An approach for combining content-based and collaborative filters. *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages* (pp. 17-24). East Stroudsburg, PA: Association for Computational Linguistics.

Oard, D. W., & Marchionini, G. (1996). *A Conceptual framework for text filtering*. Technical Report CAR-TR-830. Human Computer Interaction Laboratory. University of Maryland at College Park.

Pascoe J. (1998). *Adding generic contextual capabilities to wearable computers. Proceedings of the 2nd International Symposium on Wearable Computers* (pp. 92-99). Pittsburgh, PA: IEEE Computer Society

Schilit, W. N. (1995). *A system architecture for context-aware mobile computing.* Dissertation thesis. New York: Columbia University Press.

Schmidt, A., Beigl, A., & Gellersen, H. W. (1999). There is more to context than location. *Computers and Graphics, 23*(6), 839-901.

Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating "Word of Mouth". *Proceedings of CHI'95 Conference on Human Factors in Computing Systems* (pp. 210-217). Denver, CO: ACM Press.

W3C Consortium. (2005). *Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0*. Retrieved March 25, 2005, from http://www.w3.org/TR/CCPP-struct-vocab/

Wen J. R., Lao, N., & Ma, W. Y. (2004). Probabilistic model for contextual retrieval. *Proceed-*

ings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 57-63). ACM Press.

KEY TERMS

Context: All information about current user's situation.

Data Source: An external provider of information; the information is accessible either in passive or active way.

Distributive Profile: A part of a user's profile that defines which documents (from the relevant ones) and how should be presented to him/her in a particular time moment.

Information Filtering System: A system whose goal is to deliver to a user only this information that is relevant to her/his profile; system operates on large streams of unstructured data.

Mobile User: A user who needs an access to unstructured data anytime and anywhere.

Semantic Profile: A part of a user's profile that defines what kinds of information (topics) he/she is interested in.

Time Management System (TMS): Calendar based application that allows user to schedule her/his tasks, monitor they execution and provide various descriptions for them.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 799-804, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.18
Information Management in Mobile Environments Using a Location-Aware Intelligent Agent System

Amrish Vyas
University of Maryland, Baltimore County, USA

Victoria Yoon
University of Maryland, Baltimore County, USA

ABSTRACT

Recent rise in the level of comfort and demand to access various types of information using mobile devices can be attributed to the advancements in wireless as well as Internet technologies. This demand leads us to the new era of mobile computing. Location-based services (LBS) are engendering new passion in mobile services utilizing users' location information. Such spatio-temporal information processing entails the need for a dynamic middleware that accurately identifies changing user location and attaches dependent content in real-time without putting extra burden on users. Our work focuses on creating a distributed infrastructure suitable to support such scalable content dissemination. As a result this chapter offers a conceptual framework, location-aware intelligent agent system (LIA) in integration with publish/subscribe middleware to comprehensively address dynamic content dissemination and related issues. We discuss the operational form of our framework in terms of PUSH and PULL strategies.

INTRODUCTION

The plateau of the information superhighway keeps advancing amid the evolution of the Internet and related technologies. At the same time, popularity of mobile devices and rapid advancements in wireless technologies are making it convenient for users to access various types of information

available on the Internet over the wireless networks using their mobile devices. Moreover, as the array of mobile devices keeps expanding, users expect to be able to use different devices for accessing such information, entailing development of a research area called *mobile computing.* Although, mobile devices lack in terms of processing power, memory capabilities, display, connections to the wireless networks, and so forth, the demand for accessing dynamic content using mobile devices has grown ever more pressing (Kaasinen, 2003). On the other hand, timely and accurate data dissemination to and from various mobile devices using wireless networks and supporting technologies continues to be a progressively taxing research challenge.

Out of many challenging research issues in the mobile computing domain, a relevant challenge is context-aware computing. The term *context* refers to an application's operating environment, which consists of device location, device identity, user activity, time, state of other relevant devices, and so forth. Our focus in this chapter is on *location* and *time*. Location and time have a special relationship with regard to the content: *historical* (past) user locations and related content, *current* (present) user location and related content, and *future* user locations and related content. These scenarios represent the content usage as a function of location and time, giving rise to location-aware computing. Location-aware computing allows applications to be aware of a user device's physical location at any point in time. Applications can exploit this information for customizing their functional behavior and presentation. Users as well as providers of various types of mobile services can also rip the benefits of having access to this location information in a mobile environment. However, users are continuously moving along with their mobile devices, and hence, location information of the user and her/his device is temporal. Capturing invariably changing location information of mobile devices presents an intriguing challenge. The Federal Communications Commission's (FCC's) mandate that wireless carriers in the United States be able to determine the approximate location of mobile phones making emergency calls is a key enabler for development of techniques to capture such temporal information regarding the user; it also provides an incentive to the cellular service providers to adopt above-mentioned location-aware systems. Examples of application of such location-aware systems include:

- **E-Deals:** A motel sends a promotional electronic coupon to mobile users passing by who are potential customers. Not only that, the motel can send some additional information regarding nearby restaurants and nearby attractions with applicable discounts if they take advantage of the e-coupon.
- **E-Directory:** A yellow page service that gives details on nearby services; for example, a user can locate the closest gas stations to her/his driving location, along with gas prices. Some additional information such as deals available on rotation of tires, car batteries, and so forth at a nearby auto center can be passed on to the user while presenting the user required information.

In performing the above tasks, location-aware systems need to combine the functionality of location-detection technologies (e.g., Global Positioning System, GPS), wireless or cellular telephone technologies (e.g., code division multiple access, CDMA), and information technologies (e.g., the Internet) under the scope of mobile computing to lay foundations for pervasive (anywhere, anytime) environments and services. On one hand, such services have the potential to dramatically improve usability of mobile devices and applications that adapt the content and presentation of services to each individual user and her/his current context of use. On the other hand, devising such location-aware systems is a tremendously complex task. Designers of location-aware systems have to keep in mind not only the continuous movement of mo-

bile devices-related location-detection techniques, as well as connection with location-dependent content. Additionally, user perceptions about information privacy and security pose substantial challenges. System designers need to safeguard users' privacy in terms of making their location visible to the system all the time. They also need to secure users' personal information traveling through the system. Until users feel satisfied with the initiatives in this regard, the visions of innovative and powerful location-aware applications and services cannot be realized on a public network. Another challenge for system designers is the need to easily customize the existing Web content to specific geographic locations.

Location-detection techniques are one part of overall location-aware systems. There are several possible options for determining location of a mobile client, each requiring a different set of infrastructure and resulting in a different accuracy level. A few examples are: time difference of arrival (TDOA), angle of arrival (AOA), location pattern matching (LPM), Bluetooth technology, and the Global Positioning System (GPS). Out of these, TDOA, AOA, and Bluetooth are used either indoors or in limited range. GPS has emerged, recently, as not only a cost-effective, but widely acceptable locating technique that is also compatible with most wireless networks as well as information technologies. However, currently GPS carries an inability to function efficiently indoors and in urban areas. Whilst RADAR (Bahl & Padmanabhan, 2000) or Bluetooth type technologies perform better in indoor areas, they cannot perform outdoors. Hence, an effective location-detection technology has to be a combination of these technologies so that outdoor as well as indoor locations can be effectively detected.

In addition to location-detection technologies, such systems also consist of wireless communication technologies such as cellular telephone technologies as well as information management technologies. In this chapter, we focus on developing information management techniques enabled with location-detection techniques. Location-aware applications, by default, are scalable distributed systems. We advocate using a middleware as a base for these distributed systems, as it provides not only the platform and protocols for communication, but also management supports making such systems as efficient and transparent as possible. We propose a *location-aware intelligent agent (LIA) system* that integrates two already proven components: agent technology and Publish/Subscribe paradigm. Agents append intelligence to an already flexible and scalable Publish/Subscribe architecture.

Publish/Subscribe (referred to as Pub/Sub from hereon) middleware has acquired attention and approval of researchers and is becoming popular tool for data dissemination in mobile environments (Anceaume, Datta, Gradinariu, & Simon, 2002; Baldoni, Beraldi, Querzoni, & Virgillito, 2004; Chen, Chen, & Rao, 2003; Farooq, Parsons, & Majumdar, 2004; Fenkam, Kirda, Dustdar, Gall, & Reif, 2002; Jose, 2004). It can offer distinguished assistance in extending the advantages of service-based architectures to the development of location-based services. The limited growth in such service-based architecture is mostly based on a *pull model.* A pull model is a user-initiated model in which a user sends a request for information to the system and gets a response in terms of location-aware service offerings or answers. However, with advances in wireless Internet technologies and increasing competitive pressure amongst the service providers, a *push model* or service-initiated model is shaping up. Under the scope of such a push model, service providers actively push location-aware information to the users via the communication network according to users' predefined interests or historical usage data. Pub/Sub middleware is compatible with Internet technologies; however, by its nature it is not able to detect the location of the user and then connect such location information with related content while saving such information for future repeat usage.

The second component of our middleware is the innovative wave of intelligent agent technology. Software agents are beginning to play a pivotal role in our lives. Until recently, most of the research in *agent technology domain* was focused on modeling and designing agent-based systems. Researchers have recently started showcasing the applications of agent technology that can revolutionize many real-life problems. However, not only end-users but even researchers have wondered what exactly these agents are. Etzioni and Weld (1994) expressed that following the Information Superhighway metaphor, an intelligent agent can be:

- A backseat driver who makes suggestions at every turn, or
- A taxi driver who drives you to your destination, or even
- A concierge whose knowledge and skills eliminate the need to personally approach the superhighway at all.

Similarly, there are several other interpretations of the "agents" as an entity. Various definitions offered by researchers that portray the variety of interpretations of agents are represented in the Related Work section. Given this difference in interpretation of what agents are or can be, we believe that that agent technology can play an important role in a distributed computing resources domain. Especially in our case it can provide a rather more robust middleware for wireless data dissemination in conjunction with Pub/Sub middleware, more so, considering the distinctive characteristics of mobile devices and/or usage patterns.

Next, we outline the related research work done so far in the area of agent technology, Pub/Sub middleware, as well as their usage in mobile environments. Secondly, we discuss the architecture of a proposed location-aware intelligent agent system, its components, and their functions. Thirdly, we discuss application of LIA in terms of push and pull strategies to an exemplary mobile services scenario, and finally conclude our discussion, along with some future research directions and a list of references.

RELATED WORK

Dynamic streams of information such as auction/stock trackers, traffic/weather alerts, and so forth communicated using mobile-distributed computing resources rely on an up-to-date view of information that can change rapidly and unpredictably. Much of this content may even be location dependent. Dissemination of such dynamic and location-dependent information to mobile clients has been a research challenge that researchers have been intrigued by for some time now. Some notable research efforts below have created direction for future research in this domain.

The first and an important aspect of location-aware systems is location-sensing techniques. Ladd et al. (2002) designed a location sensing system based on robotics using a wireless Ethernet. They began the design of the system by determining the position inside a building from measured RF signal strengths of packets on an IEEE 802.11b wireless Ethernet network. Using such a system they have tried to achieve reliable localization using general methods from Robotics following the Bayesian approach to localization. The system, however, was adoptable only within indoor locations. Smith, Balakrishnan, Goraczko, and Priyantha (2004), on the other hand, investigated the problem of tracking a moving device, which is a focal-point issue of location-aware computing. This investigation took place within the context of active as well as passive mobile architecture. They developed a real-world test-bed (Cricket) to evaluate the performance of location detection in both architectures. However, the Cricket system works indoors only, and as discussed above for a mass scale acceptance of location-aware systems, it needs to be as effective in outdoor locations as

well. In addition to these research efforts, there are several other systems developed that have paved the path for further research. The active badge system is a classic example of such systems.

It is also valuable to predict future locations that the user will travel to, in order to make location-aware systems more useful, as well as wanted. Karimi and Liu (2003) focused their attention on a predictive location model for location-aware services. They submitted that under the scope of location-aware systems, users' future locations carry far more weight than usually assumed. There are additional benefits for the users and service providers when future locations can be predicted. They devised a PLM model that will predict users' future locations so that information dependent on such locations can be transferred to a user in advance to help with planned decision making. Within the model they used historical trajectories and a probabilistic matrix related to road-level granularity of data for coming up with a prediction of future locations.

Kaasinen (2003) studied the need for location-aware mobile services from the user's point of view. The author drew conclusions regarding key issues in location-aware mobile services related to user needs based on user interviews, and laboratory and field evaluations with users and experts. They presented user needs under five main themes: topical and comprehensive contents, smooth user interaction, personal and user-generated contents, seamless service entities, and privacy issues. All these themes contribute to improving the overall usability of mobile services, applications, and in turn devices. Based on their findings, they suggested some guidelines for location-aware systems. For example, the services should be easy to find, and it should be easy for users to access an overview of the available services as well as their coverage. Also, services should be easy to take into use, and use thereafter. They also pointed out a few important aspects such as information personalization and user-generated contents.

Personalization in location-aware services provides a boost to usability of those services by providing the user with the most essential information according to the hierarchy of their personal preferences. However, the author acknowledges that if the user preferences are different in different locations, configuring the system for all available locations becomes a major task for users as well as system designers. On the other hand, they advocate active participation of users in information creation, rather than being just passive information consumers. Users' opinions, ratings, or recommendations could enhance many services. User participation is the cornerstone of development of the World Wide Web and other Internet services like Weblogs (blogs). The reason such paradigms became popular is because, among other factors, ordinary users can provide information to others. However, it is not only difficult for users to participate, but most times users lack the motivation to do so. Our system (LIA), on the other hand, helps save the Pub/Sub middleware, and the agent framework supports such ordinary user participation.

A location-aware application consists of terminal and server components (Jarvensivu, 2004). The terminal component of LIA is the location-detection as well as device communication part of an agent framework that resides on user (end-users as well as service providers) devices. The server component on the other hand resides on fixed network resources and consists of Pub/Sub as well all other parts of the agent framework including the system-wide repository. A proper coordination amongst these components enables smooth performance of the location-aware system. Scalability of these location-aware systems also determines their performance. Mokbel, Aref, Hambrusch, and Prabhakar (2003) defined scalability of location-aware systems as the system's ability to provide real-time responses to a large number of continuous concurrent spatio-temporal queries coming from the users to the central system. Mokbel et al. (2003) divided the spa-

tio-temporal queries into *snapshot queries* and *continuous queries*. Snapshot queries are queries that can be answered using data that is already collected and saved on fixed computing resources. Continuous queries on the other hand depend for response on data progressively accumulating into the fixed resources. They also propose a sharing mechanism for these queries among various entities of the system in order to achieve the optimum scalability of the system.

In addition to sensing a user's current and precise location, as well as predicting possible future locations, there are other challenges for researchers. Schilit et al. (2003) discussed a few of the current challenges in the location-aware systems domain: low-cost, highly convenient position-sensing technology, making users comfortable with respect to their location privacy, and having existing Web content easily customized to geographic locations. They initiated the Place Lab initiative to make location-aware computing a reality on a mass scale. The Place Lab initiative is meant to bootstrap location-aware systems through low-cost positioning technology in conjunction with a broad community-building effort that will create the large collection of location-enhanced Web services needed to catalyze business models.

On the other hand, passing information to users on their mobile devices has challenges of its own. For example, if the users' mobile device is turned off or disconnected from the network, how will the information be delivered to the user? Having a middleware that provides management support to the overall system is a solution to such practical problems. ELVIN (Carzaniga, Roseblum, & Wolf, 1998) is a notification system with limited support for mobile users. This system addresses the issues of message queuing, but important issues such as message distribution and location management are not addressed. CEA (Fenkam et al., 2002) and JEDI (Cugola, Nitto, & Fuggetta, 2001) are Pub/Sub middleware systems for mobile communications. Both these systems also address the queuing problem, specifically in cases of disconnections. The work of Huang and Garcia-Molina (2001) provided the operational guidelines for extension of Pub/Sub systems to a mobile environment. They analyzed the adaptation of a centralized and distributed Pub/Sub architecture with mobility. The ideas presented in Caporuscio, Carzaniga, and Wolf (2003), Chen et al. (2003), Huang and Garcia-Molina (2001), and Podnar, Hauswirth, and Jazayeri (2002) have provided us with motivation for our research efforts.

As noted above, a middleware is needed to effectively combine the functionalities of location-sensing technologies as well as wireless technologies. We use Pub/Sub middleware as a platform to share such information amongst various technologies we use in our system. Pub/Sub middleware brings information providers and consumers together on a single platform. Publishers publish the information to the Pub/Sub system to be delivered to pre-identified interested receivers in the form of events or messages. This communication exhibits the following characteristics:

- **Anonymous:** The parties exchanging information do not have to identify themselves, nor do they have to know each other in order to send/receive the information (loosely decoupled).
- **Asynchronous:** The sender and receiver do not have to be connected to the system at the same time in order to exchange information.
- **Multicasting:** One publisher can publish the same information to many subscribers, and one subscriber can receive information from many publishers.
- **Stateless:** The system does not persistently store messages for all subscribers, rather messages are sent directly to those who have subscribed.
- **Implicit:** Receivers of information are chosen implicitly based on their subscriptions and cannot be altered or controlled by the publishers.

- **Dynamic:** Publishers can publish the most recent dynamic data they have available, and the Pub/Sub system will pass on the data to the recipients.

Congruent with our thinking, many researchers have acknowledged the potentials of applying a Pub/Sub system to wireless platforms. A few highlighting studies in this area include Caporuscio et al. (2003), Cilia, Fiege, Haul, Zeidler, and Buchmann (2003), Chen et al. (2003), Farooq et al. (2004), Heimbigner (2000), Huang and Garcia-Molina (2004), and Muhl (2004). Much past research in this domain has focused either on development of Pub/Sub infrastructure or performance-oriented aspects of it. Caporuscio et al. (2003) proposed implementation of Pub/Sub systems with distributed access points where clients can connect with the system. This research focuses on enhancing the operational effectiveness of Pub/Sub-based infrastructure in a wireless environment. Farooq et al. (2004) have strictly studied the performance-related aspects of Pub/Sub middleware in mobile wireless networks. They studied publisher throughput, subscriber throughput, message loss, and handoffs, as well as workload parameters such as bandwidth, message and connection load, message size, and so on. However, most of above research has also acknowledged the fact that a platform only for sharing information amongst the service providers and service consumers is not enough; some management support is required to enhance the overall effectiveness of the middleware. We suggest integration of Pub/Sub with agent technology and harvest the benefits of both techniques to address such needs. Podnar et al. (2002) presented a multi-layered architecture that offers efficient content dissemination service targeting mobile users that motivated our work. However, their model lacked applicability to mobile environment, specifically in terms of location management, content adaptation, and information security. In our case, these features are addressed by use of intelligent agents. Our approach also differs from that of Podnar et al. (2002) in terms of component responsibilities, framework organization, and management of dynamic information processed throughout the proposed framework.

We avow that specific characteristics of intelligent agent technology can offer notable benefits within the scope of providing real-time services to mobile clients. In the past literature many researchers have attempted to define agents, but ultimately highlight the variety of interpretations that exist in defining and/or describing agents.

Nwana (1996) defines 'agents' as: "A component of software and/or hardware which is capable of acting exactingly in order to accomplish tasks on behalf of its user." This definition follows the traditional definition of agency, agent accomplishing an assigned task by and for its owner. Bradshaw, in his book (1998) *Software Agents*, defines the term 'agents' in a more blunt way: "An agent is that agent does." This definition is more in line with the trend of naming agent entities based on their functions or usage. Maes (1994) defines:

Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed.

For the purposes of our research, we define "agent" as: "A knowledge-based, self-learning software entity that autonomously accomplishes tasks on behalf of its beneficiary(s)." The definition of agents, however, encompasses few but not all agent characteristics described as follows:

- **Autonomous:** Autonomy on the part of an agent means that an agent is able to take initiative and exercise a non-trivial degree of control over its own actions.
- **Goal-Oriented:** Agents have specific goals assigned either by the users (explicitly or tacitly) or by the designers at the time of

design. They relate every action to the overall goal and have an ability to modify user requests, ask specific questions pertaining to user requests if they are of a destructive nature, or do not coincide with the ultimate goal.

- **Collaborative:** Agents cooperate and collaborate with other entities in the electronic environment. These entities include humans, other agents, other entities, and/or programs in the environment.
- **Flexible:** Agents' actions are not predetermined in many cases. They deal with a set of incomplete and unprocessed information that often keeps changing over a period of time. Ideal agents will be designed to not only accumulate and relate to past knowledge, but constantly gather new information and design a response mechanism accordingly.
- **Self-Starting/Proactive:** Unlike other software programs which are explicitly invoked by user actions or any other event in the environment, agents can be proactive and exhibit a goal-directed behavior on their own.

Personalized/Customized

An agent can be personalized to a specific user, task, or environment. It can go through a complex process of customization on its own or by the user to be assigned a specific set of tasks or a general goal or both.

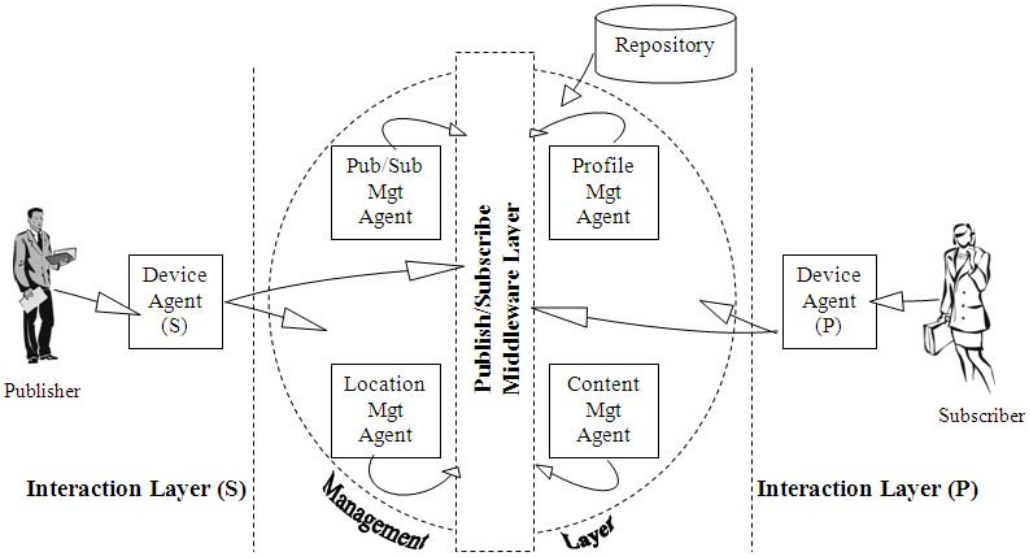
- **Reactive:** This refers to agents' ability to selectively sense and act. Agents sense the changes occurring in their environment (the physical electronic world) and act accordingly, keeping in mind its goal.
- **Adaptive:** Agents are able to learn from their own actions, users' preferences, and various elements of the environment, and fine tune their actions accordingly. This is based on the learning mechanisms built into agent design.
- **Communicative:** Agents can engage in simple to complex communications with other agents, users, and other applications in order to accomplish personal or common goals. Communications with users are carried out using natural languages, whereas the application uses machine languages and other communication protocols.
- **Mobility:** Agents can be built with mobility, transporting not only from one machine to another, but also across different systems, architectures, and platforms.
- **Inferential Capability:** Agents have models of self, users, and so forth based on which they exhibit a certain level of inferential capability. This means that agents not only have knowledge, but can infer upon that knowledge for taking specific actions.
- **Temporal Continuity/Long-Lived:** An agent is a continuously running program or process. As Etzioni and Weld (1994) describe, "It is not a 'one-shot' computation that maps a single input to a single output, then terminates." In addition it follows some level of persistence in terms of its identity and state of being over a long period of time.

We argue that the above discussed qualities of intelligent agent systems in integration with Pub/Sub middleware make our system more robust, intelligent, and location aware in a mobile environment. To this end, we propose the *location-aware intelligent agent system* in integration with Pub/Sub middleware for mobile environments.

SYSTEM ARCHITECTURE

Figure 1 depicts the proposed multi-layered architecture of LIA. Because of various limitations we provide a distributed architecture which is a surrogate type of agent platform that provides

Figure 1. Architecture of LIA



management support to all other entities in the system. Hence, only device agents on the subscriber or publisher side are mobile, while the rest of the agent framework resides on fixed network resources. We chose this type of agent platform because of memory and processing capability constraints of mobile devices. The major components of our framework are: *Publishers, Subscribers,* and *LIA.*

Publishers

Publishers are service providers who have an interest in reaching the consumers of the content they will publish in the form of events or messages. When a publisher is ready to deliver the content, it does so via Pub/Sub middleware. Assuming that the content matches the subscriber's interests as well as device specifications, the publisher defines the message/event to be published and transfers the content to the device agent-publisher (PA), who in turn sends a publish request to Pub/Sub. The publisher does not usually hold references to the subscribers, neither do they know how many of the subscribers the content will reach. We will learn below, however, that the PA keeps a log of published content as well as content requests that came from subscribers, keeping their decoupling intact. Publishers do not have to be connected to the system all the time or even at the same time when subscribers are connected to the system. Publishers could be host sites on a fixed network or can be mobile clients themselves. The message from one publisher can reach more than one subscriber, and also one advantage of our system is that a publisher does not have to keep publishing the same content over and over again.

Subscribers

Subscribers are content consumers. They are interested in receiving up-to-date dynamic information about their subjects of interest. Subscribers register their interests with the profile management agent (PMA) via device agent-subscriber (SA). The PMA shares these specific interests of the subscribers with the Pub/Sub middleware. This subscription information is not passed on to the publisher so that decoupling could be maintained. In addition to this, subscribers are also supposed

to register with the PMA a number of devices they will be using for receiving services such as a cell phone, PDA, notebook, or desktop computer. This will help the management layer locate the user device and recommend an appropriate list of location-aware services. Subscribers mainly use the Internet to connect, disconnect, and then reconnect to the system through various different access points based on the device they are using. It is noteworthy here that most mobile users will be faced with frequent disconnections and thus present a complicated challenge for the system for efficient message queuing, as well as delivery procedures.

LIA

LIA is an agent-based framework that is location aware as well as intelligent. LIA is composed of four layers: *Interaction Layer (P)* which is the interaction layer between LIA and the publishers, *Middleware Layer* which is the Pub/Sub middleware, *Management Layer* which provides the management support to the entire framework and is composed of four different types of agents, one system-wide repository, and finally *Interaction Layer (S)* which is the interaction layer between the subscribers and LIA.

Interaction Layer (P)

This interaction layer contains device agent-publisher (P). *Device Agent-P (PA)* resides on fixed network resources as long as the publisher is located in a fixed network as well. PA mainly communicates with the publishers and assists them in publishing their content to the service/information consumers. PA also assists publishers to dynamically offer location-aware dynamic information. The intelligent part of this service is how it helps the publishers dynamically configure and reconfigure the list of services in accordance with subscriber preferences. Not only that, based on subscription requests from the subscribers, PA might suggest that the publisher offer a specific service in the close vicinity of physical or logical location.

Publish/Subscribe Middleware

Pub/Sub provides an interaction platform for publishers and subscribers. Publishers and subscribers can connect to the Pub/Sub middleware through their device agents (PA, SA). Pub/Sub middleware has a distributed structure that allows it to be scalable. Pub/Sub gets input from publishers, subscribers, and management layer. Publishers publish their content proactively or upon the request of subscribers to the middleware via PA. As such, Pub/Sub does not play an active role in adapting the content to users' interests; neither does it save any information related to the subscribers. However, Pub/Sub makes sure that only relevant information is published to subscribers. Pub/Sub enables multicasting by directing the same content to more than one subscriber. Pub/Sub middleware supports multicasting, asynchronous-anonymous communication, as well as time, space, and synchronization decoupling.

Management Layer

This layer provides management support to the overall architecture and is situated on fixed computing resources making all of its components stationary. It is composed of four intelligent agents with a distinct set of responsibilities—(1) *P/S Management Agent (PSMA),* (2) *Location Management Agent (LMA),* (3) *Profile Management Agent (PMA), and* (4) *Content Management Agent (CMA)*—and a system-wide repository.

P/S Management Agent

PSMA is responsible for managing Pub/Sub operations, specifically, subscriptions and messages. It helps identify publishers/subscribers by locating identifier tags in either type of communication.

A highlighting role PSMA plays is intelligent queuing of subscriptions or messages in a flexible way. In case of mobile device disconnection from the network, PSMA queues the subscription or message with time stamp and identifier information until reconnection is established. Also, when the content is delivered, it purges the queue. It requests and keeps a copy of all subscriptions or messages from the originating party to address message loss. It also generates knowledge by accessing publisher expertise as well as subscriber interests from the repository and makes proactive suggestions to both parties for better matching mechanism.

Location Management Agent

This component of agent framework helps locate publishers or subscribers who move along the network and connect through various access points. LMA is responsible for identifying all active devices registered to single user. Not only that, it is supposed to identify the user's geographical location while supporting several connect spaces (IP address, telephone numbers, DNS entries, etc.). There are many location-sensing technologies available: GPS, e911, Bluetooth, Active Badges, Cricket, and so forth. Although no single technology is likely to become dominant, as there are too many dimensions along which location-sensing mechanisms can vary. For its universal acceptance we use a GPS system. LMA works as a location server, and both SA and PA are receivers of geospatial information about the users with an inbuilt GPS receiver on their devices. These GPS receivers keep sending geospatial information for the user at every *x* minutes. Changes in users' direction, speed of traveling (assuming its constant), and all other pertinent information is recorded based upon this information.

As the GPS system transmits data in terms of coordinate values, a trajectory of traveling path is created based upon which information resources are organized. Once the momentary location and traveling path are identified, all information sharing is focused within the scope of such geospatial data. However, end-to-end control for such geospatial data is provided by the management layer considering the existent privacy and security concerns over users' information. The management layer filters already published information and seeks new information for the current as well as predicted future location (PFL) from repository as well as publishers. In addition to this, locations that users have visited in the past are also saved in the repository in case of any revisits. Supporting efficient rendering and transmission of geospatial representation will require attention to interaction issues associated with database access and knowledge discovery, which is supported by the management layer.

Profile Management Agent

PMA is responsible for organizing and communicating details regarding subscribers and publishers. PMA saves subscribers' interests, publishers' expertise, subscribers' registered devices, recent locations subscribers have been to, and so forth. In other words, every detail regarding publisher or subscriber is channeled through PMA to various components of the LIA. Both parties can choose a device (from a list of registered devices) and time for receiving/publishing the content, and register these interests with PMA.

Content Management Agent

CMA keeps a log of all messages and subscriptions that travel through the system. CMA deals with customizing the content as required. One of the most important responsibilities CMA carries out is to define associative rules for related or complementary information to be presented in a way that is preferred by the user. CMA keeps record of each transmission, and when request for similar information comes from any other subscriber, CMA pushes the information to PSMA

who in turn passes it on to the subscriber. CMA enforces content adaptation with respect to the active user device following its specifications. CMA also enforces information security using public-key protocol with all other components of the infrastructure. We assume that certification authorities can be built that can provide public key certificates prior to any communication between any components of the framework. We will also deploy a threshold secret sharing scheme offered in Shlolz (2002), where an (m,n) threshold scheme permits a message to be projected onto n shares such that any m of them can be combined to reconstruct the original message or subscription, but less than m of them can not. This will maintain the message security.

Interaction Layer (S)

This layer consists of device agent-subscriber (SA). Device agents are designed to interact directly with the end-users. SA resides mainly on the devices of the users; in this sense they are stationary agents. However, users also have them installed on their mobile devices; in that context they are mobile agents. SA communicates with users as well as the rest of the entities of the framework. Subscribers communicate their preferences and interests to the PMA through SA. SA keeps a log of all the services received by the subscribers on various devices. SA carries its "State" completely and folds into a mobile code when the user switches the device of active use and transfers itself onto the current active device. SA dynamically reconfigures the list of location-aware services in association with other agents in the system and suggests any location-aware services users might be interested in while passing by that specific physical location.

Both publishers and subscribers can register more than one mobile device, and LIA carries its "State" completely and folds into a mobile code (transferred back to central architecture) when the users switch the device of active use and transfer itself onto the current active device.

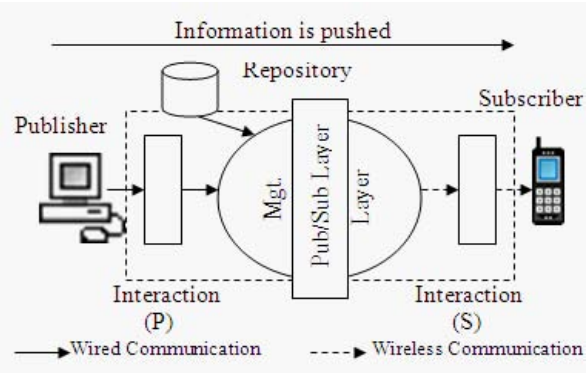
APPLICATION STRATEGIES

We outline the framework performance in view of an application scenario: Manni is driving on route 81-South near Salem, Virginia. She has her cat (Missy) with her. She is tired and wondering about availability of a specific type of room in a relatively inexpensive motel in the surrounding area. Fortunately, she is carrying her PDA with pre-installed device agent-subscriber (SA) as well as a location-sensing GPA receiver, which are components of LIA. SA is her gateway to LIA. We discuss two major strategies/models (Push, Pull) in light of this application scenario for location-aware dynamic data dissemination.

Push Strategy

Manni's PDA is turned on and is online. Quality Inn, located in Salem, Virginia, is participating as a publisher with LIA. PA on QI's side senses existence of Manni's device in the range and pushes offers from QI along with types of rooms available to Pub/Sub middleware. Pub/Sub shares this information with the management layer. LMA identifies the precise location of Manni on I-81. PMA determines the fact that Manni will

Figure 2. PUSH strategy



be using her PDA to receive her services today. CMA receives this input from LMA, PMA and processes the content to match the location and adapt to fit the specifications of Manni's device. PSMA identifies the message with its identifier and queues the final message for Manni's device. Pub/Sub contacts the SA at this point and intimates that there is a message in queue waiting for Manni to read. Along with this information the management layer generates a list of additional services available for Manni. For example, CMA collects that Denny's restaurant nearby offers early-bird discounts for Breakfast before 9 a.m., a golf club nearby is currently offering huge discounts to leisure guests, an area-wide directory service provides information on sightseeing places, car mechanics, grocery stores, departmental stores, shopping malls, and so forth. After receiving up-to-date information about all of these services, CMA bundles this information with the original message in a user-friendly manner.

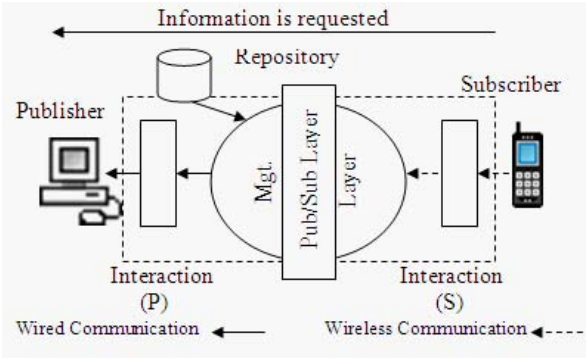
Pull Strategy

On the other hand, let us assume that Manni has forgotten Missy's (the cat's) food. Manni wants to find out if there are any specialty pet stores in the Salem area. Manni communicates this requirement and additional preferences to SA who shares the query with Pub/Sub. CMA accesses this information and explores the repository to find out if there is already existing information in this regard. Assuming that there is no information available already, the management layer intimates the Pub/Sub to post a message to relevant publishers requesting such information. In addition, the management layer recognizes, with the help of PMA, that Manni likes to eat at Friendly's and has registered this interest with LIA. Pub/Sub locates service providers for the relative content and intimates them with the help of PA for generation of appropriate content for Manni. The service provider provides the content to Pub/Sub, who shares the information with the management layer. Through the same channel, information is communicated to Manni. By this time, let us assume that Manni has checked into a room at the motel and communicates her preference to receive the content on her laptop. PMA identifies her device, and CMA helps configure the content in a publishable form. LMA locates and identifies the notebook computer; CMA personalizes and reformats the information for an appropriate display on Manni's notebook computer. Figure 3 depicts the application of this strategy.

CONCLUSION

Dynamic content dissemination, particularly to mobile clients, is gaining an increasing amount of popularity. We have provided a framework which exhibits that agent technology, in association with Pub/Sub architecture, can make the system autonomous, intelligent, mobile, and secure, while keeping the benefits of the Pub/Sub paradigm. These are highlighting additions that agent technology can make to the research agenda at hand which are not documented to be delivered by other competitive technologies. Ours is the first effort to amalgamate contributions of agents and Pub/Sub in one system for effective performance in the mobile services domain. The synergetic effects of these two paradigms are also unprecedented.

Figure 3. PULL strategy



Pub/Sub is a well-accepted solution for asynchronous and anonymous communications in a mobile environment. However, past literature left issues like system scalability, content adaptation, location awareness, and generation of dynamic content unanswered. Our framework, LIA, not only addresses all of the above issues, but extends the overall architecture to make it more robust. We believe that LIA will mark the future research direction for many intelligent and location-aware applications in the mobile services domain.

On the other hand, vulnerability of intelligent agents, especially in mobile environments, has been well documented. Although we have tried to address information security issues, we believe that it remains an open issue for future research. In addition, privacy issues relating to user-specific information requires further investigation.

In this chapter, we have chosen not to focus on the technical aspects and applicability issues of LIA due to the limitation of space and other resources. However, we believe that such discussion will be a driving factor in the acceptance of LIA, and we aim to encompass such discussion in our continuing research efforts. Another important limitation of our presentation so far is that we assume that the infrastructure required for successful operation of LIA is not only existent, but at par with the performance measures. Integration of such infrastructure, including all three components—publishers, subscribers, and LIA—will mark the direction for ongoing research in this domain.

REFERENCES

Anceaume, E., Datta, A., Gradinariu, M., & Simon, G. (2002). Publish/Subscribe scheme for mobile networks. *Proceedings of the Workshop on Principles of Mobile Computing,* Toulouse, France (pp. 74-80).

Bahl, P., & Padmanabhan, V. (2000). RADAR: An in-building RF-based user location and tracking system. *Proceedings of the 19th Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM),* Tel Aviv, Israel (Vol. 2, pp. 775-784).

Baldoni, R., Beraldi, R., Querzoni, L., & Virgillito, A. (2004). Subscription-driven self-organization in content-based Publish/Subscribe. *Proceedings of the International Conference on Autonomic Computing,* New York (pp. 332-333).

Bradshaw, J. (Ed.). (1998). *Software agents.* Boston: The MIT Press.

Caporuscio, M., Carzaniga, A., & Wolf, A. (2003). Design and evaluation of a support service for mobile, wireless Publish/Subscribe applications. *IEEE Transactions on Software Engineering, 29*(12), 1059-1071.

Carzaniga, A., Roseblum, D., & Wolf, A. (1998). *Design of a scalable event notification service: Interface and architecture.* Technical report, Department of Computer Science, University of Colorado at Boulder.

Chen, X., Chen, Y., & Rao, F. (2003). An efficient spatial Publish/Subscribe system for intelligent location-based services. *Proceedings of the 2nd International Workshop of Distributed Event-Based Systems,* San Diego, CA.

Cilia, M., Fiege, L., Haul, C., Zeidler, A., & Buchmann, A. (2003). Looking into the past: Enhancing mobile Publish/Subscribe middleware. *Proceedings of the 2nd International Workshop on Distributed Event-Based Systems,* San Diego, CA.

Cugola, G., Nitto, E. D., & Fuggetta, A. (2001). The JEDI event-based infrastructure and its application to the development of the OPSS WFMS. *IEEE Transaction on Software Engineering, 27*(9), 827-850.

Etzioni, O., & Weld, D. (1994). A softbot-based interface to the Internet. *Communications of the ACM, 37*(7), 72-76.

Farooq, U., Parsons, E., & Majumdar, S.(2004). Performance of Publish/Subscribe middleware in mobile wireless networks. *Proceedings of the 4th International Workshop on Software & Performance*, Redwood City (pp. 278-289).

Fenkam, P., Kirda, E., Dustdar, S., Gall, H., & Reif, G. (2002). Evaluation of a Publish/Subscribe system for collaborative and mobile working. *Proceedings of the 11th IEEE International Workshops Enabling Technologies: Infrastructure for Collaborative Enterprises,* Pittsburgh, PA.

Ge, Z., Ji, P., Kurose, J., & Towsley, D. (2003). Matchmaker: Signaling for dynamic Publish/Subscribe applications. *Proceedings of the 11th IEEE International Conference on Network Protocols, Atlanta, GA* (pp. 4-7).

Heimbigner, D., (2000). *Adapting Publish/Subscribe middleware to achieve Gnutella-like functionality.* Technical Report, Department of Computer Science, University of Colorado at Boulder, USA.

Huang, Y., & Garcia-Molina, H. (2001). Publish/Subscribe in a mobile environment. *Proceedings of the International Workshop on Data Engineering for Wireless and Mobile Access,* San Diego (pp. 27-34).

Jarvensivu, R., Pitkanen, R., & Mikkonen, T. (2004). Object-oriented middleware for location-aware systems. *Proceedings of the 19th Annual ACM Symposium on Applied Computing,* Nicosia, Cyprus (pp. 1184-1190).

José, R., Moreira, A., & Rodrigues, H. (2003). The AROUND architecture for dynamic location-based services. *Mobile Networks and Applications, 8,* 377-387.

Kaasinen, E. (2003). User needs for location-aware mobile services. *Pers Ubiquit Comput, 7,* 70-79.

Karimi, H., & Liu, X. (2003). A predictive location model for location-based services. *Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems,* New Orleans, LA (pp. 126-133).

Ladd, A., Bekris, K., Rudys, A., Marceau, G., Kavraki, L., & Wallach, D. (2002). Robotics-based location sensing using wireless Ethernet. *Proceedings of the 8th Annual International Conference on Mobile Computing and Networking,* Atlanta, GA (pp. 227-238).

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM, 37*(7), 31-40.

Martin-Flatin, J. P. (1999). Push vs. pull in Web-based network management. *Proceedings of the 6th IFIP/IEEE International Symposium on Integrated Network Management (IM'99),* Boston (pp. 3-18).

Mokbel, M., Aref, W., Hambrusch, S., & Prabhakar, S. (2003). Towards scalable location-aware services: Requirements and research issues. *Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems,* New Orleans, LA (pp. 110-117).

Mühl, G., Ulbrich, A., Herrmann, K., & Weis, T. (2004). Disseminating information to mobile clients using Publish-Subscribe. *Data Dissemination on the Web, IEEE Internet Computing, 8*(3), 46-53.

Nwana, H. (1996). Software agents: An overview. *Knowledge Engineering Review, 11*(3), 1-40.

Padovitz, A., Zaslavsky, A., & Loke, S. (2003). Awareness and agility for autonomic distributed systems: Platform-independent Publish-Subscribe event-based communication for mobile agents. *Proceedings of the 14th International Workshop on Database and Expert Systems Applications,* Prague, Czech Republic (pp. 669-673).

Podnar, I., Hauswirth, M., & Jazayeri, M. (2002). Mobile push: Delivering content to mobile users. *Proceedings of the 22nd International Conference on Distributed Computing Systems,* Vienna, Austria (pp. 563-370).

Schilit, B., LaMarca, A., Borriello, G., Griswold, W., McDonald, D., Lazowska, E., Balachandran, A., Hong, J., & Iverson, V. (2003). Challenge: Ubiquitous location-aware computing and the "Place Lab" initiative. *Proceedings of the 1st ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots,* San Diego, CA (pp. 29-35).

Schlolz, J., Grigg, M., Prekop, P., & Burnett, M. (2003). Development of the software infrastructure for a ubiquitous computing environment—the DSTO iRoom. *Proceedings of the Workshop on Wearable, Invisible, Context-Aware, Ambient, Pervasive and Ubiquitous Computing,* Adelaide, Australia.

Smith, A., Balakrishnan, H., Goraczko, M., & Priyantha, N. (2004). Tracking moving devices with the Cricket location system. *Proceedings of the 2nd International Conference on Mobile Systems, Applications and Services,* Boston (pp. 190-238).

Sutton, P., Arkins, R., & Segall, B. (2001). Supporting disconnectedness—transparent information delivery for mobile and invisible computing. *Proceedings of the IEEE International Symposium on Cluster Computing and the Grid*, Brisbane, Australia (pp. 277-285).

Wood, M., & Glade, B. (1996). Information servers: A scalable communication paradigm for WAN and the information superhighway. *Proceedings of the 7th Workshop on Systems Support for Worldwide Applications*, New York (pp. 305-310).

Vyas, A., & Yoon, V. (2004). Location-aware intelligent agent system (LIA) for Publish/Subscribe middleware in mobile environments. *Proceedings of the 3rd Workshop on E-Business (Web'04),* Washington, DC (pp. 75-81).

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar , pp. 1-17, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.19
Topology for Intelligent Mobile Computing

Robert Statica
New Jersey Institute of Technology, USA

Fadi P. Deek
New Jersey Institute of Technology, USA

INTRODUCTION

We discuss an *interconnectivity framework* for data and content delivery to *mobile devices* that allows data of higher priority to reach the mobile unit in the *shortest time* possible. Two possible scenarios are presented; one that connects the servers in an *N-cube* configuration network, and another that shows the same N servers connected in a grid type network. The goal is to minimize the rate of *data jumps* from server to server until it reaches the mobile device. As the mobile user travels, the mobile device registers itself with the next server and the session is migrated from the old server to the new one without interruptions, in an analogous way, cell phones move from one cell to another. Starting with the idea that all data is not equal (in importance/priority), this article suggest a framework topology for intelligent mobile computing that guarantees data will reach the mobile device in a minimum amount of time, assuring at the same time the privacy of transmission. The integration of this type of technology into the 3rd Generation (3G), and 4th Generation (4G) *mobile computing* is also discussed.

Pervasive computing is rapidly emerging as the next generation of computing with the underlying premise of simplicity (of use), minimal technical expertise, reliability, and intuitive interactions. As technology continues to advance and mobile devices become more and more omnipresent, the aim towards achieving easier computing, more availability and prevalence is becoming a given. Through the clever use of advanced technologies, the new generation of intelligent mobile computing has the opportunity to serve user needs via prevalent computing devices that are ever more transportable and connected to an increasingly ubiquitous network structure. *Mobile communication* is changing as the trends of media convergence including the Internet and its related electronic communication technologies and *satellite communications* collide into one.

A change is being ushered by the 3G (3rd Generation) mobile technology with the usability and usefulness of information delivered to mobile

devices taking on added features. For example, *multimedia messaging*, as opposed to *voice transmissions*, being delivered to cell phones has rendered such mobile devices an integral part of people's lives and a core part of how they conduct their daily business rather than an add on tool (Buckingham, 2001).

The *3G mobile phone* system aims at unifying the disparate standards of current second generation *wireless systems*. The idea is to eliminate the different types of *global networks* being adopted with a single standard network. This will allow for the delivery of multimedia content and propagation through the network without the need for conversion from one standard to another. 3G systems need smaller cells thus the need for more base stations (mostly due to their operating frequency, power requirements, and modulation) and in many cases will not be feasible to install them in areas where population is not so dense (i.e., rural areas) (Garber, 2002). Because of these requirements and conditions, a better way to deliver the communication must be established. However, global access to such mobile devices will create *data delivery* challenges and servers can become clogged with unwanted communication, like that of wired Internet access. The need for moving relevant data to mobile devices in the shortest time possible becomes of utmost importance.

BACKGROUND

As the evolving functionalities of mobile computing take on primary roles at both the individual and the organizational levels, researchers and developers move to further enhance the technology. Bettstetter, Resta and Santi (2003) offer a random waypoint model for wireless ad hoc networks suggesting that the spatial distribution of network nodes movement, according to this model, is in general nonuniform and impairs the accuracy of the current simulation methodology of ad hoc networks. They present an algorithm that looks at the generalization of the model where the pause time of the mobile modes is chosen arbitrarily in each waypoint and a fraction of nodes remain static for the entire simulation time. They further show that the structure of the resulting distribution is the weighted sum of 3 independent components: the static, pause, and mobility (Bettstetter et al., 2003)

Xie and Akyildiz (2002) address the problem of excessive signaling traffic and long signaling delays in mobile IP. They argue that it is possible to have a distributed and dynamic regional location management scheme for Mobile IP where the "signaling burden is evenly distributed and the regional network boundary is dynamically adjusted according to the up-to-date mobility and traffic load for each terminal". This is suggested for minimizing the cost of content delivery over mobile IP networks (Xie & Akyildiz, 2002).

La Porta (2002) describes mobile computing as "a confluence of communication technologies (particularly the Internet), computing devices and their components, and access technologies such as wireless." He argues that a mobile computing environment will include not only real-time mobility of devices, but also mobility of people across devices, stressing the fact that the environment must include a wide range of devices, applications and networks (La Porta, 2002).

Zimmerman (1999) states that "The proliferation of mobile computing devices including laptops, personal digital assistants (PDAs), and wearable computers has created a demand for wireless personal area networks (PANs)" showing at the same time the fact that the mobility of such devices places considerable requirements on PANs not only for connectivity, cross-platform and networks but also for content delivery in minimum time (Zimmerman, 1999). This article further addresses the subject of data and content delivery to mobile devices with a keen interest in time and cost issues.

MOBILE COMMUNICATION SYSTEMS

There are four major categories in which data can be classified. These are real-time data, daily data, occasional data, and junk data:

1. **Real-Time Data (RTD):** Both hard and soft real-time is data that needs to be sent/ received as soon as possible regardless of cost.
2. **Daily Data (DD):** Data that is sent only once or twice a day at a predetermined time (status reports, weather forecasts, etc.).
3. **Occasional Data (OC):** Data that is sent from time to time (software updates, customer service reports, etc.).
4. **Junk Data (JD):** Data that is considered useless (spam).

One way to speed up the data delivery is called data shorthand, where properly configured computers can send chunks of data based on data changes so only the changed data is sent to the mobile device (Ungs, 2002). But this type of data exchange requires mobile devices to store each of the exchange. Caching transmissions can be used successfully for non real-time data communication (like browsing the Internet, checking e-mail, etc.). For real-time message exchanges, caching cannot be used and other methods of speeding up delivery become necessary.

Convergence between broadband wireless mobile devices and access is currently a significant issue in wireless communications. With the recent technological advances in digital signal processing, software-definable radio, intelligent antennas, and others, the next generation of mobile wireless systems is expected to be more compact, with limited hardware and will feature flexible and intelligent software elements (Rao, Bojkovic, Milovanovic, 2002). Wireless mobile Internet (WMI) is a key application of the converged broadband wireless system where the actual device will be compatible with mobile and global access services, including wireless multicasting and wireless trunking. Some of the characteristics of these mobile devices will be: at least 90% of the transmission traffic will be data, voice recognition functions will be operational for every command, the mobile device will support multiple users and various service options, the mobile device will be adaptive and upgradeable, and the entire transmission will be encrypted for ensuring privacy of communication (encryption will be done in hardware for faster processing).

Mobile wireless communication implies support for user's mobility and the overall communication infrastructure needed to handle movements within the home network cell/servers map but also outside the home network in situations where communication is provided by other providers (Agrawal & Zeng, 2003; Rao et al., 2002). A mobile station (MS) should be able to communicate without session interruptions as it travels anywhere using local wireless infrastructure facilities. Because of this, session handoff between cells and mobile switching centers (MFCs) of various wireless service providers should be supported. As a MS travels from a location to another, it has to register itself with the next cell/server that serves that particular area. Each of the servers maintains a visitor location register (VLR) that is an index of the MS IDs that are in its active area. As the MS leaves a cell/server, an entry is made in the home location register (HLR) of the home network so the current location is known at all times. Based on these registers, data can be sent over the network to reach the mobile device. Our work is concerned less with the way the handoff of the communication session takes place, but more with how many times the data has to jump before it reaches the mobile device as is described below.

NETWORK INTERCONNECTION FRAMEWORK

First, we look at a possible N-cube configuration for mobile switching centers/servers. Each imaginary cube will have eight servers (one server in each corner of the cube). The worst-case scenario will be a maximum of three jumps inside a cube. Any server can be reached from any other one with 1, 2, or 3 jumps. The 2^{nd} cube will also have eight servers and so on all the way to N cubes. The total number of servers will be N [cubes] × 8 [servers/cube]. Each of the MSCs will only have knowledge of their neighbors MSCs as well as the corresponding MSCs in the adjacent cubes. If we consider that for the real-time data (RTD) a maximum of 0.001 s propagation time between two MSCs inside the same virtual cube, and considering the propagation speed of the signal to be at 2/3 × C (where C is the speed of light in vacuum=300,000 [Km/s]) then the maximum distance between two adjacent MSCs should not exceed: D = (2/3 × C) x 100 ms = (2/3 × 3 ×10^8 m/s) × 0.001 s = 200 [Km]. This will make the worst-case scenario to be three jumps inside any given virtual cube that means a 3 × 0.001 [s] delay = 0.003 [s].

If the mobile station (MS) is not in the area covered by a cube, then a jump is needed from a cube to the next cube. If the MS is not registered with a cube, all eight servers/switching stations know that from the VRL list, so it will send the transmission to the next adjacent cube (1 jump only). If we consider the adjacent virtual cubes to be with in 200 [km] of each other, then the jump will not take longer than 1 micro second (0.001 s). In the case of N cubes, we have a maximum of Ln (8N) / Ln (2) jumps to reach the MSC that has the MS registered. In that case considering that every jump introduces a 0.001 [s] delay, in the worst-case scenario we will have 0.001 [s] × Ln (8N) / Ln (2). So if, for example, an MS travels, say, 5000 [km], that means that the area can be covered with the maximum 5000 [km] / 200 [km] = 25 virtual cubes.

If data needs to propagate from the HS (home station) to the position where the MS station is (5000 Km away), then the total number of jumps (in the worst-case) will be given by: Ln (8 × 25)/Ln (2) = 7.65 → 8 jumps. At 0.001 (s) delay for each jump, then the MS will be reached in minimum 8 × 0.001 s = 0.008 (s). Of course this would represent the best-case scenario ignoring the overhead and delays introduced by the switching equipment itself. If we use a safety factor of 2 (to cover the overhead delays) then in the worst-case scenario the MS at 5000 km away will receive the data 0.016 (s) after it was sent. This would apply to real-time data (RTD) that cannot be cached and needs to reach the MS the fastest way possible. For the data that is not real time and/or has smaller priorities, if channels in MSCs are available, it can be sent the same way as real-time data. But if the MSCs are busy routing real time transmissions, the lower priority data will be cached and put in a FIFO queue for later delivery.

Another possible interconnection of MSCs would be having the servers in a grid type network. Keeping the same constraints like for the virtual N-cube configuration, then the distance between each of the servers would be 200 [km]. Considering that each of the MSCs/servers can only communicate with a maximum of two other MSCs (no diagonal communication), and the grid has a square profile (m × m) with n number of MSCs, then the total number of jumps inside the grid would be 2(m-1). Considering the same example as above, with the MS traveling at 5000 [km] away. That distance can be covered by a 25 × 25 grid with each located at MSC at 200 [km] away from its neighbors. That would imply a number of jumps equal to 2(25-1) = 48. At 0.001 [s] delay per jump, the total delay in the case of a grid network would be 0.001 [s] × 48 jumps = 0.048 [s]. If we consider that the system has some residual delays, and if we use the same safety factor as in the N cube network, then we have a delay of 0.096 [s] which is six times worse than the virtual N cube configuration delay.

For the case when the MS travels outside an area covered by any of these two types of interconnections, a satellite service (MSAT-mobile satellite service) must be used to relay the data. MSAT are communication satellites in geostationary (Keplerian) orbits (35,786 km) and operate in the frequency range of 1626.5-1660.5 [MHz] for uplink and 1550-1559 [MHz] for downlink and currently serve only the US and Canada (Roddy, 2001). The periodicity of such satellite is 23 hr 56 min 4 sec which matches the Earth's periodicity. Some relevant points regarding this type of satellite: the satellite must move Eastward at the same rotational speed as the Earth, the orbit must be circular (and must maintain the same distance from the Earth), and the inclination of the orbit must be zero degrees (Roddy, 2001). Also, there is only one geosynchronous orbit thus communication via those satellites is still expensive for private use (for a two-way communication).

Satellite communication will introduce an additional delay due to the time it takes for the signal to reach the satellite and come back. At 35,768 [Km] away, a signal takes a time T= 35,768 [km]/300,000 [km/s] = 0.119 [sec] each way (Statica, 2002). So the round trip delay would be 2 × 0.119 [s] = 0.238 [s]. If we use the same safety factor of 2 (for delays due to sending acknowledgments, control signals, etc.), then the delay introduced by the satellite would be 0.238 [s] × 2 = 0.476 [s]. This time would be added to the time it takes to propagate through the terrestrial N cube or grid network when the signal arrives from the satellite.

Because the home location register (HLR) knows the exact location of the MS when is registered with any of the VLRs, when the value for a particular MS is not known by the HLR, two possibilities are considered: the MS is turned off (thus unable to communicate with any of the MSCs and register itself in a VLR) and the MS traveled outside an area covered by the terrestrial MSCs. At that point, if communication arrives at the HS for a MS that doesn't have the position known (registered) with any VLRs, then the HS must buffer the communication. When the MS comes on line and registers with a local MSC, even if that MSC is not directly connected to any N cube or grid network (for example is in another country where the HS is located), the MSC that has registered the MS will send a registration request via the satellite to the HS. At that time, the HS will know where to send the data for that MS and the transmission is relayed via the geosynchronous satellite. Of course satellite communication is not suitable for real-time data, due to the delays it introduces, but the MS can still get the data if needed (but not in real time).

FUTURE TRENDS

The push for advanced technology and the high demand for reliable, secure, low cost, high speed wireless connections as well as the high demand for access of data anywhere and anytime has revolutionized the wireless industry and moved the wireless systems from the secondary means of communications to a primary means crossing the line and merging at the same time the personal communication systems with business systems. Third generation (3G) wireless systems are in place, but there are already faster systems that will make the 3G network technically obsolete. Also, the fact that smaller cells are required for 3G systems would create a disadvantage when the technology is deployed over large and less populated rural areas. These systems considered to be the 4^{th} generation (4G) aim at delivering high-definition video signals at rates 10 times faster than the 3G systems (and also cheaper). 4G systems will focus on integrating wireless networks and become the platform for mobile systems. Also, 4G systems will be IP-based multimedia services in a seamlessly integrated network that will allow users to us any system at anytime and anywhere.

The new generation of wireless communication will be multi-everything (multi-mode, multi-band, multi-functional) and is aiming to provide the best connection to users (Kim et al., 2003). 4G systems will also incorporate automatic switching functions for mobile networks, mobility control, fast hand-offs and rapid routing of packets. Because of these capabilities, it is estimated that the 4G systems will improve the coverage in highly populated areas and it will carry more traffic by utilizing various technologies and methodologies to deliver the best possible services in the most efficient way. Important advantages of 4G over 3G are higher transmission rate (almost double), higher capacity, higher frequency band (higher than 3GHz), single mobile device, increased area of coverage, and lower system cost.

As these new systems continue to be developed and improved, the need to be able to deliver data as fast as possible (with minimum delays) becomes a major factor that needs to be incorporated so the new systems live up to their promise of delivering real-time video and data to mobile units.

CONCLUSION

As the integration trend of voice and data transmissions and making them available on the same mobile device grows, it becomes important to route data from the home station (HS) to the mobile station (MS) in the shortest time possible. Moreover, real-time data (RTD) must have the highest priority, should not be queued, and ought to be delivered to the MS also in the shortest time possible. The framework discussed in this article presents a possible N cube configuration and a grid type interconnection of the Mobile Switching Centers (MSCs), which are actually data servers, so the data can jump a minimum amount of servers and reach the MS as quickly as possible. In the case when the HS does not have a direct connection with the network that the MS is registered with (while traveling), a possible satellite communication is suggested using a mobile satellite service (MSAT) network.

In an N cube configuration, it is shown that the worst-case delay would be given by the equation: 0.001 [s] × Ln (8N) / Ln (2), where N is the number of cubes the data has to jump. The number of cubes N is calculated based on the distance the MS is at, considering a distance of maximum 200 [km] between cubes (and between each of the MSCs inside a cube). In the square grid configuration it is shown that the worst-case delay is given by the equation: 0.001 [s] × 2(m-1), where m is the side of the square. Using the MSAT satellite will introduce a minimum of 0.476 [s] delay, only due to the propagation of the signal.

We found that having an N cube interconnection of MSCs will give the smallest time in data delivery to a MS. With relatively small number of cubes, we can cover a relatively large area of service. For example, for a 5000 [km] travel of MS we need around 25 cubes. The proposed topology fits well in the design of the 4G systems and will help in achieving the design parameters of the 4th generation wireless networks by allowing data to reach the end user the fastest (and cheapest) way possible.

REFERENCES

Agrawal, D., & Zeng, Q. (2003). *Wireless and mobile systems.* Pacific Grove, CA: Thompson Learning.

Bettstetter, C., Resta, G., & Santi, P. (2003). The node distribution of the random waypoint mobility model for wireless ad hoc networks. *IEEE Transactions on Mobile Computing, 2*(3), 257-269.

Buckingham, S. (2001). *Yes 2 3G.* New York: Mobile Streams Inc.

Garber, L. (2002). Will 3G really be the next big wireless technology? *IEEE Computer Magazine, 35*(1), 26-32.

Kim, Y., Jeong, B. J., Chung, J., Hwang, C., Ryu, J. S., Kim, K. et al. (2003, March). Beyond 3G: Vision, requirements, and eEnabling technologies. *IEEE Communications Magazine, 41*(3), 120-124.

La Porta, T. F. (2002). Introduction to the IEEE transactions on mobile computing. *IEEE Transactions on Mobile Computing, 1*(1), 2-9.

Prakash, R. (2001). A routing algorithm for wireless ad hoc networks with unidirectional links. *ACM/Baltzer Wireless Networks Journal, 7*(6), 617-626.

Rao, K. R., Bojkovic, Z., & Milovanovic, D. (2002). *Multimedia communications systems*. Upper Saddle River, NJ: Prentice Hall.

Roddy, D. (2001). *Satellite communications* (3rd ed.). New York: McGraw-Hill.

Spohrer, J. C. (1999). Information in places. *IBM Systems Journal, Pervasive Computing, 38*(4), 602.

Statica, R. (2002). *Multimedia satellite communications.* Research Report, CCS-NJIT.

Ungs, K. (2002). *Mobile networking.* Everett, WA: Intermec Technologies.

Xie, J., & Akyildiz, I. F. (2002). A novel distributed dynamic location management scheme for minimizing signaling costs in mobile IP. *IEEE Transactions on Mobile Computing, 1*(1), 163-175.

Zimmerman, T. G. (1999). Wireless networked digital devices: A new paradigm for computing and communication. *IBM Systems Journal, Pervasive Computing, 38*(4), 566-574.

KEY TERMS

Antenna: The part of a transmitting or receiving device that radiates or receives electromagnetic radiation (electromagnetic waves).

Bandwidth: The difference in Hertz between the limiting (upper and lower) frequencies of a spectrum.

Broadband: Refers to systems that provide the user with data rates in excess of 2MBps and up to 100 Mbps.

Cellular Network: A wireless communication network in which fixed antennas are arranged in a special pattern (hexagonal pattern) and mobile stations communicate through nearby fixed antennas.

Channel Capacity: The maximum possible information rate through a channel subject to the constraints of that channel.

Downlink: The communication link from a satellite to an Earth station.

Frequency: Rate of signal oscillation in Hertz.

Geostationary: Refers to geosynchronous satellite angle with zero inclination. So the satellite appears to hover over one spot on the Earth's equator.

Packet: A group of bits that includes data (payload) plus source, destination address and other routing information (in the header).

Transmission Medium: The physical path between a transmitter and receiver (can be wired (guided medium) or wireless (un-guided medium).

Uplink: The communication link from an Earth station to a satellite.

Wireless: Refers to transmission through air, vacuum or water by the means on an antenna.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour , pp. 1070-1074, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.20
Robust Intelligent Control of Mobile Robots

Gordon Fraser
Institute for Software Technology, Graz University of Technology, Austria

Gerald Steinbauer
Institute for Software Technology, Graz University of Technology, Austria

Jörg Weber
Institute for Software Technology, Graz University of Technology, Austria

Franz Wotawa
Institute for Software Technology, Graz University of Technology, Austria

ABSTRACT

An appropriate control architecture is a crucial premise for successfully achieving truly autonomous mobile robots. The architecture should allow for a robust control of the robot in complex tasks, while it should be flexible in order to operate in different environments pursuing different tasks. This chapter presents a control framework that is able to control an autonomous robot in complex real-world tasks. The key features of the framework are a hybrid control paradigm that incorporates reactive, planning and reasoning capabilities, a flexible software architecture that enables easy adaptation to new tasks and a robust task execution that makes reaction to unforeseen changes in the task and environment possible. Finally, the framework allows for detection of internal failures in the robot and includes self-healing properties. The framework was successfully deployed in the domain of robotic soccer and service robots. The chapter presents the requirements for such a framework, how the framework tackles the problems arising from the application domains, and results obtained during the deployment of the framework.

INTRODUCTION

An appropriate control architecture is a crucial premise for successfully achieving truly autono-

mous mobile robots. The architecture should allow for a robust control of mobile robots during the execution of a wide range of different tasks. Moreover, it should be flexible enough to facilitate different control strategies and algorithms. In addition, the architecture should be adaptable in order to handle more complex tasks and to be able to operate in different environments.

Finding an appropriate architecture for a specific purpose is a challenging task. In fact, no single architecture can be sufficient for all purposes. There is always a trade-off between general applicability and usability. The issue of determining an appropriate architecture suitable to robustly control an autonomous mobile robot can be divided into several sub-problems:

- The first and easier one is the question of which control paradigm to choose. In this chapter the different control paradigms are introduced. It is then motivated why a hybrid paradigm is the most appropriate for applications where robots carry out complex and non-trivial tasks.
- The second problem is more related to software engineering and concerns the software architecture. The software architecture determines how the functionality of the software is physically organized. Several projects working on an architecture sufficient for the needs of mobile robots are introduced. For an in-depth discussion of the issue of choosing or implementing an appropriate software framework, the reader is referred to Orebäck (2004). Finally, an example of a successful solution to this issue is illustrated.
- A more or less strong, deliberative component is part of every hybrid control paradigm. The use of symbol-based abstract decision making has two major drawbacks. First, in general, planning techniques are insufficiently reactive for unpredictable and highly dynamic environments. A solution to this problem is presented, which enables the deliberative component to react more quickly to such effects. Second, if an abstract deliberative component is used, then some kind of connection between the quantitative world of the sensors and actors and the qualitative world of planning and reasoning is necessary. If sensors and actors are prone to uncertainties, then this abstraction of knowledge is difficult. Unfortunately, such uncertainties are nearly always adherent to sensors and actors. Therefore, a novel symbol grounding mechanism is presented, which significantly relaxes this problem.
- The final problem is especially important in the area of autonomous mobile robots. Tolerance of the robot and its control system against faults is crucial for long-term autonomous operation. It is shown how a model-based fault-diagnosis and repair system improves the overall robustness of the control architecture.

Consideration of all these features and requirements has resulted in a control architecture that serves as a platform for research in several areas in autonomous mobile robots, for example, RoboCup robot soccer and service robots. This robust and flexible architecture will serve as a running example and as a guideline throughout this chapter.

SOFTWARE FRAMEWORKS FOR MOBILE ROBOTS

This section addresses the problem of software frameworks for autonomous mobile robots. For this, the applicable control paradigms are introduced. Control paradigms describe how control is organized. Then, general requirements for software architectures in order to be usable for autonomous robots are identified. Finally, popular publicly available software frameworks are reviewed.

Robot Control Paradigms

A robot control paradigm guides the organization of the control of a mobile robot, which enables the robot to carry out given tasks. It structures how the robot maps its sensor readings to actions via a more or less intelligent decision-making module.

One of the first attempts to structure control was the *sense-plan-act* (SPA) paradigm, depicted in Figure 1. This paradigm was inspired by the research on artificial intelligence (AI) of the late '60s and was first successfully used by Nilsson in the robot *Shakey* (Nilsson, 1984). It was guided by the early view on artificial intelligence. The paradigm divides the control into three different functionalities. *SENSE* is responsible for the perception of the robot's internal state and its environment. The data provided by the robot's sensors are interpreted and combined in a central abstract model of the world. Based on the information contained in the world model, a description of the capabilities of the robot and the goal of the task the *PLAN* module tries to find a plan (i.e., a sequence of actions) that will lead to a given goal. The *ACT* module executes this plan in order to achieve the goal. Although the SPA paradigm is powerful and flexible, it suffers from a set of drawbacks. First of all, planning takes a lot of time even on very powerful computers. Therefore, the reaction to dynamic environments is slow. Another drawback is caused by the fact that planning algorithms generally work on a qualitative and abstract representation of the world. The design of such a representation and the transformation of quantitative sensor data into this representation is far from trivial.

In contrast, the reactive *sense-act* (SA) control paradigm provides a different organization of control. Figure 2 depicts the SA paradigm. The paradigm is biologically inspired by the mechanism of reflexes, which directly couple the sensor input with the actor output. Such a reflex of the robot is commonly called a behavior. More complex behaviors emerge through the combination of different reflexes.

A system that follows this paradigm was first proposed in the mid 1980s with the subsumption architecture by Brooks (1986). This architecture achieved significant progress in the research on mobile robots and is still popular and widely used. Brooks argued that abstract knowledge about the world and reasoning is not necessary for the control of a mobile robot. The paradigm is able to control a robot also in dynamic environments because the reaction time is very quick due to the encoding of the desired behavior into reflexes and the tight coupling of the sensors and actors. Although relatively complex behaviors can be achieved by blending different reflexes, the paradigm is prone to fail for more complex tasks. This arises from the fact that neither explicit information about the internal state of the robot nor about the world, nor additional knowledge about the task, are used. Therefore, for complex tasks a goal-driven approach seems much more appropriate than a simple instinct-driven one.

Even though the choice of an appropriate control paradigm sometimes seems to be rather a question of faith than of science, there is a relatively clear commitment within the robotics research community that the most appropriate architecture is hybrid architecture (see Figure 3). Hybrid systems combine the advantages of the planning approach and the reactive paradigm while avoiding most of their drawbacks. Such systems use reactive behaviors where reactivity is needed (e.g., avoiding a dynamic obstacle)

Figure 1. The sense-plan-act control paradigm

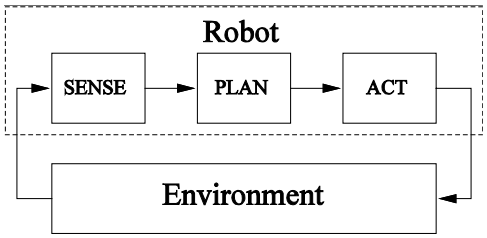
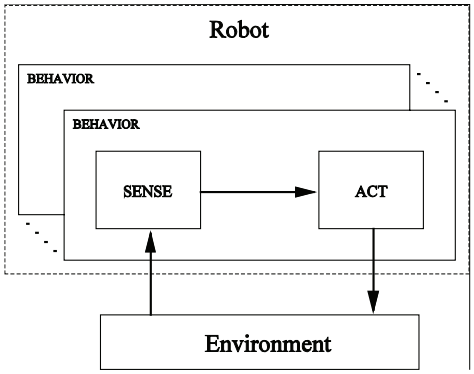


Figure 2. The reactive sense-act control paradigm

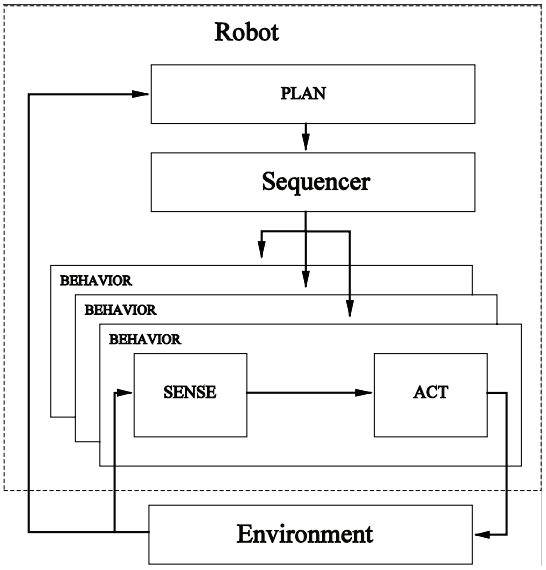


and use planning and reasoning where complex decisions have to be made and a limited amount of delay is not critical.

Commonly, hybrid systems consist of three layers. The *reactive layer* uses reactive behaviors to implement a fast coupling of the sensors and the actors in order to be reactive to a dynamic environment. Often this layer implements the basic skills of a robot, that is, basic movement primitives and obstacle avoidance. The *deliberative layer* has a global view on the robot and its environment, and is responsible for high-level planning in order to achieve a given goal. Typical functionalities that are located in this layer are mission planning, reasoning, localization, path planning, and the interaction with humans or other robots. The *sequence layer* is located between the reactive and the deliberative layer and bridges the different representations of the two layers. The sequencer generates a set of behaviors in order to achieve a sub-goal submitted by the deliberative layer. It is also responsible for the correct execution of such a set of behaviors and should inform the higher layer if the sub-goal was successfully achieved or the execution failed for some reason.

A more detailed introduction to the different control paradigms can be found in the book by Kortenkamp, Bonasso, and Murphy (1998) and the book by Murphy (2002).

Figure 3. The hybrid control paradigm



Requirements for the Software Architecture of an Autonomous Robot

The control paradigm guides the functional decomposition of the robot control on a more abstract view. The software architecture, on the other hand, guides the modular decomposition of the system into different components and the implementation of such components. Furthermore, it concerns the encapsulation of different functionalities into manageable modules.

A well-designed software architecture should provide, among others, the following features:

- Robustness
- Flexibility
- Sensor and actor interface abstraction
- Easy exchange and reuse of components
- Reliable communication between components
- Easy adaptation of the system for new purposes
- Easy portability to other hardware platforms

- Support of a defined development process
- Support for test and evaluation

For a long time the above requirements and principles have been neglected during the development of many prototype control systems. This is because the issue of software architecture design is not tightly coupled to pure robotic research. As a result, most of the research software is hard to maintain and to adapt and therefore lacks general usability. Fortunately, many of the best-practice principles and processes from the software development community are now widely accepted by the robotic research community. These principles are, amongst others, object-oriented design, the incorporation of well-known design patterns, the reuse of established libraries, the compliance to widely accepted standards, and the use of test and evaluation frameworks. This leads to higher quality and to improved adaptability of the software. Furthermore, a great pool of software frameworks for robotic research has been developed. Most of these frameworks can be used out of the shelf and fulfil most of the requirements proposed by robotic research.

Existing Frameworks for Mobile Robots

There exists a number of popular frameworks for mobile robot research. Some of them are more general and flexible than others while some of them are closely related to specific robots or tasks. Examples for such frameworks are: task control architecture (TCA) (Simmons, 1994), Saphira (Konolige & Myers, 1998), *Carnegie Mellon Robot Navigation Toolkit* (Carmen) (Montemerlo, Roy, & Thrun, 2003), and Player/Stage (Gerkey et. al, 2001). A very good overview and a more formal and detailed evaluation of existing frameworks is given in (Orebäck & Christensen, 2003). In the next section, we will discuss the framework named *Middleware for Robots* (Miro), on which our own developments are based.

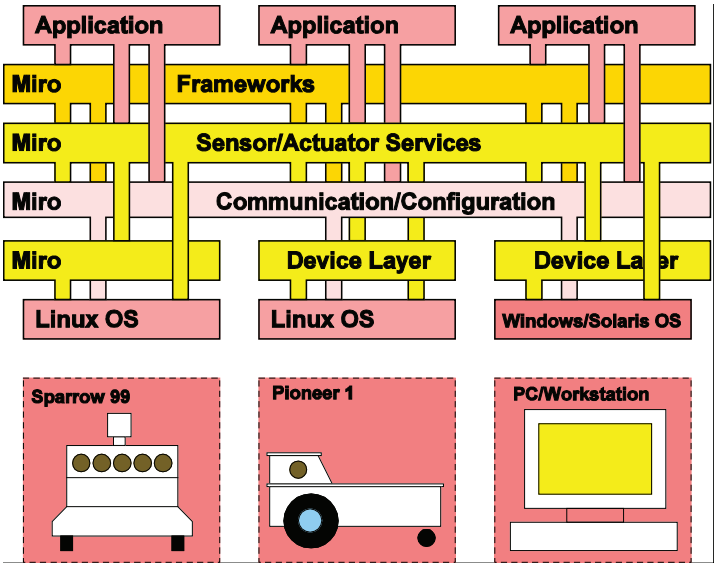
Miro is a distributed object-oriented software framework for robot applications. It has been developed at the Department of Computer Science at the University of Ulm (Utz, 2005; Utz, Sablatnüg, Enderle, & Kraetzschmar, 2002). The aim of the project is to provide an open and flexible software framework for applications on mobile robots. In general, Miro is a software framework and not a ready-to-use robotic application. The goals for the design of Miro are:

- Full object-oriented design
- Client/server system design
- Hardware and operating system abstraction
- Open architecture approach
- Multi-platform support, communication support, and interoperability
- Software design patterns
- Agent technology support

Miro achieves these goals by adopting an architecture that is divided into three layers. Figure 4 depicts the architecture of Miro. The use of the *Adaptive Communication Environment* (ACE) and CORBA for the communication between the layers and other applications enables a flexible, transparent, and platform-independent development. Miro uses *the ACE Object Request Broker* (TAO) (Schmidt, 2000) as CORBA framework. The implementation of this object-oriented framework is completely done in the C^{++} language.

The *Miro device layer* provides object-oriented interface abstractions for all sensory and actuatory facilities of a robot. This is the platform-dependent part of Miro. The *Miro communication and service layer* provides active service abstractions for sensors and actuators via CORBA *interface definition language* (IDL) descriptions and implements these services as network-transparent objects in a platform-independent manner. The programmer uses standard CORBA object protocols to interface to any device, either on the local or a remote robot. Miro provides a number

Figure 4. The architecture of Miro (Figure from Utz, 2005)



of often-used functional modules for mobile robot control, like modules for mapping, self-localization, behaviour generation, path planning, logging and visualization facilities.

A complete description of Miro and many useful examples can be found in The Miro Developer Team (2005). The use of Miro as a basis for further development has the following advantages:

- **Object-oriented design:** The design of the framework is fully object-oriented, elaborated, and easy to understand. Moreover, there are a whole bunch of ready-to-use design patterns that provide, for example, multi-threading, device reactors, and so forth.
- **Multi-platform support and reuse:** Miro comprises a great number of abstract interfaces for numerous different sensors and actors, for example, odometry, bumper, sonar, laser, and differential drives. Moreover, for all of these interfaces Miro already provides implementations for many different robot platforms. Due to the clear design and the use of CORBA and IDL, the implementation of interfaces for a new robot platform and the integration of new interfaces is straightforward. Miro currently supports many different common robot platforms like the B21, the Pioneer family, and some RoboCup MSL (Middle-Size League) robots.
- **Communication:** For the communication between different components of the robot control software, Miro provides two main mechanisms:

Direct CORBA method calls are used in a client/server manner for a 1-1 communication of components. This mechanism is commonly applied to actor and sensor interfaces. Due to the use of CORBA, the user does not have to deal with the internals of such a communication, for example, marshalling or memory management. Furthermore, the communication is completely transparent even if the client and the server run on different computers or use different programming languages.

The *event channel*, on the other hand, provides 1-n communication. The event channel follows the producer/consumer paradigm. The producer

simply pushes an event of a certain type to the channel. All consumers who are subscribed for this event are automatically informed when this specific event is available. Although this mechanism has a lot of advantages, it has to be mentioned that a heavy use of this mechanism leads to a poor run-time performance because of the computational overhead in the event channel.

- **Behaviour engine:** Miro contains a complete module for the modeling and the implementation of reactive behaviours. The *behaviour engine* follows the behavioural control paradigm introduced by Brooks (1986). The module uses a hierarchical decomposition of behaviours. On the base of the hierarchy, there are different atomic *behaviors* like, for example, "dribble ball" for a soccer robot. These behaviors can be grouped in *action patterns*. Such action patterns may be comprised of, for example, a dribble action and a local obstacle avoidance behavior. Different action patterns can be combined to a *policy*.

Once the behaviours are implemented, action patterns and policies are built up by describing them in an XML-file. Therefore, experiments with different action patterns and policies are easy and straightforward.

Unfortunately, Miro does not provide any paradigms and implementations for a deliberative layer. Therefore, extension of the Miro framework by a planning system is described in the third section.

A MODULAR ARCHITECTURE FOR AUTONOMOUS MULTI-PURPOSE ROBOTS

In order to fulfil as many of the requirements for an appropriate framework stated in the second part of the second section as possible, a novel design approach for mobile autonomous robots was developed (Fraser, Steinbauer, & Wotawa, 2004). The approach is based on a consequent modularization of both the robot's software and its hardware.

It is possible to distinguish between the functional view on the software and the software architecture. The functional view provides a decomposition of functionality into layers with increasing levels of abstraction. Therefore, the functionality is organized into different layers ranging from an abstract top layer with planning and reasoning capabilities down to a layer with direct hardware access. The software architecture represents a physical view on the software system.

The framework presented in this section is built upon the Miro framework and was deployed and evaluated in the RoboCup middle-size league team (MSL) of the Graz University of Technology. Besides using the robots for soccer games, the robots are also used for research in the area of service robotics.

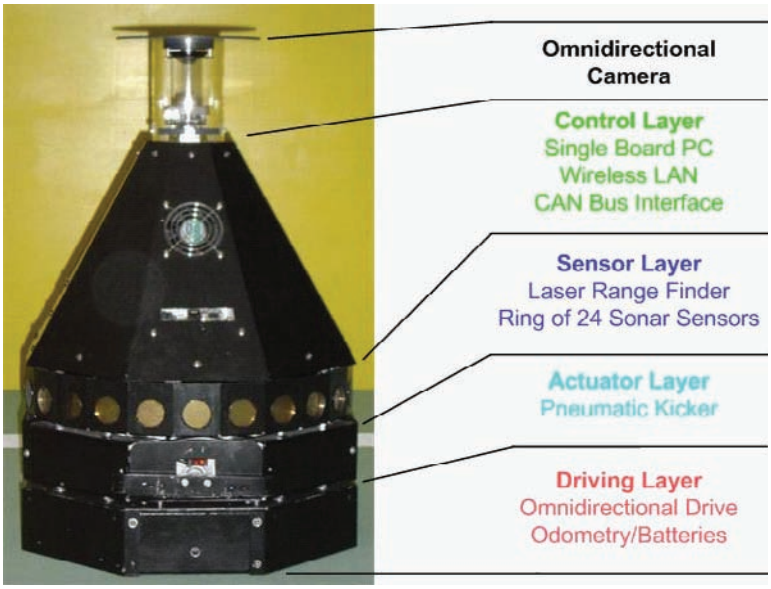
Figure 5 shows the modularized robot platform, which was developed and used as a target system for the proposed control architecture. The hardware of the robot is divided into five layers. Each layer provides one particular skill. The layers are stacked to build up the robot platform.

Functional View on the Software

The functional view on the software is guided by a consequent modularization. The functionality of the software is divided into three layers with an increasing level of abstraction. The layers are shown in Figure 6. Note that the functionality of a layer fully relies on the layer below.

The idea of functional layers with different levels of abstraction is similar to the idea of cognitive robotics (Castelpietra, Guidotti, Iocchi, Nardi, & Rosati, 2002). As mentioned above, a combination of reactive behaviors, explicit knowledge representation, planning, and reason-

Figure 5. The modularized robot platform



ing capabilities promises to be more flexible and robust. Furthermore, such an approach will be able to fulfil far more complex tasks. Note that this functional design is inspired by the hybrid control paradigm.

Hardware Layer

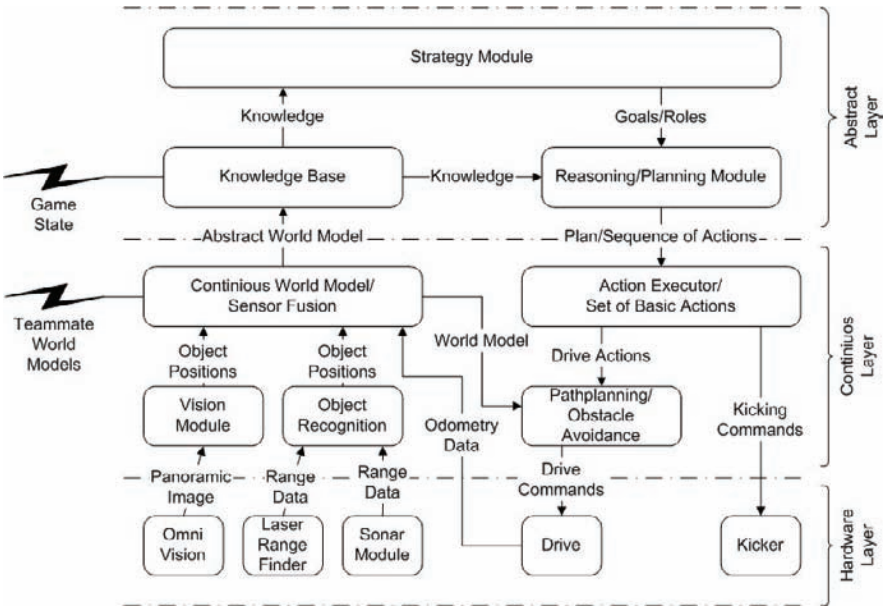
The hardware layer implements the interfaces to the sensors and actuators of the robot. This layer delivers raw continuous sensory data to the next layer and performs a low-level control of the actuators.

Continuous Layer

The continuous layer implements a numerical representation (quantitative view of the world) of the sensing and acting of the robot. This layer performs the processing of range data and the image processing. This processing computes possible positions of objects in the environment, including the robot's own *pose*. A pose consists of position and orientation of an object. These positions, together with the motion information from the odometry, are fused into a continuous world model by Kalman Filters (other objects) (Dietl, Gutmann, & Nebel, 2001) or Monte Carlo methods (own pose) (Fox, Burgard, Dellaert, & Thrun, 1999). Of course, all sensing and acting of a real mobile robot is afflicted with uncertainty. Therefore, sensor fusion is done using the above probabilistic methods. The world model represents the continuous world by estimating the most likely hypothesis for the positions of objects and the position of the robot itself.

Furthermore, this layer is responsible for the low-level execution of actions. Execution is based on a set of actions implemented as patterns of prioritized simple reactive behaviors. To take an example from robot soccer, suppose that the abstract layer (see next section) chooses to execute the high-level action DribbleBall. This action could be implemented by the following reactive behaviors: *dribble*, *kick*, and *obstacle_avoidance* (*oa*). *Dribble* will have the lowest priority, *oa* the highest. That is, if an obstacle is detected, only *oa* is executed in order to avoid a collision with the

Figure 6. Functional view of the software (robot soccer example)



obstacle. However, as long as there is no obstacle, this behavior will be inactive, and so behaviors with lower priorities can be executed. *Kick* is inactive most of the time; it becomes active only when the robot is in a proper distance and angle to the opponent goal. Then this behavior will emit a kick command, which is sent to the hardware layer, and become inactive again. *Dribble* is executed at those times when the other behaviors are inactive. This action execution forms the reactive part of the hybrid control paradigm and is similar to Brooks' subsumption architecture (Brooks, 1986).

Abstract Layer

The abstract layer contains a symbolic representation (qualitative view of the world) about the knowledge of the robot and a planning module for the decision making. A detailed description of the planning system can be found in Fraser et al. (2005). A similar approach also has been proven to work in the RoboCup MSL domain (Dylla, Ferrein, A., Lakemeyer, 2002). The abstract layer allows for an easy implementation of a desired task by specifying the goals, actions, and knowledge as logical sentences.

The core of this layer is the knowledge base. It contains a symbolic representation of the entire high-level knowledge of the robot. This knowledge consists of predefined domain knowledge, of a qualitative world model, which is an abstracted representation of the continuous world model, and of an abstract description of the actions the robot is able to perform.

The qualitative world model is represented by a set of logical propositions. The knowledge about actions is represented using a STRIPS-like representation language (Fikes & Nilsson, 1972) enriched by quantifiers from first-order logic. That is, an action is described by means of a precondition and an effect, where precondition and effect are represented by conjunctions of logical propositions. An action can be executed only if its precondition holds, and the effect states which conditions hold after the action has been executed successfully. Note that those propositions that are not included in the effect are not changed by the action.

Based on the agent's domain knowledge, the strategy module chooses the next goal the robot

has to achieve for fulfilling the long-term task. In a simplified view, the strategic knowledge could be regarded as a set of condition-goal pairs, where the condition is a logical sentence, and the goal can only be chosen if this condition is fulfilled. The planning module generates a plan that is supposed to achieve this goal. Any planning algorithm can be used within the planning module. Currently a simple regression planner (Russell & Norvig, 2003) or the more effective Graphplan is used (Blum & Furst, 1995). It has to be mentioned that the use of planning suffers from two drawbacks. First, planning takes time. Decisions that are made by planning are not feasible for time-critical tasks where a tight schedule have to be maintained. Furthermore, planning needs a significant amount of computational resources. Therefore, planning is only feasible for systems with enough resources or for systems where the time needed for the decision making is not that important.

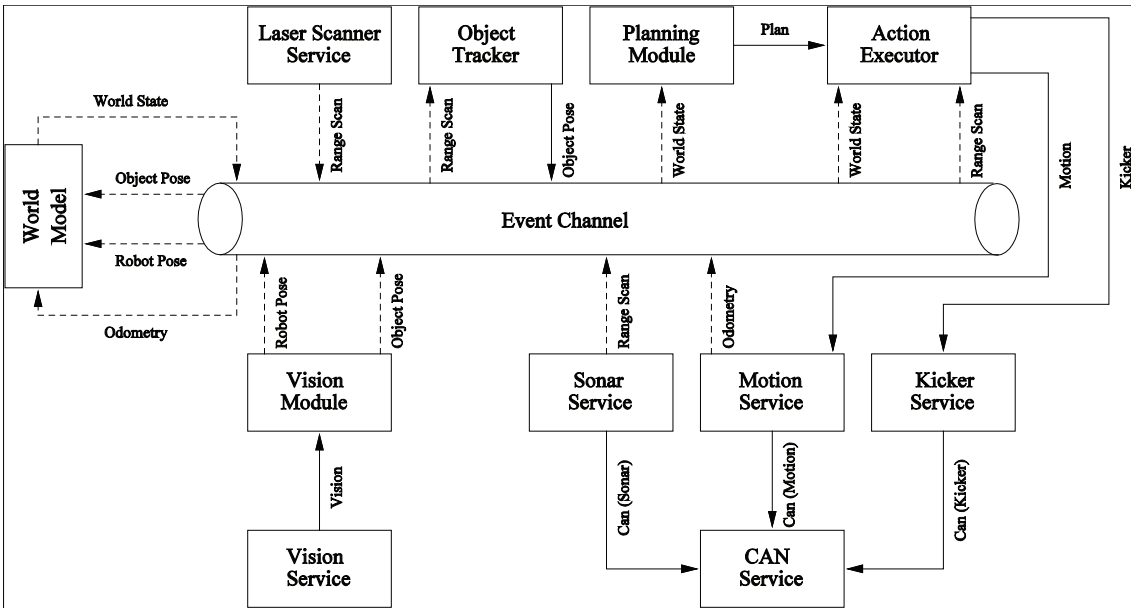
The calculated plan is communicated to the action executor, which implements the actions by means of simple behaviors. Note that a plan is monitored permanently for its validity during execution. The plan is canceled or updated if preconditions or invariants of the plan or its actions are no longer valid. This concept will be explained below.

Software Architecture

The software architecture, which is based on Miro (Utz et al., 2002), is shown in Figure 7. All software modules are implemented as autonomous services. Each service runs as an independent task. The communication between services primarily employs two mechanisms, the CORBA-interfaces and the event channel, which were described in the second section.

Hence, the services are independent of each other and an adaptation of software modules or the integration of new services is very easy and transparent to the rest of the system. It has to be mentioned that the event channel has a major drawback beside its advantages. The problem is that the Event channel needs significantly more

Figure 7. Software architecture. Solid connections represent remote invocations to CORBA IDL interfaces. Dashed connections denote an event-based communication.



computational resources than, for instance, a simple CORBA-interface. This fact is caused by its more complex communication mechanism and results in greater communication delays. Therefore, the use of the event channel is limited to only the necessary communications. Furthermore, time-critical communications like the connection between the action-execution with its behaviours and actuators are implemented by the faster CORBA method-calls.

ROBUST PLAN EXECUTION

In this section, an extension to hybrid architectures is presented that improves the quality of the plan execution in unpredictable and dynamic environments. First, it is illustrated how a plan is created, which fulfils a chosen goal. Then issues related to the plan execution and monitoring are discussed. Finally, it is shown how a quantitative world model can be transformed to a qualitative symbolic representation and which problems arise. Examples in this section are based on a RoboCup scenario, which is the original intent of the architecture presented in the previous section.

A Simple Example

Suppose that the following predicates are used for a qualitative representation of the world:

InReach(*x*)	true iff object *x* is in reach, i.e., within a small distance
KickDistance(*x*)	true iff the object *x* is close enough to be reachable by a ball which is kicked by this robot
HasBall	true iff the robot has the ball
IsAt(*x*,*y*)	true iff object *x* is at position *y*

where *x* and *y* can be one of the object constants *Ball* or *OppGoal* (opponent goal).

A classical AI-planning problem (Fikes & Nilsson, 1972) is defined by an initial state, a goal state, and the available actions. In the example, the initial state refers to the current state of the world, that is, the state of the environment and the robot itself. Suppose that initially all predicates are false and that the agent wants to achieve that a goal is scored. Furthermore, the knowledge base contains four high-level actions described in first-order logic, which the robot is able to perform.

1. Initial state $s_{init} = \neg InReach(Ball) \wedge \neg InReach(OppGoal) \wedge \neg HasBall \wedge \neg \ldots$
2. Goal state $g = IsAt(Ball,OppGoal)$
3. Actions schemas: $\Lambda^A = \{GrabBall, MoveT, Score, DribbleTo\}$. The schema definitions are given below.

As already explained, an action is defined by a precondition and an effect. The precondition denotes which conditions must hold before the action can be started, and the effect states how the qualitative world state is changed by the action. Consider the following example actions, where *target* is a placeholder for object constants. See Box 1.

A planning algorithm (Weld, 1999) searches for a sequence of actions that fulfils a given goal. In this example, the algorithm can find the following plan that achieves the goal *IsAt*(*Ball*,*OppGoal*):

1. *MoveTo*(*Ball*)
2. *GrabBall*
3. *DribbleTo*(*OppGoal*)
4. *Score*

Figure 8 depicts the qualitative states that appear during the execution of this plan. The nodes of the graph represent states, while the labels of the edges are actions. Only those predicates that are true are shown in the states; all other predicates are false. The state on the top of the graph is the initial state, in which all predicates are false. The last state at the bottom satisfies the goal *g*. In the initial state, only the action *MoveTo*(*Ball*) can be executed, as the preconditions of the

Box 1. A set of example actions

action:	precondition:	effect:
GrabBall	$\neg$*HasBall* $\wedge$ *InReach*(*Ball*)	*HasBall*
MoveTo(*target*)	-	*InReach*(*target*)
Score	*HasBall* $\wedge$ *KickDistance*(*OppGoal*)	$\neg$*HasBall* $\wedge$ $\neg$*InReach*(*Ball*) $\wedge$ *IsAt*(Ball,*OppGoal*)
DribbleTo(*target*)	*HasBall*	*KickDistance*(*target*)

other actions are not fulfilled in this state. After the execution of this action, the ball is in reach. Thus, action *GrabBall* can begin. If this action succeeds, *DribbleTo*(OppGoal) can be executed, and finally *Score* is supposed to achieve the goal. Please note that the term "goal" is twofold in the robot soccer domain.

Plan Execution and Monitoring

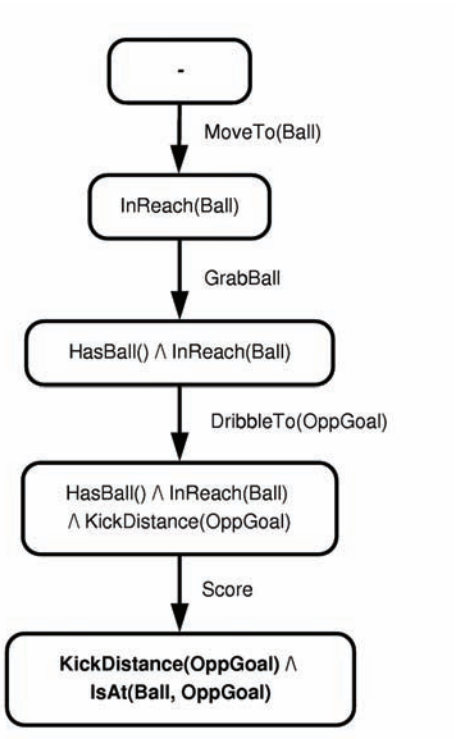
The classical AI planning theory makes some simplifying assumptions. Actions are considered atomic, their effects are deterministic, and they do always succeed. Furthermore, all world changes are caused by actions of this agent only, therefore the environment is fully deterministic. Finally, it is assumed that the environment is fully observable.

While these assumptions are convenient for the planning theory, they often do not hold in real-world applications. In practice, robots operate in dynamic and uncertain environments. Actions are not atomic, as their execution needs time, and unpredictable things can happen during the execution. Thus, actions may fail or their effect may be non-deterministic. Another issue is to determine when an action is finished, that is, when the effect is achieved. Moreover, the world is not only influenced by actions of this agent but also by other agents, for example, by other robots, by humans, or by other influential factors. In addition, due to the limitations of the sensors, the environment is not fully observable.

This discussion shows that permanent execution monitoring is required in real domains. The execution monitoring has to deal, among other things, with the following issues:

- Has the current action already achieved its effect?
- When is it necessary to abort the ongoing action due to unforeseen changes of the environment?
- Can the next action of the plan be executed in the current world state? It might be the

Figure 8. An example plan that aims at scoring a goal



case that its precondition is no longer fulfilled due to a change of the environment.

- Is the plan still valid? Under which conditions is it necessary to abort a plan?

In order to be able to perform a permanent monitoring of the plan execution, the qualitative world state has to be re-computed each time a continuous world state is created. After the computation of the truth values of the predicates, the preconditions and effects of the action in the plan can be re-evaluated. Action invariants are also used, that is, conditions that must hold as long as the action is executed. In most cases, the action invariant is equal to the action precondition. Let us come back to the example plan in Figure 8. Suppose the first action, *MoveTo*(*Ball*), has succeeded, and therefore the robot executes *GrabBall*. The invariant and the effect of this action are permanently re-evaluated.

Traditional approaches to plan execution and monitoring only take the preconditions of actions into account. In Fraser, Steinbauer, and Wotawa (2005), the use of plan invariants is proposed as a means to supervise plan execution. When a violation of the plan invariant is detected, the plan is aborted immediately. A plan invariant is a condition that refers to the plan as a whole and thus can include conditions that are not present in the action preconditions and invariants. The use of plan invariants is proposed in order to define conditions such that a violation indicates that it is no longer desirable to pursue this plan, although the plan is feasible according to the action preconditions and invariants, and its goal could be achieved. This allows the agent to adapt its behaviour to unforeseen changes in the environment.

In Fraser et al. (2005) the *extended planning problem* was defined, which extends the original definition in Fikes and Nilsson (1972). It consists of:

1. An initial state s_{init}
2. A goal state g
3. A set A that contains all possible actions
4. The plan invariant *inv*

A plan invariant can be defined for each goal and thus is assigned to plans that are created for this goal. In the example, one could define that the goal *IsAt*(*Ball*,*OppGoal*) should no longer be pursued when two or more players of the opponent team have moved between the ball and the opponent goal, no matter whether or not this robot already owns the ball. The idea is that in this case it is not reasonable to attempt to score a goal, as the chances of success are low, and that it may be better to change the tactics. Therefore, a new predicate *Blocked*(*Ball*,*OppGoal*) is introduced, and its negation is part of the plan invariant.

This robust method of plan execution has been deployed and successfully evaluated in the RoboCup middle-size league.

From the Quantitative World to a Symbolic Representation

As already explained, the perceptions from different sensors are fused in order to create the most probable continuous model of the real environment. The resulting model contains the robots' own pose and the positions of objects in the world.

This purely quantitative model is transformed to an abstract world model, which consists of logical predicates that are true or false. The example predicate *InReach*(*x*) has already been introduced, which is true if and only if the object *x* is within a certain distance of the robot. The truth value of a predicate is re-computed each time a new quantitative world model is available. For each type of predicate, an evaluation function is defined, which determines the truth value. For example, the value of *InReach*(*x*) is computed as follows:

```
COND_InReach(x, m): boolean
  return (dist(r, x) < t)
```

where *m* is the continuous world model, *r* is the robot, and *t* is a threshold.

This approach has, compared to reasoning based on continuous data, many advantages. Amongst others, a qualitative model has only a finite number of possible states, and qualitative models are implicitly able to cope with uncertain and incomplete knowledge. Another reason is the fact that the programming of the robot is simplified and can also be done by human operators who have no programming skills. The knowledge and the strategy of a robot can be neatly expressed in logical formulas on a more abstract level. Such programming appears more intuitive.

The mapping from a quantitative model to symbolic predicates in a dynamic and uncertain environment leads to two major problems: First, the truth value of predicates is calculated using thresholds, that is, there are sharp boundaries. Thus slight changes of the environment can cause truth value changes and result in abortion of plans due to a violation of the invariant, even if the plan still could be finished successfully. The consequence is instability in the high-level decision-making process. A longer-lasting commitment to a plan, once it is chosen, is desired. Second, sensor data is inherently noisy. Hence, due to the sharp boundaries, sensor noise leads to unstable knowledge, that is, to undesired oscillation of truth values, even if the environment does not change.

In Steinbauer, Weber, and Wotawa (2005a) a predicate hysteresis was proposed as an attempt to reduce the problems described above. The term hysteresis is well known from electrical engineering. It means that the current state is influenced by a decision that has been made previously. This concept was adapted in order to improve the robustness of the decision making process. The basic idea is that once a predicate evaluates to a certain truth value, only significant changes of the environment can cause a change of this truth value.

In an example, an improved evaluation function for *InReach*(*x*) is introduced:

```
COND_InReach(x, m, l): boolean
 if l then
    return (dist(r, x) < t + h
 else
    return (dist(r, x) < t – h
```

where the variable *l* represents the current truth value of *p* and *h* is the hysteresis size.

Such an evaluation function has the advantage that at the boundary only a significant change in the quantitative world causes a change of the truth value. This leads to a stabilization of the evaluation. However, it has to be mentioned that hysteresis is always a trade-off between stability and reactivity.

RUNTIME FAULT DETECTION AND REPAIR

Even if the control software of a mobile robot is carefully designed, implemented, and tested, there is always the possibility of faults in the system. Generally, faults are the deviation of the current behaviour of a system from its desired behaviour. Carlson and Murphy (2003) presented a quantitative evaluation of failures on mobile robots. The situation gets even worse if one thinks about autonomous robots, which operate for a long time without the possibility of human intervention, for example, nuclear inspection robots, space probes, or planetary rovers. Therefore, control architectures that are robust and fault-tolerant are crucial for truly autonomous robots.

Because faults are not totally avoidable, it is desirable that mobile robots are able to autonomously detect and repair such faults. When a permanent fault, that is, a broken hardware component, is identified, then the robot should

at least provide basic functionality or should be able to switch to a safe state. These requirements can be fulfilled if a dedicated diagnosis system is attached to the robot control software. Usually, a diagnosis system is comprised of three modules: (1) a monitoring module, (2) a fault detection and localization module, and (3) a repair module. The first module observes the actual behaviour of the hardware and software of the robot system. The fault detection uses observations and a model of the system's desired behaviour to detect deviations between them. A deviation is equivalent to a detected fault. However, in practice the detection of a fault is not enough. The module should also identify the hardware or software component that caused the fault. If a fault and its location are identified, the repair module tries to resolve the fault. This could happen by a restart or reconfiguration of the affected components.

There are many proposed and implemented approaches for fault diagnosis and repair in autonomous systems. The Livingstone architecture by Williams, Muscettola, Nayak, and Pell (1998) was used on the space probe *Deep Space One* in order to detect failures in the probe's hardware and to recover from them. The fault detection and recovery is based on model-based reasoning. Model-based reasoning uses a logic-based formulation of the system model and the observations. Verma, Gordon, Simmons, and Thrun (2004) used particle filter techniques to estimate the state of the robot and its environment. These estimations together with a dynamic model of the robot were used to detect faults. Rule-based approaches were proposed by Murphy and Hershberger (1996) to detect failures in sensing and to recover from them. Additional sensor information was used to generate and test hypotheses to explain symptoms resulting from sensing failures. Roumeliotis, Sukhatme, and Bekey (1998) used a bank of Kalman filters to track specific failure models and the nominal model.

In this section, a solution for real-time fault diagnosis and repair of control software of autonomous robots is presented. The fault diagnosis follows the model-based diagnosis paradigm (Reiter, 1987). It is based on observations of the current behaviour of the control system's components, a model of the desired behaviour of the control system's components, and the dependencies between them. A monitoring module continuously observes the behaviour of the control software. If a deviation of the desired behaviour is observed, a diagnosis kernel derives a diagnosis, that is, a set of malfunctioning software components explaining the deviation. Based on this diagnosis and on a model of the software components and their connections, a repair module executes appropriate repair actions to recover the system from the fault. The proposed diagnosis system has been implemented and tested as part of the proposed control architecture on RoboCup MSL robots within the robotic soccer scenario (Steinbauer & Wotawa, 2005).

Robot Control Software

Figure 9 shows an overview of the robot control software framework, the features of which have been presented in the preceding sections. In Figure 9, remote CORBA calls are shown as solid lines directed to the server. The data-flow between server and client is shown as chain dotted lines. The figure also shows the dependencies between services. Remote CORBA calls are called *strongly dependent* because a fault in the server directly affects the client. Connections using the event channel are shown as dashed lines, and the data-flow direction is always directed from the producer to the consumer. Connections that rely on this communication mechanism impose a *weak dependency*, that is, the services are loosely coupled and a failure of the server does not cause a failure of the client. The distinction between strong and weak dependencies is later important for the diagnosis and repair process.

The above described structure of the control software, the different types of connections, and

the dependencies between the services are used to build a model of the desired behaviour of the control software. In the next section it is described how this model together with various observations of the behaviour of services and connections are combined to form a diagnosis system for the control software of the robots.

Diagnosis System

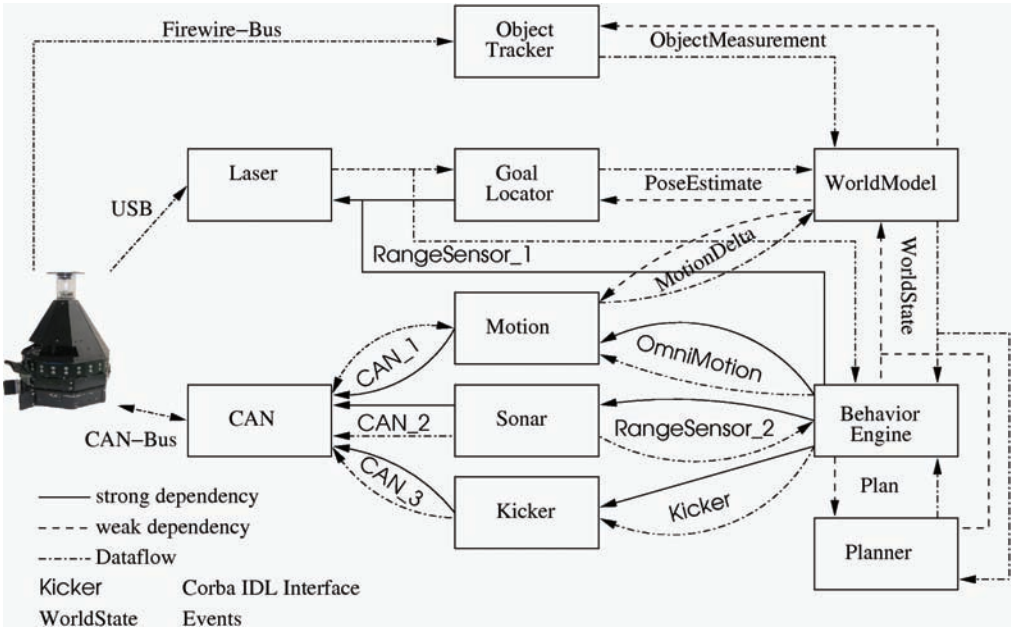
Monitoring

The task of the monitoring module of the diagnosis system is to observe the actual behaviour of the control system. For this purpose, the concept of observers was introduced. An observer monitors the behaviour of a service or the communication between services. The observer determines a misbehaviour of the control system if the observed behaviour deviates from the specified behaviour. In the current implementation, the following observers are used:

- **Periodic event production:** This observer checks whether a specific event *e* is at least produced every *n* milliseconds.
- **Conditional event production:** This observer checks whether an event e_1 is produced within *n* milliseconds after an event e_2 occurred.
- **Periodic method calls:** This observer checks whether a service calls a remote method *m* at least every *n* milliseconds.
- **Spawn processes:** This observer checks whether a service spawns at least *n* threads.

There are several requirements for the monitoring module. First, if observers are used, there should be no, or at least only a minimum necessity for, changes in the control system. Furthermore, the monitoring component should not significantly reduce the overall performance of the control system. Both requirements can be met easily by using mechanisms provided by CORBA (Hen-

Figure 9. Dependencies and data-flow within the robot control software



ning & Vinoski, 1999) and the Linux operating system (OS). The first two observers are implemented using the CORBA event channel. The third observer is implemented using the CORBA portable interceptor pattern. The last observer is implemented using the information provided by the *proc* file-system of the Linux OS. For all these observers, no changes are necessary in the control system. Furthermore, the computational power requirements for all the observers are negligible.

Diagnosis

A fault is detected if an observer belonging to the monitoring module recognizes a deviation between the actual and the specified behaviour of the system. However, so far one does not know which misbehaving service causes this fault. The model-based diagnosis (MBD) paradigm (de Kleer & Williams, 1987; Reiter, 1987) is used to locate the faulty service.

Figure 10 shows an example for the diagnosis process in case of a malfunctioning CAN-Service. First, an abstract model of the correct behaviour of the CAN sonar and motion service is built. Therefore, two predicates are introduced: $AB(x)$ becomes true if a service x is abnormal, meaning x is malfunctioning. $Ok(y)$ becomes true if a connection y, either a remote call or an event, shows a correct behaviour. The model for the correct behaviour of the example could contain the following clauses:

1. $\neg AB(CAN) \rightarrow ok(CAN\_1)$
2. $\neg AB(CAN) \rightarrow ok(CAN\_2)$
3. $\neg AB(Sonar) \wedge ok(CAN\_1) \rightarrow ok(RangeSensor\_2)$
4. $\neg AB(Motion) \wedge ok(CAN\_2) \rightarrow ok(MotionDelta)$

The principles of MBD will be explained with a simple example. Lines 1 and 2 specify that, if the CAN-Service works correctly, then also the connections *CAN_1* and *CAN_2* work correctly. Line 3 specifies that if the sonar service and its input connection *CAN_1* work correctly, the connection *RangeSensor_2* has to show a correct behaviour. Line 4 specifies similar facts for the motion service.

If there is a deadlock in the CAN-Service, the motion and sonar services cannot provide new events or calls, as they get no more data from CAN. This fact is recognized by the corresponding observers and can be expressed by the clause: $\neg ok(RangeSensor\_2) \wedge \neg ok(MotionDelta)$. If a correct behaviour of the system is assumed (expressed by the clause $\neg AB(CAN) \wedge \neg AB(Sonar) \wedge \neg AB(Motion)$, then a logical contradiction is observed. This means a fault was detected.

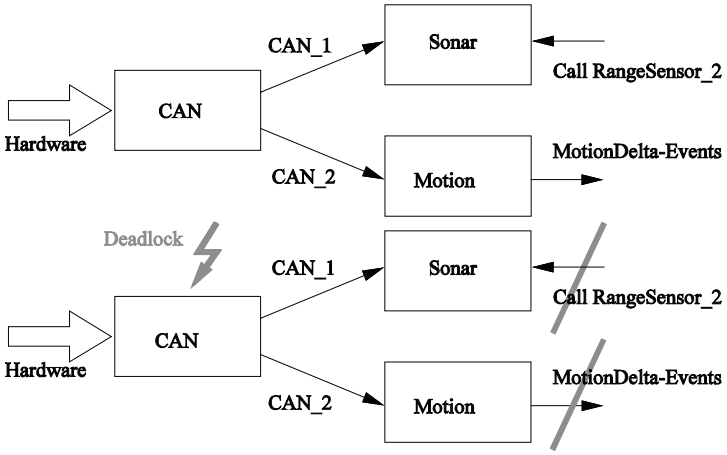
Finding the service that caused the fault is equivalent to finding the set of predicates $AB(x)$ with $x \in \{CAN,Motion,Sonar\}$ that resolves the contradiction. These sets are called *diagnoses*, Δ. Sets with minimal cardinality are preferred, that is, a single faulty service. These diagnoses are in general sufficient as multiple faults are unlikely. In this example, the set $\{AB(CAN)\}$ with only one element is able to resolve the contradiction. Therefore, the faulty CAN-Service is located.

Repair

Once the diagnosis system has found one or more malfunctioning services responsible for a detected fault, it should be able to recover the control system from this fault. Therefore, the repair module determines an appropriate repair action based on the described diagnosis and the dependencies between the services. The repair action consists of a stop and a restart of the malfunctioning services.

The appropriate repair action is derived in the following way: Put all members of the diagnosis D in a set R. The members of this set R are scheduled for restart. In the next step insert all services into R, which strongly depend on a member of R. Repeat this step until no more services are

Figure 10. Diagnosis of a fault in the Can-Service. The upper figure shows the desired behaviour. The lower figure shows the behaviour after a deadlock in the Can-Service.



added to *R*. *R* now contains all services which that to be restarted. But first of all the scheduled services have to be stopped in an ordered way. This means to first stop all services that no other service strongly depends on. Afterward, stop all services for which no more services are running that depend on them. This process is necessary to avoid additional crashes of services caused by a sudden stop of a service another service depends on. Hereafter, restart all affected services in the reverse order. Services that were restarted because of a strong dependency on the malfunctioning services should be able to retain data that it had gained so far or they should be able to recover such data after a restart. Otherwise, the control system may become inconsistent. After this repair action took place, the robots control system is again in the desired state.

CONCLUSION

In this work, a successful control architecture for autonomous mobile robots was presented. The proposed architecture is comprised of four major parts that give the architecture the power to control an autonomous mobile robot in complex tasks in difficult environments while still maintaining a great amount of flexibility. The four parts answer the questions that were posed in the introduction:

- **Control paradigm:** It has been motivated that only a hybrid control paradigm that combines the advantages of reactive and deliberative control enables a robot to perform complex tasks robustly. A three-layered paradigm was used to organize the control of the presented architecture.
- **Framework:** The experiences in building a mobile robot have shown that the adoption of modular frameworks for software and hardware can significantly reduce the development time and costs. At the same time, it increases the flexibility and robustness of the robot. By using this design approach, it was possible to develop four soccer robots from scratch with limited human and financial resources within less than one year.
  The quality and robustness of the robots were demonstrated during a number of RoboCup tournaments where the hardware and software of the robots operated in a stable manner. Moreover, the adaptability and

flexibility of the proposed control solution were impressive. For example, a RoboCup player's strategy could be modified within a few minutes on the field by simply adapting the knowledge base; no re-compilation of the software was necessary. Furthermore, the same framework is also applied successfully for the control of a delivery robot within the office domain.

However, the framework suffers from two major drawbacks. For one, the extensive use of the event channel as a flexible communication mechanism slows down the system and reduces the reactivity of the robot in dynamic environments. In addition, the system's reactivity is limited because of the incorporated classical AI planning techniques, which have a high computational complexity. However, this is necessary because only due to this abstract decision making module is the robot able to deal with very complex tasks. This drawback can be reduced by the increasing power of the used on-board computers and further achievements in the research on AI planning.

- **Robust plan execution:** The deliberative components have been enriched by a robust plan execution and symbol grounding mechanism. These extensions enhance the deliberative component, which plays a crucial part within an architecture aiming at the execution of complex tasks. The enhancements increase the applicability for unpredictable, noisy, and dynamic environments. Plan invariants and a hysteresis-based symbol grounding are used to achieve this.
- **Runtime fault diagnosis and repair:** The presented diagnosis system is capable of real-time fault detection, localization, and repair for the control software of autonomous mobile robots. The proposed system follows the model-based diagnosis paradigm. It uses a general abstract model of the correct behaviour of the control system together with observations of the actual behaviour of the system to detect and localize faults in the software. Furthermore, a repair method was presented that is able to recover the software from a fault. Because of its general methods, the proposed system is also applicable to software other than robot control software.

The presented overall control architecture solves a couple of problems that arise from the deployment of autonomous mobile robots in complex tasks in dynamic and unpredictable domains. It is an excellent example for a general flexible architecture for robust intelligent control.

ACKNOWLEDGMENT

This research has been funded in part by the Austrian Science Fund (FWF) under grant P17963-N04.

REFERENCES

Blum, A., & Merrick, F. (1995). Fast planning through planning graph analysis. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95)* (pp. 1636-1642).

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation, RA-2*(1), 14-23.

Carlson, J., & Murphy, R. R. (2003). Reliability analysis of mobile robot. *Proceedings of the 2003 IEEE International Conference on Robotics and Automation, ICRA 2003.* Taipei, Taiwan.

Castelpietra, C., Guidotti, A., Iocchi, L., Nardi, D., & Rosati, R. (2002). Design and implemantation of cognitive soccer robots. In *RoboCup 2001: Robot soccer World Cup V,* vol. 2377 of *Lecture Notes in Computer Science*. Springer.

de Kleer, J., & Williams, B. C. (1987). Diagnosing multiple faults. *Artificial Intelligence, 32*(1), 97-130.

Dietl, M., Gutmann, J. S., & Nebel, B. (2001). Cooperative sensing in dynamic environments. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'01),* Maui, HI.

Dylla, F., Ferrein, A., & Lakemeyer, G. (2002). Acting and deliberating using golog in robotic soccer - A hybrid approach. *Proceedings of the 3rd International Cognitive Robotics Workshop (CogRob 2002).* AAAI Press.

Fikes, R., & Nilsson, N. (1972). Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence, 2*(3-4), 189-208.

Fox, D., Burgard, W., Dellaert, F., & Thrun, S. (1999). Monte Carlo localization: Efficient position estimation for mobile robots. In *AAAI/IAAI*, 343-349.

Fraser, G., Steinbauer, G., & Wotawa, F. (2004). A modular architecture for a multi-purpose mobile robot. In *Innovations in Applied Artificial Intelligence, IEA/AIE, vol. 3029, Lecture notes in artificial intelligence* (pp. 1007-1014). Ottawa, Canada: Springer.

Fraser, G., Steinbauer, G., & Wotawa, F. (2005). Plan execution in dynamic environments. *Proceedings of the 18th Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE, LNAI 3029,* (pp. 208-217). Bari, Italy: Springer.

Gerkey, B. P., Vaughan, R. T., Stroy, K., Howard, A., Sukhatme, G. S., & Mataric, M. J. (2001). Most valuable player: A robot device server for distributed control. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2001)* (pp. 1226-1231).

Henning, M., & Vinoski, S. (1999). *Advanced CORBA©Programming with C++* (1st ed.). Addison Wesley Professional.

Kortenkamp, D., Bonasso, R. P., & Murphy, R. (Ed.). (1998). *Artificial intelligence and mobile robots. Case studies of successful robot systems.* MIT Press.

Montemerlo, M., Roy, N., & Thrun, S. (2003). Perspectives on standardization in mobile robot programming: The Carnegie Mellon navigation (CARMEN) toolkit. *Proceedings of the Conference on Intelligent Robots and Systems (IROS).*

Murphy, R. R. (2002). *Introduction to AI Robotics.* MIT Press.

Murphy, R. R., & Hershberger, D. (1996). Classifying and recovering from sensing failures in autonomous mobile robots. *AAAI/IAAI, 2*, 922-929.

Nilsson, N. (1984). *Shakey the robot.* (Tech. Rep. No. 325). Menlo Park, CA: SRI International.

Orebäck, A. (2004). *A component framework for autonomous mobile robots.* Unpublished doctoral thesis, KTH numerical and computer science.

Orebäck, A., & Christensen, H. I. (2003). Evaluation of architectures for mobile robotics. *Autonomous Robots, 14*, 33-49.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence, 32*(1), 57-95.

Roumeliotis, S. I., Sukhatme, G. S., & Bekey G. A. (1998). Sensor fault detection and identification in a mobile robot. *Proceedings of IEEE Conference on Intelligent Robots and Systems* (pp. 1383-1388).

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Schmidt, D. C. (2000). *TAO developer's guide* (1.1a ed.). Object Computing Inc.

Simmons, R. (1994). Structured control for autonomous robots. *IEEE Transactions of Robotics and Automation, 10*, 34-43.

Steinbauer, G., Weber, J., & Wotawa, F. (2005a). From the real-world to its qualitative representation - Practical lessons learned. *Proceedings of the 18th International Workshop on Qualitative Reasoning (QR-05)* (pp. 186-191).

Steinbauer, G., & Wotawa, F. (2005b). Detecting and locating faults in the control software of autonomous mobile robots. *Proceedings of the 16th International Workshop on Principles of Diagnosis (DX-05)* (pp. 13-18).

Steinbauer, G., & Wotawa, F. (2005c). Detecting and locating faults in the control software of autonomous mobile robots. *Proceedings of 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, Edinburgh, UK.

The Miro Developer Team (2005). *Miro Manual, 0.9.4 ed*. University of Ulm: Department of computer science.

Utz, H. (2005). *Advanced Software Concepts and Technologies for Autonomous Mobile Robotics*. Unpublished doctoral dissertation, University of Ulm.

Utz, H., Sablatnüg, S., Enderle, S., & Kraetzschmar, G. (2002). Miro: Middleware for mobile robot applications [Special issue]. *IEEE Transactions on Robotics and Automation, 18*(4), 493-497.

Verma, V., Gordon, G., Simmons, R., & Thrun, S. (2004). Real-time fault diagnosis. *IEEE Robotics & Automation Magazine, 11*(2), 56-66.

Weld, D. S. (1999). Recent advances in AI planning. *AI Magazine, 20*(2), 93-123.

Williams, B. C., Muscettola, N., Nayak, P. P., & Pell, B. (1998). Remote agent: To boldly go where no AI system has gone before. *Artificial Intelligence, 103*(1-2), 5-48.

This work was previously published in Architectural Design of Multi-Agent Systems: Technologies and Techniques, edited by H. Lin , pp. 335-355, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.21
A Neural Network–Based Mobile Architecture for Mobile Agents

Anand Kuppuswami
University of Western Sydney, Australia

ABSTRACT

Wide area network (WAN) offers advantages like providing myriad services available on globally diversified computers with reasonably simple process. The ability to dynamically create networks offers the processing powers of various processors at our command. With the advent of protocols like SOAP and Web services, the consumption of services are more organized. In spite of various advances in communication techniques, the consumption of services through mobile gadgets is still only at the research level. The major impedances in implementing such systems on a mobile network are the latency factor, abrupt disconnection in service, lower bandwidth and minimal processing power. The mobile agent's paradigm proves to be an effective solution to various issues raised. It has received serious attention in the last decade and several systems based on this paradigm have been proposed and built. All such systems have been designed for a static network, where the service providers and the requestors are connected to the server on a permanent basis. This chapter presents a new framework of managing the mobile environment and the participating nodes with active intelligent migration. The functioning of the mobile agents in such a scenario is also presented.

INTRODUCTION

Mobile devices foster a new set of applications that are receiving enormous support in the global electronic community. This is primarily because of the ability of mobile devices to connect and reconnect with each other dynamically. The ability to create a network of devices "on the fly" can be adjudged as a major advantage of mobile gadgets as compared with the previous computer networks. The ability of being able to communicate and, more importantly, process information irrespective of time and place is a very promising feature offered by mobile technologies, as also ratified by Gray, Kotz, Nog, Rus, and Cybenko (1996). With the advent of the Internet and, especially, in communication, almost all actions related to

an application have been reduced to a few mouse clicks. As is obvious in our day-to-day usage, the Internet is a vast repository of information and services. Even though both Internet and related mobile technologies are very promising, their usage by business is still full of hurdles. For example, accessing a simple HTML page in a mobile gadget requires various configurations in them. This can be attributed to low latency, physical obstruction, and network connectivity of the mobile gadgets. The mobile gadgets have low memory, processing capabilities, and are prone to sudden failure (Jipping, 2002). These factors unveil the problem of client server architecture (Gray et al., 2002b). Often, having a well-thought-out networking architecture is crucial to successful usage of this technology. For example, the huge amount of data that is required to be transported due to various multimedia applications to the client's site for further processing is not possible without a good architecture. This created the need for remote working where code could migrate to a different machine, execute at the new machine, and return with the result. The whole concept of working remotely started with the message passing and remote procedure calls (RPCs) provided by Java (Birrell & Nelson, 1984). But their application was limited, as the client could use only those services provided by the server. If the requested service is not present, then the client has to make intermediate or "Bridge" function class in order to get to the final result. This process results in wastage of resources and bandwidth. As an alternative approach, small subprograms could be written and passed on to the service provider to execute locally. Network Command Language (NCL) (Meandzija, 1986), remote evaluation (REV) (Stamos & Gifford, 1990), and SUPRA-RPC (Stoyenko, 1994) employ this idea in their architecture. All these architectures had one major concern: the code, once migrated to a different machine, cannot remigrate at the end of execution. The architecture also lacked in active coordination between different programs. The absence of inter-communication protocols leads to poor performance of the system.

The novel approach of transportable programs (Gray, 1995; Cybenko, Gray, Wu, & Khrabrov, 1994) offers a promising solution for various issues raised. Transportable agents or mobile agents, as they are called now, are autonomous programs that can migrate from one machine to another machine in the network. By migrating to the machine having the resource, the agents have the advantage of working on site where the resource is present and also use the processor's power. This eliminates all the middleware that is required for transporting the data to the client's site. The mobile agent's paradigm provides an effective solution to the problem of low latency, poor interface, and bad network conditions (Gray et al., 1996). The middleware and the communication control mechanism form the major workload in a client-server architecture; by eliminating them, we can build a better working environment and increase the efficiency of the system. The code and state of code is migrated to another machine for resuming its execution. This also eliminates the interface required for service access. The fact that there is no need for permanent connection makes it very suitable to the mobile environment. The ad hoc client-server model is overridden by the peer-peer model which matures into grid computing, where the machines can act as client or server depending on the environment. The programmer is swayed away from traditional multi-tier architecture to grid computing (Lauvset, 2001).

The majority of the mobile agents present in the literature is designed for a static network architecture. Baring a few mobile agents (Cabri, Leonardi, & Zambonelli, 2002; Kendall, Krishna, Pathak, & Suresh, 1998), intelligence is not embedded into them. We present a new set of mobile agents which work in a volatile mobile environment. Embedded with intelligence, agents can act autonomously, collaborate with other agents, and reduce the

human intervention. The agents are configured to be adaptable, learn from their experience, and mature into experienced agents.

REVIEW OF MOBILE AGENTS AND THEIR LIMITATIONS

A new software architecture for building distributed applications using the mobile agent paradigm, which also supports cross-referencing between agents using Telescript (Dömel, 1996), has been proposed by General Magic. The Telescript agents can begin their execution in any machine, migrate to a different machine, and continue their execution, with the only prerequisite being that all hosts are Telescript enabled. Telescript is an object-oriented language that is a collection of hierarchically organised classes. The Agent class, which is inherited from the Process class, has a method "go", which performs the migration process. When an agent wants to migrate to another machine, it issues the command "go". The agent continues its execution from the command following the "go" command. A server at each site authenticates (Tardo & Valente, 1996) the agents and executes them. Requirements like special hardware and support for single language limits the global implementation.

Tacoma (Johansen, Renesse, & Schneider, 1996)—developed at the University of Tromso and the University of Cornell—are agents written in TCL/HORUS, which is a version of TCL where the HORUS is used for group communication and fault tolerance. They have rear guards which restart lost agents and have features like electronic cash for services. They have protection mechanisms which ensure that there are no runaway agents. Broker services for scheduling and directory services are also implemented. In spite of several advantages, the programmer has to explicitly capture the state information and does not support interruption of the execution of program, which limits its usability.

Agent TCL—D'Agents as it is now called (Gray, Cybenko, Kotz, Peterson, & Rus 2002a)—is a mobile agent system which attempts to strike a balance between the Tacoma and Telescript systems. It uses the flexible scripting language, TCL (Isaacson, 2001) and also provides support for other languages. Various migration and communication primitives help the programmer to code without worrying about the low-level migration protocol and also gives enough granularities for building basic blocks. For security purposes, it uses the SafeTCL (Levy, Ousterhout, & Welch, 1997) protocol. In the Agent TCL (Gray, 1998; Kotz et al., 1999) environment, the agents can migrate from one machine to another, create child agents for performing sub-tasks, access resources across networks, and communicate with other agents residing in local or remote machines. TIAS (Transportable Intelligent Agent System) (Harker, 1995) is an improved version of D'Agents, and offers several patches over existing systems.

A new series of itinerant agents (Chess, Grosof, Harison, Levine, & Parris, 1995) has been proposed which offers a secure environment for remote applications in public networks. Itinerant agents are programs which are dispatched from a source computer to roam among a set of networked computers until they accomplish the required task. The need for knowledge representation and verification is also discussed in their architecture. The Messenger-based operating system (Marzo, Murhimanya, & Harms, 1994) and some work on creating new mobile agents from existing systems (Brazier, Overeinder, Steen, & Wijngaards, 2002) have led to creating mutants like TOMAS (Transaction-Oriented Multi-Agent System) developed by Busetta and Ramamohanarao (1998).

With the working knowledge of various systems, and taking into account their inadequacies, we propose to build a new framework that offers a promising edge over the existing systems. The framework's working includes various modules, like knowledge building, pattern recognition, and intelligent migration. The architectures based on

mobile agents' paradigm vary with one another in several aspects. The process of comparison requires concrete understanding of the terms like desired degree of configurability, scalability, and customizability (Fuggetta, Picco, & Vigna, 1998). The lack of such standardization subdues any work on comparison.

MANAGING MOBILE ENVIRONMENT

There are several issues that have to be addressed before a mobile environment can be implemented. For example, current location of the mobile agent needs to be clearly identifiable. This is required in situations where some parameters need to communicate to the agent. A service that is published in the UDDI directories can be easily availed with the help of Web services (Yang, Hsieh, Lan, & Chung, 2005). In certain cases where the services are not published in the required format, there needs to be some brokering before the services can be consumed (Cybenko & Jiang, 1999). These could be achieved by those brokering, which works as the middleware. One of the most important aspects that needs to be considered is the current working environment. If the nodes in the network are volatile, then the process of migration gets intricate. As the nodes are volatile. the services also offered variegates. Presence of certain nodes may result in the system offering better service, and their absence may force one to look for the second best choice. In the current scenario, there needs to be an active broker, who advises the agent on its migration. As the service offered cannot be compared on a binary basis, there need to be "human aspects" embedded into the broker service.

We propose to utilize the agent tracking feature of architectures like Agent TCL and Telescript, where a central system keeps track of the active agents and their locations. A central server keeps tracks of the service providers. The use of knowledge base is also incorporated into our system. Depending on current requirement, intelligence may or may not be required. In situations where simple binary logics will not suffice, we propose to use advanced models like neural networks; probabilistic methods offer a promising solution.

Introduction to Mobile Environment

Figure 1 presents a sample mobile network where several nodes offering various services are collaborating with each other and the central server. The services are hexa-coded using keywords, which are then registered with the server. In the static scenario all the nodes are permanently connected to the server, whereas in the mobile environment, the nodes will be connected for a shorter time span. The number of nodes connected to the server will also fluctuate. To explain the working of a mobile environment, let us assume that we have eight nodes, of which six are connected to the server. The agents are in active collaboration mode, which could be productive or counter-productive depending on the process.

The static architecture is more like a simple market environment, where we have all the shops and we visit them to consume the services offered. In some cases, the set of current service providers do not offer the solution we need, and we have to settle with the level of services offered. A more generalised architecture would make the service providers mobile. The service providers "walk in and walk out" of the market anytime, at their will. This results in a market that would be highly volatile. There could be a set of nodes that remain connected and offer their services on a more permanent basis, but other nodes could be volatile. The presence or absence of the volatile nodes can have a vast impact on the functioning of the system.

Figure 1. A sample mobile network

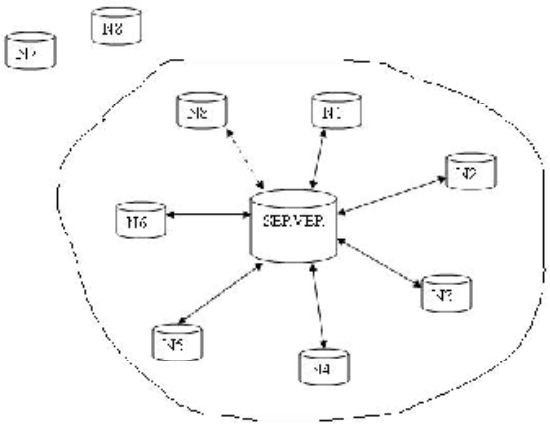


Figure 2. A sample service description header

```
<ROOT>
<START TIME>
12:15
</START_TIME>
<END_TIME>
      NA
</END_TIME>
<SERVICE>
      <NAME>
            IMAGE
      </NAME>
      <PI>
            14
      <PI>
</SERVICE>
<SERVICE>
      <NAME>
            PROCESS-
      ING
      <NAME>
<PI>
10
</PI>
</SERVICE>
<SERVICE>
      <NAME>
            NEURAL
      </NAME>
      <PI>
            11
      <PI>
</SERVICE>
<SERVICE>
      <NAME>
            NET-
      WORK
      <NAME>
<PI>
            8
</PI>
</SERVICE>
</ROOT>
```

Migration of Nodes

Each node has a header section called the Service Description Header (SDH), as shown in Figure 2, which describes the node in detail. The descriptions include the node name, the speed of connection, the quality of service factor, services offered, proximity index (PI), the operating system, and the language of coding. Whenever the node wants to enter the working environment, it sends a request to the server. The request would contain information like the START_TIME, END_TIME (if applicable), and the broad category of services. The PI would be a number ranging from 0 to 15, signifying how close the service is to the respective keyword. This would be in XML format for compatibility in heterogeneous environment.

Based on this request, the server assigns a unique ID to this node. In some cases this service may be redundant, and in that case server places the node in the queue. Once the node has been assigned an ID, the node prepares the detailed service description. The description will be in XML format and would have the structure as shown in Figure 3.

Similarly, whenever the node wants to exit the operating environment, it sends a termination request to the server. In case of termination, if any agent is being executed in that node, then the termination is postponed till the agent finishes the execution. Based on the START_TIME and the END_TIME, a latency factor and a quality of service factor are assigned to each node.

Figure 3. A sample service description in XML format

```
<ROOT>
<ID>
ABC123
</ID>
<OS>
UNIX
</OS>
<LANGUAGE>
      C++
<LANGUAGE>
<SERVICE>
      <NAME>
            IMAGE
      <NAME>
      <PI>
            14
      </PI>
<SERVICE>
<SERVICE>
      <NAME>
            PROCESSING
      <NAME>
      <PI>
            10
      <PI>
</SERVICE>
<SERVICE>
      <NAME>
            NEURAL
      <NAME>
      <PI>
<SERVICE>
<SERVICE>
      <NAME>
            NETWORK
      </NAME>
      <PI>
            8
      <PI>
</SERVICE>

<SERVICE>
      <NAME>
            HISTOGRAM
      <NAME>
      <PI>
            12
      </PI>
</SERVICE>
<SERVICE>
      <NAME>
      <PI>
            PATTERN
      <NAME>

RECOGNITION
      <NAME>
      <PI>
            0
      </PI>
</SERVICE>

<SERVICE>
      <NAME>

HIERARCHICAL
      </NAME>
      <PI>
            14
      <PI>
</SERVICE>
<SERVICE>
      <NAME>
            ADAPTIVE
      <NAME>
      <PI>
            15
      </PI>
</SERVICE>
<ROOT>
```

While assigning a new ID to the node entering the operating environment, the server checks for the existence of previous ID. If the node already had an ID, a new ID is not created and the old ID is retained. While assigning the new ID, the server checks the history for latency factor and the quality of service factor. If the values are lower than the values existing in the server's database, then the lower values are substituted; otherwise, the old values are retained. This process ensures that the quality or the latency factor is never hyped. Once the process of registration is complete, the node becomes an active member.

Challenges of Node Migration

In each instance, prior to the migration of agents between different nodes, a few factors must be considered in detail. Whenever the code migrates, it needs to take with it the working variables, so that it can continue the execution in the new environment. This brings up the question: How

much data needs to accompany the code? Let us consider a simple problem where a mobile agent is required to calculate the amount of tax that needs to be deducted from a salary. In this case, a numerical value is all that needs to accompany the code in the process of migration. This numerical value is enough for the successful execution of the code in the new environment. In this case the service provider has more information than the service requester. In a homogeneous environment, this process is further simplified. Now let us assume a situation in which the image format needs to be changed. For example, we have an image in BMP format and would need to convert it into JPEG format. Here, the mobile code has to transport the whole image to the site of service provider. This suggests that the migration of code would consume more resources than the service provider passing on the service. So we have to calculate the amount of data and the bandwidth of operation, and depending on those values, we decide on the migration.

A NEURAL NETWORK-BASED AGENT FOR INTELLIGENT MIGRATION

In the mobile environment, the nodes are volatile and hence the process of migration cannot be hard-coded. We need a run-time tool that can decide on the address of the future node. Each node varies on the service provided and has a varying level of PI associated with it. This perplexes the migration process. The migration also needs "shrewd attitude" when analysing the service factors. With these additional responsibilities thrust upon the mobile agent, the need for an advanced intelligent assistant arises. We propose an agent called the *broker agent* which performs the decision on migration of agents and acts as the "human intervention."

Request of Service and Broker Agent

When the mobile agent execution stack gets to the "go" command, the mobile agent creates a broker agent (BA). The purpose of the broker agent is to communicate with the server and identify the possible nodes which can offer the required service. The BA will communicate the agent's requirement to the server. This would result in the server identifying more than one node that can offer the required service. The BA must further drill down on the search result and track down the best node.

The list of the keywords and the service description of each node will result in varying system usability. In such a scenario a simple binary logic would not suffice. For example, if the service provider says "IMAGE EQUALIZATION" in the list of services provided and the agent is looking for "HISTOGRAM EQUALIZATION," then the simple binary logic will turn down the listed service. But in the image processing arena, both terms are used synonymously.

Ontology of Keywords and Network Architecture

As a first step, a list of keywords and their corresponding binary codes are prepared. This is presented to the neural network as inputs. The structure of the network is such that it can accommodate future unpredicted inputs. We would need multiple networks for performing multiple tasks. As each of these services does not have a maximum limit on number of keywords, we need a system which can handle this requirement. We suggest a network architecture of 16 input nodes which can accept an input between 0 to 15 for each node. This leads to an infinitely large number of a combinations of input values. The actual structure of the input is explained in the following section. Depending on the complexity of the service required, a *keyword list* (KL) is generated.

Table 1. A sample keyword list

Keyword	Comments
Image	Deals with image operations
Processing	Process images
Histogram	Performs histogram of the image
Equalization	Deals with histogram

A sample KL is shown in Table 1. The comments in the second column only aid the expert in his process of building the KL. This is not used in any part of the automated processing.

For example, let us assume that we need to perform histogram equalization. Then the list of keywords could be histogram, equalization, image, processing, and intensity. Each of them will become the input to the network. This suggests that we need five input nodes. This architecture will successfully work in an environment where the keywords are not emerging and fail otherwise.

To tackle this issue we propose to create agents having architecture which can cater for future keywords. The input to each node would be a value between 0 and 15. We also have decimal inputs which can help in getting accurate solutions. The sample input to the agent for training would be as in Table 2, where we have only shown whole numbers as the input, but decimal values are also accepted.

The input to the agent would depend on the proximity index for each service. For example, if the service provider has the keywords "image" and "processing", and their corresponding PI to be 15 and 8, this suggest that the input to the agent would be "F800 0000 0000 0000." We can map the service provider's PI to the agent input as shown in Table 2

Mapping Keywords

The process of mapping is complex and needs expertise at various stages. When mapping the PI for the known keywords, the process is relatively simple. The mapping complicates when the agent

Table 2. A sample input to the broker agent

Service Provider Keyword PI	Input to Neural Network
Image 15, Processing 8	F800 0000 0000 0000
Histogram 15, Processing 8	08F0 0000 0000 0000
Histogram 15, Equalization 10	00FA 0000 0000 0000
Image 15, Processing 15, Equalization 8	FF08 0000 0000 0000
Image 15, Processing 15, Histogram 15, Equalization 15	FFFF 0000 0000 0000

Table 3. Process of mapping

Keyword	Action
Photo	Mapped to Image
Function	Mapped to Processing
Normalization	New entry created
PCA	New entry created
Picture	Mapped to Image

encounters unknown keywords. An expert's knowledge would then be required to map the new keyword to either an existing one or to add the keyword to the KL. For example, if the service provider has "photo" as the keyword, then we can see that it is very closely related to the keyword "image". In this case, there would be no need to add a new keyword. Now let us suppose that the service provider has a keyword "PCA". Since it is not directly related to any of the existing keywords, there would be a need to alter the KL. A sample scenario is shown in Table 3.

Depending on the agent's service required, the KL may be a huge data repository in itself. This process of mapping cannot be automated at this stage of development.

Training of Network

Initially, when training the network only the five bits are used for training. The set of output values for each set of input has to be decided by an expert. This is a supervised training, which requires the knowledge and experience of a system expert. The output of the system can vary from –1 to +1, signifying how well the service provided suits the requirement. For example if the service provider has the keywords "image" and "histogram", then it has a better chance of catering to the needs when compared to a service which only has "image" as its service keyword. Depending on the PI for each keyword, the input value to the agent varies. A sample set of input and corresponding output can be formulated as shown in Table 4.

This input and output set is employed to train the network using the back propagation algorithm. After the successful completion of training, the neural network will be able to predict the outputs for unknown inputs. The node which evaluates to the highest value is selected as the service provider.

Table 4. A sample training set to the broker agent

Input to Neural Network	Output
F800 0000 0000 0000	+0.5
08F0 0000 0000 0000	+0.8
00FA 0000 0000 0000	+1.0
FF08 0000 0000 0000	+1.0
FFFF 0000 0000 0000	+1.0
F800 0000 0000 0000	-1.0

Process of Agent Learning and Maturing

After the initial phase of training, the agent is released to the network of nodes. The agent starts evaluating each node's service. The BA communicates with each node, gets its SDH, and extracts the keywords and their PIs. Depending on the keyword patterns, each node is evaluated for its service against the agent's requirements. The node having the highest match or the least conflict is selected as the service provider. In this process, if new keywords are encountered, then the agent would be unable to evaluate the performance of that node. In this case, new keywords need to be first mapped before the agent can process them. The keywords need to be mapped before the BA can proceed further. If there is no change in the KL, then without further processing the values are presented to the agent. If there is a change in the KL, the network needs to be retrained using the new sets of input. This is the process where the agent starts learning and maturing. After a series of training, the agent would have matured enough, and would not need further training and can be fully used. This process can be compared to a new staff member in an office who needs a few sessions of training before he can act independently.

IMPLEMENTATION OF INTELLIGENT MOBILE AGENTS

The system is being implemented in the .NET environment using C#, IIS as the Web server, XML, and SOAP protocol for communication. The project has been divided into four phases. In Phase I, the service request module is developed. In Phase II, the mobile environment is developed. In Phase III, the intelligent agents are developed, and in Phase IV, all the modules are integrated and fully implemented. Here we are trying to develop an application where a mobile client captures the image of a person and wants to find more information about him. This would be the case where a security person on his regular rounds finds a suspect but is not sure. In this situation we cannot expect a permanent connectivity with the server. So the ideal situation would be to capture the image and send it to the ground station for further processing. In a more generalised scenario, we can say that any person can capture the image and pass it on to the government agency for further scrutiny. Here there would be a series of activities that need to be performed before the information can be retrieved. The list of activities could range from capturing of image, communicating with the ground equipment, creating the broker agent, selecting the node for service, performing the migration, and processing the image.

All of the activities would have a list of tasks that needs to be completed. For example, image processing would need to perform tasks like image clarity improvement, image format conversion, image clipping (to get rid of background information), image histogram equalization, and image PCA. Each of these tasks would be performed by individual agents working in their environment.

In this chapter we describe the task of performing histogram equalization. Initially the agent is trained as proposed. Then the node communicates with all the nodes in the network suggested by the server. The list of keywords are analysed and the KL is updated. Depending on the KL change, the agent is retrained; this process continues until there is no further change in the KL. This would suggest that the agent has matured enough to perform the operation autonomously. Then the agent starts evaluating the nodes for its service index. The node which evaluates the highest value is selected for the service performance. Phase I has been developed, and Phase II is in the development stage.

CONCLUSION AND FUTURE WORK

We have presented a new framework of mobile agents functioning in a volatile mobile environment. The mobile environment is the generalised situation in the real world where the service providers oscillate between providing and revoking their service. We have presented a managing technique which can perform various tasks relating to managing such an environment. The agents have intelligence embedded into them in the form of neural networks, which acts in a "human way." The agents are initially trained and then let to mature by iterative retraining. This system is presently in the development stage and has several aspects that need further experimenting. Features such as the process of calculating the latency factors for each node are still in the research stages and need experimental data for analysis.

REFERENCES

Birrell, A. D., & Nelson, B. J. (1984). Implementing remote procedure calls. *ACM Transactions on Computer Systems, 2*(1), 39-59.

Brazier, M. I., Overeinder, B. J., Steen, M., & Wijngaards, N. J. E. (2002). Agent factory: Generative migration of mobile agents in heterogeneous environments. *Proceedings of the 2002 ACM Symposium on Applied Computing (SAC 2002)* (pp. 101-106).

Busetta, P., & Ramamohanarao, K. (1998). An architecture for mobile BDI agents. *Proceedings of the 1998 ACM Symposium on Applied Computing* (pp. 445-452).

Cabri, G., Leonardi, L., & Zambonelli, F. (2002). Engineering mobile agent applications via context-dependent coordination. *IEEE Transactions on Software Engineering, 28*(11), 1039-1055.

Chess, D., Grosof, B., Harison, C., Levine, D., & Parris, C. (1995). Itinerant agents for mobile computing. *IBM Research Report on Computer Science and Mathematics* (RC 20010).

Cybenko, G., & Jiang, G. (1999, July 18). Matching conflicts: Functional validation of agents. *Proceedings of AAAI 99,* Orlando, FL.

Cybenko, G., Gray, R. S., Wu, Y., & Khrabrov, A. (1994). *Information architecture and agents.*

Dömel, P. (1996). Mobile Telescript agents and the Web. *Proceedings of the 41st IEEE International Computer Conference* (p. 52).

Fuggetta, A., Picco, G. P., & Vigna, G. (1998). Understanding code mobility. *IEEE Transactions on Software Engineering, 24*(5).

Gray, R. S. (1995, December). Agent TCL: A transportable agent system. *Proceedings of the CIKM Workshop on Intelligent Information Agents of the 4th International Conference on Information and Knowledge Management (CIKM 95),* Baltimore, MD.

Gray, R. S. (Ed.). (1998). *Agent TCL: A flexible and secure mobile-agent system.* Hanover, NH: Dartmouth College.

Gray, R. S., Cybenko, G., Kotz, D., Peterson, R. A., & Rus, D. (2002a). D'Agents: Applications and performance of a mobile-agent system. *Software—Practice and Experience, 32*(6), 543-573.

Gray, R. S., Kotz, D., Nog, S., Rus, D., & Cybenko, G. (Eds.). (1996). *Mobile agents for mobile computing.* Hanover, NH: Dartmouth College.

Gray, R. S., Kotz, D., Peterson, R. A., Barton, J., Chac, D. A., Gerken, P. et al. (2002b). Mobile agent versus client/server performance: Scalability in an information-retrieval task. *Proceedings of the 5th International Conference on Mobile Agents* (pp. 229-243).

Harker, K. (Ed.). (1995). *TIAS: A Transportable Intelligent Agent System.* Hanover, NH: Dartmouth College.

Isaacson, P.C. (2001). Object-oriented programming in TCL/TK. *Journal of Computing Sciences in Colleges, 17*(1), 206-215.

Jipping, M.J. (2002). *Symbian OS communications programming.* New York: John Wiley & Sons.

Johansen, D., Renesse, R., & Schneider, F. B. (1996). Supporting broad Internet access to TACOMA. *Proceedings of the 7th ACM SIGOPS European Workshop* (pp. 55-58).

Kendall, E. A., Krishna, P. V. M., Pathak, C. V., & Suresh, C. B. (1998). Patterns of intelligent and mobile agents. *Proceedings of the 2nd International Conference on Autonomous Agents* (pp. 92-99).

Kotz, D., Gray, R., Nog, S., Rus, D., Chawla, S., & Cybenko, G. (1999). *Agent TCL: Targeting the needs of mobile computers* (pp. 513-523).

Lauvset, K. J. (2001). Separating mobility from mobile agents. *Proceedings of the 8th Workshop on Hot Topics in Operating Systems* (p. 173).

Levy, J. Y., Ousterhout, J. K., & Welch, B. B. (Eds.). (1997). *The Safe-TCL security model.* Sun Microsystems.

Marzo, G. D., Murhimanya, M., & Harms, J.R. (1994). *Messenger-based operating systems.* Technical Report No 90, Cahier du CUD, University of Geneva, Switzerland.

Meandzija, B. (1986). A formal method for composing a network command language. *IEEE Transactions on Software Engineering, 12*(8), 861-865.

Stamos, J. W., & Gifford, D. K. (1990). Remote evaluation. *ACM Transactions on Programming Language Systems, 12*(4), 537-564.

Stoyenko, A.D. (1994). SUPRA-RPC: SUbprogram PaRAmeters in remote procedure calls. *Software–Practice and Experience, 24*(1), 27-49.

Tardo, J., & Valente, L. (1996). Mobile agent security and Telescript. *Proceedings of the 41st IEEE International Computer Conference* (p. 58).

Yang, S. J. H., Hsieh, J. S. F., Lan, B. C. W., & Chung, J.-Y. (2005). Composition and evaluation of trustworthy Web Services. *Proceedings of the IEEE International Workshop on Business Services Networks* (p. 5).

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 285-296, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.22
Semantic Web Services for Smart Devices Based on Mobile Agents

Vagan Terziyan
University of Jyvaskyla, Finland

ABSTRACT

Among traditional users of Web resources, industry has a growing set of smart industrial devices with embedded intelligence. Just like humans, they need online services (i.e., for condition monitoring, remote diagnostics, maintenance, etc.). In this paper, we present one possible implementation framework for such Web services. Such services should be Semantic Web enabled and form a Service Network based on internal and external agents' platforms, which can host heterogeneous mobile agents and coordinate them to perform needed tasks. The concept of a "mobile service component" assumes not only exchanging queries and service responses, but also delivering and composition of a service provider. Mobile service component carrier (agent) can move to a field device's local environment (embedded agent platform) and perform its activities locally. Service components improve their performance through online learning and communication with other components. Heterogeneous service components' discovery is based on semantic P2P search.

INTRODUCTION

The intersection of Web Service, Semantic Web, and Enterprise Integration Technologies is recently drawing enormous attention throughout academia and industry (Bussler et al., 2003), and the expectation is that Web Service technology in conjunction with Semantic Web Services will make Enterprise Integration dynamically possible for various enterprises compared to the traditional technologies (Electronic Data Interchange or Value Added Networks).

The Semantic Web is an initiative of the World Wide Web Consortium with the goal of extending the current Web to facilitate Web automation, universally accessible content, and the Web of Trust. Tim Berners-Lee (Berners-Lee et al., 2001) has a vision of a Semantic Web, which has ma-

chine-understandable semantics of information, and trillions of specialized reasoning services that provide support in automated task achievement based on the accessible information. Management of resources in Semantic Web is impossible without use of ontologies, which can be considered as high-level metadata about semantics of Web data and knowledge (Chandrasekaran et al., 1999). DAML-S or DAML for Services (Ankolekar et al., 2002; Paolucci et al., 2002) provides an upper ontology for describing properties and capabilities of Web services in an unambiguous, computer-interpretable markup language, which enables automation of service use by agents and reasoning about service properties and capabilities. There also is a growing interest in the use of ontologies in agent systems as a means to facilitate interoperability among diverse software components (Ontologies, 2003). The problems related to that are being highlighted by a number of recent large-scale initiatives (e.g., Agentcities, Grid computing, the Semantic Web and Web Services). A common trend across these initiatives is the growing need to support the synergy between ontology and agent technology.

The key to Web Services is on-the-fly software composition through the use of loosely coupled, reusable software components (Fensel et al., 2002). Still, more work needs to be done before the Web service infrastructure can make this vision come true. Among the most important European efforts in this area is the SWWS (Semantic Web and Web Services, swws.semanticweb.org) project, which is intended to provide a comprehensive Web Service description, discovery, and mediation framework.

Usually a Web service is accessed by human users or by applications on behalf of human users. However, there already exists a growing new group of Web service users, which are smart industrial devices, robots, or any other objects equipped by embedded intelligence. There is a need to launch special Web services for such smart industrial devices. Such services will provide necessary online information provisioning for the devices, allow the heterogeneous devices to communicate and exchange data and knowledge with each other, and even support cooperation between different devices. There are many open questions to be answered within this research area.

In this paper, we discuss an approach for implementing emerging Semantic Web and Web services technologies to a real industrial domain, which is field device management. The goal of this paper is to discuss possible implementation framework to Web services that automatically follow up and predict the performance and maintenance needs of field devices.

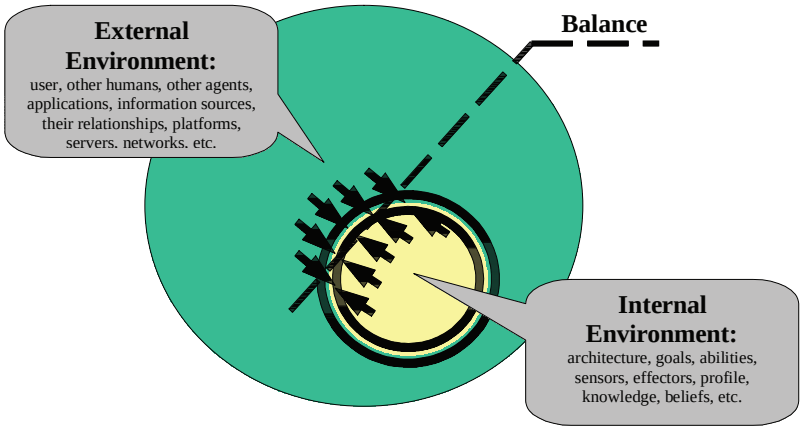
The rest of the paper is organized as follows: Section 2 briefly introduces our concepts of an intelligent agent and mobility; Section 3 presents two alternative architectures for distributed problem solving based on mobile agents; Section 4 describes the domain of field device maintenance and ways of implementing agents in it; Section 5 discusses implementation issues related to the Web service network (OntoServ.Net) of smart devices based on integration of Semantic Web services and multiagent technologies; Section 6 concludes.

AGENTS, SEMANTIC BALANCE, AND MOBILITY

In spite of the existence of many definitions for the concept of an intelligent agent, we will use our own. The definition is based on the concept of Semantic Balance (Terziyan & Puuronen, 1999). In Figure 1, the concept of internal and external environments is illustrated.

We consider Intelligent Agent as an entity that is able to keep a continuous balance between its internal and external environments in such a way that in the case of unbalance, the agent can choose a behavioral option from the following list:

Figure 1. Internal and external environments of an agent



- *make a change within the external environment* to be in balance with the internal one;
- *make a change within the internal environment* to be in balance with the external one;
- find out and *move to another place* within the external environment where balance occurs without any changes;
- *communicate* with one or more agents (human or artificial) to be able *to create a community* so that the internal environment of the community will be in balance with the external one.

This means that an agent is:

1. goal-oriented, because it should have at least one goal—to keep a continuous balance between its internal and external environments;
2. creative because of the ability to change external environment;
3. adaptive because of the ability to change internal environment;
4. mobile because of the ability to move to another place;
5. social because of the ability to *communicate to create a community.*

Thus, we see that mobility is an important adaptation ability of an intelligent agent.

"MOBILE AND DISTRIBUTED BRAINS" ARCHITECTURES

Assume that there is a certain intelligent task (i.e., remote diagnostics of a device based on sensor data) that appears somewhere on the Web. Assume also that necessary intelligent components ("distributed brains") to perform this task are distributed over the Web (e.g., in the form of Web-Services). Finally, assume that there is also an intelligent engine able to perform integration

of autonomous components for solving complex tasks.

Consider the following two architectures for this distributed problem solving:

- *Mobile Engine architecture*. To integrate distributed service components into one transaction in order to solve the task, the intelligent engine (i.e., mobile transaction management agent) makes necessary visits to all distributed platforms that host these services and provides all necessary choreography. Mobility here is an option that can be replaced by remote access to the components.
- *Mobile Components architecture*. Alternatively, the necessary components discovered for performing the task move to the platform where the engine is resized and choreography is performed locally. According to business models around the concept of a Web-service, it is natural to assume that services (intelligent components, in our case) should be "self-interested," and whenever they move, they should serve according to the interests of their creators. This means that the appropriate concept for such components is the concept of mobile agents. An agent is a self-interested entity that can act according to certain goals whenever it needs to accomplish a task.

Both architectures can be considered appropriate for implementation of an environment for distributed monitoring and remote diagnostics Web services for field devices, which is discussed in the following sections of this paper.

FIELD DEVICE MANAGEMENT AND AGENT TECHNOLOGIES

The expectations from smart field devices include advanced diagnostics and predictive maintenance capabilities. The concerns in this area are to develop a diagnostics system that automatically follows up the performance and maintenance needs of field devices and also offers easy access to this information. The emerging agent and communication technologies give new possibilities in this field. Field device management, in general, consists of many areas; the most important are:

- Selection
- Configuration
- Condition monitoring
- Maintenance

Valuable information is created during each phase of device management, and it would be beneficial to save it into a single database. This information can be utilized in many ways during the lifetime of the devices, especially since life-cycle cost (or lifetime cost) of all assets is getting more and more attention. Accordingly, the concept of life-cycle management of assets has become very popular (Pyötsiä & Cederlöf, 1999).

Field Agent is a software component that automatically follows the "health" of field devices. This agent either can be embedded to a device (Lawrence, 2003) or resized at the local network. It is autonomous, it communicates with its environment and other Field Agents, and it is capable of learning new things and delivering new information to other Field Agents. It delivers reports and alarms to the user by means of existing and well-known technologies such as intranet and e-mail messages. Field device performance has a strong influence on process performance and reliable operation in more distributed process automation architecture based on FieldBus communication (Metso, 2003; Sensodec, 2003). In this situation, easy online access to the knowledge describing field device performance and maintenance needs is crucial. There is also a growing need to provide automatic access to this knowledge, not only to humans, but also to other devices, applications, expert systems, agents, and the like, which can

use this knowledge for different purposes of further device diagnostics and maintenance. Also, the reuse of collected and shared knowledge is important for other field agents to manage maintenance in similar cases.

While monitoring a field device via a single information channel (Figure 2), one can get useful information about some dimension of the device state; then derive some useful patterns from this information, which can be considered as "symptoms" of the device's "health"; and finally recognize these symptoms using "Ontology of Patterns."

If we monitor a device via several information channels (Figure 3), then appropriate Field Agent Infrastructure allows not only deriving and recognizing "symptoms" of the device's "health," but also deriving and recognizing a disease itself using "Ontology of Diseases." In any case, history data, derived patterns, and diagnoses can be stored and used locally. However, there should be a possibility for easy access to this information and also sharing it with other agents for reuse purposes.

There are at least two cases where such distributed infrastructure is reasonable. The first one is when we are monitoring a group of distributed devices that are physically and logically disjointed; however, they all are of the same type. In this case any history of derived patterns and

Figure 2. Agent-based symptom recognition in device monitoring

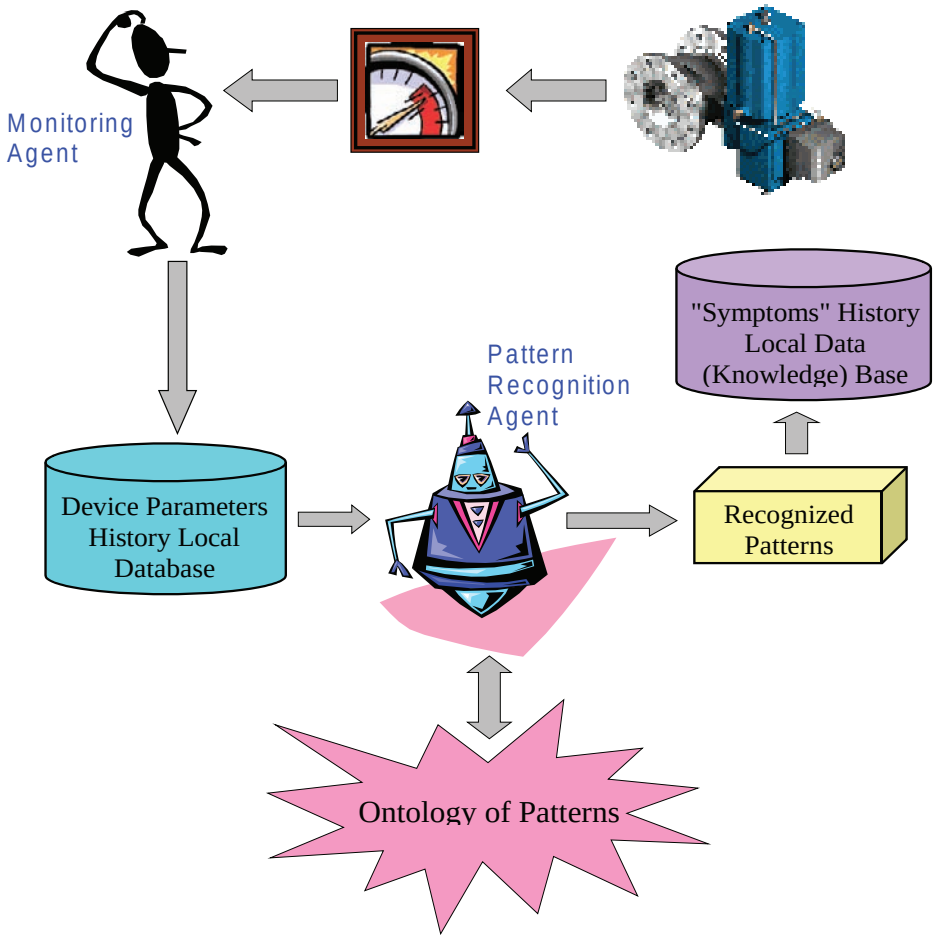
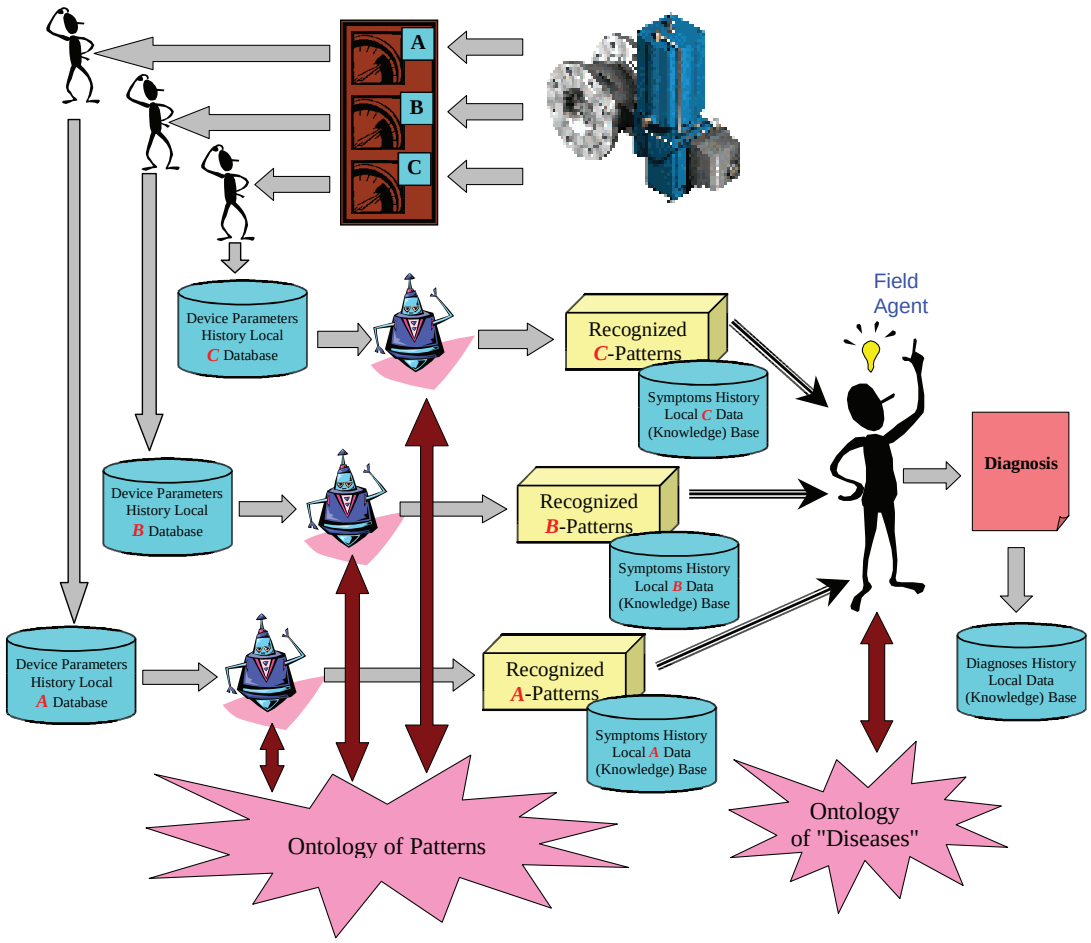


Figure 3. Agent-based diagnostics of field devices



diagnoses from one device can be useful to better interpret the current state of any other device from the group.

The second case relates to the monitoring of a group of distributed devices of different types that are considered as a system of physically or logically interacting components. In such a case, it would be extremely important for every field agent to use outcomes from other field agents as a context for interpretation of the produced diagnosis. Thus, in these two cases, appropriate field agents should communicate with each other (i.e., in peer-to-peer manner) to share locally stored online and historical information and thus improve the performance of the diagnostic algorithms. This allows even the cooperative use of heterogeneous field devices produced by different companies that share common communication standards and ontologies.

We are considering the case when (predictive) maintenance activities can be performed not only by humans, but also by embedded automatics controlled by agents. We also assume that the Semantic Web and Intelligent Web Services concepts can be applied to the problems of interoperability among field devices and will result in the essential improvement of field device maintenance performance.

ONTOSERV.NET IMPLEMENTATION ISSUES

The OntoServ.Net concept was developed by the Industrial Ontologies Group (http://www.cs.jyu.fi/OntoGroup) as a large-scale automated industrial environment for asset management. First, we consider maintenance of assets, but in general, this concept also can be applied to process control, improvement of operating efficiency, field-performance diagnostics, and the like. Better maintenance provided by OntoServ.Net considers maintenance information integration, better availability of operational data, and shift from reactive and preventive maintenance toward predictive and proactive maintenance, which means, first of all, reduced Total Life Cycle Cost of machines. OntoServ.Net is also a network of industrial partners that can share maintenance methods and information developed by working on a specific machine (i.e., device, equipment, installation). Improved locally, maintenance experience can be shared.

In addition, it is assumed that there are special commercial maintenance (diagnostics) services supported either by manufacturers of machines or by third parties. Browsing a device's internal state is extended to automatic diagnostics and recovery within a network of maintenance services or even within a network of platforms hosting several maintenance services. The role of a maintenance service is first to organize the gathering and integration of field data to learn from it, and then to support its clients (field devices) by providing remote diagnostics and maintenance services. Implementation of such large-scale environment in OntoServ.Net presents many problems to be solved. The challenge here is standardization of maintenance data across various software systems within OntoServ.Net and existing industrial systems.

Ontology-Based Standardization of Maintenance Data

We are focusing on maintenance data, which comes from field devices. For remote diagnostics (i.e., in the case of predictive maintenance), this data need to be sent to some other place beyond the local computing system (whether it is a computer or an embedded system). We assume that a maintenance network without centralized control requires some global standard for data representation in order to provide compatibility of network nodes.

We consider standardization based on ontological description of data. An ontology-based approach here stands as an alternative for development of a maintenance-specific set of standards/vocabularies/procedures for information exchange. We use the ontology concept and data representation framework, which was developed within Semantic Web activities. Ontology-based information management is more flexible and scalable, and also has the potential to become the next-generation standard for information exchange on the Internet. An ontology engineering phase includes development of upper-ontology (schema for ontology) and development of ontology itself, which includes specific data about maintenance domain (i.e., device descriptions, diagnostic methods description, etc.). Concrete data are annotated in terms of upper and common ontology. Here, ontology provides a basis for a well-understood common language to be used between devices and systems.

To consider field devices as data sources, the information that needs to be annotated includes sensor data, control parameters, and other data that presents the relevant state of the device for the maintenance process. A special piece of device-specific software (OntoAdapter) is used for translation of raw diagnostic data into standardized maintenance data. This adapter can be integrated into a legacy system used for device management or can be developed independently from existing software, if such a solution is appropriate and possible.

The type of software that uses data being described in an ontological way can vary, depending on the needs. It can be a data browser, control panel for the operator, computing system, database storage, and so forth. Because of the way data are represented, it never will be processed incorrectly, since the software can check itself on whether data semantics (as annotated) are the same or compatible with the data processing needs of the unit.

Additional benefits come from data annotation for software development, even if there is no need to deliver information outside of the original computing system. There is no more need to develop special formats of maintenance data exchange between applications, since it is already presented in a common standard by means of ontology. Software can be developed in a modular, scalable manner with support for this standard (ontology). Such commitment to the shared (upper) ontology will provide compatibility of software.

Ontology-Based Diagnostics Based on Maintenance Data

It is assumed that there are special commercial diagnostic units (maintenance services) supported either by manufacturers of machines or by third parties. Browsing a device's internal state is extended to an automatic diagnostics and recovery within a network of maintenance services. As already mentioned, the role of maintenance service is first to organize gathering and integration of field data and learning based on it, and then to support its clients (field devices) providing remote diagnostics services.

Considering the various aspects of maintenance network development, the following statements are true:

1. There are diagnostic software components (also called classifiers or diagnostic units) that perform predictive/proactive diagnostics. These diagnostic units obtain maintenance data delivered to them either locally or from a remote source, and provide diagnosis as an output.
2. Diagnosis provided by a classifier can be of several types: the class of state an observed device has; the in-depth diagnostics; and the [maintenance] actions/activities that are required.
3. Diagnostic units are specialized in certain aspects of maintenance diagnostics, so a device usually needs support from a set of different diagnostic units that operate and can be replaced independently.
4. Diagnostic units (components), in general, are not device-specific and perform similar diagnostic tasks in a variety of device monitoring systems.
5. Diagnostic units [OntoServ.Net] are platform-compatible and developed separately from maintained devices; thus, they can be used on any platform within OntoServ. Net.
6. Once a diagnostic unit possesses the ability to learn, the maintenance experience it gets will be available for diagnostics of other devices. This is done by means of running a copy of a classifier on other maintenance platforms or presenting its experience in a way that will allow reusing it by other diagnostic units. There is interest especially

in the capabilities of integration of such information obtained worldwide and applied effectively to individual devices.

7. A maintenance platform is a computing environment in which maintenance services (i.e., diagnostic units, etc.) are installed. It supports device-specific interfaces for connecting devices. On one side, it provides a maintenance-managing core with installed services and, on other side, it supports a connection to a maintenance network consisting of maintenance platforms of other devices — specialized centers for remote maintenance. Such services can be implemented, based on agent technology (Gibbins et al., 2003). Taking into account the openness of the system, the issues of security and trust management are considered as exceptionally important (Kagal et al., 2001).
8. A maintenance platform manages maintenance information exchange between devices and diagnostic units residing locally and elsewhere in the maintenance network. It also supports the search of required network resources and upgrades of installed software, and supports a mobility feature for maintenance services.

Since the available set of classifiers can vary, the type of classifier (its purpose) is specified in order to allow selection in cases, when necessary. Every classifier has its description attached — information concerning its capabilities. The description also contains information on how to use the classifier, what inputs it requires, and what output it provides.

Preliminary classification mechanism of a maintenance platform takes into account which services (classifiers) are available now, and selects those that declare themselves able to deal with the class of problems derived during the pre-classification phase. Here, pre-classification is similar to human-operator work, which can detect some abnormal device behavior and use analysis tools in order to find the source of the problem.

It is assumed that some historical maintenance data are available and are used for automatic learning of the kind of maintenance actions to be performed. This knowledge in the model supports the automated reasoning mechanism, which implements preliminary diagnostics of the device state and can identify certain deviations from normal operational state and run an appropriate diagnostic service.

The Industrial Ontologies Group is involved in the maintenance data processing, based on the descriptions of available maintenance resources (i.e., classifier services, as it was shown) and explicit representation of knowledge for initial data pre-processing.

Service descriptions allow changing of service easily. Diagnostic knowledge allows automated maintenance system activity and makes it possible to reuse the acquired knowledge (presented in the specific classification model as data, rules, etc.). A newly installed device that has no historical data yet can use the classification ontology of some other device of the same type.

We use application of Semantic Web technology for maintenance data description, service component description, and representation of classification knowledge. Those descriptive data pieces have to be in some common format for the whole maintenance system. RDF and its derivatives are a perfect basis for that.

Ontology-Based Diagnostic Services Integration

To be able to integrate heterogeneous resources and services over the Web, we have to describe them in a common way, based on common Ontology. Considering the resources in an industrial product's maintenance domain, we distinguish the following resources: smart devices, which are represented by services of their alarm or

control systems (or some software interface); a set of diagnostic services or classifiers; platforms, which are represented by clusters or a collection of elements; human, which can be represented by some special service; and large enterprises' systems. This ontology-based annotation must comprise not only a resource's description (i.e., parameters, inputs, outputs), but also many other necessary things that concern its goals, intentions, interaction aspects, and so forth. Ontology-based means that we have to create ontologies to be used by all such resources.

Each service represents a three-level resource: input parameters, black box (service engine), and output parameters. Since all services are heterogeneous, we have a need to describe each resource (service) with a common ontology and create a shell (OntoShell), which will provide common descriptions and make the service realization transparent. OntoShell is the main core of such integration environment. It has to be expanded in a scalable and modular manner, so that it may represent a mediation platform for the set of adapted, semantically annotated resources (cluster of OntoShells). OntoShell is a shell (frame) for these resources, which is a mechanism for making ontological description and providing interoperability. One of the important parts of OntoShell is an OntoAdapter for resources. If we are talking about transformation of existing resources to semantically enabled ones, then we have to develop the mechanisms for accessing the resources. These resources are developed, based on different specific standards on both content (WSDL, C/C++ (dll), Java, SQL Server, DCOM, CORBA, etc.) and transport levels (TCP, HTTP, RMI, etc.). In this case, we have to design and develop corresponding software modules (OntoAdapters) for semantic, content, and transport levels. It will be these construction blocks that will fill OntoShell, depending on the resource's description.

A set of services represents the three-level automated diagnostic system. The first level is represented by the alarm system of the device (WatchDog), which signals a change in the normal device state. The main maintenance diagnostic system (service) forms the second level of the system via preliminary (initial) diagnostics. It contains preliminary diagnosis classification based on the ontology of the device condition. The result of such initial classification is deciding on what kind of classifier (diagnostic service) is needed to make further data processing. This system initiates the third level of diagnostic for making a decision about precise diagnosis. Request for further classification is sent to respective classifiers of the centralized diagnostic system directly taking into account the probability of classifier belonging to that class of problem.

Semantic Peer-to-Peer Discovery of Maintenance Services

Within the OntoSert.Net concept, a peer-to-peer architecture of global network of maintenance Web-services is assumed. The architecture provides support for registration of maintenance service profiles based on ontology shared within the network. The profiles are registered in a local repository. The architecture includes a platform steward module, which implements a semantic search engine. The engine searches among local maintenance services, using the semantic match procedure, for one with a profile that corresponds to the query. A Steward also implements peer-to-peer functionalities like query forwarding and sending, neighbor registration, and so forth.

When the Platform Steward makes a decision about the necessity of using external Maintenance Services, it sends a formalized query to neighbor platforms. Such a necessity can occur if the requested Maintenance Service is absent on the local platform or cannot provide sufficient performance, or if the Steward needs opinions of other Classifying Services (i.e., during learning). If the neighbors can't satisfy the query, they forward it to their own neighbors, and so on. Thus, the

query can roam through many platforms in the Global Network, which increases the probability of finding the required service.

When a platform receives a query that it can satisfy, it sends the response to the query initiator about its location. The query initiator can collect a number of such responses — a list of potential partners.

To increase the efficiency of a search of Maintenance Semantic Web-services and its automated nature (initiators of the search are Smart Devices), the system should inherit the concepts of Semantic Web. That means:

1. Development of common Ontology for the Global Network, which contains a classification of Maintenance Services in a hierarchical tree and explicit definition of relations between such classes. Another option might be to provide possibilities to manage several pre-existing ontologies (Mena et al., 2000).
2. Every platform creates a local repository of profiles of available Maintenance Services based on Ontology.
3. Each Platform Steward must have a Semantic Search Engine that will find a semantic match between the query and each profile in the local repository.
4. Each Platform Steward must have a Semantic Query Engine that composes formalized queries.

Thus, we are using a combination of centralized architecture for service discovery within the platforms of services and peer-to-peer architecture for service discovery across the platforms (Arumugam et al., 2002). A Platform Steward provides centralized capabilities when it manages service discovery within its internal platform, and at the same time, it can behave in a peer-to-peer manner when it interacts with external service platforms.

CONCLUSION

The goal of this paper is to provide a possible implementation framework for Web services that automatically follow up and predict the maintenance needs of field devices. The concept of a mobile service component supposes that any component can be executed at any platform from the Service Network, including the service requestor side. This allows delivery of not only the service results, but also of the service itself. A mobile service component carrier (agent) can move to a field device's local environment (embedded agent platform) and perform its activities locally. Service components improve their performance through online learning and communication with other components. Heterogeneous service components' discovery is based on semantic P2P search. This paper addresses the very basic challenges related to Web services for smart devices and partly the related implementation issues. More research is needed to validate some of concepts discussed in this paper.

ACKNOWLEDGMENTS

The author is grateful to Tekes (National Technology Agency of Finland) and cooperating companies (Agora Center, University of Jyvaskyla, TeliaSonera, TietoEnator, Metso Automation, Jyvaskyla Science Park) for the grant supporting the activities of the SmartResource project. I am also grateful to Dr. Jouni Pyotsia from Metso Automation for useful consultations and materials. Also, thanks to colleagues from Industrial Ontologies Group (O. Kononenko, A. Zharko, and O. Khriyenko) for useful discussions related to the implementation issues.

REFERENCES

Ankolekar, A., et al. (2002). DAML-S: Web service description for the semantic Web. *The First International Semantic Web Conference (ISWC).*

Arumugam, M., Sheth, A., & Arpinar, B. (2002). The peer-to-peer semantic Web: A distributed environment for sharing semantic knowledge on the Web. *Proceedings of International Workshop on Real World RDF and Semantic Web Applications*, Hawaii.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American*, *284*(5), 34-43.

Bussler, C., Fensel, D., & Sadeh, N. (2003). Semantic Web services and their role in enterprise application integration and e-commerce. Retrieved from *http://www.gvsu.edu/ssb/ijec/announcements/semantic.doc*

Chandrasekaran, B., Josephson, J., & Benjamins, R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 20-26.

Fensel, D., Bussler, C., & Maedche, A. (2002). A conceptual architecture of semantic Web enabled Web services. *ACM Special Interest Group on Management of Data*, *31*(4).

Gibbins, N., Harris, S., & Shadbolt, N. (2003). Agent-based semantic Web services. *Proceedings of the 12th International World Wide Web Conference*.

Kagal, L., Finin, T., & Peng, Y. (2001). A framework for distributed trust management. *Proceedings of IJCAI-01 Workshop on Autonomy, Delegation and Control.*

Lawrence, J. (2003). Embedded FIPA agents. *Agentcities*: a*gent technology exhibition, Barcelona*, URL. Retrieved from *http://www.agentcities.org/EUNET/ID3/documents/exh_program.pdf*

Mena, E., Illarramendi, A., Kashyap, V., & Sheth, A. (2000). OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *International Journal on Distributed and Parallel Databases, 8*(2), 223-271.

Metso (2003). Neles FieldBrowser™ system for field device predictive maintenance. *Metso Automation Tech. Bulletin.* Retrieved from *http://www.met soautomation.com/*

Ontologies (2003). Ontologies in agent systems. Retrieved from *http://oas. otago.ac.nz/OAS2003*

Paolucci, M., Kawamura, T., Payne, T., & Sycara, K. (2002). Importing the semantic Web in UDDI. *Proceedings of Web Services, E-business and Semantic Web Workshop.*

Pyötsiä, J., & Cederlöf, H. (1999). Advanced diagnostic concept using intelligent field agents. *ISA Proceedings.*

Satoh, I. (2001). Mobile agent-based compound documents. *Proceedings of the 2001 ACM Symposium on Document Engineering*, 76-84.

Sensodec (2003). Sensodec 6C for paper. *Metso Automation Tech. Bulletin.* Retrieved from *http://www.metso automation.com/*

Terziyan, V., & Puuronen, S., (1999). Knowledge acquisition based on semantic balance of internal and external knowledge. *Lecture Notes in Artificial Intelligence*, 1611, 353-361.

This work was previously published in the International Journal of Intelligent Information Technologies, Volume 1, No. 2, pp. 43-55, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.23
Towards Autonomic Infrastructures via Mobile Agents and Active Networks

Stamatis Karnouskos
SAP Research, Germany

ABSTRACT

As we move towards service-oriented complex infrastructures, what is needed, security, robustness, and intelligence distributed within the network. Modern systems are too complicated to be centrally administered; therefore, the need for approaches that provide autonomic characteristics and are able to be self-sustained is evident. We present here one approach towards this goal, i.e., how we can build dynamic infrastructures based on mobile agents (MA) and active networks (AN). Both concepts share common ground at the architectural level, which makes it interesting to use a mix of them to provide a more sophisticated framework for building dynamic systems. We argue that by using this combination, more autonomous systems can be built that can effectively possess at least at some level of self-* features, such as self-management, self-healing, etc., which, in conjunction with cooperation capabilities, will lead to the deployment of dynamic infrastructures that autonomously identify and adapt to external/internal events. As an example, the implementation of an autonomous network-based security service is analyzed, which proves that denial of service attacks can be managed by the network itself intelligently and in an autonomic fashion.

INTRODUCTION

Systems and services are becoming more ubiquitous, which calls for sophisticated solutions to be in place. As we move towards the "Internet of things" (Dolin, 2006), it can be expected that millions of devices of different size and capability will be connected and interact with each other

over IP, e.g., sensor networks (Marsh, 2004). Therefore, any approach will have to take into consideration that:

- Complexity will increase
- Heterogeneity in devices, software platforms, online services, etc., will increase
- A large proportion of end-nodes will be connected wirelessly to the backbone infrastructure (the line of wired vs. wireless systems will blur more)
- Bandwidth and computing power will increase
- Ad-hoc computing, collaboration, task delegation, and environmental adaptation will be basic necessities
- On-demand software and service deployment will be vital
- Security and its satellite services will gain importance

In such an assumed future infrastructure, autonomic systems are expected to be of considerable help, since they will be able to be at a great degree self-sustained and also react to a dynamic changing environment.

Autonomic computing (Sterritt et Al., 2005) was introduced by IBM as a means to target increasing computer system complexity, and aimed initially at automating management of enterprise computational systems. In *The Vision of Autonomic Computing* (Kephart & Chess, 2003) it is stated that the dream of interconnectivity of computing systems and devices could become the "nightmare of pervasive computing," in which architects are unable to anticipate, design, and maintain the complexity of interactions. The essence of autonomic computing is system self-management, freeing administrators from low-level task management whilst delivering an optimized system. In a self-managing system, or Autonomic System, the human operator does not control the system directly, but only defines general policies and rules that serve as an input for the self-management process. For this process, IBM has defined the following four functional areas:

- **Self-configuration:** Automatic configuration of components
- **Self-healing:** Automatic discovery, and correction of faults
- **Self-optimization:** Automatic monitoring and control of resources to ensure the optimal functioning with respect to the defined requirements
- **Self-protection:** Proactive identification and protection from arbitrary attacks

There are two strategies in achieving autonomic behavior, i.e., through adaptive learning and via integral engineering into systems (Sterritt, 2004). Our approach focuses on how to engineer such an autonomous system, while adaptive learning, or self-learning, is seen as an ad-hoc component that can be imported from the domain of intelligent agents.

AMALGAMATION OF ACTIVE NETWORKS AND MOBILE AGENTS

Active and programmable networks (Karnouskos & Denazis, 2004) introduce a new network paradigm where network-aware applications and services can be not only distributed, but also can configure the heterogeneous network to optimally respond to task-specific requirements. We are able to utilize within the network: (a) computation, as we are able to compute on data received from active nodes, and (b) programmability, as we can inject user code into the network nodes in order to realize customized computation. Being able to achieve the above, we succeed in decoupling network services from the underlying hardware, deploy fine-grained customized services, relax the dependencies on network vendors and

standardization bodies, and generally open the way for higher-level. network-based application programming interfaces.

Agents are software components that act alone or in communities on behalf of an entity and are delegated to perform tasks under some constraints or action plans (Jennings & Wooldridge, 1996). One key characteristic of agents is mobility (mobile agents), which allows them to transport themselves from node to node and continue their execution there. Additionally, autonomy, independent decision-making, goal-directed behavior, and social ability are also key characteristics agents may possess (Genesereth & Ketchpel, 1994). Mobile agent technology has established itself as an improvement of today's distributed systems due to its benefits, such as dynamic, on-demand provision and distribution of services, reduction of network traffic and dependencies, fault tolerance, etc. The number of mobile agent platforms coming from the commercial sector, as well as academia, is increasing day by day.

Active networks and Mobile Agent technology are very close to each other, sharing common ground on theoretical/conceptual and implementational levels. From the viewpoint of mobile agent research, existing active network approaches take mobile active code very close to the mobile agent paradigm:

- **Capsule:** A typical code mobility paradigm, i.e., a single mobile agent
- **Active/programmable node:** Instantiation of code on-demand

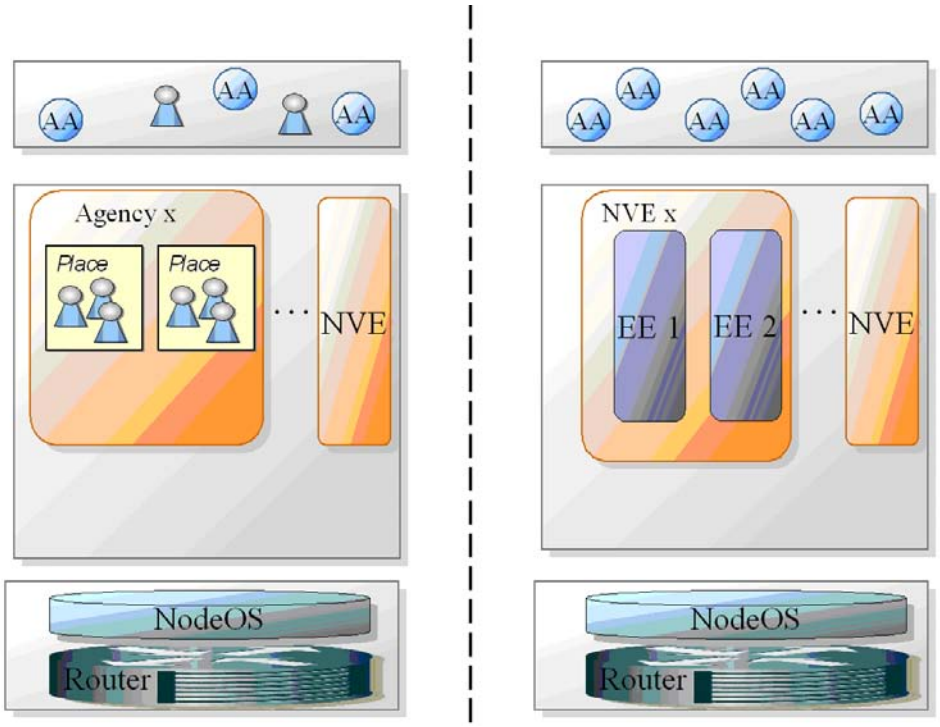
From the active network research viewpoint, the mobile agent technology is one of the possible technologies that can be used to build active networks. Mobile agents are regarded as specific types of active code and a MA-based node as a specific type of active network node. Due to the fact that the MA research arena has existed more than a decade now, it is far more advanced in active code-related matters, therefore, it could provide a boost to specific AN matters at conceptual and implementation levels.

The right side of Figure 1 depicts the architecture of a legacy active node, while the mobile agent based implementation is depicted on the right side. We can clearly distinguish the following levels:

- Active applications/services which exist as a result of the execution of active code within an EE. An active code can: (a) provide a standalone service, or (b) cooperate with other active codes residing on the same EE (EE-based service), on different EEs in the same node (multi-EE service), or even on different EEs in different nodes (network multi-EE service).
- Execution environments where the active code executes. As an active node is expected to host multiple execution environments, these environments must have the ability to communicate with each other and to group in order to ease interactions. There are several functional types of EE aggregation, such as node virtual environment (NVE), node virtual environment network (NVEN), execution environment network (EEN), etc.
- NodeOS, which is an operating system for active nodes. The nodeOS provides generic services to the hosted Ees, e.g., inter-EE communication (at the EE, NVE, or Active Application level), router resource management, EE isolation, etc. The nodeOS offers these services based on several facilities, such as resource control, security, management, demultiplexing facilities, etc.

As shown in Figure 1, one of the execution environments is the agent execution environment. This is the agency as described within the MASIF (OMG-MASIF, 1998) standard. The agent system consists of Places. A Place is a context within

Figure 1.The agent-based AN node versus the legacy one



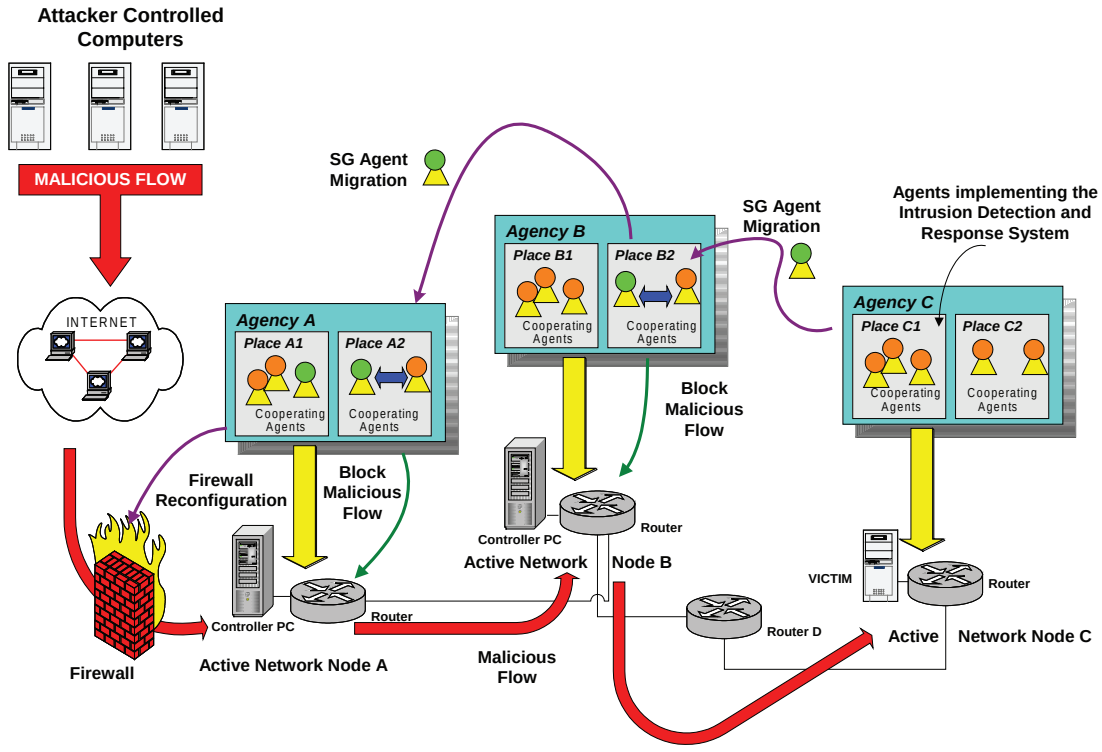
an agent system in which an agent is executed. This context can provide services/functions such as access to local resources, etc. Cooperating agents reside in the agent-based Ees, and via the facilities offered to them (re)-program the node. These can be either mobile agents (e.g., visiting agents) or even stationary intelligent ones that reside permanently in the EE, implementing various services. The integrated approach of agents and active networks allows us to apply several security techniques at the network programming level (Karnouskos, 2001) that promote service and network security. Further information on this architecture and its security issues can be found in Karnouskos (2002). The mobile agent framework is able to realize the abstract functions of the EE, NVE, etc. The AAs are considered, for implementation reasons, to be mobile agents, but could also be applications that partially depend on them.

APPLICATION SCENARIO

An autonomic computing system (ACS) is able to (re)-configure itself in response to varying environmental conditions. Such a dynamic system can deal with unknown intrusions or attacks and is event-able to recover from malfunctions or heal itself. The scenario presented here deals with a network-based security system that is able to depict at some degree the characteristics of an ACS.

Securing a network nowadays is synonymous with hardening of its services. However, this approach makes the network inflexible and blurs the line between security and usability. Furthermore, each node has its own requirements on security which may also be varying in time. Within the vision of "Internet of things," such per-node or even per-task modification of security would be impossible to manage due to the large number and complexity of devices. Furthermore, those devices will build ad-hoc, short-lived networks, where the burden of taking such actions may not be justified. Additionally, no common base exists among various security solutions available in the market. In other words, available products do not communicate with each other (interoperate), and work alone for their own and their distribution company's good, not necessarily for the user's network. A collaborative approach must be considered; however, due to the nature of the future networks, this must be done on-demand and customized to the specific context. Community-aware tactics on the other side may offer a better alternative. Adopting modeling approaches from the evolution of biological systems (Forrest et al., 1997), they are seen as networks formed from cooperating living parts that interoperate at various levels and share information. ACS systems

Figure 2. DoS threat management



seem a promising approach towards that direction, mainly due to their self-* characteristics.

We assume a typical denial of service (DoS) attack scenario. As depicted in Figure 2, the network topology consists of various active nodes (e.g., nodes A, B,and C) and legacy nodes (e.g., node D). In normal operation, the agents that implement our system reside within the agencies and filter the flow that is directed to the node. At some point, the attacker initiates the DoS attack via the compromised hosts against the AN node C. One agent in node C detects the attack. This can be a result of an attack signature recognition (if the attack is known and exists in the system database) or a result of a dynamic correlation of events received by the system. Once the attack is detected, several security guards (SG) are released within the network (dynamic lookup of the neighboring nodes) and the attack information is disseminated towards the other nodes that reside within the path of the attack. In this way, the agents continue in an autonomic way to roam the network, identify the nodes that are prone to this attack, and share information which eventually lead to a policy change and blocking of the specific malicious traffic within that node. At the end, the malicious flow is blocked towards the borders of the domain, and the network nodes are protected from this attack. Further detailed information about this approach can be found in Karnouskos (2004).

The engines behind the data analysis and event correlation, as well as decision and action management, can be standalone; however, it is much more interesting if they act in a collaborative manner. . Therefore, at each domain, central analysis points (CAP) exist which have an overview of what is happening in the domain, thereby making it easier to recognize attacks that include multiple nodes in different parts of the network. CAPs have the global view, and therefore are more efficient in attack recognition and decision making, while the action is done locally on each node; this tactic allows thin components to be deployed even to devices that do not feature high computational capabilities.

The result of this approach is that we have a network that features, at some degree, characteristics if autonomic systems. More specifically:

- **Self-configuration:** Automatic configuration of the different components that recognize the attacks is done. The agents are goal-driven and are able to reconfigure themselves based on the environmental context they act on.
- **Self-healing:** Automatic discovery and correction of faults for network parts is done. Once this is detected, the specific sub-network part can be isolated in order to avoid network misbehavior, and classical solutions to the problem can be applied.
- **Self-optimization:** Automatic monitoring and control of resources of the network can be done. In that case, early indicators can be correlated and emerging problems are easier to pinpoint.
- **Self-protection:** The network is protected from well-known attacks, including those that can be dynamically identified based on the correlation of events or even with "socializing" (i.e., information exchange) with other networks.

As presented, our approach deals with some aspects of ACS; in the future, more specific research should be invested towards a fine-grain exploitation of each of the features in detail. Self-managing mechanisms can have several instantiations: self-government, self-correction, self-organization, self-scheduling, self-planning, self-administration, self-optimization, self-monitoring, self-adjusting, self-tuning, self-configuration, self-diagnosis of faults, self-protection, self-healing, self-recovery, self-learning, self-sensing/perceiving, self-modeling, self-evolution, self-assessment of risks, etc. (Tianfeld & Unland, 2004).

CONCLUSION

We have presented an approach that is based on the amalgamation of active networks and mobile agents. We merge specific capabilities from each domain, e.g., the network programmability from active networks and the autonomic, goal-driven social characteristics of mobile agents, in order to create a powerful combination and implement a system that depicts, at some degree, behavior that characterizes autonomic systems. The approach taken is open and can be seen as a platform to further integrate research results coming from the two domains. Furthermore, we have not yet touched issues like self-learning mechanisms, which however, initially could be imported from the work already done by the research community on intelligent software agents. Security, trust, and privacy issues as identified by Cardoso and Freire (2005) need to be further examined, especially because our collaborative approach taken here heavily depends on them. Finally, other approaches that move towards the usage of agents for implementing specific scenarios also exist, e.g. (Soldatos et al., 2006).

The essence of autonomic computing systems is the creation of dynamic infrastructures that can deal in a proactive way with changing environmental contexts; this is a fact that is gaining importance as we move towards a complex heterogeneous infrastructure, e.g., as depicted in the "Internet of things," where all interconnected devices will also form ad-hoc networks for even task-specific goals. Without such approaches, large-scale complex computing systems will be unmanageable.

REFERENCES

Cardoso, R., & Freire, M. (2005, October 23-28). Towards autonomic minimization of security vulnerabilities exploitation in hybrid network environments. In P. Dini & P. Lorenz (Eds.), *Proceedings of the Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services (ICAS/ICNS 2005)*, Papeete, Tahiti, French Polynesia. Los Alamitos, CA: IEEE Computer Society Press.

Dolin, R. A. (2006, January 23-27). Deploying the "Internet of things." In *Proceedings of the International Symposium on Applications on the Internet* (pp. 216-219). Washington, DC: IEEE Computer Society.

Forrest, S., Hofmeyr, S., & Somayaji, A. (1997). Computer immunology. *Communications of the ACM, 40*(10), 88-96.

Genesereth, M. R., & Ketchpel, S. P. (1994). Software agents. *Communications of the ACM 37*, 7, 48ff.

Jennings, N., & Wooldridge, M. (1996). Software agents. *IEE Review, 42*(1), 17-21. http://www.csc.liv.ac.uk/~mjw/pubs/iee-review96.pdf

Karnouskos, S. (2001). Security implications of implementing active network infrastructures using agent technology. *Computer Networks Journal*, Special Issue on Active Networks and Services, *36*(1), 87-100.

Karnouskos, S. (2002). Realization of a secure active and programmable network infrastructure via mobile agent technology. *Computer Communications Journal*, Special Issue on Computational Intelligence in Telecommunications Networks, *25*(16), 1465-1476.

Karnouskos, S. (2004). Community-aware network security and a DDoS response system. *Annals of Telecommunications* (Annales des Télécommunications), Special Issue on Active Networks, *59*(5-6).

Karnouskos, S., & Denazis, S. (2004). Programmable networks: Background. In A. Galis, S.

Denazis, C. Brou, & C. Klein (Eds.), *Programmable networks for IP service deployment.* Artech House Books.

Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. *Computer 36*, 1(January), 41-50.

Marsh, D., Tynan, R., O'Kane, D., & O'Hare, G. (2004). Autonomic wireless sensor networks. *Engineering Applications of Artificial Intelligence, Autonomic Computing Systems, 17*(7), 741-748.

OMG-MASIF, (1998). *Mobile agent system interoperability facility, OMG.* http://www.omg.org/docs/orbos/98-03-09.pdf

Soldatos, J., Pandis, I., Stamatis, K., Polymenakos, L., & Crowley, J. (2006). Agent based middleware infrastructure for autonomous context-aware ubiquitous computing services. *Computer Communications.* Available online January 17, 2006.

Sterritt, R. (2004) Autonomic networks: Engineering the self-healing property. *Engineering Applications of Artificial Intelligence, Autonomic Computing Systems, 17*(7), 727-739.

Sterritt, R., Parashar, M., Tianfield, H., & Unland, R. (2005). A concise introduction to autonomic computing. *Advanced Engineering Informatics, Autonomic Computing, 19*(3), 181-187.

Tianfield, H., & Unland, R. (2004). Towards autonomic computing systems. *Engineering Applications of Artificial Intelligence, Autonomic Computing Systems, 17*(7), 689-699.

KEY TERMS

Active Application: This is the code that is actually executed in the Execution Environment of the node. Via its execution in the EE, the code programs the node according to the user's preferences.

Active Networks: Active networks are a communication paradigm that allows packets flowing through a communication network to dynamically modify the operation of the network.

Autonomic Computing: An initiative started by IBM in 2001. Its ultimate aim is to create self-managing computer systems to overcome their rapidly growing complexity and to enable their further growth.

DoS: Denial of service attacks result in computers consuming their resources for malicious events without being able to further process legitimate user requests

Execution Environment: This is the place where the active code executes. The EE offers access to the core node resources via a policy-controlled scheme. This can be, for instance, a mobile agent system that takes care of the execution of an agent.

Mobile Agents: A mobile agent is a composition of computer software and data which is able to migrate (move) from one computer to another autonomously and continue its execution on the destination computer.

Sensor Networks: Sensor networks are computer networks of many, spatially distributed devices using sensors to monitor conditions at different locations, such as temperature, sound, vibration, pressure, motion, or pollutants. Usually these devices are small and inexpensive, so that they can be produced and deployed in large numbers, and so their resources in terms of energy, memory, computational speed, and bandwidth are severely constrained.

This work was previously published in Encyclopedia of Internet Technologies and Applications, edited by M. Freire and M. Pereira, pp. 633-639, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.24
Mobility Management in Mobile Computing and Networking Environments

Samuel Pierre
Ecole Polytechnique de Montreal, Canada

ABSTRACT

This chapter analyzes and proposes some mobility management models and schemes by taking into account their capability to reduce search and location update costs in wireless mobile networks. The first model proposed is called the built-in memory model; it is based on the architecture of the IS-41 network and aims at reducing the home-location-register (HLR) access overhead. The performance of this model was investigated by comparing it with the IS-41 scheme for different call-to-mobility ratios (CMRs). Experimental results indicate that the proposed model is potentially beneficial for large classes of users and can yield substantial reductions in total user-location management costs, particularly for users who have a low CMR. These results also show that the cost reduction obtained on the location update is very significant while the extra costs paid to locate a mobile unit simply amount to the costs of crossing a single pointer between two location areas. The built-in memory model is also compared with the forwarding pointers' scheme. The results show that this model consistently outperforms the forwarding pointers' strategy. A second location management model to manage mobility in wireless communications systems is also proposed. The results show that significant cost savings can be obtained compared with the IS-41 standard location-management scheme depending on the value of the mobile units' CMR.

INTRODUCTION

Mobile communication networks are made possible by the convergence of several different technologies, specifically computer networking protocols, wireless-mobile communication systems, distributed computing, and the Internet. With the rapidly increasing ubiquity of laptop computers, which are primarily used by mobile users to access Internet services such as e-mail and the World

Wide Web (WWW), support of Internet services in a mobile environment has become a growing necessity. Mobile Internet providers (IPs) attempt to solve the key problem of developing a mechanism that allows Internet protocol (IP) nodes to change physical locations without changing IP addresses, thereby offering Internet users the so-called "nomadicity." Furthermore, advances in wireless networking technologies and portable information devices have led to a new paradigm of computing called *mobile computing*. According to this concept, users who carry portable devices have access to information services through a shared infrastructure regardless of their physical location or movements. Such a new environment introduces new technical challenges in the area of information access. Traditional techniques to access information are based on the assumptions that the host locations' distributed systems do not change during computation. In a mobile environment, these assumptions are rarely valid or appropriate.

Mobile computing is distinguished from classical, fixed-connection computing due to the following elements: (a) the mobility of nomadic users and the devices they use, and (b) the mobile resource constraints such as limited wireless bandwidth and limited battery life. The mobility of nomadic users implies that the users might connect from different access points through wireless links and might want to stay connected while on the move, despite possible intermittent disconnections. Wireless links are relatively unreliable and currently are two to three times slower than wired networks. Moreover, mobile hosts powered by batteries suffer from limited battery life constraints. These limitations and constraints provide many challenges to address before we consider mobile computing to be fully operational. This remains true despite the recent progress in wireless data communication networks and handheld device technologies.

In next-generation systems supporting mobile environments, mainly due to the huge number of mobile users in conjunction with the small cell size, the influence of mobility on the network performance is strengthened. More particularly, the accuracy of mobility models becomes essential to evaluate system design alternatives and network implementation costs. The device location is unknown a priori and call routing in general implies mobility management procedures. The problems which arise from subscriber mobility are solved in such a way that both a certain degree of mobility and a sufficient quality of the aspired services are achieved.

This chapter analyzes the problem of managing users' mobility in the context of mobile computing and networking environments. Mobility management implies two major components: *handover management* and *location management*.

Handover management is the way a network functions to keep mobile users connected as they move and change access points within the network. Generally, there are two types of handover: intracell handover and intercell handover. Intracell handover occurs when a user experiences degradation of signal strength within a cell. This leads to a choice of new channels with better signal strength at the same *base transceiver station* (BTS), also called *base station* (BS). Intercell handover occurs when a user moves from a cell to another. In this case, the user's connection information is transferred from the former BTS to the latter one. The following procedure occurs for both intracell and intercell handovers. First, the user initiates a handover procedure. Then, the network or the mobile unit (depending on the unit that controls the handover operation) provides necessary information and performs routing operations for the handover. Finally, all subsequent calls to the user are transferred from the former connection to the latter one.

Location management is the process used by a network to find the current attachment point of a mobile user for call delivery (Akyildiz & Wang, 2002). The first step of the procedure is the *location registration*. In this phase, the mobile

user periodically notifies the network of its new access point. The notifications allow the network to authenticate users and update their location profiles. The second step is the *call delivery*. When a call destined to a user reaches the network, a search for the user's profile is usually conducted in a local database. Then, the call is forwarded to the user according to his profile.

This chapter aims to analyze different mobility management schemes and protocols in order to state their applicability to handle some key issues related to emerging mobile environments. The main concerns include the search for efficient and cost-effective location management schemes allowing to provide services and applications to users with an acceptable quality of service. The next section summarizes background and related work. Then we propose some new location management schemes, evaluate the performance of these schemes, and present some numerical examples. Finally, the chapter concludes and outlines future research directions.

BACKGROUND AND RELATED WORK

Mobility is the primary advantage offered by *personal communication systems* (PCS). Location management is one of the most important issues of mobility management. Location management techniques essentially consist of partitioning the coverage area into many *location areas* (LAs) which are sets of cells. *Mobile Units* (MUs) within a cell communicate with a cell BS through wireless links. BSs, in turn, are connected to the wireline network through a *mobile switching center* (MSC) which serves a single LA. Each MSC is identified by a unique address. This address is stored in the memory of the MUs that are roaming in the MSC's LA and is broadcasted by the cell BSs within that particular LA. The MU compares the broadcasted address with the address stored in its memory. When these two addresses differ, the MU recognizes that it has moved to a new LA and sends a registration message to the MSC of the new LA. Then, the MSC forwards this message to the network database. In PCS, the wireline network uses the *Signaling System Number 7* (SS7) to carry user information and signaling messages between the MSCs and the location databases.

Main Standards and Basic Procedures

Two major standards are used for location management, namely IS-41 (Gallagher & Randall, 1997; TIA/EIA, 1996) and Global System for Mobile Communications [GSM] (Mouly & Paulet, 1992). This chapter only considers the IS-41 standard, which uses a two-level database architecture consisting of an HLR and some *visitor location registers* (VLRs). Each network comprises a single HLR and many VLRs. The HLR is a centralized database containing the profiles of its assigned subscribers. Most of the current PCS manufacturers implement a combined MSC and VLR with one VLR per MSC. A VLR stores the profiles of the MUs that are currently residing in its associated LA. Figure 1 illustrates the PCS architecture and signaling network.

Two main procedures are used in the IS-41 location management scheme: *location update* and *location search*. A location update occurs when an MU moves to a new LA; a location search occurs when a fixed or mobile host wants to communicate with an MU whose current LA is unknown. In IS-41, the HLR is queried for every location search or update, resulting in tremendous strain on the use of the network resources as the number of PCS subscribers increases.

We describe the location update and location search procedures as specified in revision C of the IS-41 standard (TIA/EIA, 1996), along with some additional models that were proposed to augment it (Cayirci & Akyildiz, 2002; Lin, Lee, & Chlamtac, 2002). Figure 2 illustrates the

Figure 1. PCS architecture and signaling network

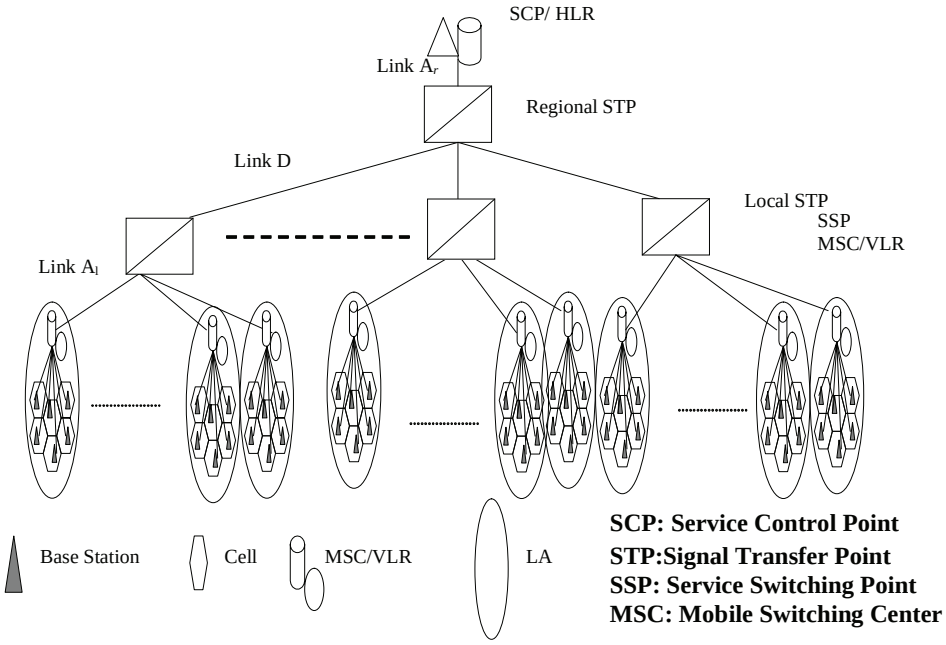
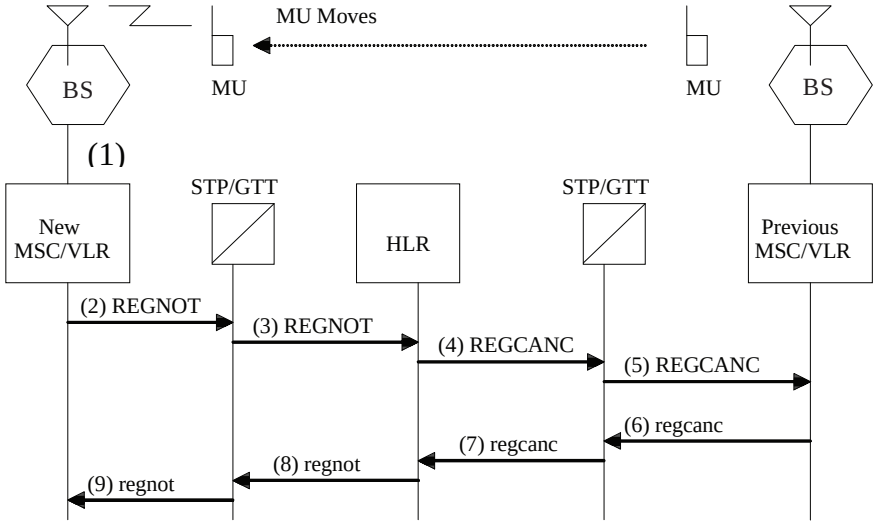


Figure 2. Location update procedure according to the IS-41 scheme

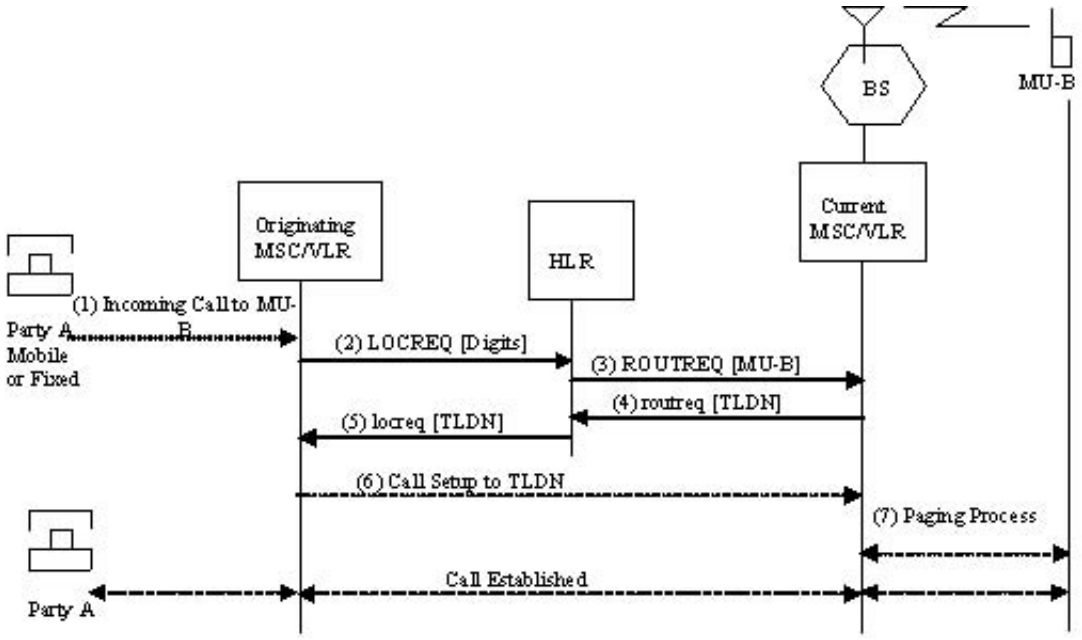


location update procedure of the IS-41 standard which can be described as follows. When an MU enters a new LA, it registers at its MSC or VLR. Then, the new MSC/VLR sends a *registration notification* message (REGNOT) to the network database HLR through an Single Transfer Point (STP) node. The STP node executes the global title translation (GTT) procedure using the MU's identification number in order to determine the HLR of the MU. Upon receiving the registration message, the HLR sends a *registration cancellation* message (REGCANC) to the MU's previous MSC/VLR. The previous MSC deletes the MU's profile in its associated VLR and sends a *cancellation acknowledgment* message (regcanc) to the HLR. The HLR acknowledges the location update by sending a regnot message that includes the MU's profile to the new MSC/VLR. When the new MSC/VLR receives the regnot message, it starts providing service to the MU.

The IS-41 location search procedure is presented in Figure 3. In this figure, a call is initiated to the mobile unit B (MU-B) through an originating MSC/VLR which sends a *location request* (*LOCREQ*) message to the HLR of the called MU through an STP node using the GTT technique. Upon receiving the request, the HLR sends a *routing request* (ROUTREQ) message to the MU-B's current MSC/VLR, which then allocates a temporary location directory number (TLDN) for the call and returns it to the HLR in a routreq message. The HLR relays the TLDN to the originating MSC/VLR in a locreq message which routes the call to the TLDN. Finally, the current VLR/MSC launches a paging process to find the MU-B's current cell. At this point, the communication has been established.

In these two procedures, the network database HLR is queried every time the MU moves to a new LA or receives a call. Several strategies have

Figure 3. Location search procedure according to the IS-41 scheme



been proposed to reduce the location update load and location search costs (Escalle, Giner, & Oltra, 2002; Mao, 2002; Suh, Choi, & Kim, 2000).

Advanced Mobility Management Schemes

A new signaling protocol was proposed by Wang & Akyildiz (2001) for intersystem roaming in next-generation wireless networks. According to this protocol, LAs that are on the boundary of two adjacent systems (*X* and *Y*) are called *peripheral location areas* (PLAs). Therefore, they are PLAs in systems *X* and *Y*. The intersystem location registration is controlled by a *boundary interworking unit* (BIU), which ensures the compatibility between the two systems and maintains a database of the roaming information of the mobile terminals (MTs) moving between the two networks. The underlying principle of this protocol is that the MT can request a location registration of intersystem roaming when it is in a PLA. As a result, it may finish signaling transformation and authentication before it arrives at the new system. Using this protocol, the HLR is not involved unless the MT goes from a PLA to a non-PLA. Moreover, Boundary Location Register (BLR) provides MTs with up-to-date location information; the incoming calls of the intersystem-roaming MTs are delivered to the serving MSC/VLR directly, rather than delivering it to the previous system. The numerical results show that the BLR protocol can reduce signaling costs, the latency of location registration, and call delivery, as well as call-loss rates for the MTs moving across different networks.

The models generally used to determine the optimal distance threshold are often based on certain simplified assumptions that do not give an accurate representation of a realistic cellular network (structured topology and configuration, geometric or exponential cell residence time distribution, symmetric random walk as a movement model, etc.). Wong and Leung (2001) overcome these drawbacks by focusing on the determination of the optimal update boundary for the distance-based location update algorithm, in a realistic environment, in order to minimize the expected total cost between call arrivals. The proposed model is applicable to arbitrary cell topologies; the call residence time can follow a general distribution and the movement history is taken into account. An implementation is described using an arbitrary cell topology. The location update is decided upon a simple table lookup.

Numerical results show that the proposed model gives a more accurate update boundary (distance threshold) in real wireless cellular environments compared with those derived from a hexagonal cell configuration with random-walk movement pattern. This is due to the fact that the network can maintain a better balance between the processing due to location update and the ratio bandwidth used for paging between call arrivals. The main drawback is the use of a Poisson distribution for the call arrival rate, although this may not be the case in real mobile environments.

The user mobility pattern (UMP) proposed by Cayirci and Akyildiz (2002) for location updates and paging where MTs keep track of their UMPs in a data structure called *user mobility pattern history* (UMPH). The UMP is a list of cells expected to be visited by the MT, and the UMPH records the mobility history of an MT. The model proposed by these authors differs from the other user profiles or history-data-based location update techniques in two main aspects. The first aspect is that according to this model, MTs are responsible to predict and register the UMPs, reducing the signaling traffic (for maintaining UMPH) and increasing the resolution of the data in UMPH. The second aspect is that an effective selective paging can be executed, based on call delivery times, by using the UMP nodes (pairs of cell identification and expected call entry time).

The performance of the UMP scheme is compared with the time-based and movement-based location update techniques, the blanket,

the selective, and the velocity paging techniques. The UMP technique creates less location update traffic than the other techniques when reasonable time intervals and movement thresholds are used. It consistently outperforms other paging performance techniques.

Cayirci and Akyildiz (2003) proposed the *traffic-based static LAD* (TB-LAD) scheme where the mobile traffic between the cells is predicted according to the characteristics of the crossing loads. Then, by using these traffic expectations, the traffic-based cell-grouping technique groups cells into LAs such that the neighbor cell with higher intercell traffic is assigned to the same LAs. Since the number of cells in an LA is a fixed parameter in TB-LAD, paging traffic is undisrupted. However, a better design of an LA reduces the number of location updates. Therefore, the TB-LAD decreases the number of location updates without increasing the number of paged cells during a call delivery.

To investigate this concept, traffic data was collected in a metropolitan area. Then, the relation of inter-LA traffic with the cell size and the number of cells in an LA was analyzed, and the performance was compared with the one of a *proximity-based LAD* (PB-LAD). Experimental results show that the TB-LAD reduces inter-LA traffic from 27% to 36% on the average over PB-LAD. TB-LAD outperforms PB-LAD when the average cell size exceeds 2,000 m and the average number of cells within an LA is superior to nine. In the metropolitan city where the experiments were conducted, the optimal solution was obtained with an LA size larger than 13 cells and a cell size larger than 2,500 m.

A trade-off cost analysis was made for the movement-based location update and paging in wireless mobile networks without using simplifying assumptions. A general framework was proposed by Fang (2003) for the study of such problems, and analytical results were derived to obtain the crucial quantity and the average number of location updates during an interservice time (used in all cost analyses under general assumptions). The analytical approach and results developed in his paper can be very useful to design an optimal mobility management scheme for future wireless mobile networks. In fact, the study shows that the total cost of the location update and paging is a convex function of the movement threshold.

Location Management with Mobile IP

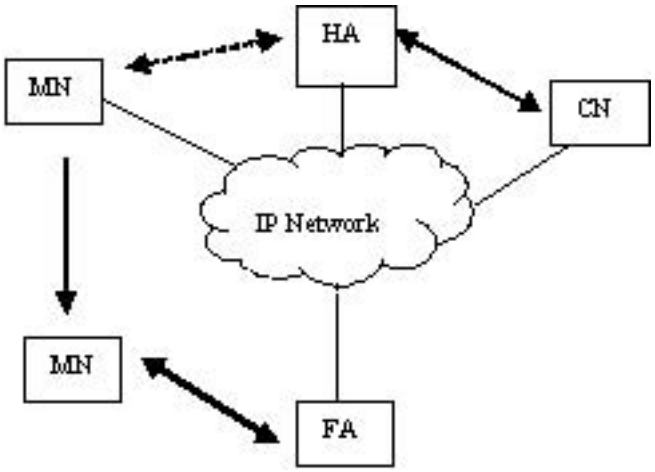
Mobility support in the IP protocol was developed by the Internet Engineering Task Force (IETF), leading to the Mobile IP protocol and all-IP networks (Chiussi, Khotimsky, & Krishnan, 2002). Currently, two versions of Mobile IP are available, Versions 4 (IPv4) and 6 (IPv6). An MN is a mobile node able to move from one subnet to another without requiring a change of IP address. The MN accesses the Internet via a home agent (HA) or a foreign agent (FA). The correspondent node (CN) is a node that connects with the MN. The HA is a local router on the MN's home network, and the FA is a router on the visited network. Figure 4 illustrates a Mobile IP network architecture.

The following operations are introduced by the Mobile IP protocol:

1. *Discovery*: How an MN finds an agent (HA or FA).
2. *Registration*: How an MN registers with its HA.
3. *Routing and Tunneling*: How an MN receives datagrams when visiting a foreign network.

Location management operations include agent discovery, movement detection, forming care-of-address (CoA), and location (binding) update. Handover operations include routing and tunneling.

Figure 4. Mobile IP network architecture



Discovery

To detect movement from one subnet to another, the MN uses two methods: advertisement lifetime and network prefix.

Advertisement Lifetime: The first detection method is based upon the lifetime field of the Internet Control Message Protocol (ICMP) router advertisement message. An MN records the lifetime indicated in any agent advertisement until that lifetime expires. If the MN has not maintained contact with its FA when the lifetime elapses, it must solicit a new agent.

Network Prefix: The second method uses the network prefix-detection movement. In some cases, an MN can determine whether or not a newly received agent advertisement was received on the same subnet by using its CoA. If the prefixes differ, the MN can assume that it has moved. This method is not available if the MN is currently using an FA's CoA. Once it has discovered a new FA and obtained a new CoA, the MN performs the location update procedure as follows:

1. The MN registers a new CoA with its HA by sending a binding update.
2. The MN notifies its CN of its current binding information.
3. If the binding update has an expiration date, the CN and the HA send a binding request to the MN to obtain the MN's current binding information.

Registration

When visiting any network away from home, each MN must have an HA. The MN registers with its HA in order to track the MN's current IP address. Two IP addresses are associated to each MN: one for location and one for identification purposes. The new IP address associated to an MN while it visits a foreign link is called its CoA. The association between the current CoA and the MN's home address is maintained by a mobility binding, so that packets destined to the MN may be routed using the current CoA, regardless of the MN's current point of attachment to the Internet.

Each binding has a predetermined lifetime period, which is negotiated during the MN's registration, after which time the registration is deleted. The MN must reregister within this period in order to continue servicing this CoA. Depending on its method of attachment, the MN sends location registration messages (when moving between two subnets or once the lifetime elapses) directly to its HA, or through an FA which forwards the registration to the HA. In either case, the MN exchanges registration request and registration reply messages based on IPv4 as follows.

1. The MN registers with its HA using a registration request message (the request may be relayed to the HA by the current FA).
2. The HA creates or modifies a mobility binding for that MN with a new lifetime.
3. The appropriate agent (HA or FA) returns a registration reply message containing necessary information on the request status, including the lifetime granted by the HA.

Routing and Tunneling

The process of routing datagrams for an MN through its HA often results in the utilization of paths that are significantly longer than optimal. Route-optimization techniques for Mobile IP employ the use of tunnels to minimize inefficient path use. For example, when the HA tunnels a datagram to the CoA, the MN's home address is effectively shielded from intervening routers between its home network and its current location. Once the datagram reaches the agent, the original datagram is recovered and delivered to the MN. Currently, there are two protocols for routing optimization and tunnel establishment: route optimization in Mobile IP and the tunnel establishment protocol.

Route Optimization in Mobile IP: The route optimization aims to define extensions to basic Mobile IP protocols that allow better routing so that datagrams can travel from a CN to an MN without first going to the HA. These extensions provide a means for nodes to cache the binding of an MN, and then tunnel datagrams directly to the CoA indicated in that binding, bypassing the MN's HA.

Tunnel Establishment Protocol: In this protocol, Mobile IP is modified in order to perform among arbitrary nodes. Upon establishing a tunnel, the encapsulating agent (HA) transmits protocol data units (PDUs) to the tunnel endpoint (FA) according to a set of parameters. The process of creating or updating tunnel parameters is called tunnel establishment. Generally, the establishment of parameters includes a network address for the MN. In order to use tunnel establishment to transmit PDUs, the HA must determine the appropriate tunnel endpoint for the MN. This is done by consulting a table that is indexed by the MN's IP address. After receiving the packets, the FA "decapsulates" the PDUs and sends them to the MN.

Mobile IP provides simple scalable mobility solutions. However, it causes excessive signaling traffic and long delays. Xie and Akyildiz (2002) introduced a distributed and dynamic, regional location management mechanism for Mobile IP to distribute traffic and dynamically adjust the regional network boundaries more efficiently. (A distributed gateway foreign agent (GFA) system architecture is proposed where each FA can function as either an FA or a GFA). This distributed system may allocate signaling load more uniformly. An active scheme is adopted by the distributed system to dynamically optimize the regional network size of each MN according to its current traffic load and mobility.

The distributed and dynamic scheme proposed by Xie and Akyildiz (2002) can perform optimally for all users from time to time, and the system robustness is enhanced. Since the movement of MNs does not follow a Markov process, a novel, discrete, analytical model for cost analysis, and also a new iterative algorithm to find out the optimal number of FAs in a regional network,

which consumes terminal network resources, was proposed. The proposed model is not plagued by constraints on the shape and the geographic location of Internet subnets.

Analytical results demonstrate that the signaling bandwidth is significantly reduced through the proposed distributed system architecture compared with the IETF Mobile IP regional registration scheme. It also demonstrates that the dynamic scheme has significant advantages under time-variant user parameters in cases where it is difficult to predetermine the optimal regional network size. The location management scheme requires that all FAs be capable of functioning as both FAs and GFAs. This increases the processing capability requirements of each mobile agent. There is additional processing load on the MTs, such as the estimation of the average packet arrival rate and subnet residence time.

Lee and Akyildiz (2003) proposed a cost-efficient scheme for route optimization to reduce the signaling costs caused by route optimization. Link-cost functions represent the network resources used by the routing path; signaling costs reflect the signaling and processing load incurred by route optimization. A Markovian decision model was used in order to find an optimal sequence for route optimization. In order to simplify the decision process, the model was restricted to intradomain handoff, which resulted in a decision rule. The optimal sequence is obtained by following the decision rule at each decision stage (minimizing the total cost). The performance of the optimal sequence is compared with the other sequences' action with route optimization (ARO) and action without route optimization (NRO). Simulation results show that the optimal sequence provides the lowest costs among the given sequences.

The recent advent of voice-over IP (VoIP) services and their fast growth is likely to play a key role in successful deployment of IP-based convergence of mobile and wireless networks. Kwon, Gerla, and Das (2002) have focused on mobility management issues regarding VoIP services in wireless access technologies. Different mobility management schemes are explored with a focus on Mobile IP and Session Initiation Protocol (SIP; Lin & Chen, 2003; Wu, Lin, & Lan, 2002); these two approaches are compared. The shadow registration concept is also presented; it aims at reducing distribution time in interdomain handoff for VoIP sessions in mobile environments. Considering the functionality of authentication, authorization, and accounting (AAA), the signaling message flow is illustrated for the two approaches in the presence and absence of shadow registration. Finally, an analytic comparison between the two approaches (Mobile IP and SIP) in terms of delay at initial registration and distribution in intradomain- and interdomain-handoff delay is presented.

Based on the previous analyses, numerical results show that the disruption for the Mobile IP handoff approach is smaller than the SIP approach in most situations. However, SIP shows shorter disruptions when the MN and the CN are nearby. Even though the smooth handoff is not taken into consideration in the disruption analyses, it is argued that it will play an important role in reducing disruption in interdomain handoff in the Mobile IP approach.

Misra, Das, Dutta, McAuley, and Das (2002) presented two enhancements to the Intradomain Mobility Management Protocol (IDMP) for IP-based hierarchical mobility management so that it can be used in a 4th Generation (4G) cellular environment. Thus, it would be highly relevant to develop IP-layer-fast handoff and paging solutions that would work across heterogeneous access technologies.

To minimize packet loss during intradomain handoffs, a time-bound localized multicasting approach is presented. By proactively informing its associated mobility agent (MA) of an imminent change, an MN enables the MA to multicast packets for a limited time span to a set of neighboring subnets. Subnet agents (SAs) buffer such multicast

packets for a short while. Then, if the MN enters its subnet, the SA is able to immediately forward these packets to the mobile, thus eliminating packet-loss delays.

Next-generation mobile and wireless networks are already under preliminary deployment and they are likely to use the current existing infrastructure for economic reasons. Global roaming in current and next-generation networks is a major issue for the integration and the interoperability of different systems. Zahariadis, Vaxevanakis, Tsantilas, Zervos, and Nikolaou (2002) proposed a hierarchical cell architecture consisting of an infrastructure that is either installed or under development. Soft horizontal-mobility management mechanisms and vertical handover (for Wireless Local Area Networks [WLAN], 2nd Generation [2G], and satellite networks) are discussed. An enhanced roaming scenario for next-generation networks is initiated by the MT and supported by an all-mobile IP network.

Mao (2002) presented an *intralocation-area location update* (intra-LA-LU) strategy in order to reduce paging traffic in mobile networks while keeping the standard location area update unchanged within the LA. The intra-LA-LU is performed whenever the MT changes its location between the anchor cell (the MT residing cell) and the rest of the cell in the LA. For call delivery, either the anchor cell or the other cells of the LA are paged to locate the MT. The proposed analytical model considers a continuous-time Markov chain to describe the MT movement.

Compared to the conventional location-tracking scheme, one can think that the proposed strategy will add an extra cost by performing the inter-LALUs. However, numerical results indicate that the savings of paging costs is much more significant than the newly added location update costs. The proposed strategy is suitable for users roaming in the LAs associated with their homes or workplaces. If the location of a mobile subscriber is uniformly distributed within an LA, the proposed strategy should not be used in order to avoid the intra-LA-LU, which is not suitable in this case.

Personal number (PN) or follow-me service allows users to access telecommunication services from any terminal in any location within the service area. In the existing systems, users have to manually register a phone number every time they enter a new area, which is an unsuitable solution.

Lin et al. (2002) proposed an enterprise approach for *automatic follow-me service* (AFS). The significance of the proposed approach is that AFS can be integrated with existing follow-me databases to automate PN services offered by different Public Switched Telephone Network (PSTN) service providers. AFS automatically connects calls to a user at any location with appropriate communications terminals. The authors showed how to implement AFS with the VoIP and Bluetooth technologies. More specifically, the AFS utilizes VoIP to communicate with the follow-me database in the public network, and Bluetooth is used to implement radio-tracking mechanisms. Then, the impact of polling frequency on power consumption and call misrouting was presented. The analysis provided shows that, based on the AFS cost functions, the optimal polling frequency can be found efficiently. One of the future extensions consists of developing automatic polling-frequency-adjustment heuristics based on the proposed analytic model.

NEW MOBILITY MANAGEMENT MODELS AND SCHEMES

This section introduces three new mobility management models: the *built-in memory model* (Safa, Pierre, & Conan, 2002), the *global location management scheme* (Safa et al.) and the Mobile IP network architecture (Diha & Pierre, 2003). Two of these models are based on the IS-41

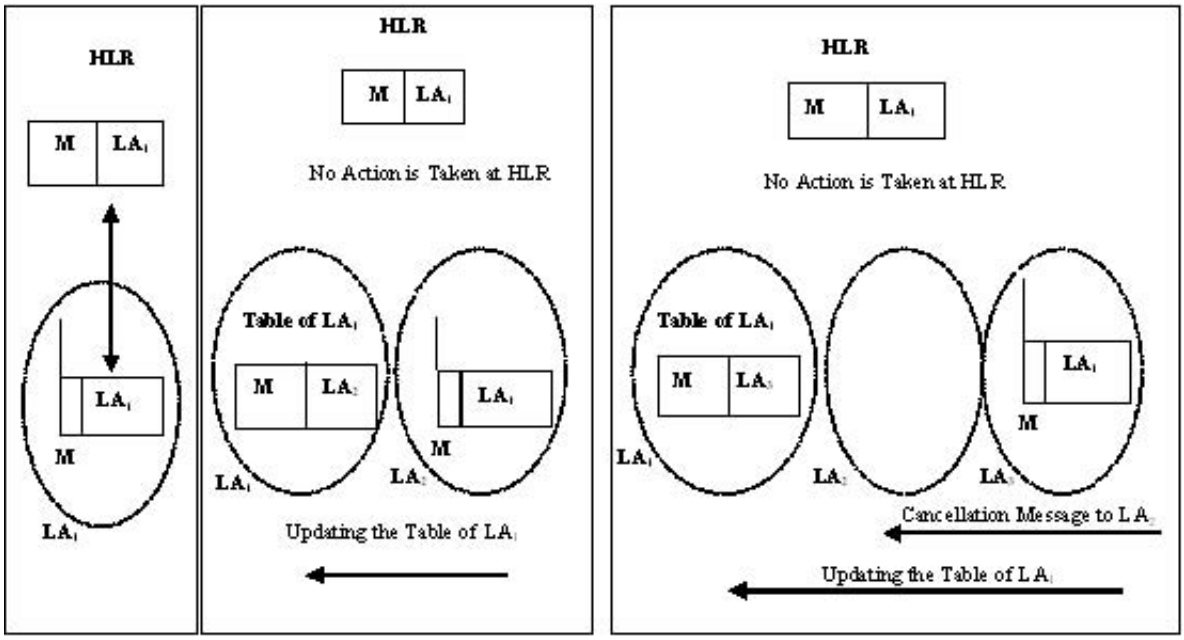
standard and aim to improve upon this standard to some extent.

Built-In Memory Model

The built-in memory model is based on the IS-41 standard with additional small, built-in MU memory and a pointer table for each LA. In this model, we define an MU's anchor LA as the LA for which the MU's data location was updated at the network database HLR. The MU built-in memory stores the address of its anchor LA, that is, the MU built-in memory data and the MU location data at the HLR are the same. The pointer table comprises two columns: the MU identification number (MIN) and the MU's current LA. The LA pointer table stores the current LA addresses of the MUs which consider this particular LA as their anchor LA. When the MU moves to a new LA, the new LA queries the MU's anchor LA to update the pointer table, that is, to create a pointer between the MU's anchor LA and the new LA. Consequently, no location update operation is performed at the HLR level. When the MU is called, its HLR is queried to determine its anchor LA. If it no longer resides in that LA, the call is forwarded to the MU's current LA by passing over a single pointer.

Figure 5 illustrates the built-in memory model. We assume that an MU joins the network in the location area LA_1 and registers at the MSC/VLR of this LA. Then, the MSC/VLR of the LA_1 location area queries the HLR to update the MU's location data (Figure 5a). Once the HLR is updated, the MU also updates its built-in memory. In other words, the location area LA_1 becomes the MU's anchor LA. Figure 5b shows how the MU moves to a new location area LA_2. Then,

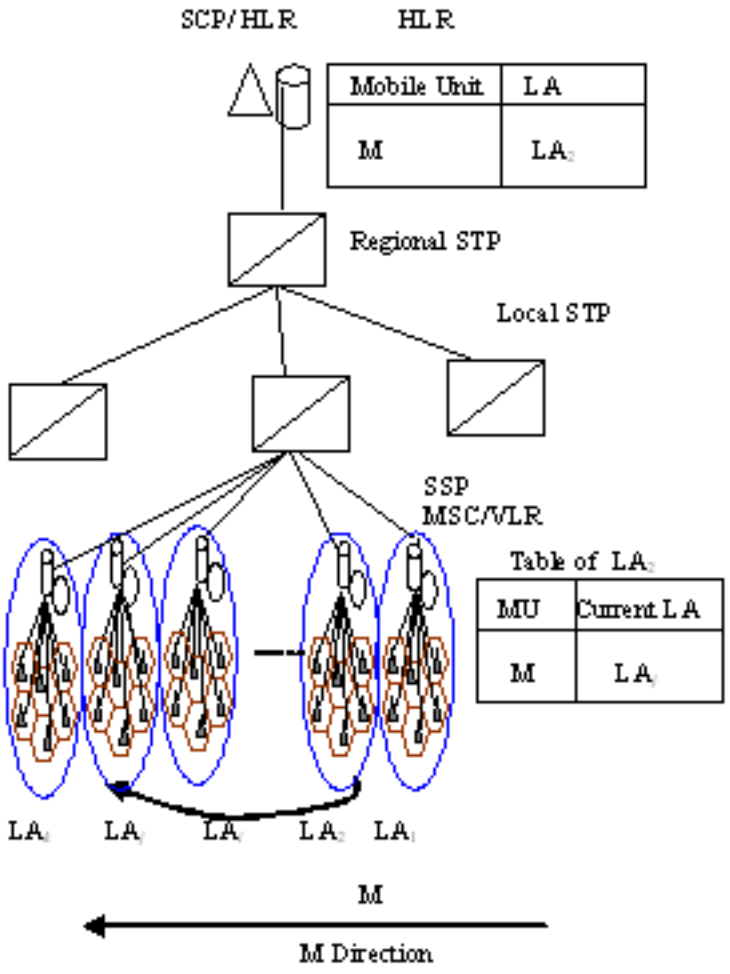
Figure 5. Illustration of the built-in memory model



(a) *(b)* *(c)*

instead of accessing the HLR, the MSC/VLR of the new location area LA_2 sends a query to the MU's anchor LA (LA_1) in order to establish a pointer between the MU's anchor LA (LA_1) and the current LA (LA_2). A pointer is established by updating any existing MU's location information in the pointer table of the MU's anchor LA, or adding this information to the table if there is none. If the MU leaves or returns to its anchor LA, the MSC/VLR of the MU's new LA sends a message to the MU's previous LA. However, when the MU's movement does not involve its anchor LA, the MSC/VLR of the MU's new LA sends two messages: a cancellation message to the MU's previous LA and a pointer updating message to the MU's anchor LA. This scenario is shown in Figure 5c. After the MU moves to the location area LA_3, the MSC/VLR of this LA sends a cancellation message to the MSC/VLR of the location area LA_2 and an updating message to

Figure 6. Example of the updating procedure



the MU's anchor LA (LA_1) in order to establish a pointer from LA_1 to LA_3.

The location update procedure in the memory built in model is shown in Figure 6. It is assumed that an MU joins the network in the location area LA_i, registers at the MSC/VLR of LA_i, and updates its location data in the HLR and its built-in memory, thus location area LA_i is the anchor LA of the MU.

A New Global Location Management Scheme

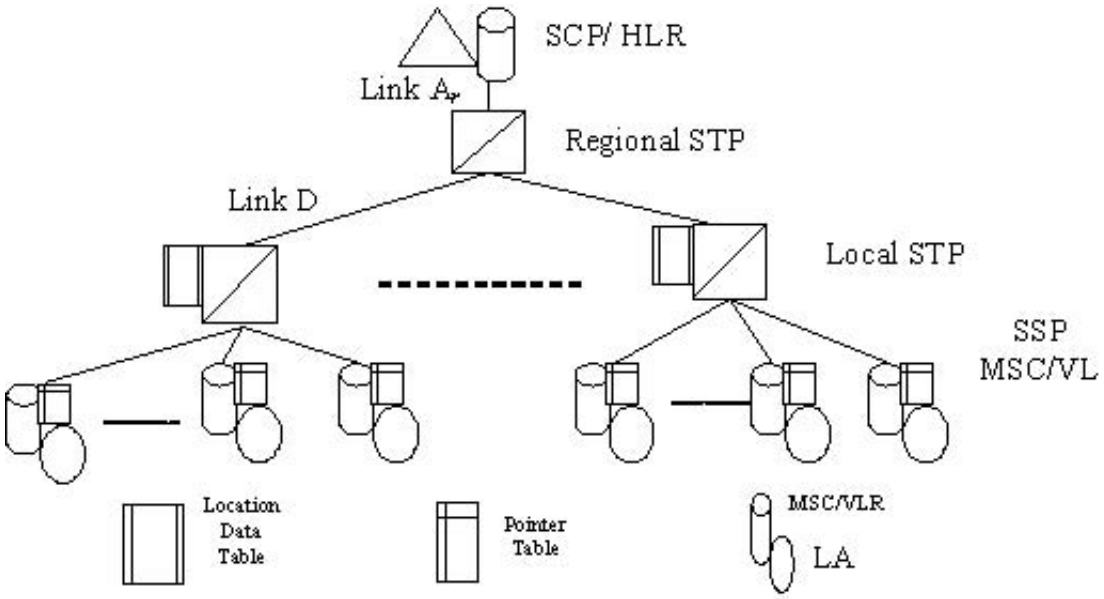
A second approach, called the global location management scheme, adds two tables to the conventional signaling network architecture (Figure 1), which are respectively identified as *location data* and *pointer tables.* A location data table is stored on an Local Signal Transfer Point (LSTP) node to serve all LAs connected to this LSTP. It contains the location information of some selected MUs, generally the ones which are frequently called from these LAs. This can significantly reduce the call-delivery procedure costs when the called MU has a profile in the location data table. Moreover, if this is not the case, there are no extra costs. In general, the use of a location data table serving many areas presents the following advantages:

- reduces network traffic by minimizing the number of updating queries sent to the data tables or to the HLR,
- saves table-installation costs,
- reduces data-storing redundancies and memory space wastes, and
- increases the frequency of locating an MU without accessing the network database. (For example, consider the case where an MU is often called from location area LA_1 but rarely from a neighboring location area LA_2. Since the same location data table may serve the two LAs, the called MU will be found locally even though it is called from LA_2.)

A pointer table is added to each LA. In order to explain the usefulness of the pointer table, we introduce the concept of the "anchor location area" of an MU as the LA in which the MU's location information is updated at the HLR. A pointer table of an anchor LA contains a pointer to the current LA of all the MUs having this LA as anchor LA. At this point, we must distinguish between two kinds of MU movements: an intra-STP move such that the new and the old LAs are connected to the same LSTP, and an inter-STP move such that the new and the old LAs belong to two different LSTPs.

We assume that each MU has a built-in memory that stores the address of its anchor LA and the addresses of the STP nodes which store the MU location information in their location data tables. This built-in memory is updated in either of two cases: The MU location information in the network database HLR is updated or the MU location information is added to a location data table. When the MU's movement is intra-LSTP, the pointer table of its anchor LA is queried to update the pointer to create a pointer between the anchor LA and the new LA. Hence, no location update operation is performed at the HLR. When the MU's movement is inter-LSTP, its new LA becomes its anchor LA and all information about the MU is then deleted from the previous anchor LA. The detailed location update and location search procedures of the proposed scheme will be presented further. They operate according to the signaling network architecture shown in Figure 7. It is implicitly assumed that when an MU's location information is added to a location data table, the MU is informed of this fact when it is called from any LA served by that location data table. Consequently, this operation does not require any additional costs.

Figure 7. Architecture of the signaling network used in the global location management scheme



When an MU moves to a new LA, a location update procedure is performed as illustrated in Figures 8 and 9, respectively, for intra-LSTP and inter-LSTP movements. For intra-LSTP movements, its anchor LA is updated instead of the HLR according to the following steps.

1. The MU moves to a new LA and sends a location update message to the MSC/VLR of this new LA.
2. The MSC of the new LA registers the MU with its associated VLR and sends a cancellation message to the previous LA.
3. The new LA queries also the MU's anchor LA in order to create a pointer from the anchor LA to the new LA. In other words, no location update operation is performed at the HLR. The anchor LA is the LA which stored the MU's address in the built-in memory of the MU and in the HLR.
4. and 5. The new LA receives an acknowledgement from both the previous and the anchor LAs.

When the MU's movement is inter-LSTP, the location update procedure shown in Figure 9 is performed. The steps of this procedure are described as follows.

1. The MU moves to a new LA and sends a location update message to the MSC/VLR of this new LA.
2. The MSC of the new LA registers the MU in its associated VLR and sends a registration notification message to the HLR via the STP.

Figure 8. Updating procedure for an intra-LSTP move

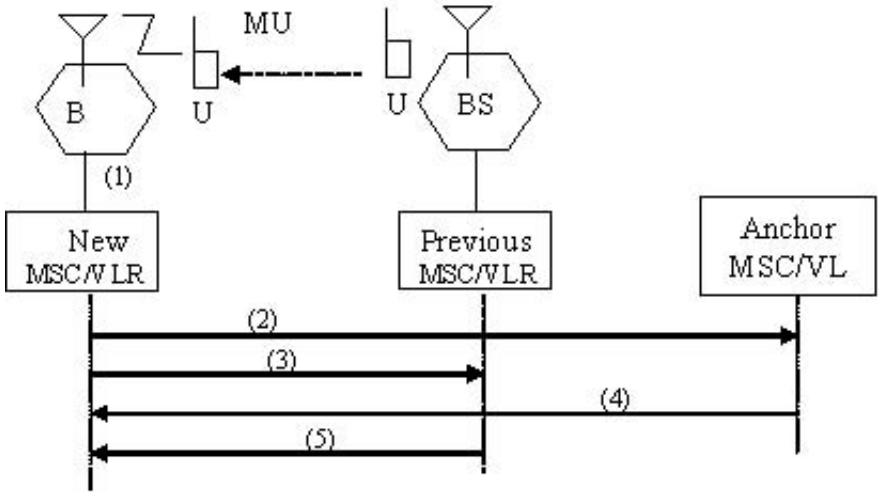
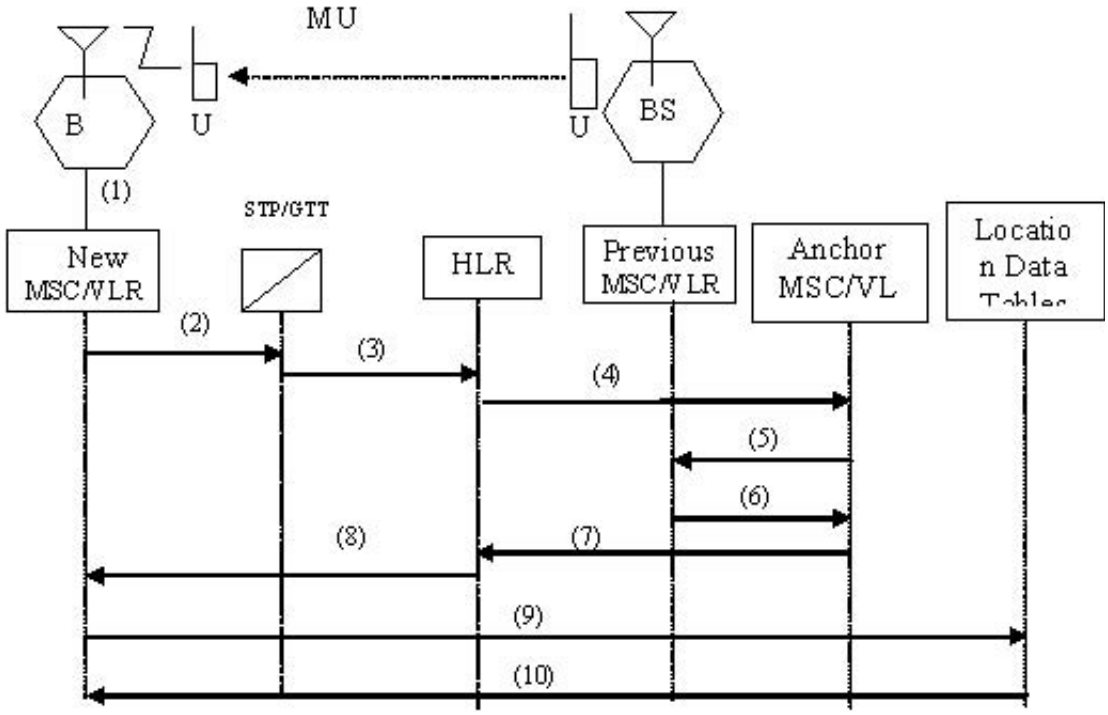


Figure 9. Updating procedure for an inter-LSTP move



3. The STP uses the MU's identification number and executes the GTT procedure to determine the MU's HLR. The registration message is then forwarded to the HLR.
4. The HLR sends a registration cancellation message to the MU's anchor LA.
5. The MU's anchor LA sends a cancellation message to the previous (old) MSC/VLR.
6. The old MSC deletes the MU's profile in its associated VLR and sends a cancellation acknowledgment message to the MU's anchor LA.
7. The anchor LA sends an acknowledgment to the HLR and deletes the MU's profile in its pointer table.
8. The HLR sends a registration confirmation message to the new MSC/VLR and provides the profile of the MU in this message. The new MSC/VLR becomes the MU's anchor LA.
9. The new MSC/VLR sends an update message to location data tables whenever necessary. (The MU provides the MSC/VLR with the addresses of those location data tables as stored in its built-in memory.)
10. After updating the MU data, the data location table sends an acknowledgement message to the new LA.

The location search procedure involves determining the current serving LA of a called MU. Figures 10, 11, 12, and 13 show four distinct, possible scenarios which must be followed by this procedure according to the proposed location management scheme.

Scenario 1: The first scenario, shown in Figure 10, addresses the case where the called MU has a record in the location data table and it is roaming in its anchor LA (i.e., the LA address stored in the location data table). The steps of Figure 10 are described as follows.

1. A call is initiated to an MU, forwarded to the MSC of the calling unit.

Figure 10. Searching procedure (Scenario 1)

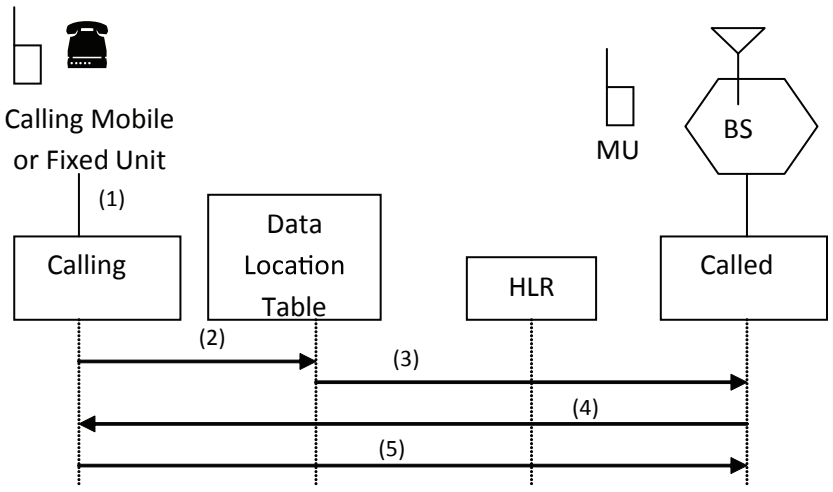


Figure 11. Location search procedure (Scenario 2)

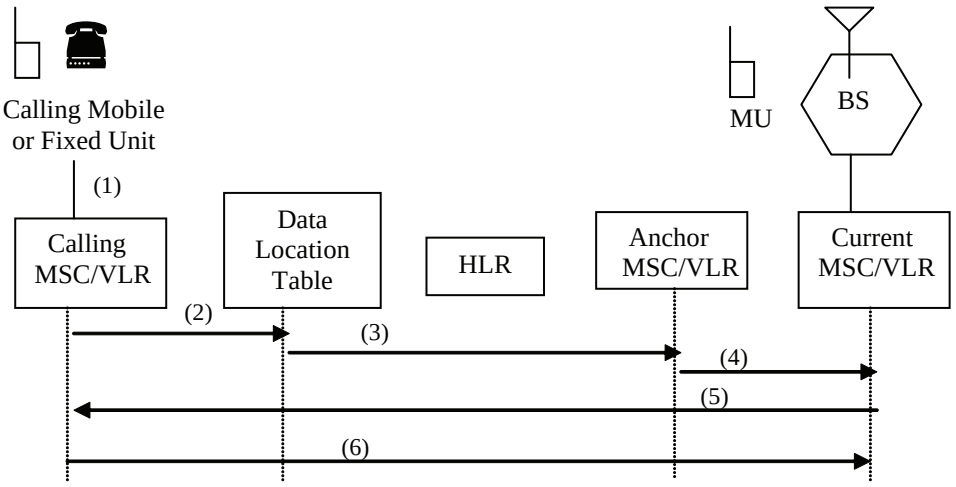
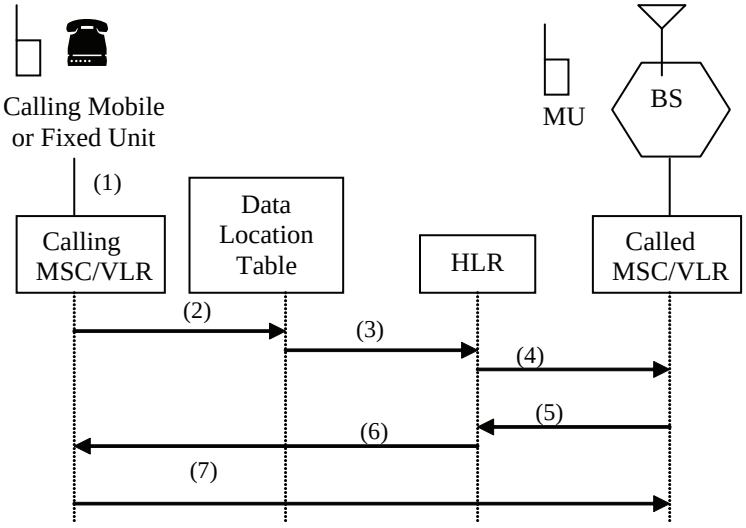
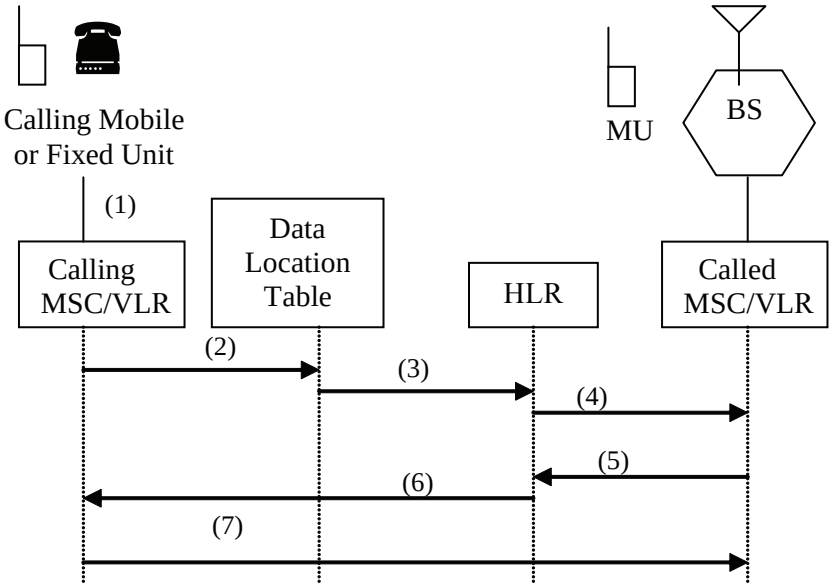


Figure 12. Location search procedure (Scenario 3)



2. The MSC sends a location request to its associated location data table which determines the anchor LA of the called MU.
3. The request is then forwarded to the anchor LA of the called MU.

Figure 13. Location search procedure (Scenario 4)



4. The called MU's MSC assigns a TLDN to the call and sends it to the calling MSC.
5. The calling MSC sets up a connection to the called MSC using this TLDN.

Scenario 2: The second scenario, shown in Figure 11, is similar to the first one. However, we assume that the called MU has a record in the location data table of the calling MU, but is not roaming in its anchor LA. In this case, a pointer should be crossed, at the destination side, to reach the current LA of the called MU. Then, the MU's current LA assigns the call a TLDN and sends it to the calling LA, which establishes a connection to the called MSC using this TLDN.

Scenario 3: The third scenario, shown in Figure 12, is the IS-41 call delivery scenario. In this case, the called MU does not have any records in the location data table, and the HLR is queried in order to determine the current LA of the called MU TLDN.

Scenario 4: The fourth scenario, shown in Figure 13, illustrates the situation where the called MU does not have any record in the location data table and is not roaming in its anchor LA. In this case, a pointer should be traversed to reach the current LA of the called MU. Then, the MU's current MSC/VLR assigns a TLDN to the call and returns it to the HLR, which forwards it to the calling MSC/VLR before the connection is established.

A New Mobile IP Network Architecture

Figure 14 illustrates a new Mobile IP network architecture proposed by Diha and Pierre (2003). The architecture introduces the following main features.

- Multiple connections of MNs and CNs to an FA or HA with different arrival rates in the network.
- The different procedures associated with an MN (registration, discovery, tunneling, and routing) represent different tasks with a specific priority.
- Multiprocessor agent (HA or FA). In this chapter, HA is emphasized. Also the HA is redundant to allow failure recovery. A main processor dispatches the different tasks arriving to an agent. A set of faster processors is defined to handle high-priority tasks.
- A processor can breakdown with a probability *p* and restarts with a probability 1 - *p*.

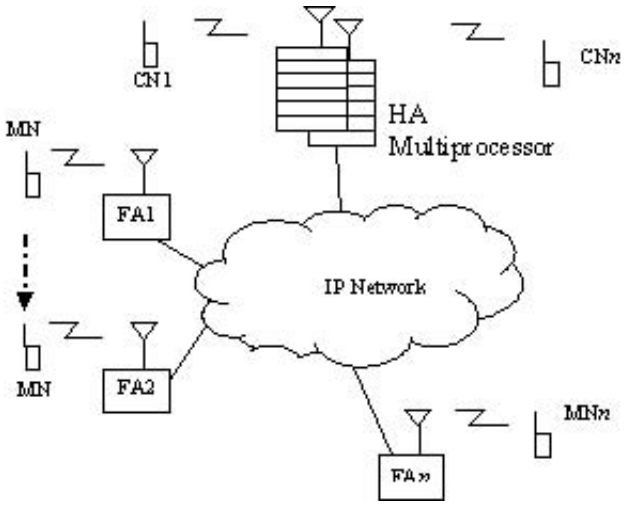
Based on this architecture, a set of new algorithms was defined to manage mobility with IP in mobile networks: registration algorithm, discovery algorithm, routing and tunneling algorithm, and task-scheduling and assignment algorithm.

Registration Algorithm

The registration procedure is a task running on the HA with the highest priority. It can preempt any other mobility management task for a given user. For example, during a tunneling procedure, if a registration request is received for the same user, the tunneling process will be delayed until the registration is completed. The different stages of the algorithm are described as follows.

1. The MN sends a registration request to the HA (it may imply the FA).
2. The HA verifies if a task other than the registration is in process for the same user. If so, this task is preempted by the registration task.

Figure 14. Mobile IP network architecture



3. The HA sends a response to the MN (it may imply the FA).
4. If the request is accepted, the procedure is completed; otherwise, the MN reattempts registration by sending a new request.

Discovery Algorithm

The discovery algorithm also introduces the notion of priority and it is based on lifetime expiration. The different steps are described as follows.

1. If the lifetime has expired and the MN notices the presence of an FA, it sends a registration request to the FA.
2. The FA verifies if higher priority tasks (e.g., registration procedure) are simultaneously executed for the same MN. If that is the case, the discovery process is delayed until the high-priority task execution is completed.
3. The FA returns a response to the MN.
4. If the registration succeeds, the MN sends new location information to its HA for location update.

Routing and Tunneling Algorithm

The new routing and tunneling algorithm also introduces the notion of priority. Thus, during a tunneling procedure, if a registration procedure is

Figure 15. Task scheduling and assignment algorithm

Given n tasks TSK_1, TSK_2, …, TSK_i, …, TSK_n
with the priorities p_1, p_2, …, p_i, …, p_n.
S a set of faster processors.
p_s the threshold for a critical task
(if $p_i > p_s$ then TSK_i is critical).
$u(i)$ the utilization rate of the task TSK_i.
$U(j)$ a vector of utilization rate of current tasks of the processor P_j.

BEGIN
Initialize i to 1.
Initialize $U(j)$ to 0.
Initialize S to 1 (at least one faster processor).
Sort the n tasks by descendant utilization rate order on the main processor P_0.

WHILE i ? n **DO**
j = min{k| $U(k) + u(i)$? 1}
IF (($p_i > p_s$) &&
($\exists P_k \in S$| (P_k is not broken down
&& ? $U(P_k) < 1$))) **THEN**
Assign TSK_i to processor P_k
ELSE IF P_j is not broken down **THEN**
Assign TSK_i to processor P_j
i ? $i+1$
END

END

received for the same user, the location procedure will be suspended until registration is completed. The algorithm steps are the following.

1. The HA receives data for an MN.
2. The HA verifies if registration is requested for the same user. In that case, the HA suspends the tunneling process until registration is completed.
3. The HA verifies if the MN is in the local network. If so, the packets are delivered through a regular IP packets-delivery procedure; otherwise, the HA forwards the packets to the MN via its current FA using its CoA.

Task Scheduling and Assignment Algorithm

The scheduling part of the algorithm is based on the Earliest Deadline First (EDF) algorithm (Johnson & Perkins, 2001). The tasks are sorted according to deadlines and assigned to processors. If a task is critical, it is assigned to a faster processor. A task is assigned to a processor only if its current utilization rate is less than one. This ensures that a processor is not used at its full capacity while others are not used. The algorithm verifies that the targeted processor is not broken down before assigning it a task. Figure 15 shows the task scheduling and assignment algorithm.

PERFORMANCE ANALYSIS

This section analyzes empirical results associated with some mobility management models proposed in the previous section. Then, their performance is evaluated by comparing them with other schemes.

Built-In Memory Model

To evaluate the performance of the location update scheme in the built-in memory model, we use a timing diagram similar to the one used by Jain and Lin (1995). The steady-state case between two consecutive phone calls was considered. In this analysis, MUs are classified by their CMR. The CMR is defined as the average number of calls to an MU per time unit, divided by the average number of times the MU changes LAs per time unit (or mean arrival rate/ mean mobility rate). If we assume that the incoming calls to an MU has a mean arrival rate λ, and the time that the user resides in an LA has mean 1/u, then, the CMR may be expressed by the following:

$$\text{CMR} = \frac{\lambda}{\mu}.$$

The objective is to determine the classes of MUs for which the memory-built-in model yields net reductions in signaling traffic and database loads. We define a(K) as the probability that the MU moves across K LAs between two phone calls. In order to evaluate a(K), the timing diagram shown in Figure 16 is used, where t_c denotes the interval between two consecutive phone calls to a mobile unit M. We suppose that the MU resides in a location area LA_0 when the first call arrives. After the first call, the MU visits another K LAs and remains in the location area LA_j for a period T_j ($0 \le j \le K$). Let t_i ($0 \le i \le K$) be the moment when the MU enters a new LA. T_i is then the interval between t_i and t_{i+1}. Let t_m denote the interval between the arrival of the first call and the time when the MU moves out of location area LA_0. Let T_i ($0 \le i \le K$) be independent, identically distributed, random variables with a general distribution $F_M(T_i)$, the density function $f_M(T_i)$, and the mean $E[T_i]$. The Laplace transform of T_i is then:

$$f_M^*(s) = \int_{t=0}^{\infty} e^{-st} f_M(t)dt. \quad (1)$$

Let $d_m(t)$ be the density function of t_m. Based on the random observer property (Mitrani, 1987), we can show that:

$$d_m(t) = \frac{1}{E[T_i]} \int_{x=t}^{\infty} f_M(x)dx = \frac{1}{E[T_i]}[1 - F_M(t)]. \quad (2)$$

Furthermore, we assume that the MU's residence time in an LA is exponentially distributed with parameter *m*. Hence, the density function of the MU residence time random variable T_i is $f_M(t) = \mu e^{-\mu t}$, and the expected residence time of an MU at an LA is

$$E[T_i] = \frac{1}{\mu}.$$

If the call arrivals to an MU are Poisson processes with mean arrival rate l, then, the interarrival time between two calls t_c is exponentially distribute with density function

$$f_c(t) = \lambda e^{-\lambda t} \text{ and } E[t_c] = \frac{1}{\lambda}.$$

Thus, the Laplace transform of te distributio of t_m is:

$$d_m^*(s) = \int_{t=0}^{\infty} e^{-st} d_m(t)dt = \int_{t=0}^{\infty} e^{-st} \mu[1 - F_M(t)]dt$$

$$= \frac{\mu}{s}[1 - f_M^*(s)]. \quad (3)$$

The probability a(K) that the MU moves across K LAs between two phone calls can be derived using Equations 1, 2, and 3:

$$\alpha(K) = \frac{\mu}{\lambda}[1 - f_M^*(\lambda)]^2[f_M^*(\lambda)]^{k-1}. \quad (4)$$

Let N denote the average number of location update operations performed among K moves. From Equation 4, N can be derived as follows:

$$N = \sum_{i=0}^{\infty} i\alpha(i) = \frac{\mu}{\lambda} = \frac{1}{CMR}. \quad (5)$$

Let M and L denote, respectively, the costs of IS-41 location update and location search procedures. Let m be the cost of the location update operation in the proposed built-in memory model. Let T denote the cost of traversing a link (pointer) between two LAs. We denote total costs between two consecutive calls for various operations used in this analysis as follows:

- U_{IS41}: Total cost of location update operations using the IS-41 scheme
- S_{IS41}: Total cost of location search operations using the IS-41 scheme
- $Total_{IS41}$: Total cost of location update and location search operations using the IS-41 model
- U_{BM}: Total cost of location update operations using the built-in memory strategy
- S_{BM}: Total cost of location search operations using the built-in memory model
- $Total_{BM}$: Total cost of location update and location search operations using the built-in memory model

The average number of location search operations executed between two consecutive phone calls is one. In the worst-case scenario, the cost of the location search procedure in the built-in memory model equals the cost of location search in the IS-41 scheme, plus the cost of traversing

a pointer. Then, the total costs can be easily calculated with the following:

$$U_{IS41} = \frac{M}{CMR}, \quad (6)$$

$$Total_{IS41} = U_{IS41} + S_{IS41} = \frac{M}{CMR} + L, \quad (7)$$

$$U_{BM} = \frac{m}{CMR}, \text{ and} \quad (8)$$

$$Total_{BM} = U_{BM} + S_{BM} = \frac{m}{CMR} + L + T. \quad (9)$$

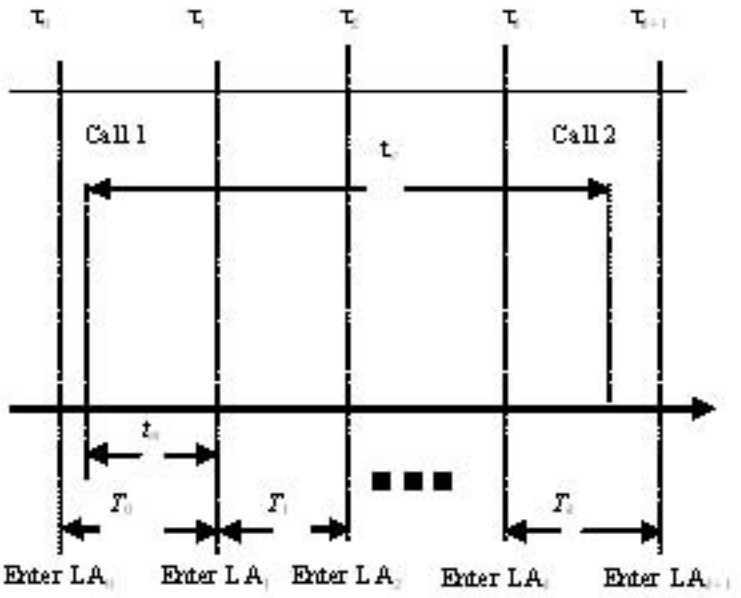
As a first approximation, we consider the values of the operation costs as follows. We observe that the location update procedure and the location search procedure in the IS-41 scheme involve the same number of messages between HLR and VLR databases. Therefore, we assume that $M = L$. Using the same reasoning, we observe that the number of messages in the built-in memory location-update procedure is about 4 times the number of messages required to cross links between two LAs served by the same LSTP, that is, we set $T = m/4$. Finally, we normalize $M = 1$. Then, from Equations 6, 7, 8, and 9, we obtain:

$$\frac{U_{BM}}{U_{IS41}} = \frac{m}{M} = m, \quad (10)$$

$$\frac{Total_{BM}}{Total_{IS41}} = \frac{4(m + CMR) + m * CMR}{4(M + CMR)}, \text{and} \quad (11)$$

$$\frac{S_{BM}}{S_{IS41}} = \frac{L + \frac{m}{4}}{L} = 1 + \frac{m}{4}. \quad (12)$$

Figure 16. Timing diagram



For simulation purposes, we assume approximate values for *m*. Figure 17 shows the total location update cost in the built-in memory model compared to the total location update costs in the IS-41 model when $m = 0.2$ and $m = 0.5$. We observe that for $m = 0.2$, the built-in memory model results in a cost reduction of 80% while the extra cost paid to locate an MU is, in the worst case, 5%. For $m = 0.5$, the total reduction of the location update cost is 50% and the extra cost required to locate the MU is 12.5%. The reduction level depends on the *CMR* of each MU. When *CMR* is low, significant cost savings are obtained with the location update in the built-in memory model. However, when *CMR* tends to infinity, the location update cost in both models tends to zero. This can be explained as follows. As *CMR* tends to infinity, the MU never moves out of an LA, so location updates are not performed.

Figure 18 shows that the reduction obtained in the total cost varies between 22% and 78% when $m = 0.2$, and between 8% and 52% when $m = 0.5$. To gain a better understanding of these results, we analyze the lower and upper bounds for the relative cost $Total_{BM}/Total_{IS41}$ given in Equation 11. We note that the lower and upper bounds of this performance measure occur as $CMR \to 0$ and $CMR \to \infty$, respectively. Then, we can write:

$$m \le \frac{Total_{BM}}{Total_{IS41}} \le 1 + \frac{m}{4}. \qquad (13)$$

Equation 13 can be explained as follows. When the *CMR* is low, the mobility rate is high compared to the call arrival rate, and the total costs of both schemes are dominated by the cost of the location update procedure. Since the built-in–memory, location-update procedure results in significant cost reduction, it outperforms the IS-41 scheme. Conversely, when the *CMR* is high, the call arrival rate is high compared to the mobility rate, and the cost of the location search procedure

Figure 17. Total location update cost: IS-41 versus built-in memory when (a) m = 0.2 and (b) m = 0.5

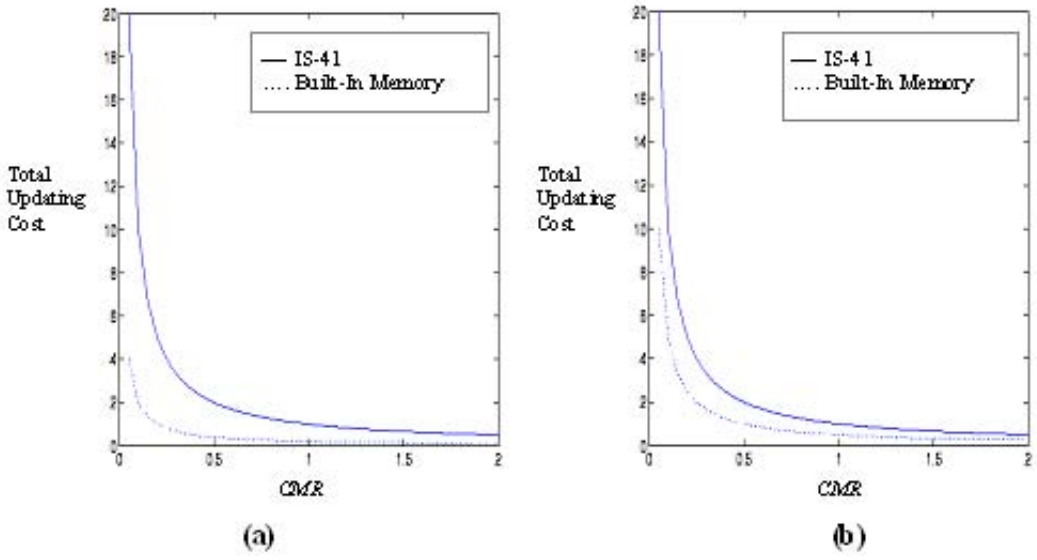
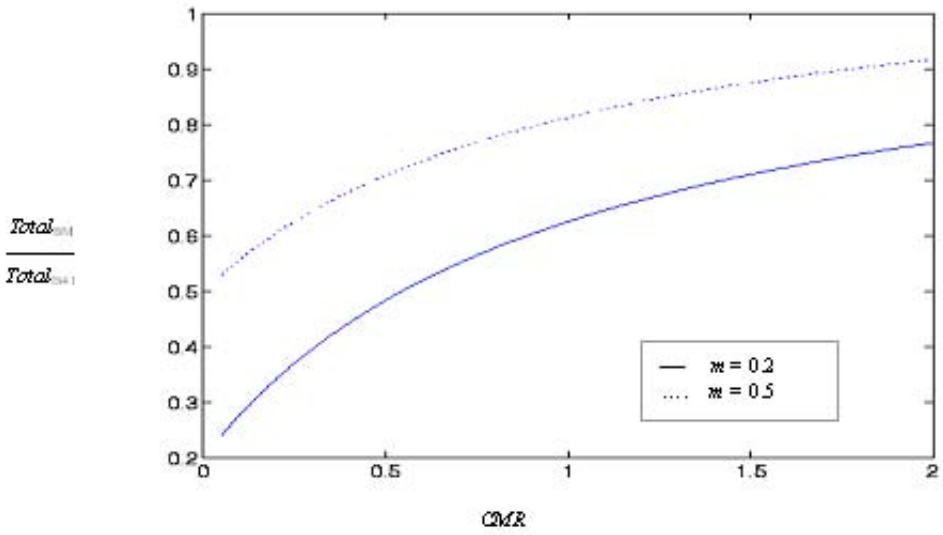


Figure 18. Reductions obtained with the built-in memory model



dominates. As the average number of incoming calls to an MU increases, the $Total_{BM}$ value approaches the $Total_{IS41}$ value. When the $CMR \rightarrow \infty$, the MU never moves out of its LA. In this case, it is likely to update the HLR which will point to the MU's current LA. Consequently, the cost of the location search procedure will be equal in both schemes and the upper bound of $Total_{BM}/Total_{IS41}$ will be 1. Otherwise, the $Total_{BM}/Total_{IS41}$ upper bound will be $1 + m/4$ where $m/4$ represents the cost associated with moving between two LAs in the memory-built-in scheme.

The Global Location Management Scheme

This section investigates classes of users for which the global location management scheme yields a net reduction in signaling traffic and database loads. All users can be classified according to their CMR. For each target MU, the following quantities are defined.

l: Average number of calls to a target MU per time unit
m: Average number of times the user changes LA per time unit
1/m: Average LA residence time for a target MU
p: Probability that the called MU has a profile in the location data table
q: Probability that the new LA (VLR/MSC) is served by the same LSTP as the previous VLR (intra-LSTP movement)
r: Probability that the called MU is found in its anchor LA

The costs of the various operations used in the proposed architecture are denoted as follows.

U_{intra}: Location update operation costs when the MU's move is intra-LSTP

U_{inter}: Location update operation costs when the MU's move is inter-LSTP

U_{global}: Estimated cost of a location update operation

S_1: Cost of a location search operation using Scenario 1 (i.e., when the called MU has a record in the location data table and is found in its anchor LA)

S_2: Cost of a location search operation using Scenario 2 (when the called MU has a record in the location data table and is not found in its anchor LA)

S_3: Cost of location search operation using Scenario 3 (when the MU has no record in the location data table and is found in its anchor LA)

S_4: Cost of location search operation using Scenario 4 (when the MU has no profile in the location data table and is not found in its anchor LA)

S_{global}: Estimated cost of a location search operation

C_{global}: Total cost for location search and location update operations

The estimated cost for the location update procedure is given by:

$$U_{global} = qU_{intra} + (1 - q)U_{inter}. \quad (14)$$

The estimated cost for the location search procedure is given by:

$$S_{global} = p(rS_1 + (1 - r)S_2) + (1 - p)(rS_3 + (1 - r)S_4). \quad (15)$$

The total cost per time unit for location search and location update is given by:

$$C_{global} = mU_{global} + lS_{global}. \quad (16)$$

In order to compute the cost of the location update procedure based on the global scheme using the reference network architecture (Figure 7) and the two procedures presented in Figures 8 and 9, we define the following costs for crossing various network elements.

A_l: Cost of transmitting a message on A-link between Service Switching Post (SSP) and LSTP

D: Cost of transmitting a message on D-link between LSTP and Regional Signal Transfer Point (RSTP)

A_r: Cost of transmitting a message on A-link between RSTP and Service Control Point (SCP)

L: Cost of processing and routing a message by LSTP

R: Cost of processing and routing a message by RSTP

C_H: Cost of a database update or query at the HLR

C_v: Cost of a database update or query at the VLR

Based on the location update procedure shown in Figure 8, the location update cost for an intra-LSTP move is given by:

$$U_{intra} = 8A_l + 4L + 2C_v. \quad (17)$$

For an inter-LSTP move, the cost of a location update operation equals the cost of the location update according to the IS-41 standard, plus the cost of updating the location data tables that store the location data of the moving MU, and the cost of updating the MU's previous anchor LA. According to Figure 9, this cost is given by:

$$U_{inter} = 4(A_l + L + A_r + D + R) + 4A_l + 2L + 2C_V + C_H + \sum_{k \in E} U_k, \quad (18)$$

where E is the set of location data tables to be updated after an inter-LSTP move and U_k is the cost of updating a location data table:

$$U_k = 2(A_l + L + R + 2D).$$

The estimated cost of the location update procedure can be derived using Equations 14, 17, and 18 as:

$$U_{\text{global}} = q(8A_l + 4L + 2C_v) + (1 - q)(4(A_l + L + A_r + D + R) + 4A_l + 2L + 2C_V + C_H + \sum_{k \in E} U_k)$$

$$= 8qA_l + 4qL + 2qC_v + 8A_l + 6L + 4A + 4D + 4R + 2C_V + C_H + \sum_{k \in E} U_k - 8qA_l - 6qL - 4qA_r - qD - 4qR - 2qC_V - qC_H - q \sum_{k \in E} U_k,$$

which can be simplified as:

$$U_{\text{global}} = 8A_l + 4L + 2C_V + (1 - q(2L + 4A_r + 4D + 4R + C_H + \sum_{k \in E} U_k). \quad (19)$$

Let t be the probability that the LAs of the called MU and calling unit are connected to the same LSTP. Since updating the location data table and the pointer table involves a simple access to a local memory, we assume that there are no additional costs to update or query this kind of table.

Scenarios 1 and 2 are applied when the called MU has a record in the location data table of the calling MU. The cost of Scenario 1, which is used when the called MU is found in its anchor LA, is given by:

$$S_1 = t(4A_l + 2L + C_V) + (1 - t)(4A_l + 4L + 2R + 4D + C_V)$$
$$= 4tA_l + 2tL + tC_V + 4A_l + 4L + 2R + 4D + C_V - 4tA_l - 4tL - 2tR - 4tD - tC_V$$
$$= 4A_l + 4L + 2R + 4D + C_V - 2tL - 2tR - 4tD. \quad (20)$$

When the called MU is not found in its anchor LA, Scenario 2 is applied. The cost of this scenario equals the cost of Scenario 1, plus the cost of passing over a pointer from the anchor LA of the MU to its current LA. This cost is given by:

$$S_2 = S_1 + 2A_l + L$$
$$= 6A_l + 5L + 2R + 4D + C_V - 2tL - 2tR - 4tD. \quad (21)$$

When the called MU does not have its location information stored in the location data table, Scenarios 3 and 4 are applied. The cost of Scenario 3, which is used when the called MU is found in its anchor LA, equals the cost of the location search in IS-41, given as:

$$S_3 = 4(A_l + L + A_r + D + R) + C_V + C_H. \quad (22)$$

When the called MU is not found in its anchor LA, Scenario 4 is applied. The cost of this scenario equals the cost of Scenario 3, plus the cost of passing over a pointer from the anchor LA of the MU to its current LA. As in Equation 21, this cost is:

$$S_4 = S_3 + 2A_l + L. \quad (23)$$

The total cost per time unit for locating an MU can be expressed as follows:

$$S_{\text{global}} = p(rS_1 + (1 - r)S_2) + (1 - p)(rS_3 + (1 - r)S_4). \quad (24)$$

Using Equations 21, 26, 27, 28, and 29, this cost can be rewritten as follows:

$$S_{\text{global}} = 6A_l + 5L + 4A_r + 4D + 4R + C_V + C_H - p[2t(L + R + 2D) + 4A_r + 2R + C_H] - r(2A_l + L). \quad (25)$$

The total cost per time unit for location update and location search using the global architecture is obtained using Equations 16, 19, and 25.

$$C_{\text{global}} = m[8A_l + 4L + 2C_V + (1 - q(2L + 4A_r + 4D + 4R + C_H +)] + \sum_{k \in E} U_k,$$
$$l\{6A_l + 5L + 4A_r + 4D + 4R + C_V + C_H - p[2t(L + R + 2D) + 4A_r + 2R + C_H] - r(2A_l + L)\} \quad (26)$$

For comparison purposes, we need to evaluate the costs of the original IS-41 scheme. We denote costs for various operations used in the IS-41 scheme as follows:

U_{IS41} : Cost for a location update operation
S_{IS41} : Cost for a location search operation
C_{IS41} : Total cost per time unit for location search and location update operations

The total cost per time unit for location update and location search under the IS-41 scheme is:

$$C_{\text{is41}} = mU_{\text{is41}} + lS_{\text{is41}}, \quad (27)$$

where

$U_{\text{is41}} = 4(A_l + L + A_r + D + R) + 2C_V + C_H$, and
$S_{\text{is41}} = 4(A_l + L + A_r + D + R) + C_V + C_H$.

Defining the relative cost of the global location management scheme as the ratio of the total cost per time unit for the global scheme to that of the IS-41 scheme, $C_{\text{global}}/C_{\text{is41}}$, we get as a function of the user's CMR:

$$\frac{C_{\text{global}}}{C_{\text{is41}}} = \frac{U_{\text{global}} + (\lambda/\mu)S_{\text{global}}}{U_{\text{is41}} + (\lambda/\mu)S_{\text{is41}}}. \quad (28)$$

Relation 28 uses the four probability terms: *p, q, r,* and *t*. Both *p* and *r*, which were defined above, can be used to classify users.

In order to quantify *q* and *t*, we assume that an LSTP consists of $x * x$ LAs arranged in a square, and each LA is itself a square. MUs are assumed to be uniformly distributed throughout the LSTP area and each MU exhibits the same arrival call rate at every VLR/MSC. Furthermore, each time an MU leaves an LA, one of the four sides is crossed with equal probability. Then, the probability that the MU's move is inter-LSTP is equal to the probability that the MU is in a border LA, multiplied by the probability that the MU's next move is to an LA belonging to a different LSTP. Define:

P_1 = Prob[MU lies in a border LA of the LSTP] = $4(x - 1)/(x * x)$
P_2 = Prob[MU's next move is to an LA belonging to a different LSTP] = 1/4
Þ Prob[MU's move is inter-LSTP] = $P_1 * P_2$ = $(x - 1)/(x * x)$.

Hence,

q = Prob[MU's move is intra-LSTP] = $1 - (x - 1)/(x * x)$.

Also, let us assume that all of the network SSPs are uniformly distributed among *n* LSTPs, and each SSP corresponds to a single LA. For example, take the case of the public, switched telephone network that includes 160 local access transport areas (LATAs) across the seven Regional Bell Operating Company (RBOC) regions (Bellcore, 1992), assuming one LSTP per LATA and the average number of LSTPs is 160/7, or 23 per region. Given that there are 1,250 SSPs per region, the number of SSPs per LSTP is 1250/23. Hence,

$$x = \sqrt{\frac{1250}{23}} \approx 7.4 \Rightarrow q \approx 0.88.$$

Under the conditions stated above, the probability *t* that both calling and called users are found in the same LSTP equals $1/n$ Þ $t = 1/23 = 0.043$. Further quantitative results associated

to the performance of the four scenarios can be found in Safa et al. (2002).

CONCLUSIONS AND FUTURE WORK

In this chapter, we analyzed and proposed some mobility management models and schemes by taking into account their capability to reduce search and location update costs in wireless networks. The first model proposed is called the built-in memory model; it is based on the architecture of the IS-41 network and aims to reduce the HLR access overhead. The performance of this model was investigated by comparing it with the IS-41 scheme for different CMRs. Experimental results indicate that the proposed model is potentially beneficial for large classes of users and can yield substantial reductions in total user-location management costs, particularly for users who have a low CMR.

The built-in memory model appears promising when the higher elements of the network constitute the network performance bottleneck. The results show that the cost reduction obtained on the location update is very significant, while the extra costs paid to locate an MU simply amount to the costs of crossing a single pointer between two LAs. The built-in memory model was also compared with the forwarding pointers' scheme. The results show that this model consistently outperforms the forwarding pointers' strategy.

A second location management model to manage mobility in wireless communications systems was also proposed. According to this scheme, two tables are added to the IS-41 network architecture. A pointer table is added to each LA, and it tracks the MUs that moved out of this LA by setting a single pointer from this LA to the current LA. The location data table is located on an LSTP node and contains the data location of the MUs that are frequently called from the LAs connected to this LSTP. The results have shown that significant cost savings can be obtained compared with the IS-41 standard location management scheme, depending on the value of the MUs' CMR.

Finally, we presented the Mobile IP network architecture and mobility management algorithms in a real-time context. Compared to some conventional architecture and algorithms, the implementation of the architecture and algorithms produced better results for the location update and tunneling average times, as well as the CMR.

Many investigations are ongoing in real-time mobility management for Mobile IP networks. Such investigations address the implementation of real-time algorithms in real networks and suggest new algorithms and architectures. Also, since the current protocols are designed for micromobility, the global roaming area remains a highly challenging research domain.

REFERENCES

Akyildiz, I. F., & Wang, W. (2002). A dynamic location management scheme for next generation multi-tier PCS system. *IEEE Transactions on Wireless Communications, 1*(1), 178-190.

Bellcore (1992). *Switching system requirements for interexchange carrier interconnection using integrated services digital network user part (ISDNUP)*(Tech. Ref. No.TR-NWT-000394).

Cayirci, E., & Akyildiz, I. F. (2002). User mobility pattern scheme for location update and paging in wireless systems. *IEEE Transactions on Mobile Computing, 1*(3), 236-247.

Cayirci, E., & Akyildiz, I. F. (2003). Optimal location area design to minimize registration signaling traffic in wireless systems. *IEEE Transactions on Mobile Computing, 2*(1), 76-85.

Chiussi, F. M., Khotimsky, D. A., & Krishnan, S. (2002). Mobility management in third-generation all-IP networks. *IEEE Communications Magazine, 40*(9), 124-135.

Diha, M., & Pierre, S. (2003). Architecture and algorithms for real-time mobility management in mobile IP networks. *Proceedings of the Second International Conference on Ad-Hoc, Mobile, and Wireless Networks (ADHOC-NOW),* (pp. 49-59).

Escalle, P. G., Giner, V. C., & Oltra, J. M. (2002). Reducing location update and paging costs in a PCS network. *IEEE Wireless Communications, 1*(1), 200-209.

Fang, Y. (2003). Movement-based mobility management and trade off analysis for wireless mobile networks. *IEEE Transactions on Computers, 52*(6), 791-803.

Gallagher, M. D., & Randall R. A. (1997). *Mobile telecommunications networking with IS-41.* New York: McGraw-Hill.

Jain, R., & Lin, Y. B. (1995). An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS. *Wireless Networks, 1*(2), 197-210.

Johnson, D. B., & Perkins, C. (2001). *Mobility support in IPv6* [Internet draft]. Internet Engineering Task Force. Retrieved from http://users.piuha.net/jarkko/publications/mipv6/drafts/mobilev6.html

Kwon, T. T., Gerla, M., & Das, S. (2002). Mobility management for VoIP service: Mobile IP vs. SIP. *IEEE Wireless Communications, 9*(5), 66-75.

Lee, Y. J., & Akyildiz, I. F. (2003). A new scheme for reducing link and signaling costs in mobile IP. *IEEE Transactions on Computers, 52*(6), 706-712.

Lin, Y. B., & Chen, Y. K. (2003). Reducing authentication signaling traffic in third-generation mobile network. *IEEE Transactions on Wireless Communications, 2*(3), 493-501.

Lin, Y. B., Cheng, H. Y., Cheng, Y.H., & Agrawal, P. (2002). Implementing automatic location update for follow-me database using VoIP and Bluetooth technologies. *IEEE Transactions on Computers, 51*(10), 1154-1168.

Lin, Y. B., Lee, P. C., & Chlamtac, I. (2002). Dynamic periodic location area update in mobile networks. *IEEE Transactions on Vehicular Technology, 51*(6), 1494-1501.

Mao, Z. (2002). An intra-LA location update strategy for reducing paging cost. *IEEE Communications Letters, 6*(8), 334-336.

Misra, A., Das, S., Dutta, A., McAuley, A., & Das, S. K. (2002). IDMP-based fast handoffs and paging in IP-based 4G mobile networks. *IEEE Communications Magazine, 40*(3), 138-145.

Mitrani, I. (1987). Modelling of computer and communication *system*. New York: Cambridge University Press.

Mouly, M., & Pautet, M. B. (1992). *The GSM system for mobile communications* (pp. 434-465). Telecom Pub: Alexandria, VA.

Safa, H., Pierre, S., & Conan, J. (2001). An efficient location management scheme for PCS networks. *Computer Communications, 24*(14), 1355-1369.

Safa, H., Pierre, S., & Conan, J. (2002). A built-in memory model for reducing location update costs in mobile wireless networks. *Computer Communications, 25*(14), 1343-1353.

Suh, B., Choi, J., & Kim, J. (2000). Design and performance analysis of hierarchical location management strategies for wireless mobile communication systems. *Computer Communcations Journal, 23,* 550-560.

TIA/EIA (1996). *Interim Standard IS-41-C.* Cellular Radio-Telecommunications Intersystem Operations. Retrieved 2004, from http://www.cdg.org/technology/cdma_technology/a_ross/Standards.asp

Wang, W., & Akyildiz, I. F. (2001). A new signaling protocol for intersystem roaming in next-genera-

tion wireless systems. *IEEE Journal on Selected Areas in Communications, 19*(10), 2040-2052.

Wong, V. W. S., & Leung, V. C. M. (2001). An adaptive distance-based location update algorithm for next-generation PCS networks. *IEEE Journal on Selected Areas in Communications, 19*(10), 1942-1952.

Wu, C. H., Lin, H. P., & Lan, L. S. (2002). A new analytic framework for dynamic mobility management of PCS networks. *IEEE Transactions on Mobile Computing, 1*(3), 208-220.

Xie, J., & Akyildiz, I. F. (2002). A novel distributed dynamic location management scheme for minimizing signaling costs in mobile IP. *IEEE Transactions on Mobile Computing, 1*(3), 163-175.

Zahariadis, T. B., Vaxevanakis, K. G., Tsantilas, C. P., Zervos, N. A., & Nikolaou, N. A. (2002). *Global roaming in next-generation networks. IEEE Communications Magazine, 40(2), 145-151.*

This work was previously published in Wireless Information Highways, edited by D. Katsaros, A. Nanopoulos, and Y. Manalopoulos, pp. 213-250, copyright 2005 by IRM Press (an imprint of IGI Global).

Chapter 2.25
Location Area Design Algorithms for Minimizing Signalling Costs in Mobile Networks

Vilmos Simon
Budapest University of Technology and Economics, Hungary

Sándor Imre
Budapest University of Technology and Economics, Hungary

ABSTRACT

In the next generation, IP-based mobile networks, one of the most important QoS parameters, are the delay and the delay variation. The cell handover causes incremental signalling traffic, which can be critical from the point of view of delay variation. It worsens the quality parameters of the real-time services, which are the backbone of next generation mobile commercial services. We have designed and implemented two algorithms: a location area forming algorithm (LAFA) and a cell regrouping algorithm (CEREAL), which can help us to guarantee QoS parameters in the next generation mobile networks. We used our realistic mobile environment simulator to generate input statistics on cell changes and incoming calls for our algorithms and by comparing the values of the cost functions proposed by us, we recognized that significant reduction was achieved in the amount of the signalling traffic; the location update cost was decreased by 40-60% in average.

INTRODUCTION

Signalling delay and the delay variation are very important service quality parameters of the next generation, IP based mobile networks. The cell handovers in mobile networks causes an incremental signalling message overhead (Akyildiz, Mcnair, Ho, Uzunalioglu, & Wang, 1998), which affects the delay variation and it is critical in the case of timing-sensitive real-time media applica-

tions. The signalling overhead is caused because the location information of a mobile is maintained by registration (Wong & Leung, 2000), where the mobile terminals update their location area information to their home agents. The determination of the location of the user is also important because the demand of mobile location dependent information services (LDIS) has fuelled in recent years. Jayaputera and Taniar (2005a) proposed a new approach to generate a query result for location-dependent information services. Another scope is when the users location moves from one base station to another and the queries cross multi-cells. Jayaputera and Taniar (2005b) gave an approach of mobile query processing in these situations.

The determination of the optimal number of cells in each location area (LA) is a very important task, but the optimal partition of cells into LAs is an NP-hard problem. There was an important contribution in the determination of the optimal number of cells in an LA (Saraydar, Kelly, & Rose, 2000), but they were not focusing on the selection of the optimal set of cells for each LA. Therefore, we propose a solution to obtain the optimal partition of cells for every LA.

The location area structure means that we can join several cells into one administrative unit—so-called location—and in this way, the cell border crossings inside this domain will be hidden for the upper hierarchical levels. Signalling overhead will be produced only when we cross a domain border, but that is rarer than a cell handover, thus, the traffic of signalling messages will be reduced (Cayirci & Akyildiz, 2003).

The question arises: What size the LA should be? Both increasing and decreasing the size have their own benefit. On the one hand, if we join more and more cells into one LA, then the number of LA handovers will be smaller, so the number of location update messages sent to the upper levels will decrease. However in the case of numerous cells belonging to a single LA, an incoming call will cause lots of paging messages (Zhang, Castellanos, & Campbell, 2002) since we must send one to every cell to find where is the mobile user inside that LA. That will increase the load of base stations.

On the other hand, if we decrease the number of cells, then we do not need to send so many paging messages (hereby we will load less links and the processing time will decrease, too), but then the number of LA changes will increase.

Accordingly, we must search for the optimal compromise between these two conflicting aspects (Li, Kameda, & Li, 2000).

The LA management is classified according to its use of time, distance, movement profile information in its paging, and location update procedures. The location update can be performed due to the time elapsed since the last registration process (Jun & Ho) or the number of cell boundary crossings measured since the previous update (Tsai & Hsiao, 2001). Wong and Leung (2001) recommend a distance-based scheme where the location update will be performed when a mobile user moves a threshold number of cells away from the cell where the last registration process was carried out. The hybrid of distance-based and zone-based is studied by Casares-Giner and Mataix-Oltra (2002). Bar-Noy, Kessler, and Sidi (1995) have compared time-, distance-, and movement-based schemes in terms of location management cost and they have shown that the distance-based one performs best. However, its implementation is hard since the distance of the mobile terminal has to be computed dynamically as it moves from cell to cell.

In this article, we propose a zone-based LA solution since they are used in all the deployed cellular mobile systems.

Our aim was to decrease the amount of administrative messages, so we designed an LA forming algorithm (LAFA) based on the statistical probabilities of the moving directions (Simon, Huszák, Szabó, & Imre, 2003) chosen by the mobile users. We have implemented this graph algorithm using the cell border crossing prob-

abilities as input. We propose a mobility simulator developed by us for the generation of a realistic border crossing and incoming call pattern as an input for our algorithm. Furthermore, we propose a cell regrouping algorithm (CEREAL), too, for a refinement optimization using an error function defined for the LA partitions, which was produced by the LA forming algorithm. The implemented program calculates the cost functions (see Section 5) of the random and of our LA forming algorithm, in the function of the paging and location update importance weights.

This article is organized as follows: the mathematical description of our paging and location update cost function is introduced in the second section. The LAFA is presented in the third section. In the fourth section, we give the CEREAL for the optimal refinement, while in the fifth section our results are shown and discussed. In the sixth section, we draw conclusions.

Cost Structure

Most of the references (Madhow, Honig, & Steiglitz 1995; Xie, Tabbane, & Goodman, 1993) related to the location area design are focused on how to determine the optimal number of cells for an LA. In this article, we presented an algorithm, which can give us the optimal partition of cells into LAs.

Although there had been earlier attempts to optimize location update cost or paging cost or a combination of them in Abutaleb and Li (1997), Akyildiz, Ho, and Lin (1996), and Merchant and Sengupta (1995), however, the aspect of minimizing the location update cost with a heuristic algorithm (LA forming algorithm) first, and after using that basic partition as an input to a regrouping algorithm, which will minimize the aggregated cost function, was not considered there. So, we have split this complex problem into two sub problems to optimize the location update cost and the aggregated cost, one after another. We defined the location and paging cost functions differently than in the previously mentioned related works because of the handling of the two sub problems.

The Paging Cost Function

On the arrival of an incoming call, the mobile switching center sends a paging message to every base station under its control in order to find out the called mobile terminal (MT) (Bhattacharje, Saha, & Mukherjee, 2004). So, each cell in the given LA will carry all the paging traffic associated with the called MTs within that LA. In order to characterize a network configuration, we define a paging cost function for the l^{th} LA by which we can describe the bandwidth seized by the paging operations in a given interval:

$$C_{p_l} = \sum_{i=1}^{K} N \cdot \lambda_i \cdot B_p \qquad (1)$$

where

- N is the number of cells in the given l^{th} LA.
- λi is the incoming call rate to the given i^{th} MT.
- B_p is the paging cost.
- K is the number of MTs in the l^{th} LA.

With Eq. (1), we can determine cost of the traffic induced by paging messages for a given LA, generated by the incoming calls in the given interval. The total paging cost for the LAs in our system:

$$C_p = \sum_{l=1}^{M} C_{p_{ll}} = \sum_{l=1}^{M} \sum_{i=1}^{K} N \cdot \lambda_i \cdot B_p = \sum_{l=1}^{M} N \cdot B_p \cdot \sum_{i=1}^{K} \lambda_i \qquad (2)$$

where M is the number of LAs in our system.

Location Update Cost Function

We define a location update cost function for our network, which will help us to determine the LA

handovers caused by the mobile users crossing the LA boundary (Chiussi, Khotimsky, & Krishnan, 2002), which generates additional location update traffic; namely they need to inform their home agents about their new location (Akyildiz et al., 1999).

The location update cost for the k^{th} LA:

$$C_{lu_k} = B_{lu} \cdot \sum_{j=1}^{B} q_j, \quad (3)$$

where

- B_{lu} is the cost required for transmitting a location update message.
- q_j is the intensity of cell boundary crossings on the j^{th} boundary.
- B is the number of the exterior cell borderlines.

The total location update cost for the LAs in our system:

$$C_{lu} = \sum_{l=1}^{M} C_{lu_k} = \sum_{l=1}^{M} B_{lu} \cdot \sum_{j=1}^{B} q_j, \quad (4)$$

where M is the number of LAs in our system.

Our final goal is to maximize the intra-domain traffic because in this way we can decrease the number of the LA handovers, and therefore, the total amount of administrative messages. We can reduce the number of handovers by joining the cells along the dominant moving directions. This LA forming principle will be introduced in the third section.

To evaluate the algorithms we need. the minimized expectation value of the aggregated cost function, with variable weight factors, which takes into consideration both aspects of forming LAs:

$$\min E\{w_1 \bullet Cp + w_2 \bullet C_{lu}\}. \quad (5)$$

On the basis of the importance of paging or rather location update cost, we can use different weights, and in that way we can dynamically vary the sizing based on the actual point of view and the required QoS parameters.

Because of the expectation value is a homogeneous linear operator, the Eq. (5) expression becomes:

$$\min(w1 \bullet E\{Cp\} + w_2 \bullet E\{C_{lu}\}). \quad (6)$$

The Location Area Forming Algorithm (LAFA)

We model our network with the $G(V,E)$ graph where the cells are the graph nodes $v \in V$ and the cell border crossing directions are represented by the edges $e \in E$ of the graph (see Figure 1).

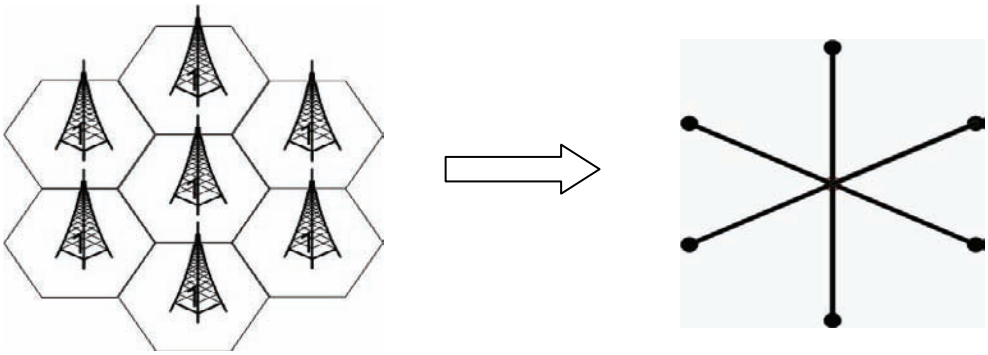
Definitions

- If $\{v_1,v_2\} \in E$ then the cells represented by v_1 and v_2 are adjacent.
- If $\{v_1,v_2\}$ and $\{m_1,m_2\}$ are the end points of $e,f \in E$, and $\{v_1,v_2\} \cap \{m_1,m_2\} \neq 0$, then the cell border crossing directions e,f are adjacent.
- If the set of nodes of graph F is consistent with the set of nodes of graph G and the edges of graph F are an acyclic subset of edges of graph G, the graph F is the spanning tree of the graph G.
- If the weight function $c:E \rightarrow \Re$ is defined on the edges and the sum of its edge weights is maximal among the spanning trees of graph G, the graph F is the maximum weight spanning tree of the graph G.

The Algorithm

A moving direction matrix can be defined to every cell, which contains the statistical probabilities of the moving directions chosen by the mobile us-

Figure 1. The representation of the mobile system by a graph



ers when they step across the cell borders. This mobility pattern database can be obtained by measurements, which is attainable for the mobile operators. In our case, we have developed a mobility simulator (see the fifth section), which serves as a realistic cell boundary crossing and incoming call pattern as an input to our algorithms.

We define weights to the edges of the graph G, not negative real numbers in the range [0,1], based on the probability matrix, namely the weight of the edges is consistent with the cell border crossing probabilities.

We must divide graph $G(V,E)$ into subgraphs $G_i(V,E)$, so that the subgraphs contain the maximum weight spanning tree. The set of edges of those maximum weight spanning trees will give us the cell groups, which compose the LAs.

Starting from the s=1 initial point, we choose from edges joint to the node s, the one, which has the biggest weight (c_{mac}), if there is more than one biggest weight, then we choose one of them randomly and include it into the set of edges $L_1(L_1=\{e_1\})$ (Figure 2). The two nodes connected by this edge are included into set $U_1=\{s_1,v_1\}$. In the next step, we search for the second largest weight (if there is more than one, we choose it in the same way as in the first step), and we examine those two nodes belong to the U_1set. If both are in the set V/U_1, then the edge, which connects them, is included into the L_2 set of edges ($L_{21}=\{e_1\}$) and the two belonging nodes into the setU_2. If one of them is in the set U_1, then we must make an evaluation step.

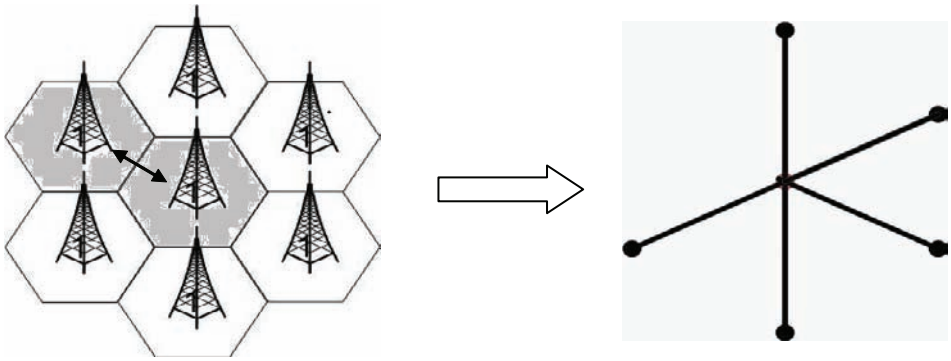
We must check if inequality

$$\frac{c_m}{\frac{1}{n}\cdot\sum_{i=1}^{n} c_i} > K \quad , \tag{7}$$

is satisfied, where c_m is the weight of the examined edge, and c_i are the weights of edges in the U_1 set. K is the lower bound, given by us. If the inequality is satisfied, the edge can be included into this set ($L_1=\{e_1, e_2\}$). If the inequality is not satisfied, this edge can not be included into this set, namely the cell which is represented with this node, can not be joint into this LA. Another upper bound can be used, we can give the maximum number of cells in one LA.

In this way, we can join the cells, which are in the same dominant moving directions, so the number of domain handovers can be decreased (highways, footpaths, etc.).

Figure 2. Joining the two nodes (cells) into one set (LA)



We run this algorithm until we get the U_i partition of nodes V, so this partition will give us the groupings of cells into clusters, namely the D_i location areas.

THE CELL REGROUPING ALGORITHM (CEREAL)

The partition of cells is an NP-hard problem; therefore, we propose a heuristic algorithm. We have developed a regrouping algorithm, which would help us to refine the cell grouping in cases when it is necessary.

Definitions

- For cell i, $Y_i=(Y_1, Y_2)$ is a two dimensional probability variable, where $Y_{i1}=q_{iF}$, q_{iF} is the border crossing intensity of the cell i and $Y_{i2}=\sum_{j=1}^{N}\lambda_j$, where λ_j is the expectation value of the incoming call distribution, and N is the number of the mobile terminals in the cell i.
- The LAFA gives us a partition of cells $\{D_1, D_2,... D_M\}$ in location areas, where M is the number of areas in our mobile system. Then $Y=\bigcup_{i=1}^{M} Di$ and $D_i \cap D_j = 0$, if $i \neq j$.

The Algorithm

If $|D_i|$ is the number of cells in the given D_i location area, then we can define the center of the area by

$$\overline{D_i} = \frac{1}{|D_i|}\sum_{Y_l \in D_i} Y_l \qquad (8)$$

We can define the distance of the Y_j cell from the D_i location area as

$$d\left(Y_j, D_i\right) = d\left(Y_j, \overline{D_i}\right) = \left(\sum_{l=1}^{2}\left(Y_{jl} - \overline{D_{il}}\right)^2\right)^{1/2}. \qquad (9)$$

A very important parameter in our regrouping algorithm will be the error function of our location area system

$$W\left(D_1,...D_M\right) = \sum_{i=1}^{M}\sum_{Y_j \in D_i} d^2\left(Y_j, D_i\right). \qquad (10)$$

Our goal is to minimize the error function by transposing the cells into adjacent location areas, and by this, we can reduce the distances among them, what will result in a significant reduction of location update and paging costs.

Steps:

1. The calculation of the initial area centers and the initial error function ($\overline{D_i}$,$W(D)$).
2. For the first cell (Y_1) and the adjacent location areas (D_i), given by our location area forming algorithm, we calculate:

$$\Delta(D_i,Y_1)=\frac{|D_i|\cdot d^2(Y_1,D_i)}{|D_i|+1}-\frac{|D(Y_1)|\cdot d^2(Y_1,D(Y_1))}{|D(Y_1)|-1} \quad (11)$$

where $D(Y_i)$ is the location area, which contains the Y_i cell.
We can prove that if we re-group the Y_i cell from the $D(Y_1)$ location area to the D_i area, the error function of our location area system will change by exactly $\Delta(D_i,Y_1)$.
So if

$$\min_{\substack{1<i<M',\\ D_i\neq D(Y_1)}} \Delta(D_i,Y_1)=\Delta(D_k,Y_1)<0, \quad (12)$$

where M^1 is the number of the location areas, which are adjacent with cell Y_1, then we transpose the cell Y_1 from location area $D(Y_1)$ to D_k.
Calculate the new location area centers and add the $\Delta(D_k,Y_1)$ to the former error function value.
3. Iterate the 2nd step for every Y_i.
4. If there are no more cells to transpose, we can stop, otherwise repeat step 2.

This regrouping algorithm will give us the final LA partition, which will minimize the inter LA movement, and by this the signalling load, too.

QUANTITATIVE ANALYSIS

Optimal partition of cells into LA-s is an NP-hard problem, so we could give only a quantitative analysis of the problem and evaluate our algorithm with a mobility simulator in two different mobility environments.

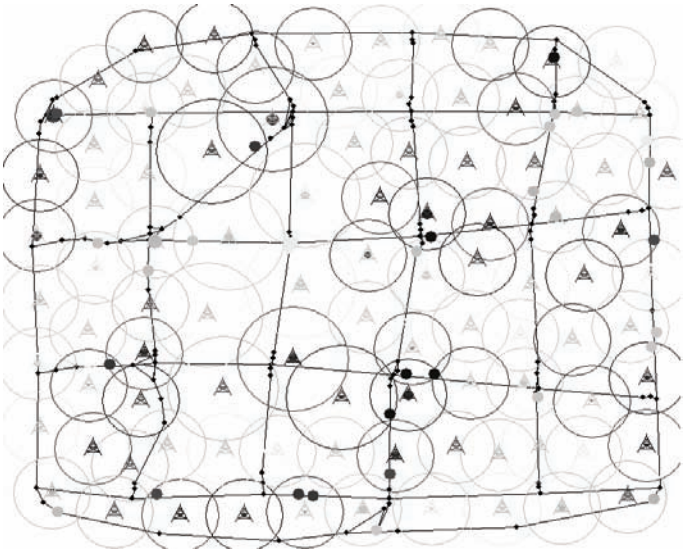
In this section, we will present our cost evaluation results using the mobility simulator as an input generator for our algorithms. The simulator produces cell changing and incoming call statistics, which are used to run our algorithms and to calculate the cost functions.

The Mobility Simulator

We developed a simulator (see screenshot in Figure 3), which will give us a realistic cell boundary crossing and incoming call database in a given mobile system, as an input to our algorithms. In the simulator, we can give an arbitrary road grid covered by cells of different size (for example WLAN, UMTS, GSM cell). We can choose between mobile terminals of different velocities, and we can give the incoming call arrival parameter to every mobile.

This way we can design different types of mobility environments (rural environment with highways or a densely populated urban environment with roads and carriageways), and grids of cells adapted to these environments. The mobile terminals will move on that road grid choosing randomly a point on the road, just like in a real life. Because in everyday action, the mobile users typically move to manage their duty tasks or entertainment (for example workplace, school, cinema, bank), and they want to arrive there in the shortest time, so we implemented the Dijkstra's algorithm to find the shortest path of mobile terminals to their wanted destination. For every mobile terminal, an incoming call arrival parameter is defined. When a call arrives to the mobile, the program assigns it

Figure 3. A designed urban environment in our mobility simulator, with a road grid covered by cells



to the cell where the mobile is in that moment. It is the same case when a mobile terminal changes a cell, the simulator register the cell's identifier in the base station transition matrix. On the end of the simulation, we get a cell boundary crossing and an incoming call distribution for every cell in our system.

With this database, which is a good representation of the mobility patterns in real life, we can run our LA forming algorithm, and evaluate it by the defined location update cost function (4).

Simulation Results

We compared the performance of a random LA partition and the LA forming algorithm by using two typical mobility environments.

We designed a rural and an urban mobility environment in our mobility simulator, the first one is rarely populated, but on the belonging highways, a big number of mobile terminals are moving with high speeds, while the second environment is densely populated, with mobile terminals moving with smaller velocities. In the rural environment, the average cell size is larger then in the urban, accordingly there is a smaller number of cells. We run the simulation on a moderate sized example network; the rural mobile system consisted of 99 base stations, while in the urban system it was about 123 base stations. Then we stored the output of the simulation, namely a cell boundary crossing matrix (base station transition matrix) and the incoming call distribution for every cell. This database was the input of our LA forming program, which designed a LA partition for both mobility environments.

1. **Employing only the LAFA:** Based on these LA partitions, we computed the location update cost (in total number of handovers between the LAs in the simulation period), and we computed the same cost for a ran-

domly designed LA partition where we can give the number of cells in one LA, and then the program designs an LA partition. It is important to point out that the randomly designed LA do not mean that the cells are joined by a random way; the cells joined to the same LA are on the dominant moving directions, so the random LA-s can be considered as a planed partition, but not the optimal one. So the results are compared to these planed structures, not to haphazard partitions.

For the rural environment, the results of the simulation are given in Figure 4 where the upper bound of LA axis represent the number of cells given in the random design or rather the additional upper bound of cells which was given in the LAFA, when Eq. (7) is satisfied. In Figure 4, the location update cost decreases as the upper bound becomes higher, but then the paging cost will increase, too. So in the random partition we can decrease the location update cost if we increase the size of the LAs, but then the paging cost will be a serious problem. A significant advantage of the LAFA is that it reduces the location update cost very significantly, by not increasing the number of cells in one LA, so the paging cost can be kept on a lower level. In the domain of 5-10 cells in an LA, our new scheme reduces the inter LA traffic by 40-60 percents on the average.

In Figure 5, the results for the urban environment can be seen. It is very similar to the rural; however, the location update cost is remarkably reduced without increasing the paging signalling load. The decrease is very significant in the interval of five to eleven cells. In the interval between 12 and 15, does mean that by increasing the number of cells the location update cost does not decrease, but the Eq.(7) is not satisfied so the algorithmic LAs are already finished, so the cost does not change.

Another simulation was done to investigate the effect of the value of parameter K (7), which can be dynamically modified, depending on the model we want to deploy. It changes in the (0,1) interval, as it converges to 1, the size of the LA-s are decreasing, so we expected that the location update cost will increase significantly as the parameter K increases. Figure 6 shows the results obtained in the rural environment as the value of parameter K is increasing, the location update cost of the random partition is increasing too, the cost of the LAFA is not increasing till K=0.5 value, but

Figure 4. The location update cost in rural environment, in random and algorithmically partition

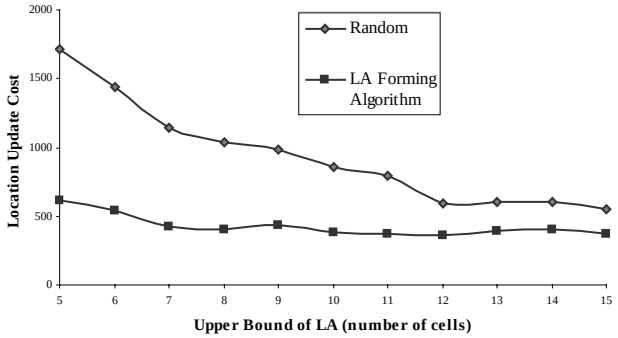


Figure 5. The location update cost in urban environment, in random and algorithmically partition

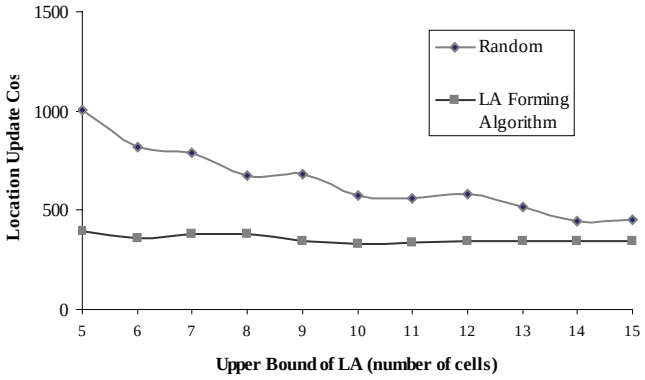
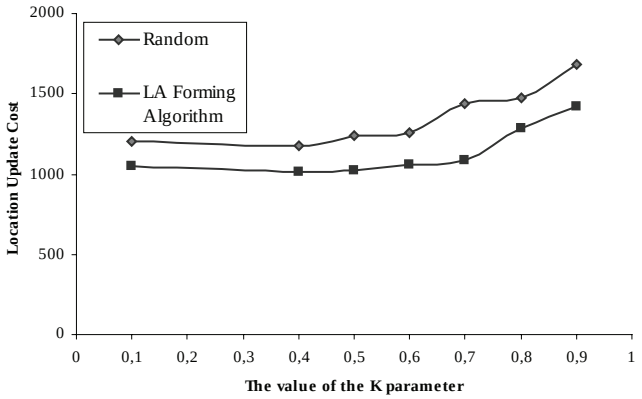


Figure 6. The location update cost vs. the value of parameter K (rural)



after it follows the cost of the random partition. So the algorithm outperforms the random partition in every value of parameter K, especially for the lower values that are characteristic for our algorithm. The minimum of the location update cost is reached for the value 0.4 of parameter K.

In the urban environment the results are more effective, the cost of the LAFA is lower by 30-40% (see Figure 7), depending on the value of parameter K. The minimum is reached in at value of 0.5 of parameter K.

Figure 7. The location update cost vs. the value of parameter K (urban)

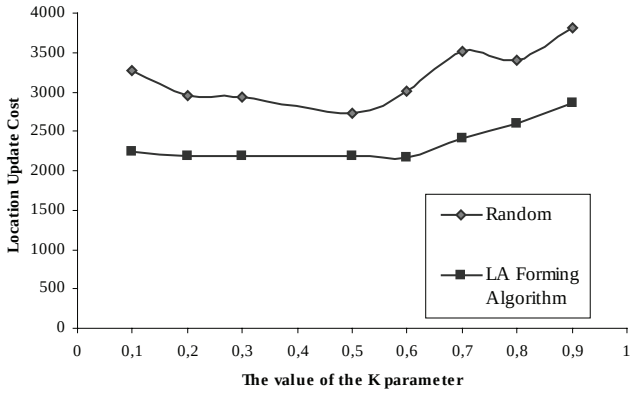
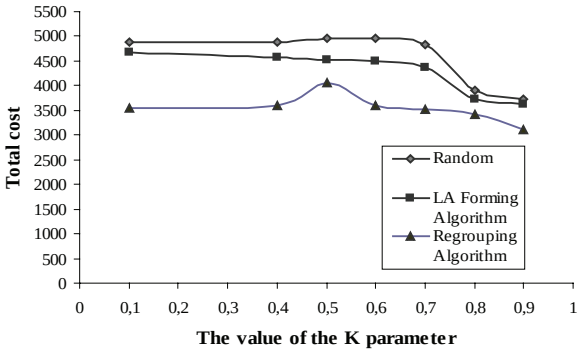


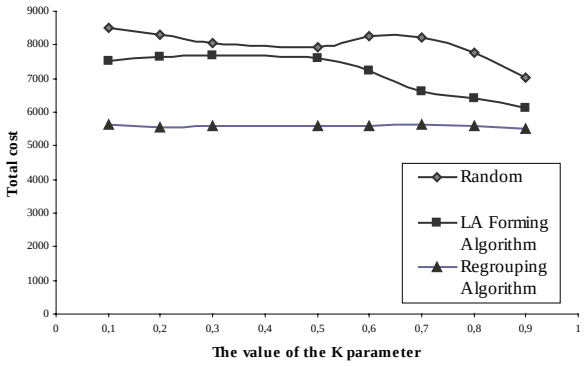
Figure 8. The total cost vs. the value of parameter K (rural)



2. **Employing both algorithms:** We examined what will happen if we employ our regrouping algorithm on the initial partitioning obtained by the LAFA. We measured the total cost, the sum of the location update, and paging cost vs. the value of parameter K, which was employed in the LAFA.

Figure 8 shows the total cost in the rural environment. It can be seen that in the case of the total cost the effect of increasing the value of parameter K is just the opposite of that in case of the location update cost. The reason is that by increasing the value of parameter K, the size of the LA-s is getting smaller, so the paging cost is

Figure 9. The total cost vs. the value of parameter K (urban)



decreasing significantly. So the LAFA is not so effective anymore, but the CEREAL is still outperforming the other two partitioning methods significantly, in some cases over 50%.

Figure 9 shows the total cost vs. K parameter in the urban environment. It is the same case, the CEREAL decreases the amount of the total cost by 30-40%, and it does not depend on the value of parameter K.

Depending on our objective, we can deploy either the LA forming algorithm or the regrouping algorithm. If we want to decrease the location update cost, the LA forming algorithm is the solution, however if we want to decrease the total cost, we need to employ the regrouping algorithm.

CONCLUSION

An important benefit of optimized LA planning is preventing needless radio resource usage (Demirkol, Ersoy, Caglayan, & Delic 2004), but the most important is that we can support global QoS parameters, like signalling delay and delay variation, which can be critical in time sensitive services of the next generation mobile systems.

It can be achieved by reducing the signalling cost, which means that the inter LA movement must be minimized. The input of this algorithm was obtained by a mobility simulator developed by us that produces network information (base station transition matrix, incoming call distribution to every cell) in a realistic manner. We also proposed a cell regrouping algorithm, which uses the LA partitions obtained by our algorithm, like an initial step.

To evaluate the performance of our new scheme, we designed a rural and an urban environment in the mobility simulator, and with this database, we run the LA forming algorithm. Then we compared it with a randomly designed LA structure, examining the relation of the location update cost with an upper bound on the number of cells.

The simulation results show that the LA forming technique reduces the location update cost by 40-60%. The regrouping algorithm performs well if we want to decrease the total cost of our

system, it can reduce the total cost by 30-40%, sometimes over 50%. We recognized, by comparing the random algorithm and our LA forming one, that a significant reduction was attained in the signalling traffic that causes delay and delay variation, helping us improving QoS parameters in general.

Due to the difficulty of the problem, our future research plan is to use simulated annealing to obtain the optimal LA partitions, in reasonable running times. In this case, we will examine the minimization of the location update cost, subject to the paging cost, as an inequality constraint (using constrained optimization). Another research direction is to develop an optimization algorithm based on the inter-LA movements, which will help us to plan a hierarchical mobile structure, which results in the minimal signalling traffic.

REFERENCES

Abutaleb A., & Li , V. O. K. (1997). Paging strategy optimization in personal communication systems. *Wireless Networks, 3,* 195-204, Amsterdam, The Netherlands: Baltzer.

Akyildiz, F., Mcnair, J., Ho, J., Uzunalioglu, H., & Wang, W. (1998). Mobility management in current and future communications networks. *IEEE Network Magazine,* July/August.

Akyildiz, I. F., Ho, J. S. M., & Lin, Y. B. (1996). Movement-based location update and selective paging for PCS network. *IEEE/ACM Transaction Networking, 4*(4), 629-638.

Akyildiz, I. F., Mcnair, J., Ho, J. S. M., Uzunalioglu, H., & Wang, W. (1999). Mobility management in next generation wireless systems. *Proceedings of the IEEE, 87*(8), 1347-1385.

Bar-Noy, A., Kessler, I., & Sidi, M. (1995). Mobile users: To update or not to update? *Wireless Networks, 1*(2), 175-185.

Bhattacharje, P. S., Saha, D., & Mukherjee, A. (2004). An Approach for location area planning in a personal communication services network (PCSN). *IEEE Transactions on Wireless Communications, 3*(4), 1176-1187.

Casares-Giner, V., & Mataix-Oltra, J. (2002). Global versus distance-based local mobility tracking strategies: A unified approach. *IEEE Trans. Veh.Technol., 51,* 472-485.

Cayirci, E., & Akyildiz, I. F. (2003). Optimal location area design to minimize registration signalling traffic in wireless systems. *IEEE Transactions on Mobile Computing, 2*(1), January-March.

Chiussi, F. M., Khotimsky, D. A., & Krishnan, S. (2002). Mobility management in third-generation all-IP networks. *IEEE Communications Magazine, 40*(9), 124-135.

Demirkol, I., Ersoy, C., & Caglayan, M. U., & Delic, H. (2004). Location area planning and cell-to-switch assignment in cellular networks. *IEEE Transactions on Wireless Communications, 3*(3), 880-890.

Jayaputera, J., & Taniar, D. (2005a). Query processing strategies for location-dependent information services. *International Journal of Business Data Communications and Networking, 1*(2), 17-40.

Jayaputera, J., & Taniar, D. (2005b). Data retrieval for location-dependent query in a multi-cell wireless environment. *Mobile Information Systems: An International Journal, 1*(2), 91-108, IOS Press.

Jun, L.D., & Ho, C. D. On optimum timer value of area and timer-based location registration scheme. *IEEE Commun.Letters, 5,* 1106-1110.

Li, J., Kameda, H., & Li, K. (2000). Optimal dynamic mobility management for PCS networks. *IEEE/ACM Trans.Networking, 8*(3), 319-327.

Madhow, U., Honig, M. L., & Steiglitz, K. (1995). Optimization of wireless resources for personal

communications mobility tracking. *IEEE/ACM Transaction Networking, 3*(6), 698-707.

Merchant, A., & Sengupta, B. (1995). Assignment of cells to switches in PCS networks. *IEEE/ACM Transaction Networking, 3*(5), 521-526.

Saraydar, C. U., Kelly, O. E., & Rose, C. (2000). One-dimensional location area design. *IEEE/ACM Transaction Networking, 49*(5), 1626-1632.

Simon, V., Huszák, Á., Szabó, S., & Imre, S. (2003). Hierarchical mobile IPv6 and regional registration optimization. *International Conference on Parallel and Distributed Computing* (Vol. 2790, pp. 1137-1140). Euro-Par 2003, August 26th-29th, Klagenfurt, Austria, Springer, Lectures Notes in Computer Sciences.

Tsai, J. T., & Hsiao, H. H. (2001). Performance of movement-based location update and one-step paging in wireless networks with sparsely underlaid microcells. In *Proceedings of IEEE GLOBECOM* (pp. 642-647), San Antonio, TX.

Wong, V., & Leung, V. (2001). An adaptive distance-based location update algorithm for next generation PCS networks. *IEEE J. Select. Areas Commun., 19*, 1942-1952.

Wong, V. W. S., & Leung, V. C. M. (2000). Location management for next-generation personal communications networks. *IEEE Network Magazine*, 18-24, September-October.

Xie, H., Tabbane, S., & Goodman, D. (1993). Dynamic location area management and performance analysis. In *Proceedings of the 43rd IEEE Vehicular Technology Conference* (pp. 533-539), May.

Zhang, X., Castellanos, J., & Campbell, A. (2002). Design and performance of mobile IP paging. *ACM Mobile Networks and Applications. Special Issue on Modeling Analysis and Simulation of Wireless and Mobile Systems, 7*(2), March.

This work was previously published in the International Journal of Business Data Communications and Networking, edited by J. Gutierrez, Volume 3, Issue 2, pp. 36-50, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.26
Market Configuration and the Success of Mobile Services:
Lessons from Japan and Finland

Jarkko Vesa
Helsinki School of Economics, Finland

ABSTRACT

This chapter introduces a novel analytical framework called Mobile Services Matrix (MOSIM), which is used as the basis of a comparative analysis between the Japanese and the Finnish mobile services markets. The results indicate that as the mobile industry shifts from highly standardized voice services towards more complex mobile data services, the vertical/integrated market configuration (i.e., the Japanese model) appears to be more successful than the horizontal/modular configuration (i.e., the Finnish model). A brief overview of the UK market shows that the leading UK mobile network operators are transforming the industry towards a more vertical, operator-driven market configuration. The role of national regulatory framework in this industry evolution process is discussed.

INTRODUCTION

There is a paradox in the mobile services industry in Europe today: Even though the industry will be experiencing a major transformation during the next few years, for the time being the business is going too well for the senior management of many telecom operators to take action to redefine their business models and strategies in order to be well positioned in the new era of mobile multimedia services (aka mobile Internet, mobile data services, or nonvoice services).

Take Finland, for instance. Once the Internet bubble burst and business risks that were built into the growth strategies of the leading telecom operators were realized, the growth-oriented senior management was replaced with new management teams with a strong focus in cost cutting and downsizing. Financial markets expected operators to clean up their balance sheets after notorious UMTS licence auctions and other

unsuccessful attempts to become serious players in the international mobile market. As a result of this development, Finnish operators have delayed their investments in new services and networks (for instance, the first UMTS service for commercial use was launched by TeliaSonera in October 2004, even though Finland was one of the first countries in Europe to allocate spectrum for 3G!) and sold those parts of their operations that are not considered to be their core business. Against this background it is easy to understand why the leading Finnish network operators opt for the current situation where they keep on making nice profits in the saturated market instead of actively seeking to change the competitive arena. Although this approach is understandable from individual companies' point of view, recently there has been much discussion as to whether this kind of risk-avoiding strategy will turn Finland into a yesterday's hero when it comes to actively building the brave new world of mobile services. Operators can rest on their laurels for a few more years, but the longer they neglect developing their business models for the competition of the future, the more painful process lies ahead of them. Some people within the industry have realized the destructive nature of the current price-driven competition in the Finnish market: a representative of TeliaSonera expressed his concern that many of the players in the mobile market are there to cannibalize the market with their aggressive pricing schemes, not to develop the market in order to secure healthy business also in the coming years (Tietoviikko, 2004). The industry seems to ignore the fact that the worst is yet to come, as new disruptive technologies such as voice-over-WLAN, WiMax, and free Internet telephone services such as the Skype service become increasingly popular in the coming years, especially now as eBay acquired Skype in order to enhance their online auction platform.

Unfortunately mobile operators are not the only ones sticking to old voice-centric business paradigm and earnings logic. Even the national regulatory authorities fail to see the need to adjust the regulation of mobile markets to the changes in technology and in the business environment in general. While the mobile phone usage has slightly increased (i.e., the minutes of use), in mature markets like Finland, the decrease in call tariffs has led to a situation where average revenue per user has remained flat or even decreased. Although 2004 was regarded as highly exceptional due to the introduction of mobile number portability in Finland, there is a widely shared view that call tariffs will continue to fall at a rate of 20–30%.

The current development in the traditional mobile voice market has made the leading operators to turn their eyes on nonvoice services. However, so far the European operators have not managed to turn mobile multimedia services into a similar success story as their Japanese counterparts. This raises the question why, despite the similarity in services and content offered, have nonvoice mobile services not taken off as expected? Based on a comparison of two very different mobile markets, namely Japan and Finland, this chapter argues that the lack of success of mobile Internet services in Europe is more a result of wrong business models and industry structure than it is about quality of individual services or products. As mobile services evolve from highly standardized and commodized voice-based communication services (i.e., person-to-person communication) towards the personalized and complex world of digital content and services (i.e., mobile multimedia or person-to-content type services), new challenges emerge also for the creation and delivery of mobile services. I argue that mobile data services represent a "complex good," which Mitchell and Singh (1996) define as "an applied system with components that have multiple interactions and constitute a nondecomsable whole." In a complex system like this, the overall performance depends on component performance, as a chain is only as strong as the weakest link. In the closely integrated and interrelated world of mobile Internet services, not only must all components meet users'

requirements, but all the elements of the service offering (i.e., networks, handsets, services, and content) must also work seamlessly together. As the analysis of the two mobile markets presented in this chapter will demonstrate, this is where Europe has performed much worse than during the previous paradigm shift in the mobile telephony industry, which was the transition to digital mobile networks in the beginning of the 1990s.

During the past 5 years, both Japanese and European mobile markets have been studied extensively. However, the focus of these studies is often technical (e.g., how various technologies used in handsets and networks have evolved during different generations of mobile telephony, or how technical standards have been adopted in various markets), or they have focused on macroeconomic issues (e.g., mobile phone penetration rates or how well competition works in different markets). In this chapter I will present a different approach that raises the level of analysis above the discussion of superiority of competing technical standards, which all too often overshadows more important issues. In fact, I argue that the European way of introducing new technologies to the market suffers from two mind-sets that are not optimal in the context of mobile services. The first one could be described as business reductionism. As Albert-László Barabási points out in his book *Linked*, there appears to be a widely accepted belief that "once we understand the parts, it will be easy to grasp the whole" (Barabási, 2003). By following the ideals of reductionism, some researchers have tried to understand the differences in the success of mobile markets by comparing various technologies used: What kind of markup language has been used? Are the mobile networks circuit or packet switched? How many pixels do cameraphones support? Unfortunately it looks like the dynamics of complex networked industries, such as the mobile services industry, cannot be explained simply by analyzing all the pieces of the puzzle separately. In all fairness I have to admit that this was the way I was thinking when I sat on a plane on my way to Tokyo to meet some of the key players of the Japanese mobile industry in October 2002. However, it did not take long to realize that there was something bigger than just technologies, protocols, and standards behind the success of mobile services in Japan, as we will find out later in this chapter.

The other ideal causing problems for the Finnish mobile market is that open standards seem to be the only goal worth pursuing, regardless the maturity of products and services offered. A good example of the "open-market thinking" that prevails in Europe was a conversation between a representative of J-Phone and some of my colleagues at the Helsinki School of Economics. When our Japanese guest presented the J-Phone business model where the operator controls all the key components of mobile services, several people in the audience raised the question whether this kind of model is acceptable. Their argument was that as the customers of J-Phone are not free to use other operators' networks, services, or content, this kind of model is not as elegant as the European open-market approach. Our Japanese guest asked why we Europeans always emphasize so much this question of openness and the freedom to move between every possible service provider's offering? Is it not possible that a consumer or business customer in some cases would prefer to deal with only one service provider, if that service provider offered high-quality services and content delivered in a seamless, easy-to-use way at a reasonable price? While being aware of the strong argumentation in favor of open markets and standards in economic theories such as network externalities, at least I could not help finding the Japanese concept attractive. Perhaps one reason for this reaction was the fact that I was still suffering from mental trauma caused by unsuccessful use of wireless application protocol (WAP) services. Without going into details in what went wrong in the launch of WAP in Europe, one could argue that the key players of the industry in Europe had probably too much faith in the power

of open standards, while in reality the products and services that were built on the WAP protocol were everything but standardized.

By combining the two assumptions that lie behind the widely accepted notion of what would be an optimal mobile market from the European perspective, we end up with something that could be described as "standards-based technical reductionism," which aims at free and efficient competition. It is easy to see that this kind of market would look very much like today's PC business (albeit there we are talking about de facto standards, thanks to the dominance of Intel and Microsoft). Although there are many excellent qualities in the market structure and product architecture of the PC industry, the analysis of mobile services market will demonstrate several reasons why simply copying the PC business model does not work for mobile services—at least not in the current phase of evolution. One of the key findings of this analysis of the mobile industry is that the optimal industry structure and product architecture is a function of time, that is, one should not make the mistake of presuming that a product/industry configuration that has turned out to be successful in some industries at some point of time in history would necessarily be the right alternative for another industry or another point of time, despite the fact that they may share some similar characteristics.

As the previous discussion indicates, the objective of this chapter is to offer an alternate approach into the analysis of the structure and dynamics of the mobile services markets. I will first present a new analytical framework that has been derived from the Double Helix model by Charles Fine (1998). This framework will then be used in the analysis of two very different markets—Japan and Finland. As the reasons for choosing these two markets will be explained in detail later in this chapter, it is simply stated here that Japan and Finland represent the two extremes of the continuum of contemporary mobile services business models. However, in order to better understand the dynamics of mobile market evolution, I will also present a brief analysis of the UK market, which represents an intermediate or hybrid model between the two extremes. The analysis shows that the UK market is moving towards a similar vertical and integrated market configuration as in Japan, whereas the mobile services market in Finland is stuck with a horizontal business model due to regulatory roadblocks. The results indicate that in a regulated industry such as mobile services, the operators are not always allowed to implement business models that would make sense business-wise. Furthermore, recent development in Finland shows that sometimes the key players of a given mobile services market are reluctant to drive the transformation of their industry due to strategic behavior. It is argued here that national regulatory authorities (NRAs) must have powerful vision of the kind of market structure and dynamics they are striving for.

These findings raise several questions: Is there anything the key players in a given market can do to change the market configuration towards a more favorable one? What are the implications of a given industry structure and product architecture for service providers and consumers? What can we learn by comparing various markets if they all are very different by definition (e.g., culture, history of mobile telephony, size of the market)? Hopefully this chapter will give answers to some of these questions—or even better, raise new questions about the interplay between mobile service development and delivery, and the structure and dynamics of the mobile industry within which these services are implemented.

The structure of this chapter is as follows. I will begin by presenting a novel model for the analysis of the two case markets. In section 3, I will apply the model to the Japanese mobile industry, and in section 4 the Finnish mobile market is analyzed by using the analytical model. Section 5 presents a comparison between the current status and future developments of the two case markets. Section 6 discusses the impact of market configuration on

the success of mobile services in a given market, and section 7 concludes the chapter.

THE MOBILE SERVICES INDUSTRY MATRIX

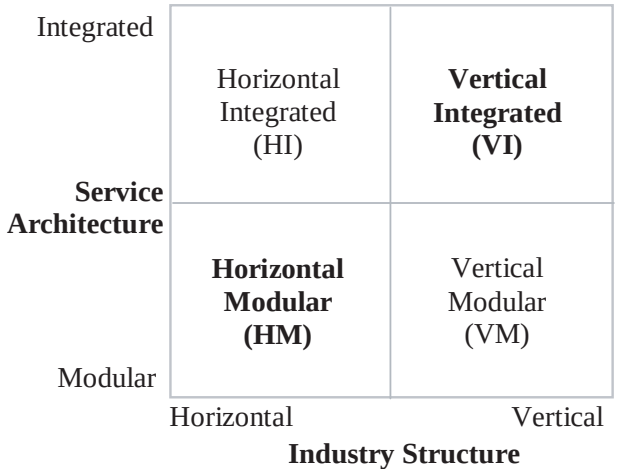
In recent years, a wide range of tools and methods have been used in the analysis of the structure and processes of the mobile and wireless industries. Various types of value chain analyses, cluster analyses, layer models, and business ecosystems approaches have become familiar to those of us who have followed academic research or market research in this field. While each of these models have certain good qualities, when applied in the context of mobile services industry they seem to be deficient in one respect—they are static by nature, that is, they provide a snapshot of an industry without giving much indication of what kind of development one can expect to see in the industry. I argue that given the magnitude of change currently taking place in the mobile industry (e.g., moving from a simple voice services to complex mobile multimedia services, and moving from closed circuit-switched networks to open packed-based IP-networks), new approaches are needed also in the research of this fast-moving industry.

One attempt to apply a new kind of approach to the research of the mobile industry was presented by Vesa (2003, 2004a) who applied the Double Helix model developed by Charles Fine (1998) in the context of the mobile services industry. Although the Double Helix model appeared to capture nicely the current trends in the mobile market, the model itself was criticized for lacking solid theoretical foundations. Some critics have even pointed out that the name "double helix" is not correct—a more appropriate name for the model would be the "Double Donut" or the "Pretzel Model"! Nevertheless, because the use of the Double Helix model in the context of mobile services has received also very much positive feedback, the matrix model presented in this chapter (see Figure 1) builds similar constructs as Fine is using in his model. However, in the Mobile Services Industry Matrix (MOSIM) the Product Architecture dimension has been replaced with Service Architecture in order to emphasize that despite of all the exciting technologies, the mobile services industry really is what its name claims it to be, a SERVICE industry. This line of reasoning is supported by recent statements within the service research discipline emphasizing the "nested relationship between service and goods" (Vargo & Lusch, 2004).

Let us take a closer look at the MOSIM presented in Figure 1. The matrix consists of two dimensions that are called Service Architecture and Industry Structure. The two extremes of the Service Architecture dimensions are defined as integrated service architecture and modular services architecture. Likewise, the Industry Structure of a given market represents either horizontal industry structure or vertical industry structure—or something in between. This last point highlights the fact that in both dimensions there can be, and very often are, various types of hybrid or intermediate industry structures or product structures (see Vesa, 2005, for an in-depth analysis of factors influencing product architecture and industry structure).

The matrix identifies four possible service architecture/industry structure configurations: (i) *Horizontal Modular* (HM), (ii) *Vertical Integrated* (VI), (iii) *Vertical Modular* (VM), and (iv) *Horizontal Integrated* (HI). The first two configurations, the Horizontal Modular and the Vertical Integrated, are the two extremes between which an industry cycles in "an infinite double loop" (Fine, 1998, p. 43). However, the MOSIM allows us to identify also two other possible combinations, that is, Horizontal Integrated and Vertical Modular configurations. In the following section, the MOSIM will be used in the analysis of the Japanese mobile services industry.

Figure 1. The Mobile Services Industry Matrix (MOSIM)



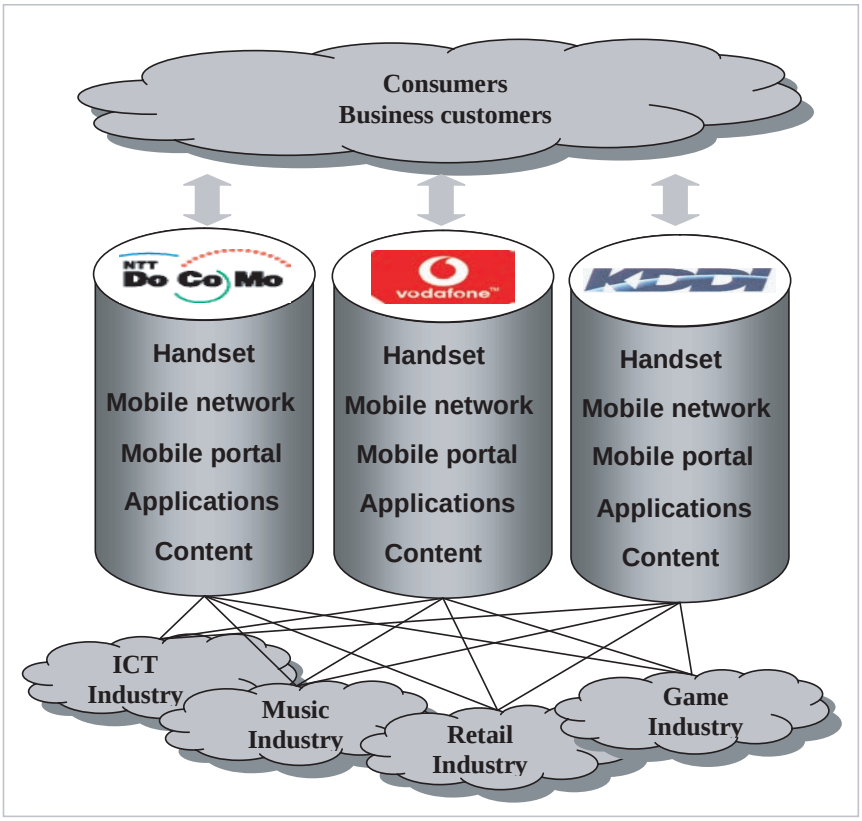
THE MOBILE SERVICES INDUSTRY IN JAPAN

There is a huge amount of academic and business research covering the various aspects of the Japanese mobile market (e.g., Baker & Megler, 2001; Funk, 2004; Kodama, 2002; Matsunaga, 2001; Natsuno, 2003; Vesa, 2003, 2005). Japan is considered to be one of the leading markets in the field of mobile Internet services: approximately 86% of the mobile phone subscribers (as of December 2003) also subscribe to the mobile Internet services, and over 20% of mobile operators average revenue per user (ARPU) comes from data access fees. For NTT DoCoMo the figures are even more impressive: almost 90% of DoCoMo's over 47 million subscribers are using the i-mode service. Furthermore, for FOMA user (the 3G service of DoCoMO), data ARPU is about 32%, which translates into over US$30 worth of data services per user each month (NTT DoCoMo, 2004). In addition the data access fees, the operators take 9–12 % of the revenue generated by content and services offerings by third parties.

In Japan, the industry structure is vertically integrated (see Figure 2) which means that mobile operators control directly or indirectly all different levels of value chain. Carriers act as wireless Internet service providers, access providers, mobile phone providers, retailers, and content aggregators (Baker & Megler, 2001). The Japanese mobile industry structure resembles the computer industry in 1975–1985 when the three largest companies (i.e., IBM, DEC, and HP) were highly integrated vertically (Fine, 1998). There is, however, one major difference: big part of the components used in the total service offering of the Japanese mobile operators comes from their partners.

In the Japanese mobile industry the competition is between business networks or ecosystems where the focal companies are the mobile opera-

Figure 2. Japan: Vertical integration with integrated product architecture



tors (see, e.g., Natsuno, 2003; Vesa, 2003, 2005). An excellent depiction of the Japanese mobile industry is the NTT DoCoMo case study by Kodama (2002), which describes the creation of a broad business network around DoCoMo's highly successful i-mode service. As Figure 2 demonstrates, the Japanese mobile operators are linking together several different industries such as the music industry and the game industry.

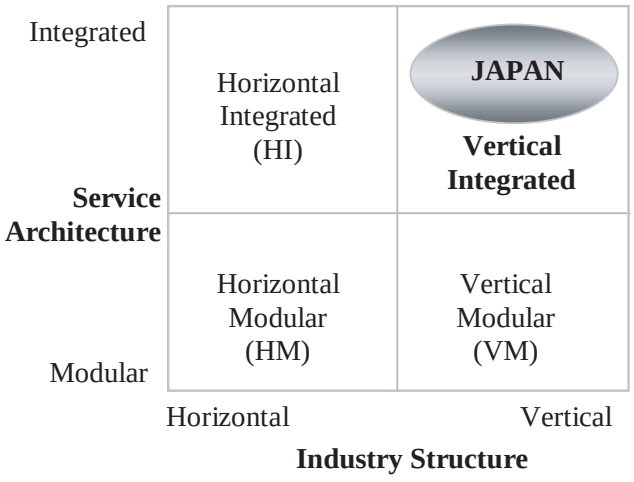
There seems to be an excellent match between the requirements of the next-generation mobile multimedia services and the traditional Japanese way of doing business. According to Hoshi, Kashyap, and Scharfstein (1991), one key component of the Japanese business environment is the concept of *keiretsu*, which is an industrial group that "coordinates the activities of member firms and finances much of their investment activity" (p. 34). Berger, Sturgeon, Kurz, Voskamp, and Wittke (1999) have named the Japanese business networks "captive value networks" that rely on dominant lead firms, i.e. suppliers of various elements of the mobile services (e.g., handsets, network technology, applications, content) are highly dependent on one or a few key customer firms. Lead firms often "urge affiliated suppliers to adopt specific

technologies" (Berger et al., 1999), as is the case in the Japanese operator-specific mobile networks and handset specifications.

The service architecture of mobile services is highly integrated in Japan. Mobile service offering consists of a handset, mobile network, and mobile portal (i.e., "a window" or "a door" to different kinds of mobile content and services) that are all closely integrated. Mobile phones are sold under an operator's brand and each phone model is designed to be used only in that operator's network. All three operators have their own mobile portal, which plays an important role in the mobile Internet business. Due to the limited screen size of a mobile phone and because of the huge amount of content available in mobile portals, it is very important how the links to various services are presented to the users when they access mobile Internet services. The last components of the Japanese mobile services business model are applications and content. Mizukoshi, Okino, & Tardy (2001) note that operators do not buy content from content owners or aggregators, nor do they create their own content, but they do control the content business through their certification process and billing service.

There are some disadvantages in the vertically integrated industry structure and integrated service architecture in Japan that have been identified in previous research. Perhaps the biggest disadvantage is that operators have to subsidize the price of the mobile phone—and the subsidy may be as high as 90% of the end-user price. Furthermore, mobile phone manufacturers are forced to develop and manufacture mobile phones that can only be used in one operator's network. According to Baker and Megler (2001), this increases the R&D and manufacturing costs in the Japanese system. It is important to keep in mind, however, that even the European "open and standardized" GSM world has been criticized for unfairness in the distribution of benefits resulting from standardization. As the CEO of a leading European mobile operator pointed out in his presentation in the 3GSM World conference in Cannes in February 2004, the only one to benefit from the fact Nokia manufactures over 100 million phones per year is the company itself—the benefits of

Figure 3. Positioning the Japanese market in the Mobile Services Industry Matrix



standardization remain within the walls of the mobile phone giant!

Due to the large domestic market of over 80 million mobile phone users, the Japanese operators have managed to reach a critical mass of users for their services, despite the lack of or limitations in the interoperability of the three commercial mobile networks in Japan (even though one could argue that at least from the technical point of view there are more than three mobile or cellular networks in Japan).

As Figure 3 demonstrates, the Japanese mobile services industry is positioned in the "Vertical/ Integrated" quadrant of the MOSIM. This concludes our brief overview of the Japanese mobile services market. Next we will examine what the Finnish mobile market looks like.

THE FINNISH MOBILE MARKET

Since the time Nokia introduced their first analog mobile phones and the state-owned telecom operator Telecom Finland (the predecessor of Sonera, which later merged with Swedish telecom operator Telia) opened their NMT (Nordic Mobile Telephony) network, Finland has been one of the world leaders of mobile telephony. The role of Finnish mobile market was also important during the launch of the digital GSM networks. Against this background it is easy to understand why the Finnish mobile market (despite the population of only 5 million people) is still considered to be an interesting subject for the research of the structure and dynamics of the mobile industry. The regulatory framework of the Finnish mobile market in particular receives much attention in different parts of the world, for reasons we will discuss later in this section.

The Finnish mobile services market is almost the opposite of the Japanese market. The industry structure in Finland is horizontal: the competition is taking place on each of the horizontal layers of the market, that is, operators are competing against each other, mobile phone manufacturers are competing against each other, and so forth

Figure 4. The horizontal and modular structure of the Finnish mobile market

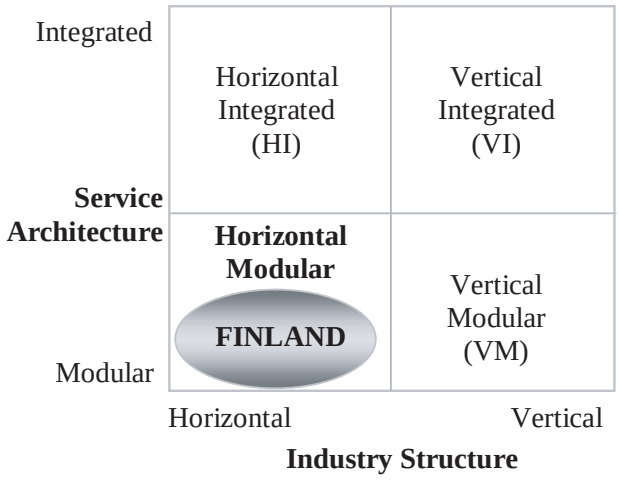
(see Figure 4). Finland is one of the few countries in Europe (along with Italy) that does not allow the use of so called SIM-lock, which prevents the users from switching to another operators network without paying penalties. In addition, the legislation in Finland prohibits mobile operators from bundling mobile phones and subscriptions—or more accurately, customers' decision to buy subscription must not affect the price of the mobile phone, in case subscription and mobile phone are sold at the same time. One of the implications of the contemporary regulatory framework is that mobile phones and mobile subscriptions are sold separately—and sometimes through different channels. An important element of the mobile business in Finland is so-called "finders fee," which operators pay to independent retail stores and chains (such as specialized stores selling only mobile phones, departments stores, and electronics and household appliance resellers) for each new subscriber. Therefore, it can be argued that the regulatory environment has a direct impact on the business models of the mobile services industry in Finland.

Handset subsidies and the bundling of mobile subscription and handset will be allowed for 3G handsets during a period of two years starting in the beginning of 2006. However, for the 2nd generation GSM handsets, bundling remains prohibited. The implications to the structure of the Finnish mobile services industry remain to be seen . The second largest mobile operator, Elisa (which has cooperative agreement with Vodafone), and handset manufacturer SonyEricsson would have preferred "the Central European" model that allows handset subsidies. The legislative authorities decided to maintain the existing legislation; in other words, operators are not allowed to subsidize the handset price. According to a representative of MINTC, authorities are prepared to reevaluate the situation if, for instance, 3G services, which were launched at the end of 2004, does not take off as expected.

As Figure 5 demonstrates, the Finnish mobile services industry is positioned in the "Horizontal/Modular" quadrant of the MOSIM.

This concludes our brief overview of the Finnish mobile services market. In the following

Figure 5. Positioning the Finnish market in the Mobile Services Industry Matrix



section we will focus on the differences between these two markets that represent the two extremes in the dichotomy introduced in the Double Helix model by Fine (1998) that has inspired the development of the MOSIM.

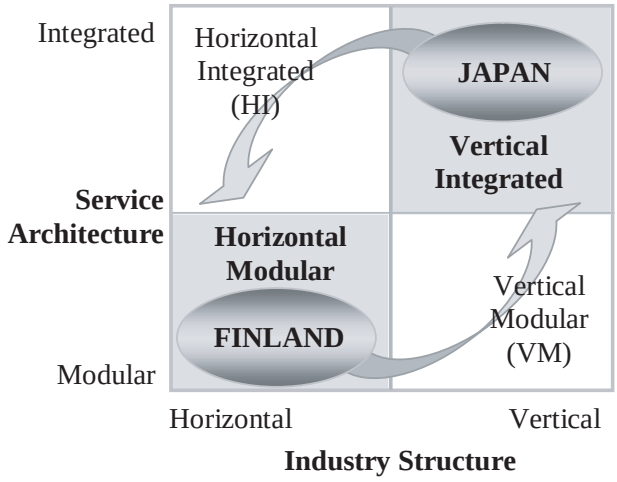
DIFFERENCES BETWEEN THE TWO MOBILE MARKETS

Next we will compare the structure and dynamics of the mobile services industries in Japan and Finland based on the MOSIM presented in section 2, and the analyses of the two markets presented in the previous two sections.

By now it has come very clear that the Japanese market (i.e., vertical/integrated configuration) and the Finnish mobile services market (i.e., horizontal/modular configuration) could not be farther apart from each other than what they are. This may not, however, be any permanent condition for neither of the markets. According to Fine (1998), neither vertical/integral nor horizontal/modular industry configurations are very stable as the forces of integration and disintegration are causing industry structures to oscillate between the two extremes in an infinite double loop. This led Fine to introduce the double helix of DNA as a metaphor when he wanted to describe the cyclic nature of industry evolution.

In Figure 6 the cycle between vertical/integrated and horizontal/modular is demonstrated by using the MOSIM. One of the benefits of the matrix approach when compared with the Double Helix loop is that it identifies the two types of intermediate market or industry configurations between the two extremes. The mobile services industry in a given country may evolve from horizontal/modular (HM) to vertical/modular (VM), for instance, if an operator chooses to use open and standardized product and service components, but wants to achieve more control in the market by integrating vertically upstream or downstream in the industry value chain, or even to different but related industries. Likewise, after moving from proprietary technologies to open standards, an industry may enter a configuration where industry structure becomes horizontal, but at the same time operators try to maintain an integrated service architecture. Later in this

Figure 6. The cycle between vertical/integrated and horizontal/modular configuration



chapter we will test the matrix model by adding a few more markets into the matrix.

Based on the analysis presented in this section, we can conclude that the industry structures of Japan and Finland are like day and night, or Yin and Yan. Next, we will try to see what is the impact of a given market configuration on the success of mobile services in the respective markets.

HOW SUCCESSFUL IS A GIVEN MARKET CONFIGURATION?

One of the objectives of this chapter is to study the relationship between an industry configuration (i.e., industry structure vs. service architecture), and the success of mobile services in a given market. In order to do this, we will next compare the key performance indicators (KPIs) of Japan and Finland. The assumptions behind this comparison are as follows:

- If industry configuration has a role in the success of mobile services, then there ought to be a significant difference in the KPIs of Japan and Finland because the industry configurations are practically the opposite.
- Success of mobile services can be measured by using the KPIs of the markets and the leading operators in the markets.

Table 1 presents some of the key performance indicators of the two case markets. The values presented here are not exact values but estimates

Table 1. Comparison of the Japanese and Finnish mobile services markets

Key Performance Indicators	**Japan**	**Finland**
Mobile phone subscribers (1,000)	84,3131 [1)]	4,880 [2)]
Mobile Network Operators	DoCoMo, KDDI, Vodafone	TeliaSonera, Elisa, Finnet
Industry structure	Vertical	Horizontal
Service architecture	Integrated	Modular
% of users using mobile Internet (excl. SMS)	***89%*** [1)]	**5%** [7)] (estimate)
ARPU US$ / user/ month	US$ 89[1)]	US$ 48 [3)]
Data ARPU % (incl. SMS)	***22 - 32%*** [4)]	**11-12%**
Non-SMS Data Revenue (Person-to-content)	14.3% [5)]	1% [6)]
Minutes of use (MoU)	219	157[3)]
Churn	1 – 1.5%[1)]	20 – 30 %[3)]

[1] *Source: NTT DoCoMo, 2004.*

[2] *Source: Helsingin Sanomat, 2004.*

[3] *Source: Results of Q1–Q3 of FY2004, Elisa Oyj and TeliaSonera Oyj.*

[4] *Source: Vodafone, June 2004; NTT DoCoMo's October 2004 (FOMA service).*

[5] *Vodafone Japan, data revenue in quarter ended 30 June, 2004 (excl. messaging).*

[6] *Estimate by a representative of a leading telecom operator in November 2004.*

[7] *According to Statistics Central of Finland, 80% of Finnish mobile phone subscribers used SMS services in 2002.*

indicating the magnitude of a certain KPI (as all operators do not provide all the KPIs presented in this comparison).

What can we learn from the comparison of the KPIs of the Japanese and Finnish mobile markets presented in Table 1? Let us focus first on three items that are particularly interesting for the purposes of this chapter. As Table 1 shows, a huge majority of Japanese mobile phone subscribers use also mobile Internet services. For instance, over 89% of NTT DoCoMo's customers subscribe to the i-mode service (NTT DoCoMo, 2004). Unfortunately there is no such information available for the Finnish market, but according to some estimates the respective figure in Finland is less than 10%. What makes this comparison somewhat challenging is that in Finland SMS is used not only for person-to-person communication but also as a means of ordering ring tones, logos, wallpapers, and so forth, digital content to mobile phones. Premium rate SMS is also an important billing mechanism as operators do not offer any other micropayment services.

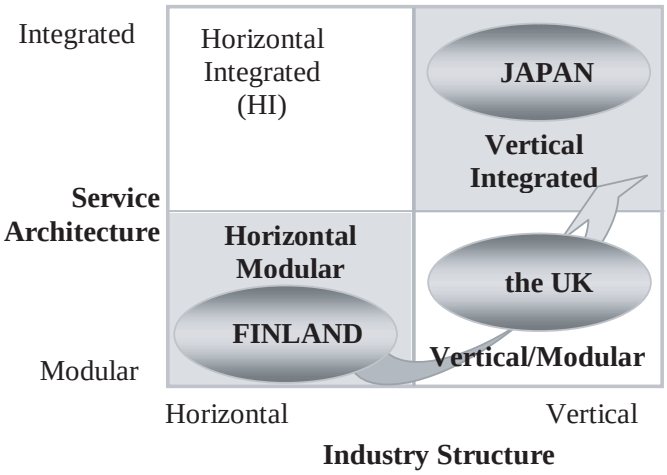
The second interesting item in Table 1 is the amount of data ARPU (average revenue per user) as a percentage of the total ARPU. In Japan data ARPU was 22% for Vodafone Japan's subscribers (mainly 2.5G users) and 32% for the subscribers of NTT DoCoMo's 3G services FOMA. In Finland data ARPU for the leading mobile operator Sonera was around 11–12% in Q3 2004. So despite the extensive use of SMS in Finland (for instance, the second largest operator's, Elisa's, subscribers sent on average 34 SMS messages per month; for Sonera this figure is below 30 messages per month), the Finnish operators are far behind the Japanese counterparts in making money with mobile data services.

The third item under scrutiny in this context is the amount of non-SMS data revenue, that is, the use of content over mobile networks (i.e., person-to-content services). Vodafone Japan reported in the end of the second quarter of 2004 that messaging represented 7.6% and data 14.3% of the total monthly ARPU. According to an estimate by a representative of a leading Finnish telecom operator, non-SMS data revenue in Finland is about 1% of the total ARPU.

Although this simple comparison of KPIs is not statistically relevant, it demonstrates the differences in the success of nonvoice mobile services in these two markets. Interestingly, both markets are considered to be advanced mobile markets—Japan because of the extensive use of mobile Internet services, and Finland due to high penetration rate of mobile phones and also because of the very low mobile phone call tariffs. It is also interesting that WAP services were introduced in the Finnish market roughly at the same time as NTT DoCoMo launched the i-mode service; however, while i-mode turned out to be a huge success story, WAP services failed miserably (although WAP technology is now starting to be an essential element of most mobile Internet services in Europe). Mobile networks in both countries were technically about on the same level although packet-based networks where first introduced in Japan. So if it was not about technology, timing, or people's willingness to adopt mobile technologies and services, why are mobile services much more successful in Japan and the gap seems to be growing? I argue that the vertical/integrated configuration of the Japanese mobile services industry has been more successful when the business paradigm is shifting from traditional voice-centric messaging-type services to content driven mobile Internet services.

Let us add one more market into our comparison. The UK market represents an interesting intermediary configuration between the Japanese and the Finnish markets. In the UK leading operators, such as Vodafone, mm02, T-Mobile, and Orange, have been building a new type of business model where mobile operators are orchestrating service delivery (i.e., moving from horizontal to more vertical industry structure) while maintaining more modular product architecture than the Japanese operators. If we add the UK market

Figure 7. The UK market in the Mobile Services Industry Matrix



into the MOSIM, we can see that we have now occupied a new quadrant of the matrix.

What we can see in Figure 7 is that the UK market is moving towards the vertical/integrated configuration, or the "Japanese model." This development is in line with the strategic statements by several leading European mobile operators who announced at the end of year 2002 that their goal was to increase operators' role in driving the mobile industry in the coming years. One of the reasons for this was the increasingly dominance of Nokia in the industry.

So how is the intermediate model of the UK doing, if we compare its success with Japan or Finland? The UK market is not even close to Japan but it has clearly advanced during the past 2 years. For instance, at Vodafone UK, data ARPU is close to 17% and the share of non-SMS data revenue is over 2%. If we compare these figures with the Finnish KPIs we can see that something has happened that has made the UK market more successful in the field of mobile services. Once again, I argue that this can be explained by comparing the differences in the market configurations.

DISCUSSION AND CONCLUSION

The objective of this chapter was to analyze the differences between the Japanese and Finnish mobile services markets by using the MOSIM. The matrix was derived from the Double Helix model developed by Fine (1998). The analysis revealed that the Japanese and the Finnish mobile industries represent very different kinds of market configurations: the Japanese market is a textbook example of vertically integrated industry structure, whereas the mobile services industry in Finland is currently in the horizontal and modular configuration. Our brief review of the UK market showed that there are also intermediate configurations between the two extremes. The transformation of the UK mobile services industry has been very fast, which indicates that

the industry clockspeed (Fine, 1996; Mendelson & Pillai, 1999) is high in the mobile services business. What is particularly interesting from industry evolution perspective in this analysis is the role of regulation in the evolution of the mobile services industry. Both the Japanese market and the Finnish market have stuck to their existing regulatory frameworks. In Japan, the national regulatory authorities did not open up the market for competition, whereas in Finland the regulatory authorities did not allow the use of SIM-lock and the bundling of subscription and handset that are typical characteristics of a vertical/integral market configuration (Vesa, 2004b). At the same time, the leading UK operators are transforming their business model towards a more vertical and integrated operator-driven business model as the regulatory framework in the UK does not prevent them from doing this. This approach seems to increase the use of nonvoice mobile services in the UK market, as the key performance indicators of the leading mobile network operators indicate. However, year 2005 appears to be a turning point both for the Finnish and the Japanese markets. As discussed earlier in this chapter, the Finnish government is in the process of allowing 3G-handset subsidies for a period of two years. At the same time, the Japanese government is in the process of opening the Japanese market for two to three mobile operators. This development in the Finnish and the Japanese markets illustrates the kind of oscillation between different market configurations described in the Double Helix model by Fine (1998).

This paper took a very challenging approach by carrying out a "macro-level" analysis of a fast-moving service industry. The analysis of two different markets that are almost the opposite of each other gave us the opportunity to identify the key characteristics of these two extremes of the continuum between horizontal/modular and vertical/integrated configurations. Even though this approach illustrates the structure and dynamics of the mobile services industry, it suffers from a well-known limitation of macro-level industry analysis: this kind of analysis is always very descriptive by nature, and gives little practical advice on how to solve the problems identified. I believe, however, that it is important both for business people and researchers who are involved in the mobile business to pay attention also to the structure and dynamics of the mobile services market where they develop and market their services. Experiences from the Japanese and the UK market indicate that a key player with enough market power (e.g., NTT DoCoMo in Japan or Vodafone in the UK) can succeed in reshaping the entire industry towards a more favorable industry configuration. There is, however, one precondition for this: the regulatory framework must not prohibit this kind of natural evolution of the industry, otherwise there is a risk that the development of mobile market is halted. The challenge for national regulatory authorities is to have the wisdom to see beyond the wishes and demands of the dominant players that may be stuck with the business paradigms of the past. In Japan the key question is how to make sure that the regulation of the mobile services industry evolves as the mobile phone market is becoming mature and new 3G technologies are being implemented. For the Finnish authorities a major challenge is to analyze whether regulatory framework that worked well in the highly standardized GSM voice market is still viable as the business moves towards more complex mobile multimedia services.

REFERENCES

Baker, G., & Megler, V. (2001). *The semi-walled garden: Japan's i-mode phenomenon*. IBM Red Paper.

Barabas, A.-L. (2000). *Linked*. New York: Penguin Books.

Berger, S., Sturgeon, T., Kurz, C., Voskamp, U., & Wittke, V. (1999, October 8). *Globalization, value networks, and national models.* Memorandum prepared for the IPC Globalization Meeting.

Fine, C.H. (1996, June 24–25). Industry clockspeed and competency chain design: An introductory essay. *Proceedings of the 1996 Manufacturing and Service Operations Management Conference*, Dartmouth College, Hanover, NH.

Fine, C.H. (1998). *Clock speed: Winning industry control in the age of temporary advantage*. Perseus Books.

Funk, J. (2004). *Mobile disruption: The technologies and applications driving the mobile Internet.* Hoboken, NJ: Wiley-Interscience.

Hoshi, T., Kashyap, A., & Scharfstein, D. (1991). Corporate structure, liquidity, and investment: Evidence from Japanese industrial groups. *The Quarterly Journal of Economics, 106*(1), 33–60.

Kodama, M. (2002). Transforming an old economy company into a new economy success: The case of NTT DoCoMo. *Leadership & Organization Development Journal, 23*(1), 26–29.

Matsunaga, M. (2001). *Birth of i-mode: An analogue account of the mobile Internet*. Singapore: Chung Yi Publishing.

Mendelson, H., & Pillai, R.R. (1999). Industry clockspeed: Measurement and operational implications. *Manufacturing & Service Operations Management, 1*(1), 1–20.

Mitchell, W., & Singh, K. (1996). Survival of business using collaborative relationships to commercialize complex goods. *Strategic Management Journal, 17*(3), 169–195.

Mizukoshi, Y., Okino, K., & Tardy, O. (2001, January 15). *Lessons from Japan.* Retrieved November 27, 2004, from http://telephonyonline.com

Natsuno, T. (2003). *The i-mode wireless ecosystem.* John Wiley & Sons.

Vargo, S.L., & Lusch, R.F. (2004). The four service marketing myths: Remnants of a goods-based, manufacturing model. *Journal of Service Research, 6*(4), 324–335.

Vesa, J. (2003, August 23–24). The impact of industry structure, product architecture, and ecosystems on the success of mobile data services: A comparison between European and Japanese markets. *Proceedings of the ITS 14th European Regional Conference*, Helsinki, Finland.

Vesa, J. (2004a, March). *The impact of industry structure and product architecture on the success of mobile data services.* Austin Mobility Roundtable, University of Texas, Austin.

Vesa, J. (2004b, Septmber 5–7). Regulatory framework and industry clockspeed: Lessons from the Finnish mobile services industry. *Proceedings of the ITS 15th Biennal Conference*, Berlin, Germany.

Vesa, J. (2005). *Mobile services in the networked economy.* Hershey, PA: Idea Group.

This work was previously published in Unwired Business: Cases in Mobile Business, edited by S. Barnes and E. Scornavacca, pp. 253-269, copyright 2006 by IRM Press (an imprint of IGI Global).

Chapter 2.27
A Mobile Intelligent Agent-Based Architecture for E-Business

Zhiyong Weng
University of Ottawa, Canada

Thomas Tran
University of Ottawa, Canada

ABSTRACT

This article proposes a mobile intelligent agent-based e-business architecture that allows buyers and sellers to perform business at remote locations. An e-business participant can generate a mobile, intelligent agent via some mobile devices (such as a personal digital assistant or mobile phone) and dispatch the agent to the Internet to do business on his/her behalf. This proposed architecture promises a number of benefits: First, it provides great convenience for traders as business can be conducted anytime and anywhere. Second, since the task of finding and negotiating with appropriate traders is handled by a mobile, intelligent agent, the user is freed from this time-consuming task. Third, this architecture addresses the problem of limited and expensive connection time for mobile devices: A trader can disconnect a mobile device from its server after generating and launching a mobile intelligent agent. Later on, the trader can reconnect and call back the agent for results, therefore minimizing the connection time. Finally, by complying with the standardization body FIPA, this flexible architecture increases the interoperability between agent systems and provides high scalability design for swiftly moving across the network.

INTRODUCTION

Many people nowadays use mobile devices such as personal digital assistants (PDA) or mobile phones to access information through the Internet. In addition, they desire to have the ability to

participate in e-business anywhere and anytime via their mobile devices. Current e-business applications, such as business-to-consumer (B2C) or Internet-based shopping, are typically developed over the Web for human-computer interaction. These applications require that users must login the intended Web sites from their personal computers or public terminals. Also, users often need to visit lots of sites and are always involved in a time-consuming process. To address these challenges, several wired agent-based e-business systems have been proposed. Kasbah (Chavez & Maes, 1996), for example, is an electronic marketplace where buying and selling agents can carry out business on behalf of their owners. Nevertheless, these systems do not satisfy the users' mobile demand due to their lack of wireless channels.

This article proposes a feasible architecture that combines agent mobility and intelligence for consumer-oriented e-business applications. It allows a user to create a mobile, intelligent agent via a mobile device, and then launch the agent to the Internet to perform business on the user's behalf. The aspect of mobility enables our architecture to support the agent's migration and the user's mobility (the ability to conduct e-business via mobile devices anyplace and anytime). The mobile agent will migrate from market to market, communicating with different trading agents to find the most appropriate one. Once an appropriate agent is found, it will inform the user of the results. This architecture complements the current Web-based, Internet systems by adding the wireless channel of mobile agents. Our current work focuses on lightweight mobile agents which act on behalf of consumers and participate in consumer-to-consumer (C2C) e-business applications. However, the architecture can be extended to business-to-consumer (B2C) or business-to-business (B2B) applications, as discussed later in the article.

Since personal software agents essentially need to communicate with other agents (to accomplish their designated tasks), they have to comply with a set of standards concerning the agent communication language and the protocols to be used. Although there is currently no universally accepted set of standards for developing multi-agent systems, the Foundation for Intelligent Physical Agents (FIPA), which aims at providing one language commonly understood by most agent-based systems (FIPA, 2006), is obtaining a growing acceptance. With FIPA becoming a de facto standard in this field, the architectures such as JADE (Java Agent Development Environment) have become available to allow for the implementation of a FIPA-compliant multi-agent system such as our proposed architecture (Chiang & Liao, 2004).

It should be noted that mobile devices suffer not only from limited battery time, memory, and computing power, but also from small screen, cumbersome input, and limited network bandwidth and network connection (Wang, Sørensen, & Indal, 2003). The proposed architecture, by making use of mobile agent technology, offers a solution to those problems. That is, after creating and initializing a mobile agent to act on the user's behalf, a user can disconnect the mobile device from the server. The user only needs to reconnect later on to recall the agent for results, hence minimizing the use of resources. In addition, mobile agent technology also addresses such challenges as increased need for personalization, high latency, demand for large transfers, and disconnected operation (Kotz & Gray, 1999).

The remainder of this article is organized as follows: the second section introduces background knowledge and related work. The third section illustrates the proposed architecture. The fourth section shows an implementation of the proposed architecture. The fifth section discusses some existing problems and future works. The sixth section concludes the article.

BACKGROUND AND RELATED WORK

Mobile Agent Paradigm

An intelligent agent is a piece of software, which differs from the traditional one by having such features as being autonomous, proactive, social, and so on. One of these characteristics is mobility, that is, the agents' ability to migrate from host to host in a network. Mobile agents are defined as programs that travel autonomously through a computer network in order to fulfill a task specified by its owner, for example, gathering information or getting closer to the required resources to exploit them locally rather than remotely. A mobile agent is not bound to the system on which it begins execution, and hence can be delegated to various destinations. Created in one execution environment, it has the capability of transporting its state and code with it to another host and execute in the same execution environment in which it was originally created. Several mobile agent systems have been designed in recent years. Telescript (White, 1996) is the first commercial mobile agent system developed by General Magic. Telescript provides transparent agent migration and resource usage control. Aglets from IBM (Lang & Oshima, 1998) is also a mobile agent system based on the concept of creating special Java applets (named aglets that are capable of moving across the network). JADE (Bellifemine, Caire, Trucco, & Rimassa, 2006) is one of the agent development tools that can support efficient deployment of both agents' mobility and intelligence in e-business applications. As a middleware implemented in Java and compliant with the FIPA specifications, JADE can work and interoperate both in wired and wireless environments based on the agent paradigm. JADE supports weak mobility; that is, only program code can migrate while no state is carried with programs. NOMADS (Suri et al., 2000) supports strong mobility and secure execution; that is, the ability to preserve the full execution state of mobile agents and the ability to protect the host from attacks.

Recently, mobile agents have found enormous applications including electronic commerce, personal assistance, network management, real-time control, and parallel processing (Lange & Oshima, 1999). Kowalczyk et al. (2002) discuss the use of mobile agents for advanced e-commerce applications after surveying the existing research. There are many advantages of using the mobile agent paradigm rather than conventional paradigms such as client-server technology: reduces network usage, introduces concurrency, and assists operating in heterogeneous systems (Lange & Oshima, 1999).

Related Work

Mobile agents have been recognized as a promising technology for mobile e-business applications. The interest of developing mobile agent systems for mobile devices has increased in recent years. Telescript describes a scenario in which a personal agent is dispatched to search a number of electronic catalogs for specific products and returns best prices to a PDA from where it starts (Gray, 1997). An integrated mobile agent system called Nomad allows mobile agents to travel to the eAuctionHouse site (http://ecommerce.cs.wustl.edu) for automated bidding and auction monitoring on the user's behalf even when the user is disconnected from the network (Sandholm & Huai, 2000). They aim at reducing network traffic and latency.

Impulse (2006) explores a scenario in which e-business meets concrete business through a system of buying and selling agents representing individual buyers and sellers that carry out multiparameter negotiation and running on the wireless mobile devices. Impulse deploys personal agents on mobile devices to help users seek agree-

ment on purchase terms. However, these personal agents are directed to move online to participate in negotiations, and hence resulting in potentially long-time connection with the Internet. We also think that the Impulse system was designed with a single communication protocol for all agents. This presents drawbacks due to the heterogeneity of exchanged information and leads to an inflexible environment, which can only accept those agents especially designed for it. Agora (Fonseca, Griss, & Letsinger, 2001) is a project conducted at HP Labs to develop a test-bed for the application of agent technology to a mobile shopping mall. A typical scenario consists of mobile shoppers with PDAs interacting with store services while in the mall, on the way to the store, or in the store itself. The Zeus agent toolkit, developed by British Telecommunications, was used to implement all agents in the Agora project. Only the infrastructure agents speak the FIPA Agent Communication Language (ACL), causing the architecture to conform partly to FIPA, although more effort in conformance is needed. The purpose of the Agora project is to gain experience in agents' communication protocols and to realize the significance of architectural standards.

It has been shown that modern agent environments such as JADE could be easily scaled to 1,500 agents and 300,000 messages (Chmiel et al., 2004). Thus, it is now possible to build and experiment with large-scaled agent systems. Moreno et al. (2005) use JADE-LEAP (JADE Lightweight Extensible Agent Platform) to implement a personal agent on a PDA. This agent belongs to a multi-agent system that allows the user to request a taxi in a city. The personal agent communicates wirelessly with the rest of the agents in the multi-agent system, in order to find the most appropriate taxi to serve the user. IMSAF (Chiang & Liao, 2004) is an architecture designed to fulfill the requirements of Impulse-introduced mobile shopping and implemented using JADE-LEAP tools. LEAP can also be used to deploy multi-agent systems spread across mobile devices and servers; however, it requires a permanent bidirectional connection between mobile devices and servers. Considering the current expensive connection fees for cell phones, such a required permanent connection is not affordable for consumers in practice.

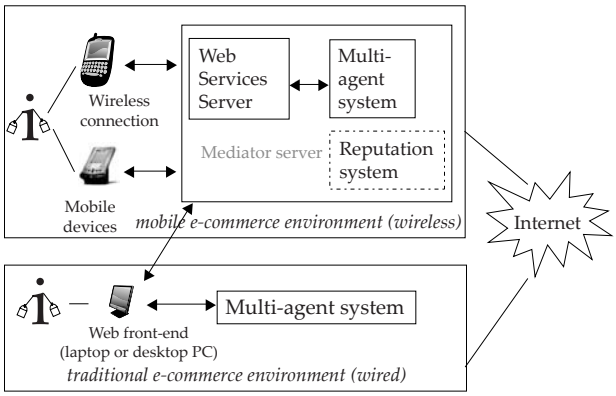
In contrast to the above works, we are motivated to propose a mediator-based architecture that attempts to enable users' wireless participation in several e-marketplaces through their mobile devices. The mobile agents can move across the network and perform trading tasks on behalf of their users when the users are disconnected from the network. We believe that it is important to consider the limitations of mobile devices, such as low-battery, low bandwidth, high latency, limited computing ability, and expensive connection fees. The fact that consumers in our physical world may need to access the worldwide markets and distributed e-business environments requires the agents to operate in heterogeneous and dynamic environments as well as to talk a common language. By complying with the FIPA specifications, the proposed architecture provides an interoperable solution to allow users dynamically to connect to the network by means of their mobile devices only when needed. Also, the mobile devices will not suffer those limitations mentioned above. In addition, the benefit of using mobile agents to a user becomes more obvious in our architecture if the user has a mobile phone and is interested in minimizing expensive connection costs. The next section explains the architecture in more detail.

SYSTEM ARCHITECTURE

Overview

Figure 1 shows a distributed C2C wireless e-business environment and a traditional wired e-business one. A consumer can connect a mobile device, such as a PDA or mobile phone, to the mediator server through wireless connection and then send a request for creating a mobile buying

Figure 1. Distributed e-business environment



or selling agent to undertake a specific business task (e.g., auction bidding) on the user's behalf. A personal agent that resides on the mobile device is needed to interact directly with the consumer and to consider the consumer's personal preferences. Considered as a true representative of the consumer, the personal agent represents the consumer's interests and allows the consumer to have a choice of dispatching either a buying or a selling agent. The mediator server sits in the fixed network and provides services such as generating mobile agents according to consumers' requests. After being created, the mobile agents will autonomously travel to multiple agent-based servers on the Internet. The agent-based servers offer places for selling and buying agents to meet and negotiate with one another. The proposed mediator server contains two main components: the Web services server, which facilitates mobile agents to interface with other agents, and the multi-agent system, which manages the agents and plays the role of a marketplace similar to an agent-based server. An additional component, a reputation system, will be necessary in our architecture. Using this reputation system, agents could sign binding contracts and check user's credit histories and reputations. The trust problem will be further studied in future research (e.g., Jøsang & Ismail, 2002 present a Beta Reputation System). Also, the mediator server provides a Web-based interface, and as shown in Figure 1, a consumer can also connect a laptop or a desktop PC to the network and launch an agent to execute on the mediator server.[1] In this article, we focus on the electronic trading of second-hand products for owners of mobile devices.

The main idea is that a consumer will request the mediator server to create a buying or selling agent and then dispatch it to agent-based servers on the Internet. The main operation that occurs in an agent-based server is price negotiation where buying agents negotiate price with selling agents. According to the consumer's preferences, the buying agent may travel to different e-market sites known by the white-page agent[2] to seek goods, when the consumer desires to conduct a global multiple markets comparison. The W3C's XML schema specification (www.w3.org/XML/Schema) provides a standard language for defining the document structure and the XML structures'

data types. The consumer's preferences can be represented in an XML format. In a real business situation, we would have to ensure that messages are reliably delivered to the mediator server from the personal agents. Although this communication protocol's reliability is not detailed in our architecture currently, we could use a reliable transport at the very least, such as Reliable HTTP (HTTPR) (Todd, Parr, & Conner, 2005), for the communication between the personal agents and the mediator server. Another consideration is to encrypt the communication. Encryption technologies can also help ensure that even intercepted transmissions cannot be read easily.

A Scenario of Our Architecture

To understand the environment best, let us consider a typical scenario taken from daily life, where two hypothetical customers, named Mary and Tom, try to participate in an eBay-like auction. Mary wants to sell her used Sony MP3 player. At her office, she initiates a selling agent from a PDA, through a wireless LAN connection with the mediator server in the building. Then this selling agent lives in the server and waits for potential shoppers. Due to some unpredictable event, Mary may have to leave her office and cannot access the selling agent via her PDA (as there may be no available wireless LAN network coverage). However, she will be able to reconnect later on.

Haphazardly, Tom enters his buying preferences into his Java-enabled mobile phone, trying to buy a second-hand Sony MP3 player under a maximum price. The personal agent on his mobile phone establishes a connection with the mediator server and asks the server to launch a mobile buying agent according to his preferences. Then Tom disconnects his cell phone from the server. The mobile agent knows where and how to migrate, as instructed in the migration itinerary. As days pass, while this buying agent is roaming around the Internet, it enters into Mary's mediator server and searches for services provided. After the negotiation between the selling agent and buying agent, they reach an agreement on the item and price. With that, the buying agent will return to its host server and send a SMS (Short Message Service)-based notification to the personal agent running on Tom's mobile phone, about the potential seller gathered from the Internet. Also the selling agent sends an e-mail-based notification to the personal agent running on Mary's PDA. Finally, things left to Mary and Tom seem to be simple and easy since they could have either the cell phone number or the e-mail address from the information reported by their personal agents, respectively. As we have seen, mobile consumers only need a small bandwidth connection twice, once for initiating a migrating mobile agent and once for collecting the results when the task is finished.

Architecture Description

We explain how the whole system works in this section. Figure 2 illustrates the system architecture and the operation process. As shown in Figure 2, mobile devices are supported by personal agents and connected to the mediator server via a wireless connection. A personal agent is a static agent running on a mobile device and offers a graphical user interface (GUI) for its user to communicate with the system. The mediator server is connected to the Internet where other mediator servers or other FIPA-compliant systems exist. In the mediator server, a servlet answers any requests from the personal agent and is linked to the behavior of a proxy agent[3] in charge of handling the requests. The proxy-agent interfaces with the servlet and constructs a bridge between the Web service server and the multi-agent system. Each instance of the behavior[4] connects not only to the AMS agent (Agent Management Service as defined in FIPA, i.e., the white-page agent mentioned above), asking for the creation of a buying or selling mobile agent in the multi-agent system as well as providing a response, but also connects to the

agent DF (Directory Facilitator as defined in FIPA, i.e., the yellow-page agent), retrieving the list of agents advertising services with the DF. In this architecture, the multithreaded-servlet server is mirrored by a multibehavior proxy agent to allow for handling multiple requests in parallel.

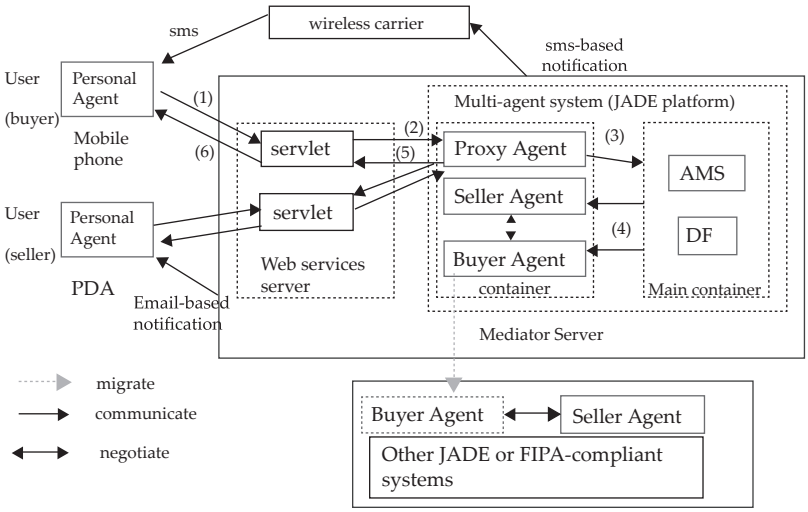
As illustrated in Figure 2, the procedures from (1) to (6) depict how a buying or selling mobile agent is created by a user according to preferences:

1. At the first step, the user configures the preferences via the personal agent (residing in the mobile device). The personal agent then sends an XML-based request to the mediator server.
2. An instance of the servlet accepts the request and communicates with the proxy agent.
3. The proxy agent cooperates with the AMS agent who lives in the main container of the JADE platform to create a buying or selling mobile agent.
4. If the buying or selling agent is created successfully in the container, it might be mobilized to other systems to undertake the user's task.
5. and (6) The personal agent receives a response from the proxy agent via the servlet and informs the user of the relevant mobile agent being created.

The above is an asynchronous process after which the user can disconnect from the network at will. Even if the user decides to disconnect from the network, the user will still receive an SMS-based notification from the mediator server via an interface with the wireless carrier, or an e-mail-based notification from the mediator server via an interface with a mail server, as long as the user reconnects to the network.

The mediator server provides the required support for the creation of mobile agents, messaging among agents, agent migration facility, collaboration, protection, destruction, and control

Figure 2. System architecture and process



of mobile agents. Mobile agent platforms such as JADE have been proposed to provide the supporting environment. Obviously, any multi-agent system can be used here as long as it provides the required support.

Different Types of Agents in Our Architecture

The following agents co-exist in our architecture: personal agents, proxy agents, buying or selling agents, yellow-page agents, and white-page agents. Among them, only buying or selling agents are mobile agents, while personal agents and proxy agents are stationary agents. Both the yellow-page and white-page agents are fixed on a component of the mediator server. Details of these agents are described as follows:

A personal agent is a stationary agent that runs on a user's mobile device and provides a graphical interface to allow the user to configure a mobile buying or selling agent (from the mobile device). When starting the personal agent on the mobile device, the user can choose either to initiate a new mobile agent or to recall a previous mobile agent. One may argue that such a personal agent is nothing more than an interface. From the agent's viewpoint, however, the personal agents are able to autonomously communicate with the proxy agent which is running in the mediator server.

A proxy agent is also a stationary agent which links the multi-agent system to the Web service server. It is one of the agents that is always up and running in the multi-agent system. The proxy agent cooperates with the AMS (white-page) agent to create a mobile buying or selling agent for each user. There is only one proxy agent per mediator server due to its unique multibehavior ability.

A yellow-page agent (such as the DF agent in the JADE platform) provides the service of yellow pages, by means of which an agent can receive information about available products or

Table 1. Attributes of a mobile agent

Attribute	Description
Agent type	The agent type that a user can select, that is, either a buying agent or a selling agent
Agent server	The configuration of the mediator server address.
User id	The user identification which can be email address, cell phone number, or IMEI (International Mobile Equipment Identity).
Quantity	Quantity of the predefined product.
Price	For a buying agent, this is the maximum price that the agent can bid: for a selling agent, this is the minimum price that the agent can accept.
Current Price Inquired	For a buying agent, this is the best price offer colledted from the Internet.
Lifetime	The total time an agent can be away before being recalled or terminated.
Mobility	Specification of whether a user desires to enalbe the agent's migration ability (i.e., in the context of a local, single or global, multiple market comparison).
Server Activity Time	The time an agent can spend on each server before migrating ot another.

find other agents providing necessary services to achieve its goal. A white-page agent (like the AMS agent in the JADE platform) represents the authority and provides naming services. It stores information about addresses and identifiers of all agents in the system. In our architecture, sellers have permission to advertise their products; and buyers are allowed to query the sellers which post the products they are looking for. Selling agents update yellow pages by publishing their services via the yellow-page agent. Buying agents query relevant services from the yellow-page agent. Both buying and selling agents update white pages by registering in or deregistering from the system. They communicate with each other via querying agent's information from the white-page agent.

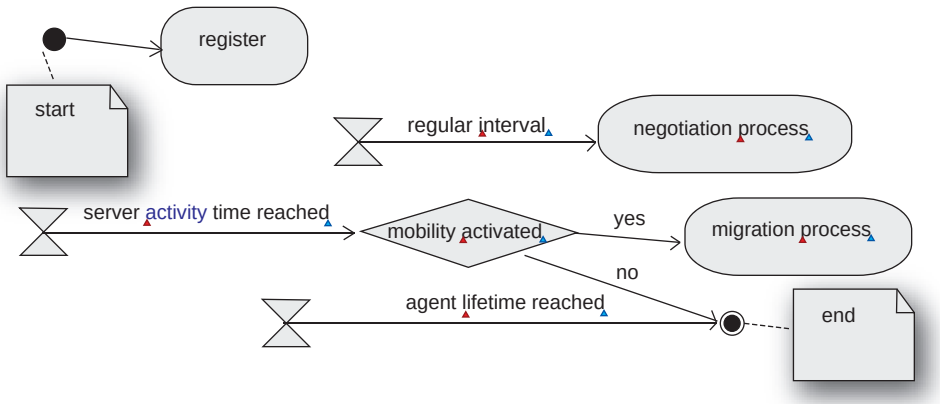
Both buying and selling agents are mobile agents, which are also called *service agents.* A service agent is the counter part of a personal agent and is involved in the migration from host to host on the Internet. A service agent first negotiates with other service agents in the same host mediator server before migrating among multiple Web sites to talk to other service agents, provided that they can talk a common language.

To demonstrate a useful mobile agent system, we present a prototype for buying and selling agents, with attributes depicted in Table 1. This means that a user will configure a mobile buying or selling agent on a mobile device, precisely according to the characteristics in Table 1.

Behaviors of Mobile Agents

As illustrated in Figure 3, a mobile (buying or selling) agent starts with its registration in the system and ends with a timeout of its lifetime. There are three time events that indicate the behaviors of a mobile agent: (1) the agent starts its negotiation process at a regular interval (e.g., every minute); (2) the agent starts its migration when activity time per server is reached; and (3) the agent ends its life cycle when its lifetime is exhausted. An argument may arise; how can one be sure that the mobile agent will be terminated according to the parameter and lifetime, as users prefer? This parameter may be changed by a third party (including the mediator server). The assumption we made is that the mobile agent can be protected from the attacks (e.g., from

Figure 3. Activity diagram of mobile agent



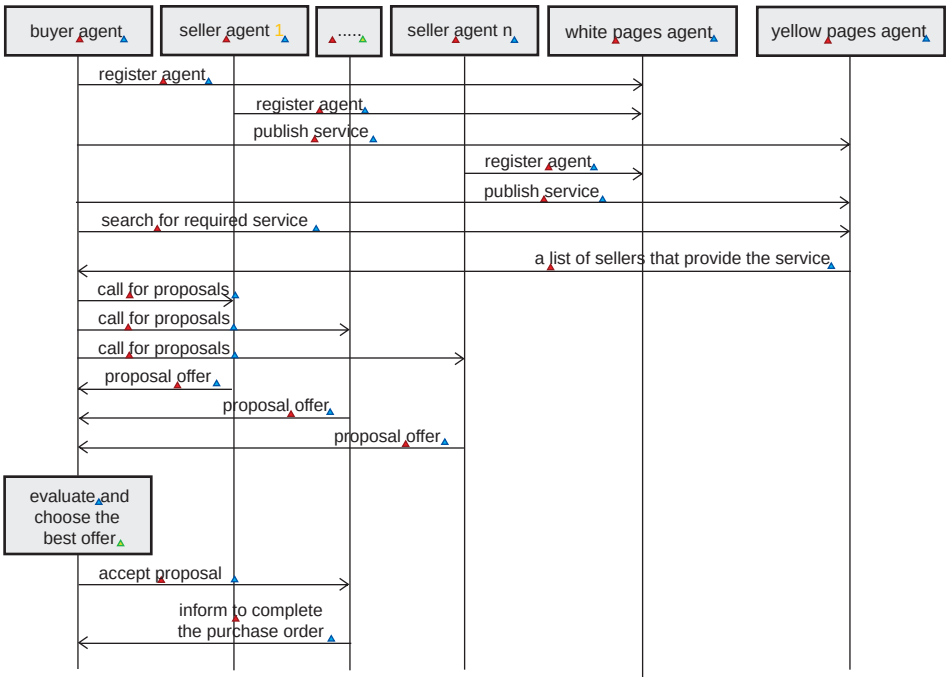
the host or other agents) once a future security mechanism is imposed on our architecture. (The security problem is discussed in the Discussion and Future Work section.)

Negotiation Process

The proposed interaction between agents complies with the FIPA-Contract-Net Protocol (FIPA, 2006). This protocol allows a buying agent (initiator) to send a call for proposals (CFP) to a set of selling agents (responders), evaluate their proposals, and then accept the most preferred one (or even refuse all of them). Both initiators and responders should register in the system before they negotiate with each other.

In this article, we consider a classical situation in which a selling agent offers a single item to the highest bidder (similar to eBay), and the simplest type of bid is an offer to buy or sell one unit at a specified price. As shown in Figure 4, the buying agent sends a CFP to all the available selling agents (obtained from the yellow-pages service). After receiving the message, a selling agent can send the buying agent a proposal with the price for the product. If the product is not available or sold, it does not need to send any proposal. The buying agent will place a purchase order if the offer price is within the maximum price that the customer has specified. Results of price negotiations are sent back to the personal agent and showed in a graphical interface to the user. Since the system is fully asynchronous, an intention to make a purchase does not have to lead to a successful transaction. By the time the offer is made, other buying agents may have already purchased the last available item.

Figure 4. Negotiation process



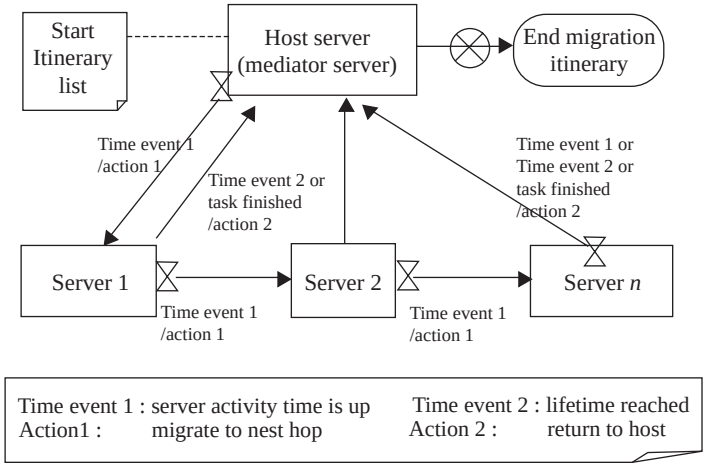
Agent Migration

The general process of migration is depicted in Figure 5. An agent starts its migration from its host server (i.e., the mediator server) with the itinerary list acquired from the host. We assume that there are *n* servers, which will be visited by the agent in sequence. In each server, two time events happen resulting in two actions respectively: if the agent reaches its lifetime, it will return to its host where it was created, and then end the migration process; if the agent exhausts its server activity time, it will migrate to the next server. Additionally, before the agent migrates to the next server, it should also make the decision if it has fulfilled the task at the current server. As we know, a task is finished when an agent receives an acceptable offer from another agent. The migration process actually describes a scenario of price comparison (finding a price less than a buyer's reservation price for buying, or searching for a price greater than a seller's reservation price for selling). The agent may access its host server repeatedly during its lifetime and updates its itinerary list every time when visiting its host server. One interesting problem here is how the mediator server maintains the itinerary list that includes a series of service-providing servers to be visited by the agent. Curbera, Duftler, Khalaf, Nagy, Mukhi, and Weerawarana (2002) state that "several individual companies and industry groups are starting to use 'private' UDDI directories to integrate and streamline access to their internal services" (p. 90). UDDI (Universal Description, Discovery and Integration) (UDDI, 2006) enables businesses to publish service listings and to discover each other. We assume that the white-page agent can interact with the UDDI server to obtain other service-providing servers' addresses (the feasibility of this function will be further studied) and therefore mobile agents can update the itinerary list during their migration. Only the mobile agents, which are originally created in this mediator server, are allowed to access this resource (a list of servers).

System Implementation

We have implemented a simple prototype to evaluate the concepts proposed in our architecture,

Figure 5. Agent migration process



using the Java programming language. Figure 6 shows the screenshots of a personal agent and a JADE-based multi-agent system, respectively. The personal agent was developed as a J2ME MIDlet[5] application that offered a graphical interface for its user to initiate or recall the mobile agent, and to dialogue with the mediator server. The mediator server played an important role in our architecture, running a Tomcat Apache Servlet Engine on a JADE platform. JADE is an open-source with good scalability, one of the best modern agent environments compliant with FIPA. As shown in Figure 6, there are two containers on the JADE system, Main-container and Container-1. Main-container holds the basic management agents defined by FIPA (AMS, DF, and RMA, which manages the GUI of the JADE platform). The proxy agent, buying agents, and selling agents run in Container-1. We can deploy the mediator-based architecture in one or several PCs.

The Web services architecture communications are based on JSR172, J2ME Web services, which include two independent parts: the JAX-RPC and JAXP. XML is chosen as the standard way for clients to interact with backend servers so as to use the remote services. J2ME JAX-RPC APIs subset solves how to access the SOAP/XML Web services and JAXP APIs subset solves how to

Figure 6. Screenshots of a personal agent and the JADE platform

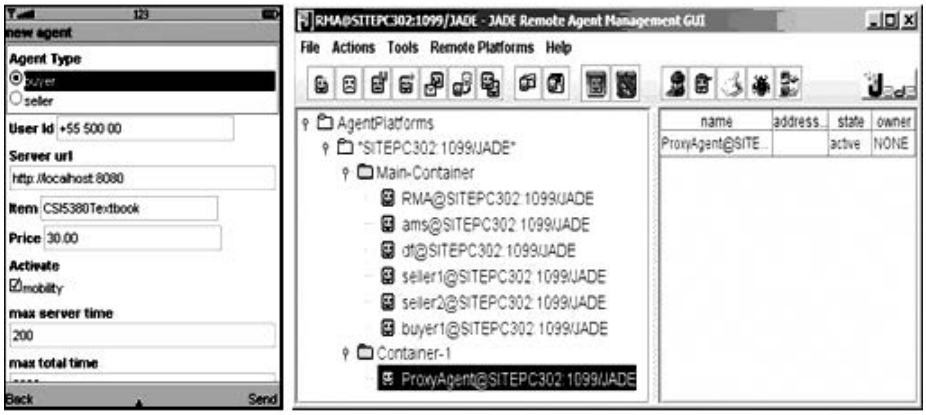
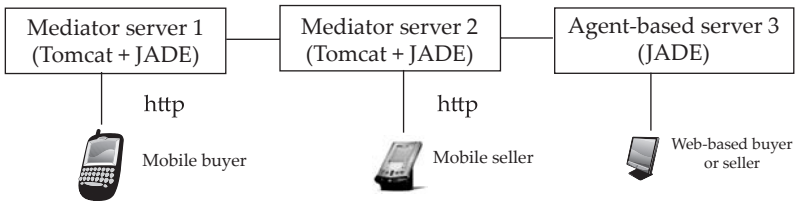


Figure 7. Experiment environment



process the XML messages. Messages exchanged by agents in the multi-agent system have a format specified by the ACL language defined by FIPA for agent interoperability.

As shown in Figure 7, we deployed three servers in the Local Area Network, installed J2ME MIDlet in two mobile phone simulators, provided one GUI for the Web-based seller, and simulated a simple used-item electronic trading scenario similar to the one we described in A Scenario of Our Architecture section previously. The mobile phone emulator is a tool provided by the Sun J2ME wireless toolkit 2.2. Both mediator servers deployed the Tomcat server and the main container of JADE platform was initialized. The third computer played the role of an agent-based marketplace on the Internet. For the buyer's emulator, the user activated the mobility of the buying agent, but it was not the case for the seller's emulator. We observed the following results:

- Mobile users can connect to mediator servers via HTTP and initiate mobile buying or selling agents in the mediator server.
- Mobile users do not need to instruct their mobile agents of what to do after configuring their preferences.
- Mobile users can add new items anytime and relevant mobile agents will be created to handle the trading of these new items respectively.
- Mobile users can kill their mobile agents to cancel their tasks by sending instruction to their personal agents.
- Mobile agents are active in their servers within the specified server activity time and then migrate to other servers.
- Buying agents can reach agreements with selling agents when the required item and price are matched. Mobile users then receive text messages from their agents, displayed on the screen of the simulators.
- Mobile agents end their life cycles when finishing their tasks.

As confirmed by the experiments, mobile users connect to their servers only when they need to add new items or to cancel their tasks. This obviously results in such benefits as reduced bandwidth utilization, increased battery life for mobile devices, and no complicated computation conducted in mobile devices. Also, mobile agents can move to various servers to negotiate autonomously, and mediator servers can accept mobile agents from outside their systems. This feature enables users to participate in multiple markets on the Internet.

In particular, we observed the migration process of a mobile agent. Mobile agents should be active in their servers within a specified time and migrate among the servers. Thus, we developed a scenario where we supposed that a buying agent started from Server1 and continued searching for the required product or service in Server2 and Server3. We set up two parameters for this mobile agent: Maximum server active time was set to 100 seconds and total lifetime was set to 850 seconds. As expected, the buying agent contacted the other agents in Server1 and then migrated to Server2 after approximately 100 seconds. Similarly, the buying agent communicated with other agents in Server2 and then traveled to Server3. The same things happened in Server3. Because there were no more sites to be visited, the buying agent migrated back to Server1, ending its first round of migration. The second round was started since the total lifetime was not reached. We assumed that no sellers offered the required product or service to this buying agent. With the time elapsed, the buying agent was in its third round and roamed into Server2. At this stage, the buying agent used up its lifetime of 850 seconds and predicted an ending of its life cycle. Therefore it migrated back to the host Server1, even though the third round trip was not finished. In another scenario, we used the same parameters for the buying agent, except that the total lifetime was enlarged to 1,000 seconds. The difference was that we dispatched a selling agent in Server2 at

the moment the buying agent was ready to launch its third round trip. This selling agent offered exactly the service that the buying agent needed. As we expected, the two agents met and reached an agreement after negotiating with each other. This experiment confirmed that after completing its task, the buying agent migrated back to Server1, regardless of its remaining lifetime that had not yet been exhausted.

DISCUSSION AND FUTURE WORK

In this article, we propose a feasible mobile agent architecture that assists users in C2C e-business. It enriches the resources for users to perform comparison shopping activities at the point of purchase. Users' mobile devices connect to the network only when needed, thus making efficient use of limited bandwidth and reducing the network traffic. In addition, it helps cell phone users save money from their expensive bills. At any time, users may add items via their personal agents and specify their preferences such as time limit and preferred price for trading. Through the negotiation process between mobile buying and selling agents, users also gain valuable information for making trading decisions.

Our proposed architecture is extensible: On the one hand, XML-based communication is used to enhance extensibility; on the other hand, the architecture could be easily extended to B2C, or even B2B business models. That is, not only individuals but also business companies can be attached to the architecture. With mobile phones and PDAs already being used as extended enterprise tools, business companies, such as retailers and suppliers, can publish their products and/or services on their servers via mobile devices. As long as these businesses take part in our architecture parties, they could benefit from the automatic discovery of other business partners. Also, it is possible for businesses, especially for retailers, to sell their products to potential buyers in the manner described in the proposed architecture as an extra way to their traditional ones. In this sense, our architecture is an integration model of C2C, B2C, and B2B e-business. Nonetheless, using mobile devices for complex tasks can be quite frustrating (e.g., difficult to enter data), so probably people will not use it. An idea is to incorporate targeted messaging or advertising into our model, where businesses could send a message to users who are physically located in their vicinity. Agents could negotiate a transaction, and the buyer would already be located nearby to complete the purchase and pick up the item.

Currently, we present a conceptual framework that needs to be refined. Using this work as a starting point, we have outlined a number of future research directions:

1. Negotiation protocols do not have to be hard-coded into the agents. Instead, mobile agents can adapt to any intelligent negotiation strategies when they arrive at a new remote location. Thus, our architecture paves the way for future research in which more general architectures can be explored to allow mobile agents to participate in a variety of negotiation protocols, such as factor negotiation (price, quality, delivery time, etc.), electronic contracting, and so on. Currently, the negotiation strategy module consists of only a purchase determined by price (agents seek a preferable price by a fixed amount). FIPA defines auction protocols (e.g., Dutch and English auctions) as well as simpler strategies such as fixed pricing, fixed pricing with a discount, and so on. We will add them into the negotiation protocols in our future research.
2. Items are described only by their names. Obviously, other attributes, such as color, age, terms of warranty and delivery should also be considered. We believe that ontolo-

gies can help to solve this problem. It should be noted that the small screen of mobile devices will bring inconvenience to users when they specify many attributes of an item. A possible solution is to make use of the persistent memory of mobile devices to store the users' preferences.

3. Mobile agent technology currently has some limitations, such as identity management, fault tolerance, protection of agents, and resource security. These limitations have brought up some concerns about the practical utilization of mobile agents. For example, in the area of security, e-business applications are often involved with money and thus users may hesitate to use mobile agents, unless mobile agents are secure enough to be trusted.

In the situation presented in this article, the mobile agents representing different buyers or sellers migrate over the Internet and then execute themselves on remote computers. These mobile agents are thus exposed to open environments and may become vulnerable. Since the mobile agents execute on unknown computers and interact with unknown agents, a reliable security infrastructure is vitally needed for the design of the system. The mobile agents must be able to deal with situations where they have been shipped off to the wrong address or to a hostile environment (Neuenhofen & Thompson, 1998). Listed below are some possible security concerns:

- Malicious mobile agents can try to access services and resources without adequate permissions. In addition, a malicious agent may assume the identity of another agent in order to gain access to platform resources and services, or to cause mischief or even serious damage to the platform.
- Mobile agents may suffer eavesdropping attack from other mobile agents. A malicious agent can sniff the conversations between other agents or monitor the behavior of a mobile agent in order to extract sensitive information from it.
- Mobile agents may suffer alteration attack from malicious hosts. To execute the agent and update its state, the host must definitely be capable of reading and writing the agent. A malicious host may steal private information from the agent or modify the agent to compute the wrong result or to misbehave when it jumps to another site.

Current research efforts in the area of mobile agent security adopt two different perspectives (Kotz, 2002): First, from the platform perspective, we need to protect the host from malicious mobile agents (such as viruses and Trojan horses) that are visiting it and consuming its resources. Second, from the mobile agent perspective, we need to protect the agent from malicious hosts. There are many mechanisms to protect a host against malicious agents. Digital signatures and trust management approaches may help identify the agent and evaluate how much it should be trusted. The malicious host problem, in which a malicious host attacks a visiting mobile agent, is the most difficult problem. We found in the literature some works on powerful techniques such as Sandboxing and Proof-Carrying Code (PCC). Sandboxing (Wahbe, Lucco, Anderson, & Graham, 1993) is a software technique used to protect a mobile agent platform from malicious mobile agents. PCC (Lee & Necula, 1997) introduces the technique in which the code producer is required to provide a formal proof that the code complies with the security policy of the code consumer. Therefore, we envisage that the security of mobile agents is an important issue that will encourage techniques and mechanisms for e-business in the future.

CONCLUSION

We propose in this article an e-business architecture that allows traders to do business at remote locations by means of mobile intelligent agents. Our architecture, which adheres to standardization efforts in the multi-agent field such as FIPA paves a possible way towards a near future when mobile buying (and selling) agents can smoothly travel among different agent-based marketplaces to carry out tasks on their users' behalves. Our purpose of presenting this idea is to improve our understanding of the value of mobility and to encourage the conceptual construction of a global community. We do not claim that buyers and sellers around the world would have to buy into this to make it work, and that worldwide C2B e-commerce would be revolutionized thereby. In practice, however, we hope that our work would be useful on a smaller scale and lead to new investigations that may result in new solutions to the problems we addressed. Our proposed architecture, aimed at providing new capabilities for advanced e-business solutions, employs an approach that integrates intelligent and mobile agents. Intelligent agents can provide automation support for decision-making tasks, while mobile agents can extend that support by allowing users to participate in several marketplaces in a networked e-business. We believe that intelligent and mobile agent technology is also a promising solution to the problems of low speed, high latency, and limited computing ability that the current wireless network is facing.

REFERENCES

Bellifemine, F., Caire, G., Trucco, T., & Rimassa, G. (2006). JADE programmer's guide. Retrieved July 7, 2007, from http://jade.cselt.it/docs

Chavez, A., & Maes, P. (1996). Kasbah: An agent marketplace for buying and selling goods. In *Proceedings of the 1st International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology,* London, United Kingdom.

Chiang, H., & Liao, Y. (2004). An agent-based architecture for impulse-induced mobile shopping, *Computer and Information Technology.*

Chmiel, K., et al. (2004). Testing the efficiency of JADE agent platform. In *Proceedings of the 3rd International Symposium on Parallel and Distributed Computing* (pp. 49-57). IEEE Computer Society Press.

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., & Weerawarana, S. (2002, March-April). Unraveling the Web services web: An introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing, 6*(2), 86-93

FIPA. (2006). Retrieved July 7, 2007, from http://www.fipa.org

Fonseca, S., Griss, M., & Letsinger, R. (2001). *An agent-mediator e-business environment for the mobile shopper* (HP Tech. Rep. No. HPL-20010157).

Gray, R.S. (1997). Agent Tcl. *Dr. Dobb's Journal*, pp. 18-26.

Impulse. (2006). Retrieved July 7, 2007, from http://agents.media.mit.edu/projects/impulse/

Jøsang, A., & Ismail, R. (2002, June). The Beta Reputation System. In *Proceedings of the 15th Bled Electronic Commerce Conference*, Bled, Slovenia.

Kotz, D. (2002). Future directions for mobile agent research. *IEEE Computer Science.*

Kotz, D., & Gray, R. (1999). Mobile code: The future of the Internet. In *Proceedings of Autonomous Agents'99: Workshop on Mobile Agents in the Context of Competition and Cooperation.*

Kowalczyk, R., et al. (2002). Integrating mobile and intelligent agents in advanced e-business: A survey. In *Proceedings of Agent Technologies, Infrastructures, Tools, and Applications for E-Services, NODe'2002 Agent-Related Workshops*, Erfurt, Germany.

Lange, B.D., & Oshima, M. (1998). *Programming and deploying Java mobile agents with aglets.* Addison-Wesley.

Lange, D.B., & Oshima, M. (1999). Seven good reasons for mobile agents. *Communications of the ACM.*

Lee, P., & Necula, G. (1997). Research on proof-carrying code on mobile-code security. In *Proceedings of the Workshop on Foundations of Mobile Code Security.*

Moreno, et al. (2005). Using JADE-LEAP to implement agents in mobile devices. Retrieved July 7, 2007, from http://www.zdnet.de/itmanager/whitepapers

Neuenhofen, K.A., & Thompson, M. (1998). Contemplations on a secure marketplace for mobile Java agents. In K.P. Sycara & M. Wooldridge (Eds.), *Proceedings of Autonomous Agents 98*, Minneapolis, Minnesota. New York: ACM Press.

Sandholm, T., & Huai, Q. (2000). Nomad: Mobile agent system for an Internet-based auction house. *IEEE Internet Computing,* pp. 80-86.

Sun. (2006). Java. Retrieved July 7, 2007, from http://java.sun.com/javame/

Suri, N., et al. (2000). NOMADS: Toward a strong and safe mobile system. In *Proceedings of the 4th International Conference on Autonomous Agents* (pp. 163-164). New York: ACM Press.

Todd, S., Parr, F., & Conner, M. (2005). An overview of the reliable HTTP protocol. Retrieved July 7, 2007, from http://www-128.ibm.com/developerworks/webservices/library/ws-phtt/

UDDI. (2006). Retrieved July 7, 2007, from http://www.uddi.org/

Wahbe, R., Lucco, S., Anderson, T.E., & Graham, S.L. (1993). Efficient software-based fault isolation. In *Proceedings of the 14th ACM Symposium on Operating Systems Principles* (pp. 203-216).

Wang, A.I., Sørensen, C.F., & Indal, E. (2003). A mobile agent architecture for heterogeneous devices. *Wireless and Optical Communications.*

White, J.E. (1999). Telescript technology: Mobile agents. In Mobility: Processes, computers, and agents (pp. 460-493). New York: ACM Press/Addison-Wesley.

ENDNOTES

1. In the experiment, we developed a GUI in the mediator server for users to launch a buying or selling agent.
2. The white-page agent maintains different service provider sites. Section 3.4 will describe this agent in more detail.
3. Detailed description of the proxy agent is provided in Section 3.4.
4. In an object-oriented context, a behavior is an inner class of the proxy agent.
5. MIDlet is a Java program generally running on a cell phone, for embedded devices, more specifically the Java ME virtual machine.

This work was previously published in the International Journal of Information Technology and Web Engineering, edited by G. Alkhatib, Volume 2, Issue 4, pp. 63-80, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.28
A Framework for Information Systems Integration in Mobile Working Environments

Javier García-Guzmán
Universidad Carlos III de Madrid, Spain

María-Isabel Sánchez-Segura
Universidad Carlos III de Madrid, Spain

Antonio de Amescua-Seco
Universidad Carlos III de Madrid, Spain

Mariano Navarro
TRAGSA Group Information, Spain

ABSTRACT

This chapter introduces a framework for designing, distributing, and managing mobile applications that uses and updates information coming from different data sources (databases and systems from different organizations) for helping mobile workers to perform their job. A summary of the state of the art in relation to mobile applications integration is presented. Then, the authors describe the appropriate organizational context for applying the integration framework proposed. Next, the framework components and how the framework is use are explained. Finally, the trials performed for testing the mobile applications architecture are discussed, presenting the main conclusions and future work. Furthermore, the authors hope that understanding the concepts related to the integration of mobile applications through the presentation of an integration framework will not only inform researchers of a better design for mobile application architectures, but also assist in the understanding of intricate relationships between the types of functionality required by this kind of systems.

INTRODUCTION

Many workers in current organizations perform their activity in mobile environments. Sellers, architects, doctors, veterinarians, and so forth perform the most part of their work outside an office, many of them in cities or at rural and remote areas. Moreover, in many cases, the information required for mobile workers comes from different information systems and databases owned by different organizations or providers, so it is necessary to provide mobile workers with devices (handhelds, pocket PCs [personal computers], tablet PCs, etc.) with software systems that employ user interfaces appropriate for this kind of devices, and with the capabilities for accessing and updating several information systems.

In order to solve this problem, an integration framework, called DAVINCI, has been defined. DAVINCI is a framework for providing mobile workers with mobile software applications to query and update information coming from different and heterogeneous databases.

The DAVINCI project was first tested with the main aim of developing a solution to help veterinarians performing in-field sanitary inspections in cattle holdings across European Union countries; DAVINCI was tested by veterinarian services from Spain, Bulgaria, Latvia, and Czech Republic.

During these trials, we identified that one of the main advantages of the DAVINCI architecture is its capability to be integrated together with different European databases for animal health controlling (for instance, EUROVET in Bulgaria or SIMOGAN in Spain registering cattle census and movements). DAVINCI also permits the development of new data warehouse systems compliant with the previously cited regulation, providing large economic costs savings. On the other hand, DAVINCI is easily adaptable to procedures in different countries, each with a singular culture and organizational structure regarding the responsibilities for livestock sanitary control.

STATE OF THE ART

Mobile computing devices (smart phones, PDAs (personal digital assistants), tablet PCs, or notebooks) increasingly include integrated wireless capabilities. Wi-Fi (wireless fidelity, 802.11) access points for wireless connectivity have appeared everywhere. Moreover, a growing number of complementary wireless networking standards, such as wireless personal area networks (802.15) and wireless metropolitan area networks (802.16), has evolved. In this sense, the users, who take their devices everywhere, expect their software applications to run as they do in the traditional network environment available at their offices.

To achieve such functionality transparently, however, these applications must meet a new set of requirements and support a specific set of capabilities related to the following.

- Provision of intelligent roaming capabilities to enable users to work without interruption, even when network connections are disrupted
- Exploitation of multiple network interfaces in a single device or the ability to select the most appropriate connection, for example, when two or more connections are simultaneously available
- Synchronization of databases by caching contents to local devices through asynchronous connections
- Access to data and applications on diverse devices through similar user interfaces
- Conservation of power at the operating-system level and maximization of performance

To implement mobile required functionality, application architects and developers have attempted to work around such problems without the benefit of development environments, application programming interfaces (APIs), or third-party

middleware solutions that are tailored for mobile environments.

The tools that are available to application architects and developers to provide this functionality are as follows.

- Mobile-application architecture guides
- Mobile-computing application servers

Mobile-Application Architecture Guides

The purpose of mobile-application architecture guides is to provide basic principles to design and develop mobile applications, facilitating the understanding of the high-level issues around mobility and mobilized software architectures.

This kind of guides identifies the primary capabilities required of mobile applications, such as efficient resource management, comprehensive context management, encoding, view consistency, extended policy and security functionality, durable storage, and reliable messaging.

Examples of mobile-application architecture guides are the following:

- *Intel® Mobile Application Architecture Guide* (Intel Corporation, 2006)
- *Mobile Applications: Architecture, Design, and Development* (Lee, Schell, & Scheneider, 2004)

Mobile-Computing Application Servers

Mobile-computing application servers are software programs that run in a server and provide the following functionality.

- Application-level logic that handles business functions involved in a particular organization and its integration with back-end database or business application systems
- Presentation services for the mobile client device (handheld computers, notebooks, PDAs, etc.). This is also called GUI (graphical user interface) in some cases, though some handhelds are more like older text terminals than PCs. It includes breaking the messages into smaller chunks, filtering redundant information, and even logically compressing the data.
- Transaction services, in some cases including multithreading for heavy volumes and persistency
- Application programming-level interfaces with specialized communications protocols

Actual implementations of an application server vary from one vendor to another. Some application servers are generic Web servers with an SDK (systems development kit) or API capability to pull data from enterprise database systems and send them to browser-based client software on a handheld device.

Depending on the heritage of the vendor and its core expertise, you can categorize application servers in the following broad classes.

- Generic application servers with a Web-based SDK, for example, Netscape, Microsoft, Sun, SilverStream, and BEA, may have support for handheld devices and wireless networks strapped onto the basic application server
- Database vendors' application servers, for example, Oracle 9i and Sybase's iAnywhere application server
- Data synchronization vendors, for example, Puma, Synchrologic, and Extended Systems
- Specialized Mobile or Web computing application servers, for example, IBM's WebSphere
- WAP-centric application servers, like Nokia's WAP Server

- E-mail-centric application servers, for example, Microsoft's mobile Information server and the EdgeMail application server

DAVINCI INTEGRATION FRAMEWORK'S MAIN CHARACTERISTICS

Our work applies many of the principles presented in the mobile-applications architecture guides to provide a framework for information systems integration in mobile working environments, paying special attention to some problems that are not well-solved by commercial mobile application servers, related to the following.

1. Provide standardized ways (independent from the vendor) to accomplish the following:
   - Define the mobile application's user interface and establish the data to be managed with this application.
   - Ship the user interface definition and data required to manage mobile applications.
   - Process the user interface definition and data to present mobile applications to the user.
2. Facilitate the development of mobile applications that integrate data coming from the following.
   - Different databases that are stored in different DBMSs (database management systems)
   - Other existing systems that have been programmed in different languages and operative systems

   This reduces the time, effort, and cost required for the development phase.
3. Optimize communications capabilities by accomplishing the following.
   - Reducing the size and number of the messages interchanged between servers and client devices
   - Increasing the possibility to adapt the mobile applications integrated to several different communication architectures
   - Recovering automatically the interrupted data messages
   - Selecting the optimal communication channel depending on the available bandwidth and the economic costs associated

Moreover, the integration framework should be deployed in different technological infrastructure coming from different vendors.

INTEGRATION FRAMEWORK WORKING CONTEXT

Before beginning with the detailed description of the integration framework, it is necessary to analyze some of the basic concepts of the working context related to information systems integration in mobile and remote settings. These concepts are the geographic area, zone of work, protocol, campaign, and workers in the field.

- **Geographic area:** This is a geographic zone that includes geographic locations (represented as geographic coordinate points) where the information management tasks are initiated. For example, in the business related to veterinarian control of pets and cattle, a geographic area could be a region of a country. In the business of the population censuses, a geographic area could be a town or a district of a large city.
- **Work zone:** This is a concrete point within a geographic area that will be visited by

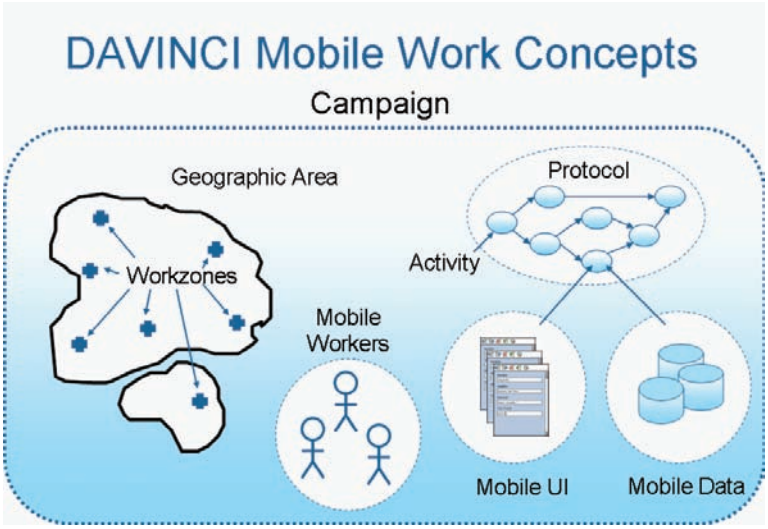
the field workers to perform their concrete tasks. For example, in the business related to veterinarian control of pets and cattle, the work zones could be each one of the cattle holdings and farms to visit. In the business of the population censuses, the work zones could be the streets or buildings of a town or a district of a large city.

- **Mobile workers:** They are the people in charge of performing concrete tasks in the assigned work zones. For example, in the business related to veterinarian control of pets and cattle, the mobile workers will be the veterinarians employed to perform the cattle and pet inspection on a farm. In the business of the population censuses, the mobile workers will be the people employed to visit each family to fill in the census forms.
- **Campaign:** A campaign is the grouping of a set of tasks of the same type that will be performed by mobile workers in several work zones and geographic areas using the same procedure.
- **Protocol:** It is the work procedure used in a work zone by a mobile worker to satisfy the objectives established in the campaign definition. In the scope of this integration framework (DAVINCI), a protocol is implemented by means of a set of applications or forms that the worker should complete or execute.

Each DAVINCI application is composed of a set of tasks or activities linked using a precedence sequence. Each one of these tasks is implemented by a form that accesses (external and/or internal) databases and updates the pertinent information through the mobile devices used by mobile workers. The forms should be defined using the standard of XForms.

The sequences of forms that define the campaign protocol configure the mobile user interface. Using this user interface, the mobile workers will

Figure 1. DAVINCI mobile work concepts



be able to manage (querying, modifying, inserting, and deleting) the data concerning each protocol task. These data will be stored in local databases, physically located in the mobile devices. In order to relate the data shown by the form and the data stored in the database, a file, named modelmapper, should be defined. These modelmappers define the relation between each one of the fields of the form and the concrete field of the local database, and the way (SQL statements) to access and update the data.

According to the concepts presented below, the integration framework works with general concepts, so this architecture is able to be applied to different business areas with very few enrichments and/or modifications.

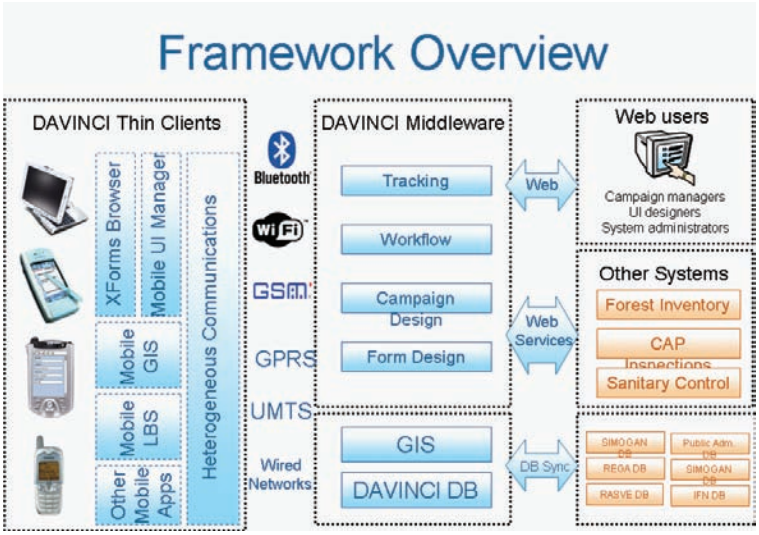
The integration framework will only store generic data on geographic areas, work zones, and mobile workers: concretely, its internal identifier, name, description, and an identifier or code assigned in an external database that contains extended information related to the concrete business area related to the tasks to be performed by the mobile workers. For example, the data to gather by a veterinarian in a work zone (cattle holding) will be different than the data collected by a questioner in a family house related to a population census.

For access to these extended data that will be stored in external databases, DAVINCI uses a specific data access module denominated DBSync.

INTEGRATION ARCHITECTURE

In order to provide support to this working context, the integration architecture suggested is based on client-server architecture, with a main module of communications and a message dispatcher, allowing the integration of any new service that is designed for a concrete area of business through its connection to the message dispatcher. Following this philosophy, DAVINCI'S integration architecture is organized in three levels, as it is shown in Figure 2.

Figure 2. Integration framework overview



a. DAVINCI provides a user interface (Web interface) for defining the process that should be used by the mobile workers and the connection of this process with the mobile software applications to be used for its consecution.
b. The core services are grouped in DAVINCI integration middleware, which permits the definition of mobile applications for performing several types of mobile work without a strong effort in software programming. These applications are defined by means of the user interface specification using XForms language (a standard for defining multimodal and multidevice user interfaces), and the specification of the information sources (server, database, table, and fields) for each user interface element (labels, text boxes, combo boxes, etc.) and the policy for selecting and updating the concrete data of a form.
   DAVINCI middleware also permits the use and adaptation of several software components for accessing and integrating heterogeneous databases that could be merged for providing the information necessary for a concrete mobile software application.
   This middleware also permits one to distribute the assignments to each mobile worker available, send the application and data for using the software application to perform the work assigned, and receive the data obtained as a conclusion of the work.
c. Moreover, DAVINCI provides some client components (currently developed for pocket PCs) that permit the following.
   - The communication with the middleware for receiving the information by the applications and data used and updated
   - The use of DAVINCI mobile applications with voice and pointing interfaces, providing access to additional capabilities related to location- and geographic-information-based services

The following sections describe these three levels in detail.

Integration Middleware

The components that permit the definition of the user interface of the final integrated user applications (Mobile UI Designer) and the components to access heterogeneous databases (DBSync) should be customizable; that is to say, in the standard version of DAVINCI, the core functionalities of these components are provided, but it is necessary to develop simple additional components to provide the full functionality needed in the final solution of the concrete sector.

Other components with full functionalities, related to the design of campaigns, the assignment of individual work to the mobile staff, and the synchronization of data and applications between the server and the mobile devices, are provided in the standard version of DAVINCI, so they do not have to be modified or enriched in any case.

Customizable Components

These components should be applied to adapt and use the integration framework in a concrete business environment.

Web Interface

The Web interface provides the required functionalities related to campaign design, the assignment of work (in terms of geographic areas or work zones) to each mobile worker, and controlling the advance of the work corresponding to a campaign.

Next, the main functionalities of the Web interface are presented in a detailed way.

- **Design of campaigns:** The design of a campaign consists of the specification of certain items.
  - The tasks to perform
  - The steps to follow for each task
  - The selection of the geographic areas and/or work zones where the mentioned tasks should be performed

  As we said previously, a protocol is implemented by an application (or set of forms) to complete or execute. Each application is composed of forms, each of them representing a task of the protocol. The forms are presented in sequence, representing in this way the precedence required among the protocol tasks. Once the forms are designed and the work assignments are planned, the forms and the related data are sent to mobile workers' devices. The above-mentioned forms are designed using the XForms standard. The

Figure 3. Campaign registration form

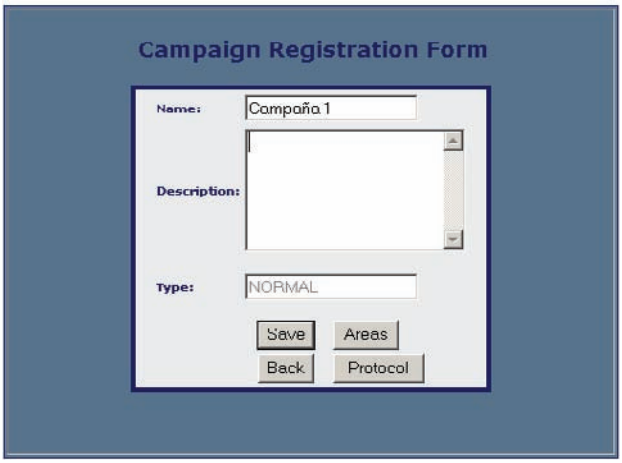
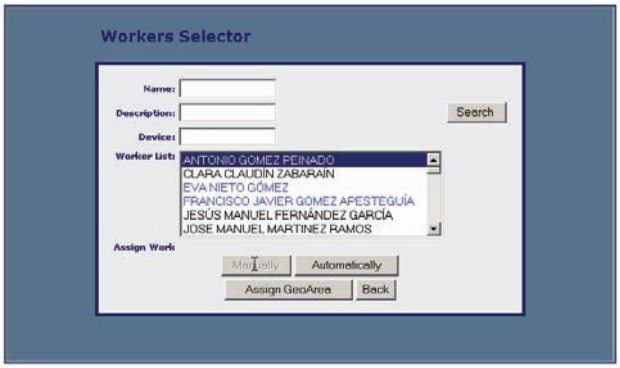


Figure 4. Workers' assignment form



query and update functions to process the data related to the forms are specified in the modelmappers files related to the forms.

- **Campaign work assignment:** Once the campaign has been designed, it is possible to begin the assignment planning. The work assignment consists of the determination of the workers assigned to a campaign and selecting the geographic areas or work zones that mobile workers have to visit to perform the campaign task. Moreover, the Web interface permits the campaign manager to fix the concrete dates to perform the tasks, configuring the agenda of the mobile workers.

  Once the work has been assigned to each available mobile worker, the campaign manager should send the assignments obtained to mobile workers; so, the integration framework sends, through a message-center server component, to the mobile workers' devices the forms definition, modelmappers

Figure 5. Work-zone selection form

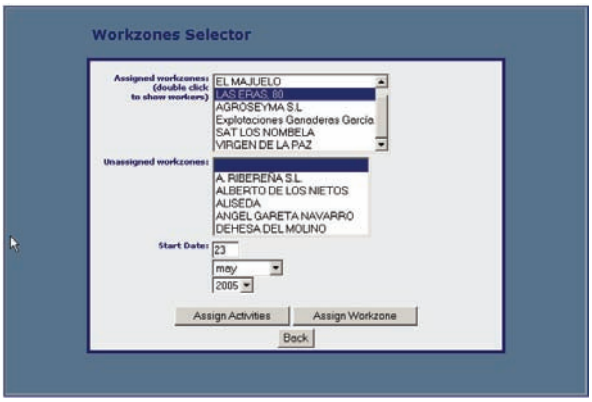


Figure 6. Campaign query form

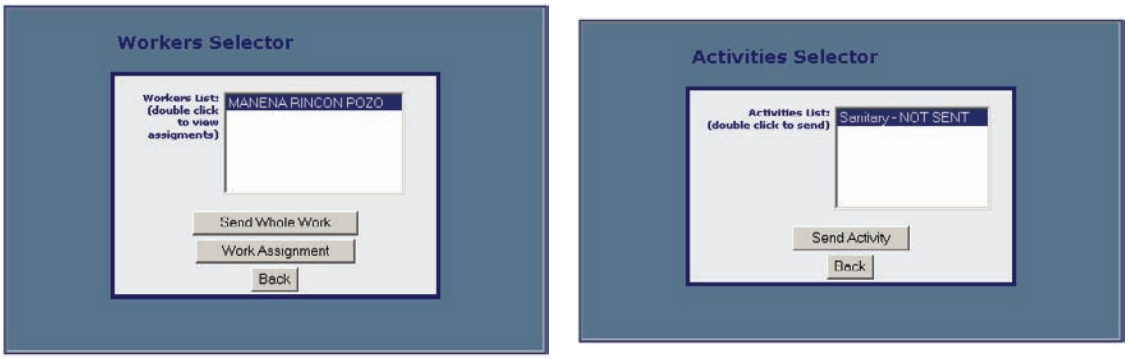
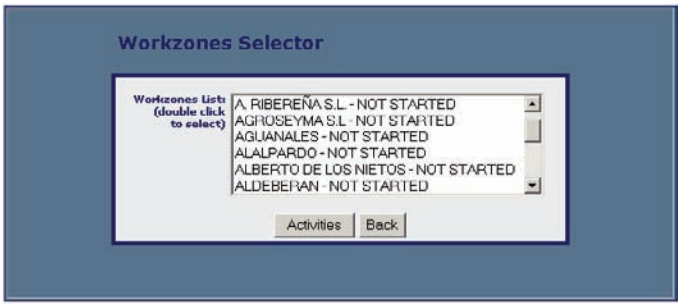


Figure 7. Work-zone selection form



files, and necessary data to perform the campaign tasks.

At any time, the allocations to the workers can be modified as long as the work in the work zone to reassign has not begun yet.

Also, the planning of a campaign could be consulted about at any time.

- **Control of the state of the work:** DAVINCI offers the required functionality to control the advance of the campaign work at any time. To achieve this aim, the work zones assigned to each mobile worker is consulted and, for each one of them, the advance degree is calculated through the information sent by the mobile applications at the completion of each protocol task in every work zone and stored in the DAVINCI internal database.

  In addition, the integration framework allows the placement of the mobile workers in order to control their availability to receive new assignments. According to the accessories installed in the mobile workers' devices and the coverage of wireless communications in the zones where they are, the location processing will be based on GPS (Global Positioning System) positioning algorithms (without cell-phone coverage) or GSM provider positioning services (with cell-phone coverage).

Mobile UI Designer

The main purpose of this component consists of helping mobile-application integrators to design mobile applications using the XForms standard, preparing them for their introduction into the DAVINCI integration framework.

This component has not been developed yet. Temporarily, we are using any XML (extensible markup language) editor that is able to process XForm schema. Concretely, during DAVINCI's deployment in the European veterinarian sector, the editor used has been XMLSpy.

The activities required for this purpose are as follows.

1. Design of the forms of the mobile applications following the XForms standard
2. Creation of the modelmapper file containing the rules of insertion, modification, deletion, and consultation of the information of each form and each data item presented in it

DBSync

This module is in charge of the communication and synchronization of data between the integration framework (DAVINCI) and any other system to connect.

DAVINCI works with a set of generic concepts that are easily adaptable to concrete concepts in external systems. For example, if we are processing work zones, which are the places where mobile workers perform their tasks, the DAVINCI integration system only stores the identifier of the zone and the extended information; the properties of the concrete business area are stored in an external system that is conveniently connected.

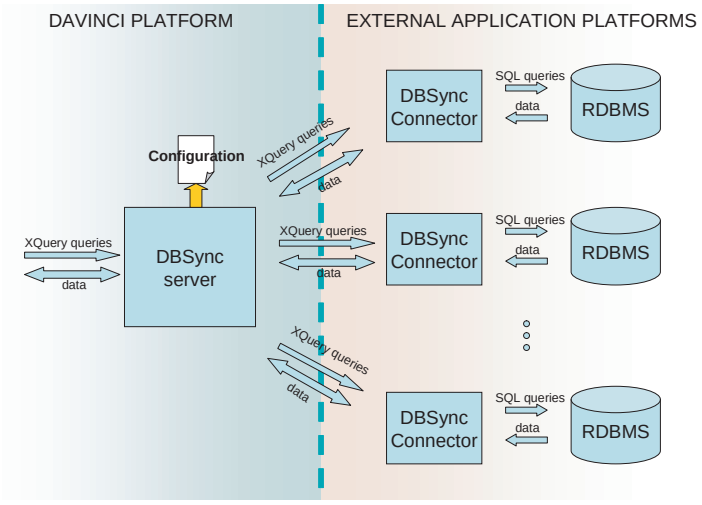
Moreover, DAVINCI applications will work with data stored in external systems and databases for recovery and modification of the business area information.

Next, DBSync's main functionalities are presented in a detailed way.

- **Get extended info:** In this use case, additional information is requested about any of the DAVINCI main subjects.
  - Work zones
  - Geographic areas
  - Campaigns
  - Workers

  The DAVINCI data model holds very little information about these elements in a generic way. Applications may need to show to the user more detailed data just for information purposes. The data obtained by using this use case is not modifiable by DAVINCI middleware, and DAVINCI does not handle them in any way but for showing it to the Web interface user in the suitable place.
- **Query data:** Several DAVINCI middleware components will need to obtain data from the data sources integrated in order to send it to the clients for showing and updating. That data are queried in a generic way by using this use case.

Figure 8. DBSync architecture



- **Update data:** Several DAVINCI middleware components will need to update or insert data in the data sources integrated in the DAVINCI platform. That data usually come from data capturing applications used by workers. A certain format is specified to transfer data from clients through DAVINCI middleware to the original data sources.
  The DBSync architecture is shown in Figure 8.
  DBSync is composed of a server that interacts with the other DAVINCI modules needing data synchronization. The DBSync server uses different DBSync connectors to access different data sources integrated with the DAVINCI platform. The data sources can be located in geographically separated points, and can make use of different relational database managing systems (RDMS). There is no limit on the number of DBSync connectors that can be used by the DBSync server.
- **DBSync server:** The DBSync server obtains the access data needed to reach the connectors from configuration, but the data can also be obtained dynamically.
- **DBSync connectors:** A DBSync connector is in charge of communicating with the DBSync server and performs the required operation regarding data querying and updating. Each connector handles one RDMS and can have its own policies about data updating and handling.

To integrate a new data source into the DAVINCI platform, it is necessary to complete the following steps.

1. *Define the data managing policies that will be used on the data source.*

The data updating and querying policies to be used on each data source are to be defined by the owner of the data source. The DAVINCI platform will provide new data and queries in the format specified above through the DBSync server, however it does not specify the way in which that data are to be updated and the queries are to be performed.

The following aspects are to be considered in order to design the data managing policies to be applied by the DBSync connector.

- Data may be overwriting when new data that has been captured by DAVINCI clients are updated in the data source. The DAVINCI platform does not perform any data storing or auditing, thus the DBSync connector should be able to manage data overrides in a consistent way with the data source. For instance, critical data should not be overwritten by new data if there is not a previous check, and historical data might be kept in order to roll back to previous states of the data source.
- Security procedures might be implemented in order to prevent unauthorized access to data or modifications. This depends on several factors such as the physical location of the data source, the connectivity of it with external threats, or application security requirements.
- There may be the need to dynamically modify the data access policies. If there is that need, configurable mechanisms can be implemented for being able to change data managing policies at run time.

2. *Implement and deploy a new DBSync connector to enable the DBSync server to access to data source.*

A new DAVINCI DBSync connector module must be implemented when a new data source is integrated into the DAVINCI platform.

The connector must expose a Web service in order to communicate with the DBSync server.

The data managing policies used by the connector to access its own data source are not restricted and are up to the particular implementation of each DBSync connector.

The connector Web service must be accessible from the DBSync server through HTTP (hypertext transfer protocol) or HTTPS. It can be deployed at any platform able to fulfill that requirement. There is no restriction about the technologies (programming language, platform operating system, applications server, etc.) used to implement and deploy the connector due to its Web service (SOAP, simple open access protocol) interface with the DBSync server.

An implementation using the same technologies as the rest of the DAVINCI middleware is provided for reference. It is built using the Java programming language and deployable on any J2EE-compliant application server; JBOSS was used during the development stage.

Box 1.

```
- <q:query xmlns:q="http://www.w3.org/XQuery">
  - <q:flwr>
    - <q:for>
      - <q:forAssignment variable="b">
          <q:pathExpr>T_WORKER</q:pathExpr>
        </q:forAssignment>
      - <q:forAssignment variable="a">
          <q:pathExpr>T_DEVICE</q:pathExpr>
        </q:forAssignment>
      </q:for>
    - <q:where>
      - <q:binaryPrefixExpr name="AND">
        - <q:binaryPrefixExpr name="LIKE">
            <q:pathExpr>b/WR_NAME</q:pathExpr>
            <q:constant datatype="xsd:string">Fran</q:constant>
          </q:binaryPrefixExpr>
        - <q:binaryPrefixExpr name="AND">
          - <q:binaryPrefixExpr name="!=">
              <q:pathExpr>b/WR_ID</q:pathExpr>
              <q:constant datatype="xsd:decimal">0</q:constant>
            </q:binaryPrefixExpr>
          - <q:binaryPrefixExpr name="=">
              <q:pathExpr>a/DV_ID</q:pathExpr>
              <q:constant datatype="xsd:decimal">0</q:constant>
            </q:binaryPrefixExpr>
          </q:binaryPrefixExpr>
        </q:binaryPrefixExpr>
      </q:where>
    - <q:return>
        <q:variable name="WR_ID" />
      </q:return>
    </q:flwr>
  </q:query>
```

The DBSync module receives Xqueries in order to collect information from the database. That kind of queries achieves a high-level abstraction over SQL sentences in such a way that the same query could be applied over different types of databases (SQL Server, Oracle, etc.).

FLWR is the subset of the Xquery standard that will be used by DAVINCI modules. This standard defines five operations that could be defined in Xquery.

- *For* creates a variable that represents the table and associates it to a variable.
- *Where* filters the query using data from tables selected in *for* expressions.
- *Return* specifies the node set of the output document with variable references. After a *for-in* expression completes the iteration, *return* delivers the query result document.

There is one more operation offered by the FLWR standard, denominated *let*. However, this operation is not necessary to offer the functionality required by server modules of DAVINCI.

An example of a FLWR expression could be as shown in Box 1.

The result of this FLWR expression through a parser would be this SQL sentence.

SELECT WR_ID FROM T_WORKER b, T_DEVICE a WHERE b.WR_NAME LIKE 'Fran' AND b.WR_ID != 0 AND a.DV_ID = 0

In this case, the FLWR parser has been developed for MS SQL Server Database in such a way that the syntax of the SQL sentence is specific for that database server.

The DBSync module has an implementation of certain capacities of the FLWR standard. Through that implementation, a new connector developed for the DAVINCI platform will be able to transform a FLWR request into an SQL sentence for SQL Server Database.

Fixed Server Components

These components should be installed in a server machine with a connection to the client devices. They provide full capabilities of the integration framework, so they do not have to be modified or enriched in any case.

Campaign Designer

This module encapsulates the functionalities for the management of campaigns and the protocols assigned to them.

Next, the campaign designer's main functionalities are presented in a detailed way.

- **Campaign design:** This functionality consists of one of the following.
  - Selecting an existing campaign to modify its geographic areas and protocols
  - Creating a new campaign and selecting the corresponding geographic areas related to the campaign in which one will work in that campaign.
- **Protocol design:** This functionality consists of one of the following.
  - Selecting an existing protocol, and adding and/or deleting activities. Moreover, for each activity, the application or form to complete is able to be changed
  - Creating a new protocol, defining the general information of each activity and the precedence between the protocol activities, and assigning a mobile application or form to each activity.

Work Planning

Next, the work planning component's main functionalities are presented.

- **Assign and unassign mobile workers to geographic areas:** The first step in the work planning component consists of determining the workers who will perform tasks in each one of the geographic areas assigned to the campaign.
- **Assign and unassign work zones to mobile workers:** Once the workers are assigned to the different geographic areas, this function allows selecting from the mobile workers of a geographic area the concrete person who is going to visit each work zone.
- **Send the assigned tasks:** Once the tasks in the concrete work zones are assigned, the integration framework provides functions to send the assigned tasks to each mobile worker. This information shipment is composed of the mobile applications to run during the tasks in the work zones assigned, the list of the mentioned work zones, and the extended information that is required to run the applications correctly.

MobileUI-Server

This component receives the orders (initiated by the user through the Web interface operations) sent by the campaign designer and work planning components, transforming them into the messages adapted for its shipment to the corresponding component in the mobile worker's device.

Moreover, MobileUI-Server processes the messages of data and requests sent by the corresponding component in the client side of this integration framework.

Next, the MobileUI-Server component's main functionalities are presented.

- **Application shipment:** This operation consists of looking for an application assigned to the worker, assembling the message to notify the client of the need of this concrete application, and sending the application to the client, which is in charge of verifying if this application is installed or not. If the application is not installed, the client will sent a new message asking MobileUI-Server for the files corresponding to the new mobile application to install.
  Moreover, this function collects the required extended data for running the application sent correctly, and sending them to the client side by means of the appropriate messages in order to be stored correctly in the corresponding mobile databases.
- **Forms request:** When a mobile device asks for the installation of a new application, MobileUI-Server sends a message with the description of the XForms documents that compose the required application.
- **Modelmappers request:** The process is similar to the comments for the forms request, being initiated by the same event of the installation of new applications in the client side, but in this case, the files sent correspond to the files that link the form fields to the local database fields.
- **Data request:** The process is similar to the comments for the modelmappers request, being initiated by the same event of the installation of new applications in the client side, but in this case, the information sent corresponds to information items to be inserted in the local database.
- **Data modification:** The purpose of this function consists of the synchronization of already updated information by the server to the client in order to permit the use of undeprecated information by the mobile workers to perform their tasks correctly.

In order to obtain this intention, whenever the server detects the modification of information shared by several mobile devices, a message to the affected devices, with the sentences to substitute the deprecated information with the valid one, is created.

eSignatureServer

This module is responsible for verifying the information signatures generated in the mobile devices to check their validity.

MessageCenter-Server

This module centralizes the communication between the different modules installed in the central integration server, considering that the communications server is the module that, in the server side, represents the mobile workers' devices.

When a server service sends a message to another server component, MessageCenter-Server redirects it considering that the information directed to a mobile device will be redirected to the communications server, which will send it to MessageCenter-Client (the message manager that is continuously running in each mobile device).

In addition, the shipment of messages between modules can be programmed, so in this case, the shipment is executed at a planned moment. In case of nonprogrammed messages, they will be sent as rapidly as possible.

MessageCenter-Server's main functionalities are as follows:

- Send a programmed message
- Send a list of programmed messages
- Send an instant message to other server component
- Send an instant message to a client component
- Start MessageCenter-Server
- Register new service

Communications Server

This module is in charge of managing the communications between the client and server in the server side, handling a queue of messages to send, reconstructing the messages received from the mobile devices, and sending them to Message-Center-Server, which is responsible for redirecting them to the corresponding server service.

The communications server's main functionalities are the following.

- Send a message to a client
- Receive message from client

Fixed Client Components

These components should be installed in each mobile device to be used by a mobile worker. The client components provide full capabilities for using and managing the mobile applications that are in the scope of the DAVINCI integration framework, so they do not have to be modified or enriched in any case.

Communications Client

This module is in charge of managing the communications between the client and server in the client side, handling a queue of messages to send, reconstructing the messages received from the server, and sending them to the message center, which is responsible for redirecting them to the corresponding component.

Next, the communications client's main functionalities are presented.

- **Send a message to the server:** When the message center has any message to send to the server, the mentioned message is introduced in a queue for its later shipment.
  The shipment will be made at any moment depending on the availability of the communication channels of the device (Bluetooth, GPRS, wired connection, etc.) and on the priority of the message to send.
  Internally, the message queue is stored in a database to avoid losses produced as a consequence of possible falls of the system.

- **Receive message from the server:** When the module of communications has received a message from the server, the messenger center is notified of this circumstance and is responsible for redirecting the message to the corresponding component.

MessageCenter-Client

This module centralizes the communication between the different modules installed in the mobile device, considering that the communications client is the module that, in the client side, represents the central integration server.

When a client component sends a message to another client component, the MessageCenter-Client redirects it considering that the information directed to the server will be redirected to the communications client, which will send it to MessageCenter-Server (the message manager in the server part).

In addition, the shipment of messages between modules can be programmed, so in this case, the shipment is executed at a planned moment. In case of nonprogrammed messages, they will be sent as rapidly as possible depending on the amount of messages in the queue and the state of communications channels.

MobileUI-Client

The MobileUI-Client component is formed by five differentiated subcomponents.

- Forms Viewer (XFormster)
- Forms Server (MobileWebServer)
- Forms Processor (XFormsModule)
- Data Access Control (MuiData)
- Applications Controller (MuiDataManager)

The dynamic collaboration between these subcomponents, with the rest of the modules of the side client, is shown in Figure 9.

The interaction begins when the DAVINCI server sends to the mobile device the applications to be installed. As it has been described previously, the messages interchanged among a concrete client and servers are provided by the communications client that is one of the services handled by the message center, as it is shown in Figure 10.

When a message of a new application is received, ClientMessageCenter redirects it to MuiDataManager, which is in charge of managing the installations of new applications, the downloading of data from the server, and the modification of the updated data to the server. The forms (written in XForms) are stored in a specific repository, the data mapping files (modelmappers) are stored in another separated repository, and finally the data are stored in a local database.

For the management of the applications' data, an independent module called MuiData, which encapsulates the logic to manage data sources, is used, so it is possible to change easily the format of the data sources to another format (i.e., text files, XML files, etc.); replacing the MuiData component with another makes it able to process the new format.

On the other side, the mobile worker, when using DAVINCI mobile applications, interacts with the forms browser, called XFormster, to navigate between the forms that compose the mobile application. This navigation implies interaction with a forms server, called MobileWebServer, which is in charge of process the forms; this is illustrated in Figure 11.

The mobile Web server works with a set of APIs that processes different types of requests created and initiated by Web pages. In this case, the corresponding API to process requests of XForm forms will be the XFormsModule. This module is in charge of gathering the data sent from the Web form to the Web server, searching the modelmapper file corresponding to the mentioned form, and updating the data in the local

Figure 9. MobileUI-Client collaboration diagram

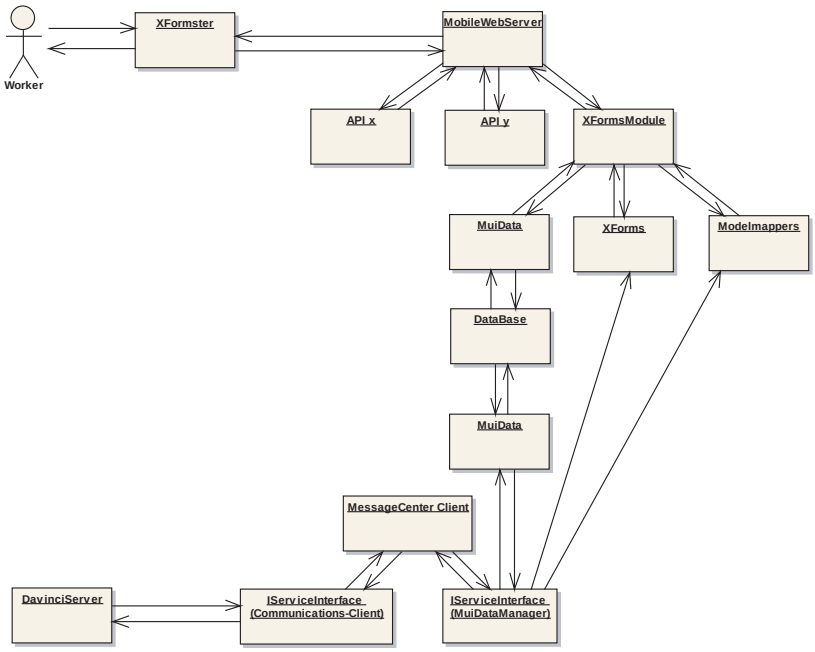
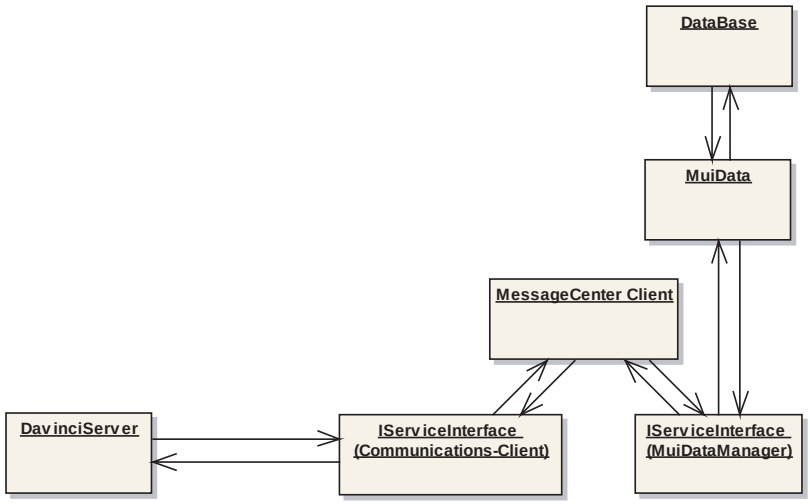


Figure 10. MuiData collaboration diagram



database according to the rules specified in the modelmapper file.

Moreover, XFormsModule is also in charge of searching for the form that should be presented to the user as a consequence of any command processes being used for this purpose, displaying the new information according to the rules specified in the modelmapper file of the new form to show.

The forms browser has the user interface displayed in Figure 12.

Navigation between the different forms will depend on the specific design and the purpose of each mobile application implemented.

eSignatureClient

This module is in charge of signing electronically the data generated by the mobile worker during the protocol application in a work zone and sending the signature to the central server for its verification and storage.

Figure 12. XFormster user interface

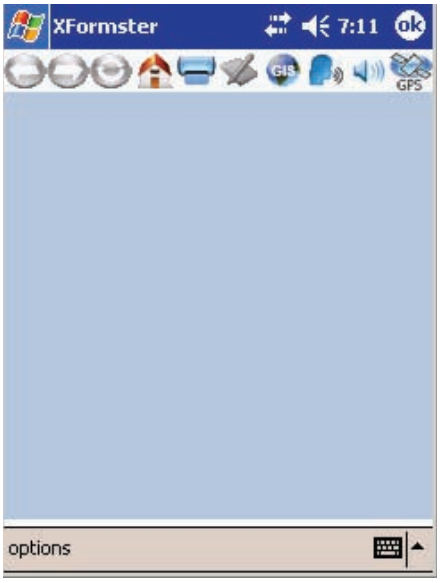
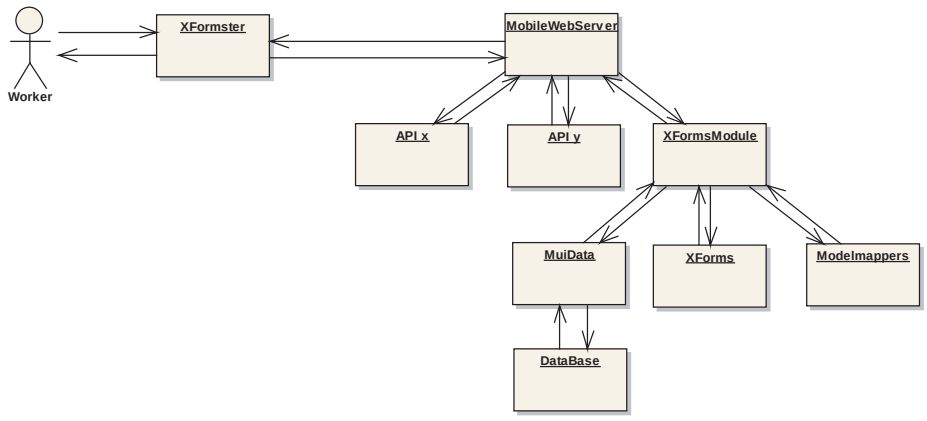


Figure 11. XFormster and MobileWebServer collaboration diagram



VoiceUI

The VoiceUI component is in charge of providing the capability to handle XFormster forms by means of voice processing; that is to say, using the VoiceUI component, mobile workers are be able to handle the DAVINCI mobile forms using commands introduced by speaking.

The voice processing interface allows the user to introduce values for form fields and execute the commands provided by command controls provided by the form, even the activation of the main functionalities offered by XFormster.

Print Manager

This module allows printing of any of the forms shown by the XFormster forms browser of the MobileUI-Client component. The print manager uses the Bluetooth port to communicate with the printer in order to send the printing jobs.

When the user decides to print through XFormster, the print manager extracts the information of the labels and fields to be printed, and then prints and transfers them to a flat text document.

Mobile Geographic Information Viewer (Mobile GIS)

This module allows the visualization of cartographic images and the presentation of information relative to parcels or enclosures defined on the shown cartographies.

Mobile GIS (Geographical Information System) is launched from the browser of MobileUI-Client's XFormster, but due to Mobile GIS's complexity and the size of the screen of some pocket devices, the cartographic information appears in a new window in order to not saturate the XFormster window; otherwise, the window to visualize GIS information would be totally unmanageable.

Positioning Component (LBS)

This component is responsible for obtaining and handling the geographic position, in terms of coordinates, of DAVINCI mobile workers. This positioning component allows one to handle geographical, cartographic, and UTM coordinates, and to work with WGS84 and ED50 data.

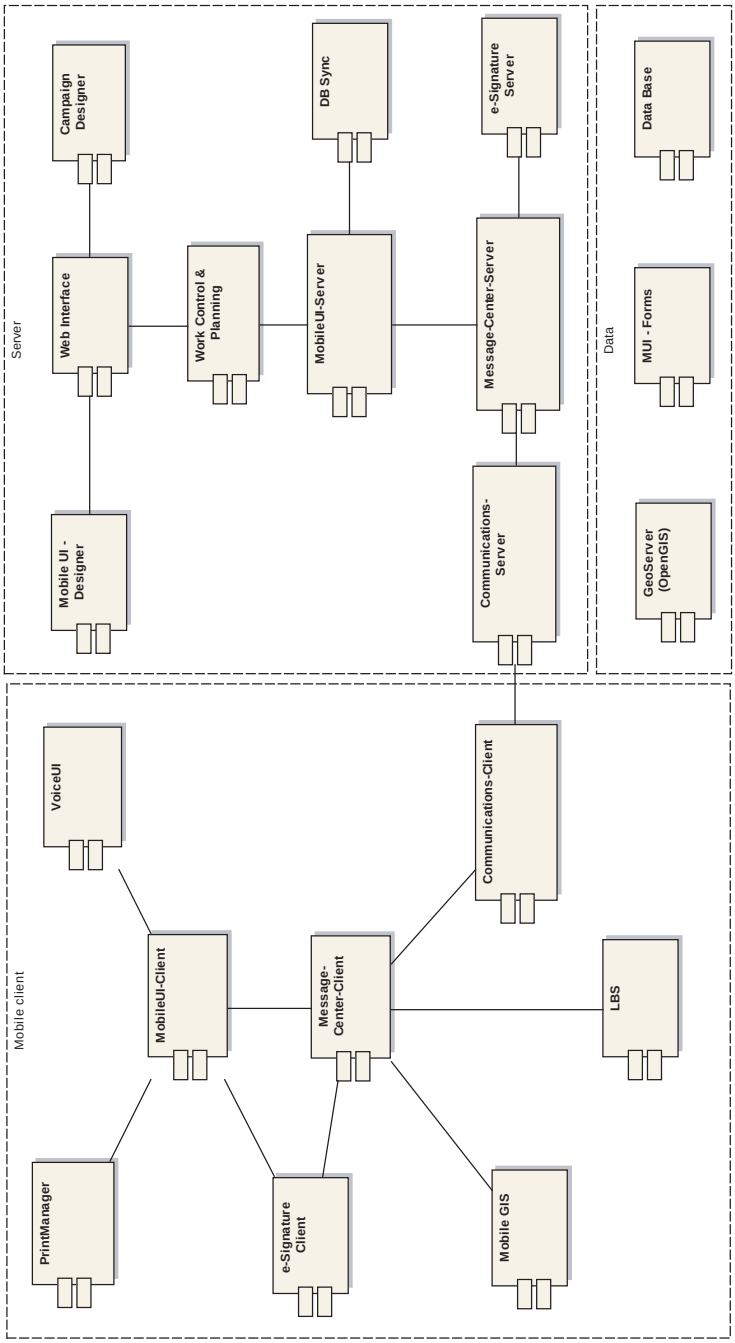
A detailed diagram of DAVINCI integration components is shown in Figure 13 using UML (unified modeling language) syntax.

STEPS TO USE THE INTEGRATION FRAMEWORK

If an organization wants to use the DAVINCI integration framework for the management of the applications used by its mobile workers, it is necessary to perform the following tasks.

1. Install and configure the server components of the integration framework in a server able to be connected to client devices with the configuration. The server machine should have installed any J2EE-compliant application server and an RDMS. Our trials were done using JBOSS and Microsoft SQL Server (this database only manages the internal data required by the integration framework).
2. Install and configure the client components in the devices that will be used by the mobile workers. The configuration is different depending on the mobile device's capabilities. This is the most complicated task because there are several exceptions and special cases related to each device vendor and model. The effort of this task could be reduced by using the same model of mobile device.
3. Define the data managing policies that will be used on the data source used to obtain

Figure 13. Integration framework architecture



and update the data managed by the mobile applications.

4. For each mobile application to deploy, it is necessary to do the following.
   - Define, design, and implement the forms of the mobile applications using the XForms standard, preparing them for their introduction into the DAVINCI integration framework. This task could be done with any XML editor that can process XForms, for example, XMLSpy.
   - Define a modelmapper that contains the information of the graphical control used to show any item of the information, and the rules to query and update the mentioned data.
   - Implement and deploy a new DBSync Connector to enable DBSync Server to access the data source.
5. Design each company campaign using the Web interface of the integration framework.
6. Define the protocols that are assigned to the campaigns, and assign a form to each protocol activity.
7. Plan the work of each campaign, assigning the job to any available mobile worker.

Then, the applications and the required data are automatically sent to the mobile workers.

As each mobile worker performs his or her job, the integration framework automatically updates the information in the data source in accordance to the data managing policies established in Step 3 of this section.

When a mobile worker has new assignments, the integration framework sends the new data and, if required, a new mobile application to enable the worker to perform the job.

SYSTEMS INTEGRATION IN EUROPEAN VETERINARIAN SECTOR USING DAVINCI

Trials Purpose

The objectives that should be satisfied by means of the realization of DAVINCI trials and technology transfer activities consist of validating the improvements and technological innovations using DAVINCI in real working environments.

The main aspects validated were as follows:

- Capability of integrating the data provided by DAVINCI with Spanish and Bulgarian cattle exploitations and livestock-movement databases. This objective allows checking the effectiveness of DAVINCI for integrating information systems of different nature.
- Reduction of costs and effort spent in relation to the new mobile applications development.
- This validation was carried out by gathering related data on the necessary time to develop mobile applications to help veterinarians perform information retrieval, and updating tasks related to the cattle sanitarian control program.
- Improvement of the ergonomics and veterinarians' comfort when performing the activities related to sanitary controls and clinical inspections. To fulfill this purpose, the effectiveness, efficiency, and ergonomic conditions for those carrying out the inspections were analyzed in detail. This analysis was performed by the veterinarians completing some evaluation questionnaires about the subjective estimation of these aspects.

Trials Scope

Trials have been performed in four European countries (Spain, Latvia, Czech Republic, and

Bulgaria) during the experimentation activities of the European-Commission-funded project called Advanced Management Services for Inspections and Sanitary Interventions in Cattle Holdings through Mobile Technologies (IPS-2001-42057).

The most efficient strategy for achieving the trial objectives was performing separate trials and technology transfer activities in Bulgaria, Spain, Latvia, and Czech Republic.

The reason for this separated approach is based on the existing differences between the infrastructures already present in each country.

In Spain, there is an information system for cattle-exploitations registering and livestock movements managed by each autonomous community and coordinated by the central government. The databases in the system satisfy the requirements in Directive Number 1760/2000 of the European Parliament. Also, in Spain there are standard GISs that are accessible by the project, so we will allow the checking of all of the project's required functionalities.

In Bulgaria, although there are databases for cattle exploitation and livestock movements, they are not standardized according to Directive Number 1760/2000 of the European Parliament. On the other hand, DAVINCI cannot use the standardized systems of the Bulgarian GIS to satisfactorily prove the GIS capacities provided by the project.

Trials in Spain

A voice recognition system was tested in a small ruminant bleeding activity because the time saved with small ruminants is by far greater than the time estimated for large ruminants. The voice recognition system will be useful for some veterinary activities on the ground, but experience shows that the time saved in a bovine bleeding job using this technology is not as profitable as in small ruminant practices. For this reason, sheep control programs were selected for this trial.

The bleeding of large animals will be done using pocket PCs. There is a reason for this choice. In the case of large animals, more activities have to be performed and the use of a pocket PC, with all the functionality that it is able to provide, facilitates the job for these activities. The objective of this demonstration is not only the use of those devices on bleeding activities, but also to become conscious of the limits and capabilities of the use of technological devices on the ground in an unclean environment.

Figure 14. Trials execution in Spain

Figure 15. Trials execution in Czech Republic

Figure 16. Trials execution in Bulgaria

Trials in Bulgaria, Czech Republic, and Latvia

Bulgaria believes that the approach to separate the demonstration from the prototypes is reasonable for a better evaluation of each one of them.

The use of a voice recognition system, no matter whether being used on small or large ruminants, is of significant interest for us, but due to some limitations on the languages that can be used in such a system, the demonstration of the technology could be limited to Bulgarian project coordinators reviewing the Spanish prototype.

Although the mobile mapping system was tested in Spanish trials, we believe that this component will be of significant importance for us during the monitoring of culicids vectors on bluetongue disease, a disease with great epidemiological significance for Bulgaria and the region. Along with that, the application of restricting measures could be tested for a geographical area in case of a disease outbreak. All those activities could be done in a lab with all the data that we have—digital maps, data from culicid traps, and

so forth—and we believe they are going to be sufficient for testing a prototype for the mobile mapping system.

Trials Activities

In order for the testing routines to be able to provide a comparable set of conclusions, a common process for the trial development has been established by DAVINCI. The necessary activities to meet the established trial objectives and the desired technology transference are the following.

1. Definition of the trials scope
2. Determination of the region and holdings to be visited during the trials, looking for concrete parameters
3. Selection of the participants in the prototype testing activity
4. Configuration of the mobile devices' hardware
5. Preparation of trial evaluation materials
6. Performance of the cattle sanitary control activities
7. Trials evaluation

CONCLUSION AND FUTURE WORK

The main success factor of DAVINCI does not lie in obtaining a system that provides a highly optimized performance, but in the automation of manual procedures, achieving acceptable performance. Moreover, it has been detected that the performance displayed is not optimal, but since it copes with minimal required performance, the fundamental objectives of the project are fulfilled.

As a summary, it may be said that the results obtained from the trials allow us to state that the main improvements brought about by DAVINCI are the following.

- The effort necessary to develop mobile applications to gather and/or consult livestock sanitary information has been reduced by up to 43.43%. The absolute value (in working hours) of this reduction will depend on the size (measured in function points) of the intended application.
  Nonetheless, it ought to be taken into account that the success mentioned has been produced in a limited context of a quite restricted set of mobile applications based on simple formularies, without complex displaying elements such as dynamic lists or tables, with simple access to databases not joining different data sources.
  Thus, this effort reduction is so far applicable to the development of simple mobile applications in terms of the insertion, modification, and consultation of data coming from a single data source.
- The economic costs needed to integrate mobile applications with existing systems and databases for livestock sanitary information management has been reduced by 9.03%.
- The average time to publish the data relative to a cattle holding has been reduced by up to 2.6 days, decreasing from 64.84 hours in the previous situation to 1.14 hours when using DAVINCI.
- DAVINCI easily adapts to the field working proceedings of different countries, with different cultures and organizational structures.
- The DAVINCI framework should adapt easily to technology changes in communications or to new mobile devices being used to perform fieldwork.
- The performance and easiness of mobile device configuration is acceptable, allowing the veterinarian to effectively carry out the field tasks.
- The communication components of DAVINCI are adaptable to several different com-

munication infrastructures, allowing one to effectively resume interrupted messaging due to uneven communication coverage. Thus, the information transmitted through the available bandwidth is optimized.

However, it is necessary to perform, in the immediate future, an optimization process of the obtained solution, introducing, among others, the following improvements.

- To extend the set of available controls to be used by DAVINCI mobile applications, allowing the usage of more complex formularies with complex information, displaying elements such as dynamic lists and tables, and allowing simultaneous access to different data sources (joins)
- To carry out, keeping in mind the improvements of the available controls for mobile applications, a larger set of studies to assess the reduction of effort and time needed to develop mobile applications within DAVINCI
- To study and define algorithms allowing the broadening of coverage time, and thus lowering the number of messages that have to be resumed due to loss of communication with the coverage currently offered in rural environments by communication providers
- To improve the ergonomics and ease of use of the central components for in-field resource management
- To improve the ergonomics and ease of use of the mobile applications developed under DAVINCI, paying attention to those aspects related to input data validation and helping with the completion of less common tasks
- To provide the veterinarian with a mobile device that allows her or him to perform the fieldwork without the need to undergo configuration changes, such as battery changes, during the activities
- To improve the performance of DAVINCI's voice recognition system so that it consumes less resources and is less sensitive to noise while not lowering the speed for voice processing

REFERENCES

Application servers. (2006). *MobileInfo.com*. Retrieved April 17, 2006, from http://www.mobileinfo.com/application_servers.htm

Baumgarten, U. (2004). *Mobile distributed systems.* John Wiley & Sons.

Boag, S., Chamberlin, D., Fernández, M. F., Florescu, D., Robie, J., & Siméon, J. (2005). XQuery 1.0: An XML query language. *W3C candidate recommendation.* Retrieved November 3, 2005, from http://www.w3.org/TR/2005/CR-xquery-20051103/

Boar, C. (2003). *XML Web services in the organization.* Microsoft Press.

Boyer, J., Landwehr, D., Merrick, R., Raman, T. V., Dubinko, M., & Klotz, L. (2006). XForms 1.0 (2nd ed.). *W3C recommendation.* Retrieved March 14, 2006, from http://www.w3.org/TR/xforms/

Intel Corporation. (2006). *Intel® mobile application architecture guide*. Retrieved April 17, 2006, from http://www.intel.com/cd/ids/developer/asmo-na/eng/61193.htm

Lee, V., Schell, R., & Scheneider, H. (2004). *Mobile applications: Architecture, design, and development.* Hewlett-Packard Professional Books.

Longueuil, D. (2003). *Wireless messaging demystified: SMS, EMS, MMS, IM, and others.* McGraw-Hill.

Mallick, M. (2003). *Mobile and wireless design essentials.* Wiley Publishing Inc.

Schiller, J. (2000). *Mobile communications.* Addison Wesley.

Siegal, J. (2002). *Mobile: The art of portable architecture*. Princeton Architectural Press.

This work was previously published in Enterprise Architecture and Integration: Methods, Implementation and Technologies, edited by W. Lam and V. Shankararaman, pp. 212-224, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 2.29
"It's the Mobility, Stupid": Designing Mobile Government

Klas Roggenkamp
Dipl. Designer Electronic Business, Germany

ABSTRACT

This chapter looks at mobility and the term government to describe influencing factors for the process of designing mobile government. A detailed review of perspectives on mobility and a subsequent examination of the government term is given to reach a better understanding of what mobile government can be. Furthermore, four questions are refined which are aimed at helping to first justify and assess a possible m-government service and second to classify this service in a given context. By focusing on mobility as a key component of mobile government, the author hopes to aid developers and researchers alike with designing new and better mobile services within the public sector.

INTRODUCTION

Mobile government as a subject of debate, research and actual services is gaining momentum within the field of electronic government. The number of mobile phone users is exceeding the number of fixed line phones and, though at a slowing pace, is still growing. With the recent roll-out of mobile broadband data services such as UMTS or Wireless LAN, one gets the idea that we are still just looking at the tip of an iceberg.

So far, e-government has allowed for a faster, more convenient and often value-added delivery of public services. It has started a regrouping and rethinking of processes in many administrations, helped to create a new and improved access to various services, and also supported citizen participation in political processes. Even where it is not obvious to "front-end users" like citizens, e-government has in many cases boosted the more efficient gathering and processing of data. Information and communication technologies in governmental organizations have reduced cost, redundancies, and errors, thus speeding up the handling of services.

Nevertheless, mobile technologies, however unknown their real value still is, will take this development to an even higher level. Not only are such mobile services promising more efficiency,

faster and less erroneous processing of data, but also an improvement of service as a whole through direct contact with citizens. "As painful as e-government transformations have been, the challenges of dealing with an always-on society and workforce will be even more daunting. Service delivery, democracy, governance and law enforcement will all be affected" (Di Maio, 2002).

Mobile government by itself can still be considered in its "infancy" (Zálešák, 2003)—when it comes to governmental organizations we can assume a "transitive state" (Kushchu & Borucki, 2004, p. 830). The services currently considered to deserve the label mobile government range from W-LAN in public buildings to stand-alone mobile applications. They depend on, utilize, or incorporate features of mobile technologies. In-between these two extremes, we can observe a vast variety of services such as mobile information via SMS, mobile tickets for parking or trains, and alike. Hence, it is rather unclear what we talk about when we actually refer to mobile government.

The definitions given in the literature vary slightly. Some describe m-government as "a functional subset of all-inclusive e-government" (Arazyan, 2002) respectively saying, that technologies used for m-government "are limited to mobile and/or wireless technologies" (Llalana, 2004) in comparison with e-government. To others, m-government is "a complex strategy for efficient utilization of all wireless devices" (Zálešák, 2003) with the goal of "improving benefits to the parties involved in e-government" (Kushchu & Kuscu, 2003).

Defining m-government as a form of mobile business, some see it as a connection between Internet and mobile communications offering context-dependent and highly individualized information and not huge amounts of data (cf. Frischmuth & Karrlein, 2002, p. 15). By ruling out mobile yet stationary interaction as well as random wireless connections, m-government is considered restricted to public mobile services that need time-critical information access (cf. Thome, 2003).

If an application fits one of these definitions or if it falls short of certain features is a debate of its own and shall not be the issue here. Whether one talks about mobile government, mobile e-government, or mobile public services is a semantic issue. The process of planning, developing, designing these service offerings is, however, important. In this process, several sets of interests need to be aligned to allow for a coherent offering in the first place.

We will look at mobility, government, and the scope of connotations to each term with the aim of describing factors to be considered in the design process of such a mobile service in the public sector. Requirements toward mobile services from both sides will be detailed and combined.

As a first step, the following second part will review the term mobility from a technological, economical, and sociological perspective. As a result, a set of questions will be derived to pinpoint issues to consider in the context of being mobile.

To gain a more comprehensive understanding of the involved parties, the third part will subsequently examine the government term. By separately looking at the organization, the actors, and the processes, the scope of this term becomes clear. To what extent these subsets and their goals can be combined will be discussed alongside a brief description of important challenges.

Following these separate considerations, part four will discuss the interaction of mobility and government and thus give an overview of factors to consider when thinking about mobile government. This will be concluded with an outlook into future developments.

MOBILITY PERSPECTIVES

Mobility in a general sense is understood as a form of *being mobile*. The adjective "mobile"

goes back to the Latin word "mobilis", meaning *movable*. In this sense, mobile objects are capable of moving or being moved.

Depending on the context, in which the term is used, its implication is variously extended. These differences lead to distinct approaches how to deal with mobility, how to become mobile, or how to support being mobile. For example, the question "Are you mobile?" does not ask for the status of someone being physically movable but for the subjects *ability to move* from one place to another in a more social context. Maybe the asked person has obligations that do not allow leaving ("I'm stuck here"), thus making him immobile, at least for a certain amount of time. It could also be asked, whether someone is mobile in a sense of *willing to move*.

The title of this article states the importance of understanding the concepts of mobility how to properly deal with them. Obviously mobility is one of the key features of mobile government of whatever kind, depending on the starting point of planning such a service, the issues considered as key problems are rather different. Between the fields of information science, economics, and sociology, we can observe very distinctive perspectives on mobility (see Hess et al., 2005; Kakihara, 2003; Urry, 2000). As a result, each comes to its own conclusions which are often not easy to align. From the technological viewpoint, dealing with mobility is primarily concerning shared computing and distributing data. Even though this also includes allowing end users to move more or less freely, this is dealt with in a very different manner compared to economical or sociological debate. Vice versa questions asked by social sciences on for example the reason to be mobile at all, or the needs while being mobile are constrained to this field. Yet again, financial aspects of mobility can hardly be found, neither in the sociological nor the technological handling of mobile topics.

Technological Perspective on Mobility

The current debate on mobile government (and more generally mobile services) on the one hand and a supposedly mobile society on the other is in part due to the availability of certain mobile technologies. No matter if it is mobile hardware (such as a PDA or a mobile phone) or mobile networks and protocols to use them (such as GSM, UMTS, W-LAN), we have to consider certain issues alongside mobility that enable us to actually create and use mobile services.

The underlying technologies are dealing with four basic concepts of mobility, of which at least three are of primary concern (Hasan, Jähnert, Zander, & Stiller, 2001):

- device mobility;
- user mobility;
- service mobility; and
- session mobility.

Device mobility deals with the continued access to services while being spatially mobile, that is, moving from one physical location to another. This access can be granted via locally limited Wireless LAN access points. Other standards include concepts of handing over connections between access points, as is the case for mobile phone networks based on standards such as GSM. The reach of the device and the general possibility to roam a broader area thus depends on available networks, and of course on the hardware itself.

Assuming that a user is mobile without physical constraints, *user mobility* from this perspective refers to location- and device-independent service access. Pre-requisite is an appropriate means of identification. Again, a common example would be the mobile phone network which is utilizing the subscriber identification module (SIM) to identify a user within the network.

With *service mobility* comes the idea of access anytime, anywhere. More appropriately one should add "anyhow" to this often cited paradigm since this concept includes the idea of service delivery regardless of device and user specific settings (see Perry, O'Hara, Sellen, Brown, & Harper, 2001). Implementations allowing for service mobility are currently hard to find—neither GSM nor UMTS or W-LAN includes the capability to provide a certain service irrespective of device and user. There are, however, approaches to offering seamless services, which automatically suit themselves (e.g., to available bandwidth of a network (FOKUS, 2000)).

The *session mobility* describes the capability of starting, pausing, and resuming a user session while switching between devices and/or services. The session itself is considered as a relation between distributed service components which integrates needed resources. None of the available communication systems allows for session mobility, yet (Hess et al., 2005).

Current mobile communication technologies particularly allow for the mobility of device and user and thus enable the growth in this field. On the path from mobile to ubiquitous computing, however, also the latter two concepts need to be fully included. As for now, some services (also in the field of government) are mobile when considering scalable and adapting front-end interfaces as a means of serving this goal. Nevertheless, offering truly mobile services in this sense can only be attained by solving technical and most of all security issues. To what extent session mobility really is an issue for mobile government has to be questioned. Secure and reliable service delivery—independent of respectively adaptive to an available device and the available network type and bandwidth—surely is of concern in the case of more complex transactional services.

Economical Perspective on Mobility

Whereas the previously discussed technological perspective on mobility takes a device- and service-based approach, the economical perspective can be described as focusing on the intersections between a business process and mobility. Within a typical business process, there are three steps to consider, each of them open for a different perspective of including mobile aspects:

- the value chain;
- market transactions; and
- mobile goods.

Looking at the influence of mobility within *the value chain* the question is raised how mobile services can contribute to the process of value creation. Mobility can affect the efficiency and effectiveness, embracing mobile technologies and thus directly offering support to employees working in a mobile setting can create added value. Typical examples can be found in the areas of field work (Rossado-Schlosser & Hacke, 2002); in a government setting this would include police forces (Bazijanec & Pousttchi, 2004). Apart from directly supporting a mobile workforce, the implementation of mobile machine-to-machine communication (e.g., radio frequency identification (RFID)) can allow for the substitution of certain human-bound sub-processes by directly connecting objects with their environment respectively the organizational information processes (Hess et al., 2005). When properly implemented, both approaches can reduce operational cost and errors while possibly speeding up processes as a whole.

As part of *market transactions*, that is, the exchange between a business (or government) organization and a buyer respectively customer (or citizen) mobility becomes an issue as part of hindering, complicating, or yet demanding mo-

bility (Khodawandi, Pousttchi, & Winnewisser, 2003). Within this setting, mobility can lead to an increase in transactional cost. Through the implementation of mobile communication, these costs can be reduced in the same way as is true for the internal value creation (cf. Kaspar & Hagenhoff, 2003). Striking examples for the effective use of mobile services in this field are, among others, services offering critical information prior to the initiation of a transaction. These may be just accessible via mobile network and device or actually allow for location-dependent information. After this initiation phase the actual transaction can be supported within a mobile context, for example, by enabling mobile payments. Following the transaction mobile services can allow for specialized CRM-methods, either simply by addressing a customer directly or by allowing for a mobile feedback channel.

Furthermore, mobility becomes an issue when dealing with or creating *mobile goods*. In the context of this chapter, this shall only span mobile information goods. Here we can observe a vast variety of goods which are merely a piece of software being delivered to mobile devices, anyhow very successful: ring tones, mobile games, and other mobile entertainment services. Considerably more complex and less intrusively marketed, another field for mobile services as a good of its own is offering adaptive, context- or location-aware information, for example, tourist guides, maps, or real-time information on public transport (see Turowski & Pousttchi, 2003, p. 181).

From this focus on business processes, one can identify general economical issues connected with mobility and ways to theoretically surpass obstacles due to it. In some cases, this perspective can actually help to identify possible new markets connected with or created by mobility, mobile services, and mobile users.

However, assuming that services will be used as planned in up-front business would mean ignoring user behavior. Not only is it inevitable to convey the gains to the prospective user. As a matter of fact, these gains need to be identified. A simple monetary equation, however promising, simply falls too short. This is all the more valid in a government setting where there are not only competing access channels or behavioral obstacles within an organization, investments have to be legitimized to a critical public and in the context of budgets constraints.

Sociological Perspective on Mobility

The *sociological perspective* on mobility is dealing with the description of mobile contexts and needs to be connected with these situations. Concerning mobile communications, we can add the more explicit question about reasons to use a certain service.

On a general level, there is a distinction made within this perspective as to what kind of mobility is applicable or rather, what types of mobility can be found. We can distinct three types:

- physical mobility;
- social mobility; and
- virtual mobility.

"Perhaps the most widely adopted usage of mobility is that of people in terms of geographical movement" (Kakihara, 2003, p. 39). This type of mobility is to be called *physical mobility*, meaning going from one place to another. To more closely define this mobility, it shall be characterized by the mode of movement respectively the overcome distance (Kristofersen & Ljungberg, 2000). Local mobility best describes "wandering" within a building or a local area. "Visiting" one place and then moving to another location extends the locally bound movement. The opposite of wandering in this sense is "traveling", describing a state of moving from one place to another by using a vehicle

of some kind. Gerstheimer and Lupp (2001) also consider the separation of mobile and fixed locations, the first matching the mode of "traveling", the latter the modes of "wandering" and "visiting" (p. 67). In general, physical mobility serves a certain purpose connected to getting from A to B. As a consequence, the process of moving is of primary concern for the moving subject, rendering other side-activities secondary. Where there are slots of attention for such activities, tools and services involved need to adapt themselves or the result of their involvement to a given setting and needs thereof. Traveling by car does not allow for interaction such as typing on a keyboard or visualizing complex content. However, voice interaction might be as well a solution as simplified displays within the sight of the driver.

Social mobility on a macro-level describes the permeability of a society between pre-defined societal levels. On a micro-level, the term refers to the ability of an individual to change roles in reaction to external influences and contexts (see Goffmann, 1967; Ling, 2000). Each physical movement leads to a change of context, if just related to surrounding people (e.g., being with friends as opposed to being with colleagues). A different social context is also attributed to physical location itself—the most general distinction being that of public and private places (Agre, 2001). To visualize this, one might think of a youth among his friends or with his parents. Assumed and fulfilled roles vary in the same way as we would for instance imagine when riding the bus or spending the evening in a comfy chair at home. With the social role played, a behavioral change can be observed, concerning language, volume of speech, even posture or clothing can be implicitly affected.

Resulting from networked and more often also location-independent communications we were able to experience the rise of what Castells (2000) calls a "network society". This new sphere is untying physical location and the range of activity of involved people, along with all due implications. This so called *virtual mobility* has a broad impact on many areas, leading to new behavioral patterns and expectations. On top of the virtuality offered by Internet communication as a whole, mobility extends the grasp of this even further. The consequences of this have been described by many authors (among others Agre, 2001; Geser, 2003; Ling & Yttri, 1999).

As a consequence of virtual mobility—especially in conjunction with mobile communications—it can be observed how, on the one hand, the social context is being more detached from the physical one. On the basis of a mobile communication channel, a mutual virtual but private space can develop between two communication partners, regardless of the fact that they might be physically located in the middle of a crowded public space.

On the other hand, just due to the option of such a virtual presence, the perception of actual physical co-presence emerges. Ling (2000) has shown how this perception develops within a youth group whose members are not all in the same location at the same time, however the absentees are perceived to be part of the actual group activity because they can be contacted wherever they are. Similar observations have been made concerning mobile workforce (Kakihara, 2003; Vincent & Haddon, 2003). Pica and Kakihara (2003) talk about a "duality of mobility". Interpersonal ties may be weakening due to the ease of having ephemeral contact which can be fit into idle times throughout a day. The same opportunities may result in more persistent communication relations for the effects of virtual mobility stated earlier, namely the perception of virtual co-presence.

Nevertheless, especially the sociological perspective leads to the conclusion that many of the topics connected to mobility should be considered as constantly evolving. Certainly, norms influencing and influenced by mobile user behavior are steadily changing. We just need to consider the general perception of using a mobile phone. A few years ago, public use of a mobile

phone was considered snobbish. These days the opposite is true: not having a mobile phone is often considered to be awkward. Usage of voice services (i.e., telephony) and other mobile services (i.e., data) is also subject to so called social shaping—not only the patterns of usage but also the perception of the use, the user, and services as a whole keep developing (see Palen & Salzman, 2002). Hence, it is rather short-sighted to plan and design services merely from the technological or economical perspective.

Bringing these three mobility perspectives together reveals a set of questions which cannot be answered by solely taking one isolated perspective into consideration.

The main goals of the different perspectives (Figure 1) are the creation of technological solutions for mobility, the economical assessment of possible business processes in regard to, above all, efficiency and value creation as well as the reflection on social preconditions and interdependencies of technology and everyday life and work.

At the intersection of technology and economy, the key questions to ask would be:

- Which are potential added values?
- (How) Can the range of old products be extended?
- What are possible new goods?

The combined perspective of economy and sociology lead to questions in the manner of:

- What are current user needs?
- How can these needs be transformed into accepted products?

When connecting sociology and technology, questions arising are:

- How technologically feasible are user wishes/needs?
- What are the interdependencies between technological solutions and usage?
- How can technology adapt to these needs/ interdependencies?

In sum, the result would be a mobility cycle with which the goal to make sense of mobility

Figure 1. Main research goals and open questions of different perspectives (Hess et al., 2005)

Perspective	Goal	Open Questions
Technological	• Development and improvement of new technologies and applications	• Business models and user behavior
Economical	• Services and business models to support business processes and development and roll-out of mobile goods	• Technologies to translate and implement business models • Forecasting user acceptance and adoption
Sociological	• Social pre-requisites for the adoption of new technologies and applications • Implications of new technologies for society	• Creation and designing new technologies and application • Economical strategies, business models, and value chains

as a feature of mobile communications can be attained. In this cycle, the possibilities of technology will be shaped into applications and services appealing to users and also offering added value to providers.

However simplified and generalized the previously-made statements are, they do show possible ways to combine the three different perspectives on mobility in general and on dealing with mobile communications in particular.

GOVERNMENT PERSPECTIVES

After having brought forth a more detailed image of mobility and dealing with mobile communications, this chapter will approach the term "government" in a similar manner. By taking into consideration the particular connotation of this term, we will derive key aspects of developing services for mobile government. Even when it seems evident what might be mobility issues, it is yet unclear whom this shall serve, who and what is crucial within the field of government when planning a certain service.

From a semantic viewpoint, government as a collective term consists of four main elements:

- the (political) system,
- organizations within the government system,
- the processes of governing,
- the actors constituting government.

In Figure 2, showing the relation between the elements, we see that the *system* is the framework in which people being administered by "a" government are located, as well as the governing organizations. Since the system itself has a rather normative character, it is of less importance when dealing with mobile government.

More important is the governing element, the *organizations* performing within the superior system, compliant to given values and norms as well as pursuing and enforcing them. The distinction between a government and administrative branch is merely functional; from the current point of view, this distinction shall also be neglected. Of interest is the single organization as a whole, supposedly acting as one instance in relation to other organizations and preliminary elements (citizen, businesses).

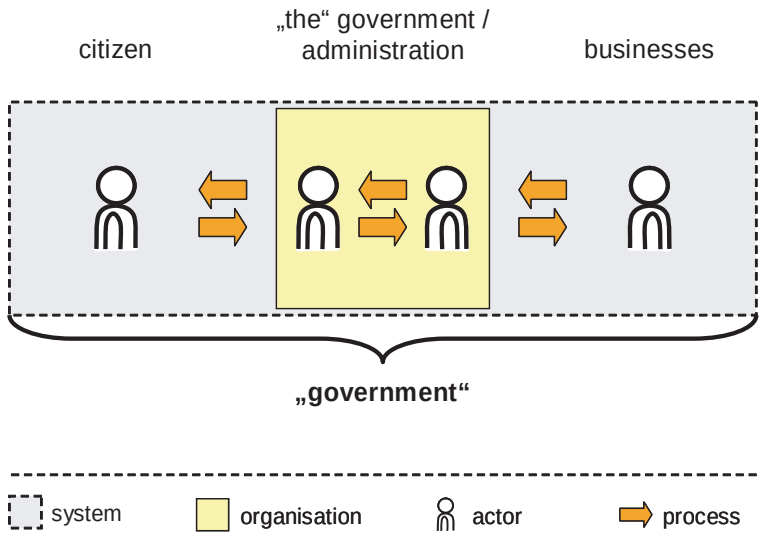
The *processes* of governing describe the actual ways of interaction between these active elements. Usually, there are certain procedural requirements which have to be met by effected participants. In general, these are available in form of laws and regulations, determining how a process is to be run, who is involved, which resources are required, and what the possible outcome is.

Involved *actors* within government constitute the governing authority, implicitly legitimizing its existence. The set of actors include the citizen as the basic part of the system, sometimes also acting as a representative of a business or as an employee or representative of a governmental organization.

In this setting, electronic government describes using ICT in the interaction of the named governmental elements, thus including m-government. To provide a rough distinction between e-government and m-government, it shall be understood here, that of course mobile government mainly extends electronic government, but at the same time, it adds mobility as a new feature, allowing for new, genuine mobile services. Whether mere extension or true innovation, m-government is dealing with mobility in the context of government, especially but not primarily with mobile technologies as a means of service delivery.

When designing a mobile government service, at least one part of the elements of government described earlier is mobile. Depending on what is being mobilized, different requirements have to be met. Of general concern are certain specifications set out by the system, possibly the availability of services to the whole public, thus often creating the problem of competing access channels. Recent

Figure 2. Elements of government



research (Pew Internet & American Life Project, 2004) reveals that multi-channel government is needed and wanted on the one hand, while on the other it decreases or at least consumes some of the efficiency gains by reducing the overall number of transactions within one channel, and requiring the support of parallel infrastructure. By adding yet another channel to the already available multiple channels (paper/forms, telephone/call centre, Internet/Web sites), it can be assumed that this issue will not be solved too easily. Thus, where there are several access channels for the same process, investment and roll-out strategies for new services have to closely monitor current capacities, and consider competition between available means of communication and effective demand. This competition issue is more relevant in the case of dealing with external parties (most likely citizen) than it is for internal exchange. In the latter case, it can be assumed that offerings are less prone to allow for diversity, due to the opportunity to simply impose standards in a top-down manner.

Since we are discussing the inclusion of mobility as a feature of a mobile government service, we have to take a closer look at the government elements who actually are mobile or to whom mobility is of importance. In the same way as we previously identified different perspectives on mobility, it will now be described how the different elements deal with mobility in the context of mobile government.

Process Requirements

The processes of government are the link between actors and organizations. In the context of mobility, one can think of many features of mobile communications which might be useful for governments. However, since government processes are rather formalized, they need to integrate many preconditions, with security being

just one. The challenge for mobile government is thus combining formal, procedural requirements with opportunities of the technologies.

From a more distant perspective, we have to first understand that a process consists of several steps. It is initiated by some kind of triggering event. This can be mobile, for example, a transaction being initiated in response to a mobile context.

Then there is the actual process, involving certain logic as to who is responsible, what resources are needed, and what decisions need to be taken. All of these can be *mobilized* as well. Decision makers can be linked to other participants via mobile communication channels, allowing for on-the-spot decisions or feedback when and where it is needed. Resources can be requested or procured in a similar manner. The final product itself may also be delivered using mobile technologies. Information as a product might be passed on to a mobile device, or the process result might trigger a new follow-up process, equally mobile.

As to whether or not a service can be made mobile is not only depending on the fantasy of a designer, but also on formal requirements imminent to the process. The question is thus a rather specific one and cannot be discussed in detail here.

Nevertheless, there are some general properties to a government process offering a generic guideline. First of all, a dominance of informational process can be assumed, that is, most of the transactions are based on information exchange and information processing. Depending on the content, especially the processing of information is more or less sensitive and time-critical. The complexity can differ, too, as well as the form of presentation.

The information focus is connected to the fact that general conditions for government processes are comparability, legal validity, and binding character of results (e.g., a decision on your taxes is based on your income, not on the mood of the decision maker; the decision is binding but can be legally challenged). The process itself is governed by law and usually processed along an organizational and functional path. Though these criteria still need to be detailed for each application, they offer assistance for the design process.

As is true for any government service, the formal requirements have to be matched with what is technically possible. A general method for such a matching process for mobile communication has been set out by Gerstheimer and Lupp (2001; see also Roggenkamp, 2004, p. 864). Notwithstanding this, describing what is feasible within the possibilities of technology and requirements of government processes is just the starting point. Designing mobile government must not be constrained to analyzing what can be mobilized. Furthermore, it is imperative to identify how and where additional value can be created, for the individual user as well as the providing organization.

Organizational Willingness

Different to a business organization, an organization in the public sector is acting in a kind of monopoly—the customer (in this context being a citizen, a business, or other governmental actor) cannot choose between different providers of public services. There usually is just one police, one local administration responsible, and so forth. In addition, a government organization is not seeking a rent of a financial kind; instead, it needs to legitimize its actions and existence towards the public, whom it is supposed to serve.

Thus, when it comes to the question as to whether, how and for what and whom an organization should rethink and re-organize its way of "doing things", a dilemma is surfacing which needs to be solved: external pressure and internal opposition. The common driver for implementing new technologies is external pressure: the demand for more and more efficient services, technologically enhanced process-handling as well as improving the public perception of an

organization respectively of the people in charge. Hence, technology is often implicitly imposed on an organization as a response to this pressure. A supposedly *new way* most often simply transfers the *old way* of doing things into the digital world. Replacing a paper-form by a Web site in fact leads to gains in productivity and cost reductions. Gora (1996) concluded that traditional organizations are insufficiently incorporating possibilities of ICT.

According to a recent survey of European public sector organizations (Net Impact, 2004) structural changes within the organization are in most cases made in response and not prior to the implementation of technology (p. 30). This results in lower gains in productivity as would have been achievable the other way around. The problem government organizations have to deal with is that "technology has evolved to a point where it is more difficult to change human behavior than it is to get the technology to do what you want" (p. 33). While there is external pressure to change, there is internal opposition hindering these changes. A description as to why this opposition occurs and how it arises has been thoroughly given by Borins (2001) and less detailed but focused on mobility issues by Kushchu and Borucki (2004). At this point it shall be of more interest how an organization can estimate the scope of this dilemma, especially in the course of designing mobile government.

While considering process requirements, it has already been roughly outlined how to identify processes. The task of an organization is to find out where stakeholders actually are confronted with mobility. From this deduction potentials for mobile services to be successful can be identified: from a productivity perspective to deal with internal opposition (by offering an enhanced mobile working environment) and with external pressure (by improving overall outcome).

Taking the previously-made statements about economical perspectives into account, we can identify three fields of action as starting points:

- internal processes (value chain);
- transaction processes; and
- government products (as results of the earlier-mentioned).

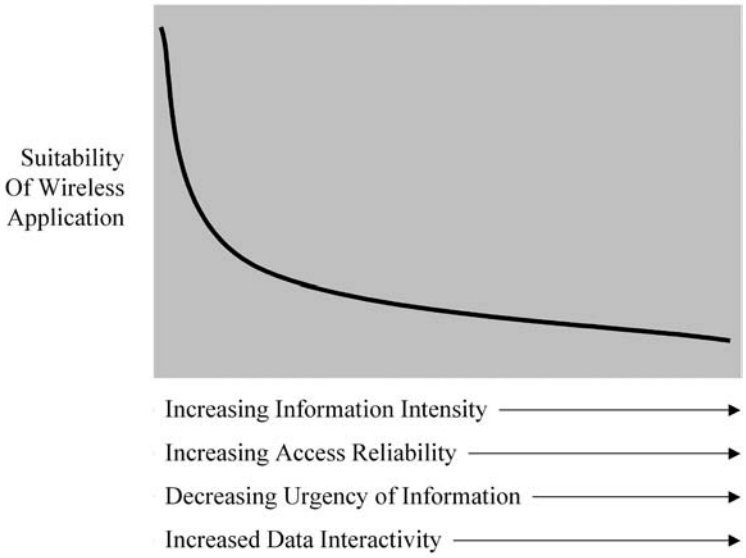
A whole set of possible services in the various fields of government has been described by the Centre for Public Service Innovation (2003) based in South Africa. Since there are already quite a few services available (see Zálešák, 2003), it should be a manageable task to identify applicable service proposals. As a method to assess whether and when an organization should consider implementing a certain service, Chang and Kannan (2002) have identified four relevant factors: the "extent of mobility in the target segment, information access needs, security/privacy requirements of the application, and technology readiness of the target segment" (p. 32). By joining the first three factors they generate an indicator for the sophistication of technology to meet certain requirements. In relation to the technology readiness, the actual question to be considered from the organizational perspective, they create a matrix (Figure 3) with which possible services can be ordered in a timely manner.

Of course, each organization which is considering mobile government services, as an option to invest in, needs to create a more detailed depiction of what it actually needs as well as an outline of why it needs a mobile service.

Concerning the fact of technology readiness—describing an attitude towards technology and not the mere competence to use it—will vary from one organization to the other, also from one department or team to another. By taking a step-by-step approach, the readiness can improve with each project. Generally speaking, as experience grows, so will the readiness to accept a new service (see also Davis, 1989, on perceptions of technologies).

When considering mobile service, it has so far been made clear that problems for an organization

Figure 3. Matrix for suitability of mobile services (Chang & Kannan, 2002, p. 21)



will most likely arise from within: in the form of internal opposition. By approaching more and more sophisticated services to deal with mobility in all fields of a government organization one step at a time failures can be avoided or at least contained.

Nevertheless, to legitimize an investment in mobile services, the value created has to be clear to the providing organization and its stakeholders. Where there is an overall strategy for service creation, criteria to assess and prioritize possible services in compliance with superior goals are easily found. In general, however, a real improvement has to be proven, for example, by pilot implementations. Mere feasibility of mobilizing the processes is one thing to be shown here, but also the feasibility of integrating the service into existing backend structures and more so into communication structures within an organization.

While the value of each service will often be easily assessed, the willingness and ability to realize potential gains through mobile services in common processes is the key from the perspective of an organization. Otherwise, even the most promising mobile added values would be just that: promising.

User Acceptance

We have so far looked at feasibility of mobile government services and discussed organizational willingness. Now we will focus on user acceptance as the main driver for the success of a service. The situation, into which mobile government will be deployed, can usually be described as competitive as there are other access channels to certain services. Thus, it is inevitable to consider why a user is lead to adopt a service.

User acceptance can be described as a product of user behavior in relation to the available technology and a given environment. Davis (1989) has described perceived usefulness and

perceived ease of use as the main influences on user acceptance of information technology (p. 320). As a consequence thereof, to achieve user acceptance the development of these perceptions need to be dealt with.

Methods in this field reach from standard quantitative approaches simply asking about how services would be or are perceived to explorative approaches of shadowing users. The latter are able to find out more about actual usage patterns and social shaping factors (see Vincent & Haddon, 2003). From the research we learn that for mobile communication the usage patterns are influenced by the "representation" (Churchill & Wakeford, 2002) of a service, meaning the image being drawn by suggesting value (based on supposed usage). This image is usually framed by the reaction of the individual environment of a prospective user, considering norms and general values.

This leads a user to the initial decision to get to know a service—to test it. During this "learning phase" previously formed expectations determine the subsequent, repeated usage, allowing for new experiences. Important at this stage is not only typical front-end usability but a more generic feeling of actually being able to control and, to a certain extent, to understand functions offered by the used technology. Palen and Salzman (2002) have described four dimensions important at this point, of which at least two are within the reach of government.

Hardware, being the first dimension, roughly covers the device used for accessing a service. On such a device, a set of *software* is controlling user flows. The *netware* is the connection of hard- and software with mobile technologies' functions. This is surrounded by the *bizware*, consisting of all sorts of customer support and customer relation.

For a government organization offering mobile services, the points to connect into this structure would mainly be the net- and bizware, offering services and supporting their usage.

Following the state of acquainting with a service, the user individually explores the opportunities offered, possibly seeking new patterns of usage, as well as feeding back his experiences to his environment. At this point, but also in the initial phase, we can observe what Vincent and Haddon (2003) call "social shaping" of mobile technology: the creation and reassessment of value expectations, changing usage and behavioral patterns as well as overall norms defining how to properly behave, use, and consider certain mobile services.

The choice to use a mobile service, as is the case for any ICT, is thus guided by the perception of use and the actual use. User needs, more so the evaluation of a service to meet these needs is crucial for its success. Yet, research has also shown (cf. Pew Internet, 2004; Vincent & Haddon, 2003) that a typical user will not only make smart choices about the technology to use, but he will often consider the appropriateness of a means from a functional (reflecting the actual service) and a social (reflecting the outside perception) point of view. Furthermore, often parallel structures are maintained, for backup purposes or to meet requirements of unexpected situations. This is related to trust on the one hand, to a somewhat haptic experience on the other (Perry et al., 2001). In both cases, the biggest challenge for mobile information access is a paper-based process, for mobile interaction services the challenge is the fixed line phone (see also Vincent & Haddon, 2003). Since government services are most often based on trust (especially when exchanging sensitive personal or corporate information) while allowing for several channels of interaction, this issue should be thoroughly considered.

Nevertheless, trust can be built (by experience), whereas complementary services as a first step will lead to substituting less helpful stationary services in the long run.

By assessing needs more closely, the impact of the previously-named issues can be softened.

In general, mobile business services consider a user in a certain situation, and come to conclusions about his needs by deconstructing the individual context (see Gerstheimer & Lupp, 2001; Roggenkamp, 2004).

For existing services to be mobilized, this is helpful, though not fully leading to conclusions. Because of this, a different method of looking at user needs shall be introduced to extend this perspective.

Before, one has to distinguish between need in the sense of wishing for something and need in the sense of demand, connected to the willingness to trade resources to satisfy this demand. Hence, the needs we will talk about now are of the latter type, considered from the supply side.

Governments offer certain services in pursuit of meeting common rules and fulfilling assumed tasks. This "supply" can be characterized (Figure 4) as to what kind of interaction it offers: information, communication, or transaction. Second, the intention of a service can be defined: whether it is aimed at following a user, thus allowing him to be mobile. Or whether it is aimed at guiding a user, thus supporting his mobility. Finally, we can consider the dependency of user and service, being connected due to time, location, or person.

With each aspect, it can be estimated as to how sensitive a service is, and how much it relates to situations which are mobile. An assessment of how a service will be perceived can be based on the results of this description.

From the initial description of influences on user acceptance, the second relevant issue for government is user experience. As has already been mentioned for organizational willingness, a step-by-step approach seems to be appropriate in the course of entering the field of mobile services. Not only can the providing organization adjust to demands of mobility, the actual users too can familiarize themselves with mobile government and explore technology, services, and functions available.

Finally, an issue not to be underestimated is the overall perception of mobile government. We have to bear in mind that mobile communications are currently most often used for purposes of social connectedness, whereas we also see mobile services most successful when offering entertainment and tools to customize the personal item "mobile phone". Against this background, it has to be proven that mobile government is more than supposedly helpful SMS or more intrusive government officials with remotely trustworthy devices that seem to reveal and collect by far too much sensitive data. Electronic government has reached a point where the extreme perceptions reaching from subversive Internet-based liberation to data-gathering "big brother"-like organizations have diffused into something convincingly useful, that can be utilized as needed. The biggest changes have occurred within the backend infrastructures.

Figure 4. Describing a service to identify resulting needs

Interaction level	Intention of a service	Dependency user/service
• Information • Communication • Transaction	• Follow user/ allow for mobility • Guiding user/ supporting mobility	• Time • Location • Role/Person

Since mobile government is, due to its mobility, doomed to take place far more in public spaces, the positive images have yet to emerge, the general perceptions yet need to be formed.

THINKING ABOUT MOBILE GOVERNMENT

When thinking about mobile government, the following four general questions should lead the way on to a definition of an aspired service. Also, they are supposed to help identifying beforehand, if and where there are flaws in the concept and actual message connected with a service and its goals.

What is Being Mobilized?

Reviewing existing m-government, it becomes obvious that the scope of services awarded this label is wide, to be distinguished by their primary function and their reach.

To answer this question, initially it needs to be defined whether the main goal is offering mobile access to some data source or if a service is supposed to be a genuine mobile application (or more likely something in-between, but tending to one side or the other).

Whereas mobile access most often means getting rid of wires, the being-mobile is part of the functionality of a full-scale mobile application. Services such as geographical or tourist information are a case in point, since being stationary bound impedes their value dramatically. The reach of a service in this context is considering geographical limitations, whether a service is limited to a single location, a region or whether it reaches even beyond. From this initial evaluation, we can derive possible bearing technologies applied for transmitting, accessing, or delivering services in question.

The second part of this evaluation dealt with needed/offered complexity, sensitivity, and criticality of a service. Referring to the earlier-identified process elements, we can look at:

- triggering events,
- control mechanisms,
- resources, and
- resulting products and recipient,

as able to be mobilized, as well as consequences thereof. Triggering a process is probably less complex than governing it. With the recipient being mobile, embedded within a certain context, and depending on the product being possibly presented mobile, there are particular problems to be dealt with for each element.

Why Mobile?

This rather broad question is supposed to lead to name driving forces behind the creation of mobile services. The case of actual, focused public demand is found rather seldom. Hence, we can conclude key motives to be efficiency, effectiveness or political goals. Furthermore, when looking at the purpose of mobilization, the mobile value being *expected* should be directly addressed. When an answer cannot be easily found here, it should be doubted that targeted users will be convinced. Less so, when they are actually obliged to cover extra cost (connection fees, infrastructure, etc.). Nevertheless, the opposite situation of being instantly able to state exact and conclusive reasons is no guarantee for success, for initial user adoption.

Are there Alternative (not Mobile) Services Available? Will Existing Services be Replaced?

Based on the already found answers, this subject should be quickly covered. The mere existence of an alternative to a mobile service puts the question of expected added values attributed to such

a service back on the agenda; especially when these alternatives are maintained.

Competing with other communication channels, the mobile channel covers certain areas of use, namely social and private ones. Therefore it is appropriate to ask whether and why the mobile alternative should be better.

Nevertheless, the data on the not-mobile services can help to assess how to improve it, and how to transfer this into a mobile context. Also, it can be better identified who would actually be served, and how often.

What is the General Agenda? Who will Benefit?

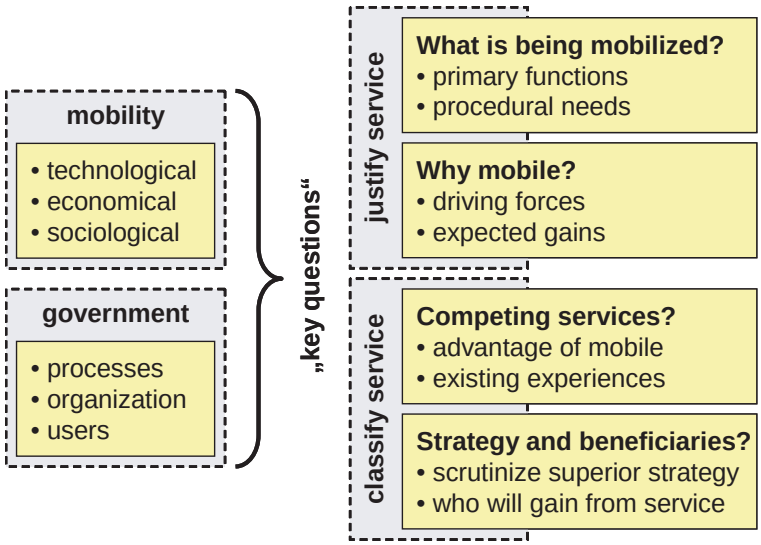
This third version of the "Why mobile?" question is asked separately, because it scrutinizes the superior strategy of a providing organization. Naming those who will benefit is especially important when a service is deployed into a kind of monopoly, as we find to be predominant in most of the public sector.

An individual cannot choose with which organization it prefers to do government business, however, in many cases it can choose how it wants to do it. By identifying those who gain advantages by a mobile service it can also be determined who will have to cover any additional cost.

CONCLUSION

In the course of this chapter, the term mobility has been described as the key component of mobile government. Three (technological, economical, and sociological) different perspectives have been shown along with goals connected to their respective view. The combination of these perspectives has helped to identify key issues to be considered when dealing with mobility. With the description of these views, it is possible to develop a mobility cycle which combines questions of technological feasibility with economical demands and user needs.

Figure 5. Key questions when designing mobile government



Focusing on the elements of the government term, it has further been described how process requirements, organizational willingness, and user acceptance can influence mobile government services. Also, a rough distinction between e-government and m-government has been provided, stating that mobile government mainly extends electronic governments, but considering mobility as a new feature, new services become possible which do not fit in a narrow view of e-government being a digital way of service-delivery.

Resulting from the previous dealing with mobility perspectives and the elements of government, four questions (Figure 5) were identified which ought to be dealt with when conceptualizing and designing mobile government.

The questions are supposed to guide developers through the process of designing mobile government services: First by helping to define and justify scope and content of a service. Second, the service can be classified and reviewed in its context, by assessing the possible competition it might face and by naming its beneficiaries.

OUTLOOK

Leading to a more generalized debate on mobile government, it is yet unclear whether mobile services in the public sector should be seen as something completely different that "traditional" services, or if we need to amend our evaluation criteria just a little.

All in all, most of the mobile services can be reduced to "old-fashioned" online services, based on existing backend infrastructure, with only significant differences being the *feature* mobility, to be understood as an add-on.

However, the debate on mobile interaction points out that the addition of "mobile" as a feature and important component of a service (including all the technological functions like localization, etc.) lead to changing behavior; the expectations toward the public sector are affected while the boundaries of public and private are diminishing.

Similar things have been said on the peak of the online hype and following the invention of the World Wide Web, just a few years ago: *virtual democracy* was expected by many commentators. However, the debate developed from euphoria to disillusionment—the drastic expectations of an Internet-based revolution did not come true. This notwithstanding can we observe changes in behavior and perception among users. Also, we have to recognize the changes on the government side, which is now looking for the creation of better services. Even the most pessimistic evaluation of mobile services in the public sector has to acknowledge the sustainable effect on users and providers alike.

Apart from the common criteria of service assessment, such as cost and quality, the potential of location-independence and other features of mobile technology have yet to be understood in all their impact.

Mobile government is not per se something new or even special. Due to the technological features and more so due to mobile users, it is an issue growing to become more and more important. And that is enabling and supporting mobility.

REFERENCES

Agre, P. (2001). Changing places: Contexts of awareness in computing. *Human-Computer Interaction*, *16*(2-4), 177-192.

Arazyan, H. (2002). *M-government: Definition and perspectives*. Retrieved August 12, 2004, from http://www.developmentgateway.org/download/143909/m-Government_Interview_2.doc

Bazijanec, B., & Pousttchi, K. (2004). Suitability of mobile communication techniques for the business processes of intervention. In D. Remenyi (Ed.),

Proceedings of the 4th European Conference on e-Government (ECEG 2004), Dublin, Ireland, June 17-18 (pp. 805-812).

Borins, S. (2001). The challenge of innovating in government. In the *Innovations in Management Series*, 02/2001. Arlington, VA: The PricewaterhouseCoopers Endowment for the Business of Government.

Castells, M. (2000). *Rise of the network society.* Sagebrush Education Resources.

Centre for Public Service Innovation. (2003). *Government unplugged – Mobile and wireless technologies in the public service.* Originally retrieved March 8, 2004; New URL [retrieved November 18, 2006 from] http://www.cpsi.co.za/contentfiles/tblFile/5_filFilePath_Government%20Unplugged.pdf

Chang, A., & Kannan, P. (2002). *Preparing for wireless and mobile technologies in government.* Arlington, VA: IBM Endowment for the Business of Government.

Churchill, E., & Wakeford, N. (2002). Framing mobile collaborations and mobile technologies. In B. Brown, N. Green, & R. Harper (Eds.), *Wireless world, social and interactional aspects of the mobile age* (pp. 154-179). London: Springer.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-339.

Di Maio, A. (2002). *Toward a wireless public sector.* Gartner Research – ID AV-18-0223.

FOKUS. (2000). Vehicular Video-Inter-car Video Transmission. Retrieved March 8, 2004, from http://www.fokus.gmd.de/research/cc/cats/projects/vehicular/flyer.pdf

Frischmuth, J., & Karrlein, W. (2002). Aktuelle Trends im Electronic und Mobile Business. In P. Blaschke, W. Karrlein, & B. Zypries (Eds.), *E-Public: Strategien und Potenziale des E- und Mobile Business im öffentlichen Bereich* (pp. 9-30). Berlin: Springer-Verlag.

Gerstheimer, O., & Lupp, C. (2001). *Zukünftige Kundennutzenpotenziale im Bereich der mobilen Datenkommunikation, Doppeldiplomarbeit im Studiengang Produktdesign, Schwerpunkt Systemdesign*. Kassel, Germany: Universität Kassel.

Geser, H. (2003). *Towards a sociological theory of the mobile phone*. Retrieved November 18, 2006, from http://www.siocio.ch/mobile/t_geser1.htm

Goffmann, E. (1967). Interactional ritual: Essays on face to face behavior. New York: Anchor Books.

Gora, W. (Ed.) (1996). *Auf dem Weg zum virtuellen Unternehmen.* Köln: Fossil.

Hasan, H., Jähnert, J., Zander, S., & Stiller, B. (2001). *Authentication, authorization, accounting, and charging for the mobile Internet.* In the IST Mobile Summit 2001, Barcelona, September 9-12.

Hess, T., Figge, S., Hanekop, H., Hochstatter, I., Hogrefe, D., Kaspar, C., Rauscher, B., Richter, M., Riedel, A., & Zibull, M. (2005). Mobile Anwendungen – eine interdisziplinäre Herausforderung. In the *Die Wirtschaftsinformatik*, Nr., 02/2005 (forthcoming).

Kakihara, M. (2003). *Emerging work practices of ICT-enabled mobile professionals.* PhD dissertation, Department of Information Systems, London School of Economics and Political Science, London.

Kaspar, C., & Hagenhoff, S. (2003). Geschäftsmodelle im Mobile Business aus Sicht der Medienbranche. In the *Arbeitsbericht*, Nr., 15/2003. Georg-August-Universität Göttingen, Institut für Wirtschaftsinformatik

Khodawandi, D., Pousttchi, K., & Winnewisser, C. (2003). *Mobile Technologie braucht neue Ge-*

schäftsprozesse. Originally retrieved September 23, 2004; New URL [retrieved November 18, 2006 from] www.wi-mobile.org/fileadmin/Papers/MBP/uni-augsburg-mobile-16-11.pdf

Kristofersen, S., & Ljungberg, F. (2000). Mobility: From stationary to mobile work. In K. Braa, C. Sørensen, & B. Dahlborn (Eds.), *Planet Internet* (pp. 41-64). Lund: Studentliteratur.

Kushchu, I., & Borucki, C. (2004). Impact of mobile technologies on government. In D. Remenyi (Ed.), *Proceedings of the 4th European Conference on e-Government (ECEG 2004),* Dublin, Ireland, June 17-18 (pp.829-836).

Kushchu, I., & Kuscu, M. (2003). From e-government to m-government: Facing the inevitable. In the *Proceedings of the 3rd European Conference on e-Government (ECEG 2003),* Dublin, Ireland, July 3-4.

Ling, R. (2000). We will be reached: The use of mobile telephony among Norwegian youth. *Information Technology and People, 13*(2), 102-120.

Ling, R., & Yttri, B. (1999). *Nobody sits at home and waits for the telephone to ring: Micro and hyper-coordination through the use of the mobile telephone*. In the Telenor Research & Dev Report 30/99.

Llalana, E. C. (2004). *mGovernment definitions and models page*. Retrieved November 18, 2006, from http://www.egov4dev.org/mgovdefn.htm

Net Impact. (2004). *Net impact: Europe eGovernment, Cisco Systems, Momentum Research Group.* Retrieved September 27, 2004, from http://www.netimpactstudy.com/pdf/NetImpact_04b.pdf

Palen, L., & Salzman, M. (2002). Welcome to the wireless world: Problems using and understanding mobile telephony. In B. Brown, N. Green, & R. Harper (Eds.), *Wireless world, social and interactional aspects of the mobile age* (pp. 134-153). London: Springer-Verlag.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001, December). Dealing with mobility: Understanding access anytime, anywhere. *ACM Transactions on Computer-Human Interaction, 8*(4), 323-347.

Pew Internet & American Life Project. (2004). How Americans get in touch with government. Retrieved from http://www.pewinternet.org/pdfs/PIP_E-Gov_Report_0504.pdf

Pica, D., & Kakihara, M. (2003). The duality of mobility: Unterstanding fluid organizations and stable interaction. In the *Proceedings of the 11th European Conference on Information Systems (ECIS 2003),* Naples, Italy, June.

Roggenkamp, K. (2004). Development modules to unleash the potential of mobile government. In D. Remenyi (Ed.), *Proceedings of the 4th European Conference on e-Government (ECEG 2004),* Dublin, Ireland, June 17-18 (pp. 857-866).

Rossado-Schlosser, A., & Hacke, M. (2002). Mobile Datendienste – Revolution der Geschäftswelt? Retrieved April 24, 2006, from http://www.digital-transformation.mckinsey.de/pdf/2889247_digital_transformation_modul5_mobdaten.pdf

Thome, R. (2003). M-government. In T. Schildhauer (Ed.), *Lexikon Electronic Business* (pp. 212-213). Munich: Oldenbourg Verlag.

Turowski, K., & Pousttchi, K. (2003). *Mobile commerce: Grundlagen und Techniken*. Berlin, Heidelberg: Springer-Verlag.

Urry, J. (2000). Mobile sociology. *British Journal of Sociology, 51*(1), 185-203.

Vincent, J., & Haddon, L. (2003). *Informing suppliers about user behaviours to better prepare them for their 3G/UMTS customers. Final Report: Assessment and Analysis of Findings.* In the UMTS Forum Report 34, UMTS Forum, Surrey, GB.

Zálešák, M. (2003). *M-government: More than a mobilised government*. Retrieved January 2, 2005, from http://www.europemedia.net/shownews.asp?ArticleID= 14482 and http://www.europemedia.net/shownews.asp?ArticleID=14495

This work was previously published in Mobile Government: An Emerging Direction in E-Government, edited by I. Kushchu, pp. 60-85, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 2.30
Design of Government Information for Access by Wireless Mobile Technology

Mohamed Ally
Athabasca University, Canada

INTRODUCTION

As the world becomes mobile, the ability to access information on demand will give individuals a competitive advantage and make them more productive on the job and in their daily lives (Satyanarayanan, 1996). In the past, government information was presented by government employees who verbally communicated with citizens in order to meet their information needs. As print technology improved, government information was, and still is in many countries, communicated to citizens using paper as the medium of delivery. Because of the cost of printing and mailing printed documents and the difficulty of updating information in a timely manner, governments are moving to electronic delivery of information using the Web. Currently, governments provide digital service to their citizens using the Web for access by desktop or notebook computers; however, citizens of many countries are using mobile devices such as cell phones, tablet PCs, personal digital assistants, Web pads, and palmtop computers to access information from a variety of sources in order to conduct their everyday business and to communicate with each other. Also, wearable mobile devices are being used by some workers for remote computing and information access in order to allow multitasking on the job. It is predicted that there will be more mobile devices than desktop computers in the world in the near future (Schneiderman, 2002). The creation of digital government will allow the delivery of government information and services online through the Internet or other digital means using computing and mobile devices (LaVigne, 2002). Also, there will be more government-to-citizen and government-to-business interactions. Digital government will allow citizens, businesses, and the government to use electronic devices in order to communicate, to disseminate and gather information, to facilitate payments, and to carry out permitting in an online environment (Wyld, 2004). Digital government will allow citizens

to access information anytime and anywhere using mobile and computing devices (Seifert & Relyea, 2004).

BACKGROUND

According to O'Grady and O'Hare (2004), mobile computing will become the major computer usage model of the future. This will be possible since the digital divide is decreasing due to wireless access, increasing use of mobile devices, decreasing cost of Internet connections and computer technology, and transparent access of computer systems. Governments need to take advantage of technology-literate citizens and design and make available information for citizens to access government information digitally from anywhere and at anytime. This is important, since citizens expect the same level of service that is being given by businesses that are providing services and information anywhere and anytime (Dawes, Bloniarz, Connelly, Kelly, & Pardo, 1999). Users need just-in-time information for the job and in the community. The use of wireless mobile devices will facilitate access of government information from anywhere and at anytime. Also, computing is becoming ubiquitous, where citizens will work from anywhere and access government information from many networks using wireless mobile devices (Huber, 2004; Perry, O'Hara, Sellen, Brown, & Harper, 2001).

Before the use of mobile devices to access government information and design of information for mobile access are discussed, it is important to examine the information processing required when citizens access digital government information. Citizens acquire government information at many levels. At the lowest level, citizens may want to be aware of what is happening in government, so they will read the information in order to be informed. For example, some citizens may want to know the changes made to tax regulations. At the next level, citizens and businesses may want to access government to apply the information to complete everyday tasks. For example, some occupations require that businesses and citizens follow approved safety procedures when completing tasks. This requires comprehension and application of the information. At the highest level, citizens and businesses may want to critically analyze, synthesize, and evaluate government information for research purposes. To achieve this, citizens will have to access government information from many sources through ubiquitous computing using mobile devices.

USE OF MOBILE DEVICES TO ACCESS DIGITAL GOVERNMENT INFORMATION

There are many benefits to the use of mobile devices to access digital government information. According to a recent report by the European Commission (2004), digital government can provide better quality public service, reduce waiting time for information and service, lower administrative costs for businesses, and allow higher productivity for the public. Using mobile devices will allow citizens to access government information from anywhere and at anytime. With the use of wireless mobile technology, users do not have to be connected physically to networks in order to access information, and the mobile devices are small enough to be portable, which allows users to take the devices to any location to send and retrieve information. For example, a worker in the field who requires specific government regulations while completing a task can use a mobile device to access the information just in time. If government regulations in a field change, the government can update the digital information to allow individuals and businesses to access the current information immediately. In addition, a worker in the field can use a mobile device to contact a government employee remotely and to request specific information for immediate use.

Mobile devices have many benefits for accessing government information; however, there are some limitations of mobile devices of which designers of government digital information must be aware when designing information for delivery on mobile devices. Some of the limitations of mobile devices in delivering government information include the small screen size for output of the information and the small input devices for accessing the information (Ahonen, Joyce, Leino, & Turunen, 2003). Designers of information must be aware of these limitations when designing government digital information for access by mobile devices and must design for ease of use. Rather than scrolling for more information on the screen, users of mobile devices must be able to go directly to the information and move back and forth with ease. Information should be targeted to the users' requests when they need it and should be presented efficiently to maximize the display of the information on the mobile device screen. The interface of the mobile device must be appropriate for individual users and the software system should be able to customize the interface based on individual user's characteristics.

Designing Government Digital Information for Mobile Devices

As the evolution of delivery medium of information changes, so does the strategy for processing the information. According to Grudin (2004), prior to writing and print, most information access and interaction were done by listening, memorizing, and speaking. With the print medium, information acquisition strategies were reading, analyzing, and writing. As government information becomes digital, acquisition strategies include searching, synthesizing, and constructing. Designers of government information for mobile devices must design for the new information acquisition and interaction strategies.

Most government information tends to be text-based, which takes longer for users to process and interpret. This is because past government information was designed for printing on paper for delivery to citizens. Designers of digital government information must use the capability of the computer to present information visually as well as textually in order to facilitate efficient processing and acquisition of the information. According to Paivio's (1986) theory of dual coding, information storage and retention is enhanced when information is represented both in verbal and visual forms. Presenting material in both textual and visual forms will involve more processing, which will result in better storage and integration of information in memory (Mayer, Fennell, Farmer, & Campbell, 2004).

In addition, because of the limited display capacity of mobile devices, government information must be designed for display using rich media such as audio, video, pictures, and graphics. Tabbers, Martens, and van Merrienboer (2004) found that for Web-based multimedia information, students who received visual cues to pictures scored higher on an information retention test compared to students who did not receive the cues for the pictures. According to cognitive psychology, information acquisition is an internal process, and the amount retained depends on the processing capacity of the user, the amount of effort expended while reading the information, the quality of the processing, and the user's existing knowledge structure (Ausubel, 1974). These have implications for how government information is designed for mobile devices. Designers must include strategies that allow the user to activate existing cognitive structure in order to conduct quality processing of the information. Mayer, Dow, and Mayer (2003) found that when a pedagogical agent was present on the screen as information was narrated to students, students who were able to ask questions and receive feedback interactively performed better on a problem-solving transfer test compared to students who only received on-screen text with no narration. It appears that narration by an intelligent agent encouraged deep processing, which

resulted in better information acquisition and higher-level information processing. This suggests that government should use audio to present government information to citizens.

Guidelines for Designing Government Digital Information for Mobile Devices

Chunk Information for Efficient Processing

Designers of government materials for mobile devices must use information presentation strategies to enable users to access and process the information efficiently because of the limited display capacity of mobile devices and the limited processing capacity of human working memory. Information should be organized or chunked in the form of information objects of appropriate and meaningful size to facilitate storage and processing in working memory (Ally, 2004).

Adapt the Interface to the User

To compensate for the small screen size of the display of the mobile device, the interface of the mobile device must be designed properly (Ally, 2004). Mobile access to government information requires interface designs for multi-mobile device access and intelligent agents to adapt the interface to the user (Nylander, Bylund, & Boman, 2004). The interface can be graphical and should present limited information on the screen in order to prevent information overload in short-term memory. Users must be able to jump to related information without too much effort. The interface must allow the user to access the information with minimal effort and to move back to previous information with ease. For interaction sessions that are information-intense, the system must adjust the interface in order to prevent information overload. Some ways to prevent information overload include presenting less information on one screen or organizing the information in the form of graphical outlines to give the overall structure of the information and then presenting the details by linking to other screens with the details. The interface also must use good navigational strategies to allow users to move back and forth between information displays. Navigation also can be automatic, based on the intelligence gathered on the user's current position in the information and the information needs of the user.

Design for Minimum Input to Retrieve and Access Information

Because of the small size of the input device on mobile devices, information access must be designed to require minimum input from users. Input can use pointing or voice input to minimize typing and writing. Because mobile devices allow access of information from anywhere at anytime, the device must have input and output options in order to prevent distractions when using the mobile devices. For example, if someone is using a mobile device in a remote location, it may be difficult to type on a keyboard or to use a pointing device when accessing government information. The mobile technology must allow the user to input data using voice input or touch screen.

Target Government Information to the User

One of the variables that designers tend to ignore when they develop information for mobile devices is the user of the devices. Different users have different styles and characteristics, and some users may be more visual, while others may be verbal (Mayer & Massa, 2003). A graphic outline of the information can be presented before the details are presented in order to cater to users who prefer to get the big picture before they go to the details

of the information. Government information must be designed with the user in mind to facilitate efficient access and processing.

Government systems must be smart and should have built-in intelligence in order to customize and target the information for individual citizens. Information must be personalized by selecting and aggregating information according to the user profile (Huber, 2004). Intelligent software systems can be built to develop an initial profile of the user based on current and previous interaction with the government information database and then present materials that will benefit the specific user, based on the user profile. As the intelligent agent interacts with the user, it learns about the user and adapts the format of the information, interface, and navigation pattern according to the user's style and needs.

Use Visual Outline to Show the Structure of the Information

A visual outline can be used to show the main ideas in the information and the relationship between the ideas rather than to present information in a textual format. High-level visual outline can be used to represent information spatially so that users can see the main ideas and their relationships (Novak, Gowin, & Johanse, 1983). Tusack (2004) suggests the use of site maps as the starting point of interaction to which users can link back in order to continue with the information.

TRENDS IN USING MOBILE DEVICES FOR ACCESSING GOVERNMENT INFORMATION

The use of mobile devices with wireless technology allows access of government information from anywhere and at anytime and will dramatically alter the way that work is conducted (Gorlenko & Merrick, 2003). For example, mobile devices can make use of global positioning and satellite systems to send and receive government information digitally. There will be exponential growth in the use of mobile devices to access government information, since the cost of the devices will be lower than desktop computers and a user can access information from anywhere and at anytime. Also, the use of wireless mobile devices by businesses and organizations will be more economical, since it does not require the building of the infrastructure to wire buildings for employees and customers. The challenge for designers of government information for mobile devices is how to standardize the design for use by different types of mobile devices. Government information systems need to have agents to deliver the right information to the user. A profile agent can be used to learn about the user and then to interact with a presentation agent in order to customize and format the information to meet the user needs (O'Grady & O'Hare, 2004).

CONCLUSION

Government information was designed for delivery on paper medium. Governments need to rethink and redesign information for delivery on mobile devices. Future development of information for mobile devices should concentrate on the user to drive the development and delivery (Gorlenko & Merrick, 2003). Mobile devices can be used to deliver government information to users, but the materials must be designed properly to compensate for the small screen of the devices and the limited processing and storage capacity of users working memory. Design principles for government information on mobile devices are the same as design principles for other applications such as education and training. The only difference is that government information tends to be one way, and there is less interaction with the information, since the purpose of most gov-

ernment information is to inform citizens. More research should be conducted on how to improve security and privacy of government information on mobile devices. Also, the type of information presented on mobile devices must match the needs and the styles of the users. Government information systems must use the power of computer technology in order to develop intelligent agents to customize the information for users and to provide context-sensitive information. According to Rist and Brandmeier (2002), more research is needed on how to flexibly translate government information from one medium into another format and how to decide which media combinations are most appropriate, considering a mobile user's style, current task, and situation. Finally, governments need to shift from print and desktop delivery of information to delivery on mobile devices in order to make the transition to becoming mobile government (m-government).

REFERENCES

Ahonen, M., Joyce, B., Leino, M., & Turunen, H. (2003). Mobile learning: A different viewpoint. In H. Kynaslahti & P. Seppala (Eds.), *Mobile learning* (pp. 29-39).

Ally, M. (2004a). Designing effective learning objects for distance education. In R. McGreal (Ed.), *Online education using learning objects* (pp. 87-97). London: RoutledgeFalmer.

Ally, M. (2004b). Using learning theories to design instruction for mobile learning devices. *Proceedings of the Mobile Learning 2004 International Conference*, Rome.

Ausubel, D. P. (1974). *Educational psychology: A cognitive view*. New York: Holt, Rinehart, and Winston.

Dawes, S. S., Bloniarz, P. A., Connelly, D. R., Kelly, K. L., & Pardo, T. A. (1999). Four realities of IT innovation in government. *The Public Manager*, *28*(1), 1-9.

European Commission. (2004). *eGovernment resource book*. Luxembourg: European Communities.

Gorlenko, L., & Merrick, R. (2003). No wires attached: Usability challenges in the connected mobile world. *IBM Systems Journal*, *42*(4), 639-651.

Grudin, J. (2004). Crossing the divide. *ACM Transactions on Computer-Human Interaction*, *11*(1), 1-25.

Huber, J. F. (2004). Mobile next generation networks. *IEEE Multimedia*, 72-83.

LaVigne, M. (2002). Electronic government: A vision of a future that is already here. *Syracuse Law Review*, *52*(4), 1-8.

Mayer, R. E., Dow, T. D., & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds. *Journal of Educational Psychology*, *95*(4), 806-813.

Mayer, R. E., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology*, *96*(2), 389-395.

Mayer, R. E., & Massa, L. J. (2003). Three facets of visual and verbal learners: Cognitive ability, cognitive style, and learning preference. *Journal of Educational Psychology*, *95*(4), 833-846.

Novak, J. D., Gowin, D. B., & Johanse, G. T. (1983). The use of concept mapping and knowledge vee mapping with junior high school science students. *Science Education*, *67*, 625-645.

Nylander, S., Bylund, M., & Boman, M. (2004). Mobile access to real-time information—The

case of autonomous stock brokering. *Personal and Ubiquitous Computing, 8*(1), 42-46.

O'Grady, M. J., & O'Hare, G. M. P. (2004). Just in time multi-media distribution in a mobile computing environment. *IEEE Multimedia*, 62-74.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: understanding access anytime, anywhere. *ACM Transactions on Computer-Human Interaction*, *8*(4).

Rist, T., & Brandmeier, P. (2002). Customizing graphics for tiny displays of mobile devices. *Personal and Ubiquitous Computing, 6*, 260-268.

Satyanarayanan, M. (1996). Accessing information on demand at any location: Mobile information access. *IEEE Personal Communications*, 26-33.

Schneiderman, R. (2002). *The mobile technology question and answer book: A survival guide for business managers.* New York: American Management Association.

Seifert, J. W., & Relyea, H. C. (2004). Considering e-government from the U.S. federal perspective: An evolving concept, a developing practice. *Journal of E-Government, 1*(1), 7-15.

Tabbers, H. K., Martens, R. L., & van Merrienboer, J. J. G. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British Journal of Educational Psychology, 74*, 71-81.

Tusack, K. (2004). Designing Web pages for handheld devices. *Proceedings of the 20th Annual Conference on Distance Teaching and Learning,* Madison, WI.

Wyld, D.C. (2004). The 3 Ps: The essential elements of a definition of e-government. *Journal of E-Government, 1*(1), 17-22.

KEY TERMS

Concept Map: A graphic outline that shows the main concepts in the information and the relationship between the concepts.

E-Government: The delivery of government information and services using electronic technologies.

Information Object: Digital information stored in chunks in a digital repository and tagged for retrieval to meet users' information needs.

Intelligent Agent: A computer application software that is proactive and capable of flexible autonomous action in order to meet its design objectives set out by the designer.

Interface: The components of the computer program that allow the user to interact with the information.

Mobile Device: A device that can be used to access information from anywhere and at anytime. The device consists of an input mechanism, processing capability, a storage medium, and a display mechanism.

Pervasive Computing: Use of computer devices to access information from interconnected networks using wireless technology.

Ubiquitous Computing: Computing technology that is invisible to the user because of wireless connectivity and transparent user interface.

User: An individual who interacts with a computer system to access information.

Wearable Computing Devices: Devices that are attached to the human body so that the hands are free to complete other tasks.

This work was previously published in Encyclopedia of Digital Government, edited by A. Anttiroiko and M. Malkia, pp. 291-295, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Section III
Tools and Technologies

This section presents extensive coverage of the technology that both derives from and informs mobile computing. These chapters provide an in-depth analysis of the use and development of innumerable devices and tools, while also providing insight into new and upcoming technologies, theories, and instruments that will soon be commonplace. Within these rigorously researched chapters, readers are presented with examples of the tools that facilitate and support mobile computing. In addition, the successful implementation and resulting impact of these various tools and technologies are discussed within this collection of chapters.

Chapter 3.1
Evaluation of Mobile Technologies in the Context of Their Applications, Limitations, and Transformation

Abbass Ghanbary
University of Western Sydney, Australia

ABSTRACT

Emerging mobile technologies have changed the way we conduct business. This is because communication, more than anything else, has become extremely significant in the context of today's business. Organizations are looking for communication technologies and corresponding strategies to reach and serve their customers. Mobile technologies provide ability to communicate independent of time and location. Therefore, understanding mobile technologies and the process of transitioning the organization to a mobile organization is crucial to the success of adopting mobility in business. Such a process provides a robust basis for the organization's desire to reach a wide customer base. This chapter discusses the assessment of a business in the context of mobile technology, describes the application and limitations of mobile technology, presents a brief history of mobile technology and outlines an initial approach for transitioning an organization to a mobile organization.

INTRODUCTION

This chapter evaluates the effects of mobile technologies on business and outlines an initial process of transitioning to mobile business. In 1874, when Alexander Graham Bell invented the telephone, he could not have imagined the significant impact it would have on number of aspects of human life. Similarly today, advancement in information and communication technology (ICT) has dramatically changed the way people live and conduct their businesses. One of the dramatic aspects of modern-day business activities is that these activities are conducted independent of location and time. For example, businesses are able to sell goods, facilitate customer enquiries, and coordinate their services through disparate

geographical and time boundaries primarily due to the wonders of communications technologies. Alter (1996) describes the ICT as tools for doing things, rather than just for monitoring performance of yesterday or last week. Thus it is quite logical to conclude that ICT has changed the very nature of the workplace.

The basis for the communications technologies in most modern business applications is the Internet. Increasingly, the required access and connection to the Internet has become very simple and ubiquitous in most developed nations. This Internet access has opened up opportunities for organizations to revolutionize their business processes. Undoubtedly, improvement of the communication technology has impacted not only our business domain, but also our socio-cultural domain. This, as per Unhelkar (2004), has resulted in the "next wave" of technologies called mobile technologies:

Mobile technologies are becoming the next technology wave as the increasing popularity and the functionality captures many hearts. Riding on the back of traditional Internet, mobile networks ensure that information is available to its users independent of a physical location.

This ubiquitous connectivity accorded by mobile networks referred to above impact has facilitated the increased communication between people. Furthermore, this mobile connectivity has also improved the ability of business processes to exchange data and conduct transactions.

This transformation of businesses has been evolutionary rather than revolutionary. For example, at the beginning of the Internet age, with the aid of its communications capabilities, businesses were transferred to e-business, and we even had the opportunity to do our daily business activities from home. Ghanbary (2003) has described the Internet as the most powerful tool that brings information to our homes through communication lines, like water and electricity that come by power lines and pipes. Powerful search engines and the capability of sharing information are the great advantages of the Internet.

With the aforementioned strengths of mobile connectivity, it is also essential to work out a process that would outline "how" an organization can transition to such an m-enabled organization. However, this potential process of transitioning an organization to a mobile organization needs to incorporate all the major advances of mobile technologies of the past decade.

This is so because the philosophy of ordinary communication has given way to more advanced and efficient communications based on mobile and wireless technologies that enable business processes to be executed independent of time and location, resulting in a better, faster, and satisfactory response to the needs of the customer.

This impact of mobility is an important element of the mobile transition process that is felt at both business and personal levels. However, as of today, this process framework remains a challenge that needs to be further researched to enable businesses to transition successfully. This need for further investigation is also ratified by Ranjbar (2002), who correctly mentions: "It is not always possible to foresee all the implications of a new technology until it is adopted by the mass of population and used for a relatively long time."

With the increase in the number of mobile organizations, the service providers realize that they need to identify their strengths as well as their weaknesses in terms of providing mobile services that provide solutions as well as rectify the shortcomings. The analyses of the weaknesses and strengths will give them an advantage to provide a convenient service and increase their customer loyalty.

BACKGROUND OF MOBILE TECHNOLOGY

The known mobile phones used today are the extension of American mobile radiotelephone. However, the distinction between such phones and two-way radios is not clearly known. The advancement on mobile technology and the relevant gadgets have been very moderate due to the limitation of this technology and the government regulations on radio transmission. The major concern of the American Communication Commission was to decide who would get what frequencies and which emergency service should have the priority on air for transmitting first.

The ordinary usage of mobile took place in the 1980s. The first generation of cellular mobile phones was used widely only by transmitting analogue signals. The large size and very high prices of mobile devices in addition to the cost of the calls are the well-known facts of this period.

In the 1990s, the second generation of mobile phones was available in the market using advanced digital technology. Very fast signals, cheaper calls and handsets, more reliable services, and the smaller handset devices are the characteristics of this period.

The uncontrollable growth of mobile technology has created new culture in the business world. The use of mobile and Internet technology has passed their boundaries to a business and social revolution. The new technology has capabilities of text, voice, and videoconferences using wireless devices as well as the ability to connect to the World Wide Web.

According to Conners and Conners (2004), the application and the use of mobile and Internet technologies are to organize people into various common interest groups. These groups can vary from harmless fun to serious military or political operations.

Information and communication technology has enlightened the business activities. The use of mobile devices is going to be another crucial factor to remain in the competitive market. Vaghijiani and Teoh (2005) explain this phenomenon of mobile technologies as an enormous opportunity for players up and down the value chain, from the device suppliers to carriers to the end users.

BACKGROUND TO A TRANSITION FRAMEWORK

New process frameworks need to build on existing work on process transitions. Electronic transitions have been studied and experimented by Ginige et al. (2002). In the electronic transitions, there has been ample focus on the effect of a dynamic environment and the rapidly evolving technology on organizations. Undoubtedly, these changes cause organizations to restructure and would introduce a new suite of business processes for them to enable them to remain in the market as well as grow by dealing with a greater number of customers.

The new business model and the use of technology in the development of these changes were the cause of the new term e-business. The term e-business might mean trade on the Internet for some managers, however it is looking at the facts in deeper methodology. The Australian e-business guide (Philipson, 2001) translates e-business as any business transaction or activity that uses the Internet. This includes not only the sale of goods and services directly over the Internet, but also the use of the Internet to promote and facilitate the sale of goods and services.

By using mobile phones or any other mobile devices, we are able to make our e-business model more accessible. The improvement and the efficiency will create more benefits, hence the productivity remains even when people are out of their offices.

The proper design of the e-business model is a necessary component for the success of the m-business model. M-business makes the practice of the e-business model easier, more effective,

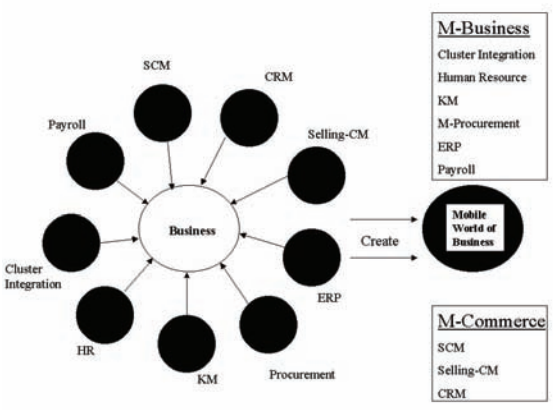
and more profitable since there is no restriction on approaching the required data. The share of internal data, providing better and more reliable customer service and better control over the organization in general, are other major benefits of mobile business.

Figure 1 shows the re-engineered and mobilized individual enterprise enabling the transformation of an ordinary business to a new and modern world of mobile business. The business and commerce sections are given in different boxes that provide a more comprehensive study of the transformation.

Figure 1 represents electronic mobile business activities in more detail.

- **Clusters Integration:** Our mobile business model must be able to integrate all the clusters of the organizations. This fact might look easy, however by careful analysis it is realised that to connect all the clusters of the organizations is a huge task. It needs more than technological advancement since some clusters might hesitate to share their information.
- **Customer Relationship Management:** The organizations must create close relationships with their customers. Customers want reliable and fast service. If the organizations provide them with what they want, they gain more business, and more business basically means more revenue. The CRM could be classified as a crucial factor of a mobile business, as it is in direct contact with people who are outside of the company. These people practically do not care how things are running as long as the great service is provided to them. A combination of technology, software, people, and re-engineered business processes are the fundamental of great customer relationship management.
- **Selling Change Management:** The right mobile business model should provide information about the available product directly to the customers. Direct interaction with the customers eliminates the retailer, and this will enable the business officials to provide detailed information about the product to the clients. As there are fewer hands involved in a purchase order, the prices offered to the customers will be dramatically low.
- **Enterprise Resource Planning:** The organizations have realised the importance of having the knowledge about back-office systems which could improve their customer order, integration of their clusters, and provide them with more sufficient information on how to run day-to-day activities by reducing the cost.
- **Supply Chain Management:** The technology is enabling the organizations to eliminate unnecessary processes to save time as well as money. The supply chain management of a business is the plan for materials to be directed to the customers as quickly as possible by cutting the inessential retailers. The information about the delivery, financial matters, as well as order transmissions are provided on any mobile gadgets. The mobile business has more power of monitoring and control over the order status.

Figure 1. Electronic mobile business model



- **Procurement:** As per Kalakota (1999), purchasing refers to the actual buying of materials and those activities associated with the buying process. Procurement on the other hand has a broader meaning and includes purchasing, transportation, warehousing, and inbound receiving.
  Organizations spend millions of dollars for procurement every year. Mobile procurement gives them the opportunity to have a better control on their inventory, better control over the purchase approval, and so on. They have better control over the cost as well as knowledge regarding their assets.
- **Human Resources:** The functionality of human resources in the mobile world could be classified as providing the title of available jobs on mobile devices and providing more sufficient information about the positions and required level of desired criteria on the Internet. Enterprise Bargaining Agreement online and other responsibilities of HR could be done while using the new technologies.
- **Payroll:** The available technology is giving the opportunity to advance the payroll that saves a great deal of time and human resources. This strategy is more beneficial to the organizations with odd clusters such as shift workers. If their roster is also automated, the payroll office could automatically generate a payroll list based on the automated roster, while only a human supervisor is required.
- **Knowledge Management:** In today's competitive business world, knowledge plays the crucial rule. The knowledge must be reliable and accessible through all clusters of the organization anywhere and anytime. Knowledge created by the supplier is about available services and products, and knowledge about the users is in the form of profiles. Imagine if a customer calls the engineering department of his electricity company to get the approval for the extension of his house. At the end the inquiry, he asks for the amount of his bill; the system should be able to provide the necessary information rather than transferring his call to another operator. The available information should support the day-to-day running of the business.

By the aid of the emerging mobile technology and re-engineering the individual departments of the organization, the new mobile organization is created. In general, the purposes of the value-added services are to impress customers and create more control over the business. By replacing business with m-business, we can reduce the cost and create more revenue by having more satisfied customers.

APPLYING MOBILE TECHNOLOGIES

By correct application of mobile technologies into the business processes, the business enterprises are likely to gain advantages such as increased profits, satisfied customers, and greater customer loyalty. These customer-related advantages will accrue only when the organization investigates its customer behaviour in the context of the mobile environment.

In general, the application of mobile technologies could be classified in two different categories, online and off-line services. The applications of online services are the executed applications when the mobile gadgets are connected to the mobile Internet.

The applications of the online wireless mobile Internet depends on situation (walking, driving), place (remote area, city metropolitan area), goal (aim of the connection), immediacy (instant action and reaction to demand), and load (how occupied the Internet provider is at the time of the connection). The congestion of the network clearly depends on the usage, which varies at different

times of the day. As expected during the business hours, the network load on the Internet service provider is heavy.

The major online mobile applications are:

- **Information:** General information about movies. Location of cinemas, hotels, hospitals, and universities. News, sports, travel, weather, and financial information.
- **E-Mail:** To send and receive mail while online using the mobile handheld.
- **Payment:** To buy the product and receive the service and pay by your mobile device and receive the payment on your mobile bill. There is trusted third party in mobile commerce regarding billing inquiries that increase the cost since another party is involved.
- **Mobile Internet Banking:** To complete the banking transaction using your mobile device while online. The participating banks decide what kind of transaction is allowed and how they provide the security for their clients.
- **Mobile Internet Shopping:** To shop online using the mobile gadgets. There are advantages and disadvantages in mobile shopping since the participants are not able to touch or smell the items they are buying unless it is first sale vs. the repeated sale.
- **Education:** Using mobile handset to download lectures, use library facilities (order book, search the library), and use laboratory.
- **Government Applications:** Election, government bulletin and broadcasting, disaster information system with the aid of location-based services, and automatically giving the priority to the broadcast.
- **Communication:** Videoconferencing, telephony, sending and receiving pictures, and international communication.
- **Leisure:** Download music, video, TV, and games, and for some particular people gambling could be classified as a leisure activity.
- **Telemetric:** Location-based services, global positioning services, and car navigation systems. Telemetric applications could be expanded to give the opportunity to your device to book a hotel room, purchase a ticket, gather your required information, and any related scenario just by a click of the button or voice order, assuming your personal mobile gadget is already holding all your personal and credit card details. These functions could be performed by connecting to your mobile Internet or just by connecting to your network provider.
- **Advertising:** Conjunction of mobile and Internet technology for advertising. Receiving an advertisement on the Internet (pop-up screens) is not something new; furthermore marketing organizations could use the same idea for mobile Internet.

The off-line applications are the services offered by related network providers and extra available features on the particular mobile devices. Networks must have infrastructure to support the fast transmission of the data, reliability of the data, the integrity of the data, and quality of service. The major off-line mobile applications are:

- **Communication:** Phone calls, SMS, messages, sending and receiving pictures.
- **Memory:** Phone book, music, different sounds, different effects (vibrate, volume), display, storing desired pictures and schedules and entertainment.
- **Expert System:** It would be possible for the organizations to use their mobile device as an expert system if their network provider has the capability to support it.
- **Remote Supervision:** To have control over the personnel while they are in an inaccessible area.

- **Traffic Information System:** Informing the drivers of traffic locations when they are driving close to the congested area.
- **M-Newspaper:** Subscribing to a newspaper if provided by the newspaper agency.
- **Advertising:** Based on the device's position, receiving the local advertisement.

Consumers' demands and corporate objectives could be different in the m-enabled world. While the application remains the same, expectation and usage are different. Usage in an m-enabled society is classified in three categories of interaction (voice, e-mail, chat, digital postcards, etc.), trading and business (banking, shopping, auctions, advertising, ticketing, etc.), and mobile-provided services (news, entertainment, driving direction, and much more).

It is very important to identify the sufficient information that is required to make the purchase decisions while the organizations are re-engineering their business processes. Ease of navigation and necessary links to other related Web sites are crucial factors for the software developers to consider while they are designing the new applications.

LIMITATION OF MOBILE TECHNOLOGY

Rising customer expectations have a direct connection to the advancement of technology. People's demand of the technology has not always been so realistic. The word "technology" has constantly fascinated human beings. Information and communication as the defining technology of the modern era have increased the expectations to an irrationally higher level. As Toffler (1980) predicted, people's dependence on technology has increased to a high level where technology has affected every aspect of human life. Mobile technology, which is an integration of communication and computer technology, has created such expectations in human behaviour that people cannot think of an era without such technology today.

It is clear that people rely on technology even when technology does not have the capability, or it is not robust enough, to support their task. There is no guarantee that I will not lose my work while writing this chapter on my computer, and the very same technology is used when human lives are involved. As an example, computers are used to take off and land airplanes.

However, mobile technology has its own characteristics and limitations which should be clearly identifiable to business enterprises. Of course these limitations will increase when mobile devices are connected to the mobile Internet.

Jamalipour (2003) explains that the access to the wireless mobile Internet is not just an extension of the Internet into the mobile environment giving users access to the Internet while on the move. However, it is about integrating the Internet and telecommunications technologies into a single system that covers all communication needs of people. He also believes that current network architectures used in either the wired Internet or the cellular networks would not be appropriate and efficient for future wireless mobile Internet, even if we assume that the cellular network will provide the major infrastructure of the mobile Internet. He concludes by saying that access to the mobile Internet is slow, expensive, and confusing.

Some limitations of mobile technology are as follows:

- **Cost:** The cost of restructuring the organization and personal devices.
- **Call Drops:** Disconnection while taking or downloading the data.
- **Connectivity:** Constant connectivity to a network is a big issue for mobile network providers. There are improvements in this area, but it should be taken into consideration when we are mobilizing the enterprise that

the network must support the expected assignment of the enterprise.

- **Lost Work:** Losing the performed work due to disconnection or dead battery.
- **Managing Technology:** Consistent maintenance of the software and the hardware.
- **Security:** Payment online, user behaviour, rules and hassles, mobile virus protection, file encryption, access control, and authentication are the most important security factors in the mobile environment, considering that mobile devices are very personal; in case of loss or theft, who is accessing the corporate or personal data?
- **Integrity:** The transmitted data is actually going to the expected individual. The message received is actually the message sent, and also the sender is the real owner of the mobile handset.
- **Privacy:** Who is accessing the corporate database and where personal details of the individuals are involved.
- **Regulations:** Government roles and regulations regarding the mobile matters.
- **Standardization:** Technical standards and compatibility of the users (business-to-business, business-to-customer).
- **Health Hazards:** By encouraging people to use mobile gadgets, are we jeopardizing their health?
- **Data Transmission Speed:** Slow transmission is very costly and ineffective.
- **Coverage:** The coverage of the network in a remote area is an identified and unresolved problem.
- **Adaptation:** Some people are resistant towards technology, and it would take time for them to adapt to the new technology.
- **Training:** The cost of training, managing mobile workforce, and controlling their activity.
- **Marketing Issues:** It would be a new era of marketing issues in mobile age such as sex, age, and so on.
- **Social Aspects:** Technology is creating a new pressure for ordinary people. People resistant to changes should be the major concern while planning the mobile transformation. Perpetual contact is another issue that is changing the face of our society. Mobile users are communicating with some other person while driving, walking on the street, and when they are in different public places. It is becoming commonly acceptable in our society to give priority to the mobile caller even when personal face-to-face conversation is getting disrupted.

Limited processing powers of handsets' microprocessors, memory size, battery life, small screen of handheld devices and their resolution, replacement costs, required ongoing support, network charges, as well as mobile Internet charges and enhancements are the other critical shortcomings of the mobile technology.

TRANSITIONING TO A MOBILE ORGANIZATION

The transformation of the organization by introducing the new and re-engineered processes is a very crucial matter. Should the enterprise revolutionize and transform as soon as possible or use the evolutionary process? Should the enterprise adapt to the new technology as soon as it is available or delay the process to see the outcome by using another organization's experience?

Considering there is no suitable answer for the above questions, there may be a need for a new approach to clarify this uncertainty. This new approach is supposed to be the combination of the revolution and evolution—revolution since the organization should not fall behind by remaining competitive in the market, and evolution to reduce the risk of not having a successful e-transformation. According to Murugeson and Deshpande

(2001), the development of an organization could be classified as the following:

The choice of a suitable development model, according to practitioners and researchers, is site (and applications), its document orientation, content and graphic design, budget and time constrains and the changing technology.

It could be quite risky if the organisations adapt to the new technology as soon as it is available in the market, as the system is definitely unknown and there might be hundreds of unresolved issues as well. Another factor at an early stage entry is the high cost involved in the introductory level of the new technology.

However, if they do not adapt during a specific period of time and their competitors do, there is a great chance of not being able to catch up with the advancements in the professional world. These are some issues that management is facing today. Their crucial decision making will determine the future of their companies. Serour (2005) clarifies that senior and middle management find it hard to proceed when there is (still) very little guidance available from real-world experience.

The organizations must allow internal and external parties involved to know that there are some changes that need to take place. In view of people's resistance to change, this will give them some time to prepare and adjust. The core of the training is for internal parties of the organization; however it is very important to provide sufficient information to external parties and advise them about the change.

The organizations must plan and manage change (cultural, technological, internal, and external) and understand the key areas associated to dangers related to their working environment that others have discovered and faced. The Australian e-business guide (Philipson, 2001) describes that implementations for e-business initiatives must be rapid and each project should be delivered in a maximum of three months. Build quickly and move to the learning stage, then build the next stage and fix the previous ones based on what you have learned.

Management must support the variation in business and market strategies, organizational restructure, and management strategies. The corporations must prepare all the existing clusters ready for change. Managing the transformation by having a reliable and calculated plan is the crucial factor for success. The transition must remain persistent alongside with detailed knowledge of the development of the individual clusters. According to Brans (2003), generally mobile transition takes place by distinguishing what kind of portable devices, networks, application gateways, and enterprise applications are required.

The benefits of mobilizing the organizations are: quick sale, closer communication within the internal departments, more strength and opportunities and less weaknesses and threats, professional façade for the organization, quick and reliable generation of customer data, and mobility at work.

CONCLUSION AND FUTURE DIRECTIONS

This chapter described some characteristics of m-business and offered a brief background of mobile phone technology.

When the Internet was introduced, nobody could imagine that this tool was going to make the next paradigm shift in all human interaction as well as business transactions. M-business is enabling organizations to increase global productivity. With the aid of mobile technology, the capability exists to operate in a very modern and extraordinary manner. The problems faced in the transformation of an organization to m-organization were identified and some solutions were recommended.

The domain of this chapter was to explain the particulars of a mobile business model, em-

phasising their significance, application, and the shortcomings of this technology in the world of business and trade. It is hoped that this chapter could convince the developers of the m-applications to spend more time in their design to fulfil the needs of the end users.

However, there are some critical issues that are unresolved in the world of mobile technology. These issues can be classified as security (integrity and privacy), national/international regulation, international standardization, security and integrity of databases on mobile devices, managing mobile workers and coordination of their activities, and the consistent maintenance of mobile hardware and software.

To create a robust and reliable mobile world, the developers could consider some other shortcomings of mobile technology. These issues are mobile payments, health hazards, the cost of restructuring the organization, damage to the handheld devices, legal liability of handheld devices (since the mobile gadgets are very personal and can keep confidential data related to the organization), and constant connectivity of mobile devices.

REFERENCES

Alter, S. (1996). *Information systems. A management perspective*. Benjamin/Cummings.

Brans, P. (2003). *Mobilize your enterprise*. Pearson Education.

Conners, J., & Conners, S. (2004). The impact of mobile technology on business planning. *Proceedings of IRMA 2004*, New Orleans, LA.

Deshpande, Y., & Ginige, A. (2001). Corporate Web development: From process infancy to maturity. In S. Murugesan, & Y. Deshpande (Eds.), *Web engineering managing diversity and complexity of Web application development* (p. 36). Germany: Springer-Verlag.

Ghanbary, A. (2003). *Effects of computers on family and leisure time*. Honour Thesis, University of Western Sydney, Australia.

Ginige, A. (2002). New paradigm for developing evolutionary software to support business. In S. K. Chang (Ed.), *Handbook of software engineering and knowledge engineering* (Vol. 2). World Scientific.

Jamalipour, A. (2003). *Wireless mobile Internet: Architectures, protocols and services*. Hoboken, NJ: John Wiley & Sons.

Kalakota, R., & Robinson, M. (1999). *E-business roadmap for success*. Boston: Addison Wesley Longman.

Murugesan, S., & Deshpande, Y. (2001). *Web engineering (publication data)*. Berlin/Heidelberg: Springer-Verlag.

Philipson, G. (2001). *Australian e-business guide*. McPherson's Printing Group.

Ranjbar, M. (2002). *Social aspects of information technology*. Sydney, Australia: University of Western Sydney.

Serour, M. K. (2005). The organizational transformation process to globalization. In Y. Lan (Ed.), *Global information society: Operating information systems in a dynamic global business environment*. Hershey, PA: Idea Group Publishing.

Toffler, A. (1980). *The third wave*. William Morrow and Company.

Unhelkar, B. (2005). Web services and their impact in creating a domain shift in the process of globalization. In Y. Lan (Ed.), *Global information society: Operating information systems in a dynamic global business environment*. Hershey, PA: Idea Group Publishing.

Vaghjiani, K., & Teoh, J. (2005). Comprehensive impact of mobile technology on business. In Y. Lan (Ed.), *Global information society: Operating information systems in a dynamic global business environment.* Hershey, PA: Idea Group Publishing.

This work was previously published in Handbook of Research in Mobile Business, edited byB. Unhelkar, pp. 602-612, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.2
Knowledge Representation in Semantic Mobile Applications

Panjak Kamthan
Concordia University, Canada

INTRODUCTION

Mobile applications today face the challenges of increasing information, diversity of users and user contexts, and ever-increasing variations in mobile computing platforms. They need to continue being a successful business model for service providers and useful to their user community in the light of these challenges.

An appropriate representation of information is crucial for the agility, sustainability, and maintainability of the information architecture of mobile applications. This article discusses the potential of the Semantic Web (Hendler, Lassila, & Berners-Lee, 2001) framework to that regard.

The organization of the article is as follows. We first outline the background necessary for the discussion that follows and state our position. This is followed by the introduction of a knowledge representation framework for integrating Semantic Web and mobile applications, and we deal with both social prospects and technical concerns. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

In recent years, there has been a proliferation of affordable information devices such as a cellular phone, a personal digital assistant (PDA), or a pager that provide access to mobile applications. In a similar timeframe, the Semantic Web has recently emerged as an extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

The goal of the mobile Web is to be able to mimic the desktop Web as closely as possible, and an appropriate representation of information is central to its realization. This requires a

transition from the traditional approach of merely presentation to *representation* of information. The Semantic Webprovides one avenue towards that.

Indeed, the integration of Semantic Webtechnologies in mobile applications is suggested in Alesso and Smith (2002) and Lassila (2005). There are also proof-of-concept semantic mobile applications such as MyCampus (Gandon & Sadeh, 2004) and mSpace Mobile (Wilson, Russell, Smith, Owens, & Schraefel, 2005) serving a specific community. However, these initiatives are limited by one or more of the following factors: the discussion of knowledge representation is one-sided and focuses on specific technology(ies) or is not systematic, or the treatment is restricted to specific use cases. One of the purposes of this article is to address this gap.

UNDERSTANDING KNOWLEDGE REPRESENTATION IN SEMANTIC MOBILE APPLICATIONS

In this section, our discussion of semantic mobile applications is based on the knowledge representation framework given in Table 1.

The first column addresses semiotic levels. Semiotics (Stamper, 1992) is concerned with the use of symbols to convey knowledge. From a semiotics perspective, a representation can be viewed on six interrelated levels: physical, empirical, syntactic, semantic, pragmatic, and social, each depending on the previous one in that order. The physical level is concerned with the representation of signs in hardware and is not directly relevant here.

The second column corresponds to the Semantic Web"tower" that consists of a stack of technologies that vary across the technical to social spectrum as we move from bottom to top, respectively. The definition of each layer in this technology stack depends upon the layers beneath it.

Table 1. Knowledge representation tiers in a semantic mobile application

Semiotic Level	Semantic Mobile Web Concern and Technology Tier	Decision Support
Social	Trust	Feasibility
Pragmatic	Inferences	
Semantic	Metadata, Ontology, Rules	
Syntactic	Markup	
Empirical	Characters, Addressing, Transport	
Physical	Not Directly Applicable	

Finally, in the third column, we acknowledge that there are time, effort, and budgetary constraints on producing a representation and include feasibility as an all-encompassing factor on the layers to make the framework practical. For example, an organization may choose not to adopt a technically superior technology as it cannot afford training or processing tools available that meet the organization's quality expectations. For that, analytical hierarchy process (AHP) and quality function deployment (QFD) are commonly used techniques. Further discussion of this aspect is beyond the scope of the article.

The architecture of a semantic mobile application extends that of a traditional mobile application on the server-side by: (a) expressing information in a manner that focuses on *description* rather than presentation or processing of information, and (b) associating with it a knowledge management system (KMS) consisting of one or more domain-specific ontologies and a reasoner.

We now turn our attention to each of the levels in our framework for knowledge representation in semantic mobile applications.

Empirical Level of a Semantic Mobile Application

This layer is responsible for the communication properties of signs. Among the given choices, the

Unicode Standard provides a suitable basis for the signs themselves and is character-by-character equivalent to the ISO/IEC 10646 Standard Universal Character Set (UCS). Unicode is based on a large set of characters that are needed for supporting internationalization and special symbols. This is necessary for the aim of universality of mobile applications.

The characters must be uniquely identifiable and locatable, and thus addressable. The uniform resource identifier (URI), or its successor international resource identifier (IRI), serves that purpose.

Finally, we need a transport protocol such as the hypertext transfer protocol (HTTP) or the simple object access protocol (SOAP) to transmit data across networks. We note that these are limited to the transport between the mobile service provider that acts as the intermediary between the mobile client and the server. They are also layered on top of and/or used in conjunction with other protocols, such as those belonging to the Institute of Electrical and Electronics Engineers (IEEE) 802 hierarchy.

Syntactic Level of a Semantic Mobile Application

This layer is responsible for the formal or structural relations between signs. The eXtensible Markup Language (XML) lends a suitable syntactical basis for expressing information in a mobile application.

The XML is supported by a number of ancillary technologies that strengthen its capabilities. Among those, there are domain-specific XML-based markup languages that can be used for expressing information in a mobile application (Kamthan, 2001).

The eXtensible HyperText Markup Language (XHTML) is a recast of the HyperText Markup Language (HTML) in XML. XHTML Basic is the successor of compact HTML (cHTML) that is an initiative of the NTT DoCoMo, and of the Wireless Markup Language (WML) that is part of the wireless application protocol (WAP) architecture and an initiative of the Open Mobile Alliance (OMA). It uses XML for its syntax and HTML for its semantics. XHTML Basic has native support for elementary constructs for structuring information like paragraphs, lists, and so on. It could also be used as a placeholder for information fragments based on other languages, a role that makes it rather powerful in spite of being a small language.

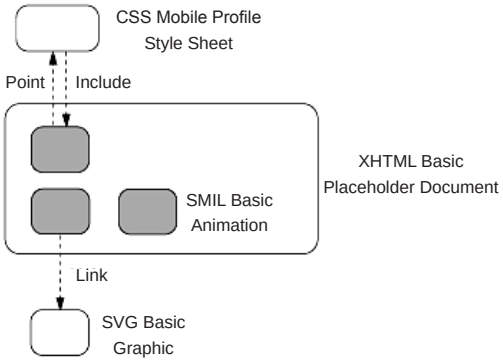
The Scalable Vector Graphics (SVG) is a language for two-dimensional vector graphics that works across platforms, across output resolutions, across color spaces, and across a range of available bandwidths; SVG Tiny and SVG Basic are profiles of SVG targeted towards cellular phones and PDAs, respectively.

The Synchronized Multimedia Integration Language (SMIL) is a language that allows description of temporal behavior of a multimedia presentation, associates hyperlinks with media objects, and describes the layout of the presentation on a screen. It includes reusable components that can allow integration of timing and synchronization into XHTML and into SVG. SMIL Basic is a profile that meets the needs of resource-constrained devices such as mobile phones and portable disc players.

Namespaces in XML is a mechanism for uniquely identifying XML elements and attributes of a markup language, thus making it possible to create heterogeneous (compound) documents (Figure 1) that unambiguously mix elements and attributes from multiple different XML document fragments.

Appropriate presentation on the user agent of information in a given modality is crucial. However, XML in itself (and by reference, the markup languages based on it) does not provide any special presentation semantics (such as fonts, horizontal and vertical layout, pagination, and so on) to the documents that make use of it. This is because the separation of the structure of a docu-

Figure 1. The architecture of a heterogeneous XML document for a mobile device



ment from its presentation is a design principle that supports maintainability of a mobile application. The cascading style sheets (CSS) provides the presentation semantics on the client, and CSS mobile profile is a subset of CSS tailored to the needs and constraints of mobile devices.

With the myriad of proliferating platforms, information created for one platform needs to be adapted for other platforms. The eXtensible Stylesheet Language Transformations (XSLT) is a style sheet language for transforming XML documents into other, including non-XML, documents. As an example, information represented in XML could be transformed on-demand using an XSLT style sheet into XHTML Basic or an SVG Tiny document, as appropriate, for presentation to users accessing a mobile portal via a mobile device.

Representing information in XML provides various advantages towards archival, retrieval, and processing. It is possible to down-transform and render a document on multiple devices via an XSLT transformation, without making substantial modifications to the original source document. However, XML is not suitable for completely representing the knowledge inherent in information resources. For example, XML by itself does not provide any specific mechanism for differentiating between homonyms or synonyms, does not have the capabilities to model complex relationships precisely, is not able to extract implicit knowledge (such as hidden dependencies), and can only provide limited reasoning and inference capabilities, if at all.

The combination of the layers until now forms the basis of the mobile Web. The next two layers extend that and are largely responsible for what could be termed as the semantic mobile Web.

Semantic Level of a Semantic Mobile Application

This layer is responsible for the relationship of signs to what they stand for. The resource description framework (RDF) is a language for metadata that provides a "bridge" between the syntactic and semantic layers. It, along with RDF Schema, provides elementary support for *classification* of information into classes, properties of classes, and means to model more complex relationships among classes than possible with XML only. In spite of their usefulness, RDF/RDF Schema suffer from limited representational capabilities and non-standard semantics. This motivates the need for additional expressivity of knowledge.

The declarative knowledge of a domain is often modeled using ontology, an explicit formal specification of a conceptualization that consists of a set of concepts in a domain and relations among

them (Gruber, 1993). By explicitly defining the relationships and constraints among the concepts in the universe of discourse, the *semantics* of a concept is constrained by restricting the number of possible interpretations of the concept.

In recent years, a number of initiatives for ontology specification languages for the semantic Web, with varying degrees of formality and target user communities, have been proposed, and the Web Ontology Language (OWL) has emerged as the successor. Specifically, we advocate that OWL DL, one of the sub-languages of OWL, is the most suitable among the currently available choices for representation of domain knowledge in mobile applications due to its compatibility with the architecture of the Web in general; and the Semantic Webin particular benefits from using XML/RDF/RDF Schema as its serialization syntax, its agreement with the Web standards for accessibility and internationalization, well-understood declarative semantics from its origins in description logics (DL) (Baader, McGuinness, Nardi, & Schneider, 2003), and provides the necessary balance between computational expressiveness and decidability.

Pragmatic Level of a Semantic Mobile Application

This layer is responsible for the relation of signs to interpreters. There are several advantages of an ontological representation. When information is expressed in a form that is oriented towards presentation, the traditional search engines usually return results based simply on a string match. This can be ameliorated in an ontological representation where the search is based on a *concept* match. An ontology also allows the logical means to distinguish between homonyms and synonyms, which could be exploited by a reasoner conforming to the language in which it is represented. For example, Java in the context of coffee is different from that in the context of an island, which in turn is different from the context of a programming language; therefore a search for one should not return results for other. Further, ontologies can be applied towards precise access of desirable information from mobile applications (Tsounis, Anagnostopoulos, & Hadjiefthymiades, 2004). Even though resources can be related to one another via a linking mechanism, such as the XML Linking Language (XLink), these links are merely structural constructs based on author discretion that do not carry any special semantics.

Explicit declaration of all knowledge is at times not cost effective as it increases the size of the knowledge base, which becomes rather challenging as the amount of information grows. However, an ontology with a suitable semantical basis can make implicit knowledge (such as hidden dependencies) *explicit.* A unique aspect of ontological representation based for instance on OWL DL is that it allows logical constraints that can be reasoned with and enables us to *derive* logical consequences—that is, facts not literally present in the ontology but *entailed* by the semantics.

We have a semantic mobile portal for tourist information. Let Mont Tremblant, Laurentides, and Québec be defined as regions, and the subRegionOf property between regions be declared as transitive in OWL (see Example 1.)

Then, an OWL reasoner should be able to derive that if Mont Tremblant is a sub-region of Laurentides, and Laurentides is a sub-region of Québec, then Mont Tremblant is also a sub-region of Québec. This would give a more complete set of search results to a semantic mobile application user.

In spite of its potential, ontological representation of information presents certain domain-specific and human-centric challenges (Kamthan & Pai, 2006). Query formulations to a reasoner for extracting information from an ontology can be rather lengthy input on a mobile device. It is currently also difficult both to provide a sound

Example 1. Ontological Inferences

```
<Region rdf:ID="MontTremblant">
 <subRegionOf rdf:resource="#Laurentides"/>
</Region>
<Region rdf:ID="Laurentides">
 <subRegionOf rdf:resource="#Qu&eacute;bec"/>
</Region>
<owl:TransitiveProperty rdf:ID="subRegionOf">
 <rdfs:domain rdf:resource="#Region"/>
 <rdfs:range rdf:resource="#Region"/>
</owl:TransitiveProperty>
```

Example 2. Device Profile

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:ccpp="http://www.w3.org/2002/11/08-ccpp-schema#"
 xmlns:prf="http://a.com/schema#">
. . .
 <ccpp:component>
 <rdf:Description rdf:about="http://a.com/HardwareDevice">
 <rdf:type rdf:resource="http://a.com/schema#HardwarePlatform"/>
 <ccpp:defaults rdf:resource="http://a.com/HardwareDefault"/>
 <prf:vendor>MyMobileCompany</prf:vendor>
 <prf:cpu>ABC</prf:cpu>
 <prf:displayHeight>200</prf:displayHeight>
 <prf:displayWidth>320</prf:displayWidth>
 <prf:memoryMb>16</prf:memoryMb>
 </rdf:Description>
 </ccpp:component>
. . .
</rdf:RDF>
```

logical basis to aesthetical, spatial/temporal, or uncertainty in knowledge, and represent that adequately in ontology.

Social Level of a Semantic Mobile Application

This layer is responsible for the manifestation of social interaction with respect to the representation. Specifically, ontological representations are a result of consensus, which in turn is built upon trust.

The client-side environment in a mobile context is constrained in many ways: devices often have restricted processing capability and limited user interface input/output facilities. The Composite Capabilities/Preference Profiles (CC/PP) Specification, layered on top of XML and RDF, allows

the expression of user (computing environment and personal) preferences, thereby informing the server side of the delivery context.

In Example 2, CC/PP markup for a device whose processor is of type ABC and the preferred default values of its display and memory as determined by its vendor are given. The namespace in XML is used to disambiguate elements/attributes that are native to CC/PP or RDF from those that are specific to the vendor vocabulary.

CC/PP can be used as a basis for introducing context-awareness in mobile applications (Sadeh, Chan, Van, Kwon, & Takizawa, 2003; Khushraj & Lassila, 2004).

One of the major challenges to the personalization based on profile mechanism is the user concern for privacy. The Platform for Privacy Preferences Project (P3P) allows the expression of privacy preferences of a user, which can be used by agents to decide if they have the permission to process certain content, and if so, how they should go about it. This ensures that users are informed about privacy policies of the mobile service providers before they release personal information. Thus, P3P provides a balance to the flexibility offered by the user profiles in CC/PP.

The Security Assertion Markup Language (SAML), XML Signature, and XML Encryption provide assurance of the sanctity of the message to processing agents. We note that an increasing number of languages to account for may place an unacceptable load, if it is to be processed exclusively, on the client side. We also acknowledge that these technologies alone will not solve the issue of trust, but when applied properly, could contribute towards it.

FUTURE TRENDS

The transition of the traditional mobile applications to semantic mobile applications is an important issue. The previous section has shown the amount of expertise and level of skills required for that. Although up-transformations are in general difficult, we anticipate that the move will be easier for the mobile applications that are well-structured in their current expression of information and in their conformance to the languages deployed.

The production of mobile applications, and by extension semantic mobile applications, is becoming increasingly complex and resource (time, effort) intensive. Therefore, a systematic and disciplined approach for their development, deployment, and maintenance, similar to that of Web engineering, is needed. Related to that, the issue of quality of represented and delivered information will continue to be important. The studies of specific attributes such as usability (Bertini, Catarci, Kimani, & Dix, 2005) and "best practices" for mobile applications from the World Wide Web Consortium (W3C) Mobile Web Initiative are efforts that could eventually be useful in an "engineering" approach for producing future semantic mobile applications.

The process of aggregation and inclusion of information in a mobile application is primarily manual, which can be both tedious and error prone. This process could be, at least partially, automated via the use of Web services where mobile applications could be made to automatically update themselves with (candidate) information. Therefore, manifestations of mobile applications through Semantic Webservices (Wagner & Paolucci, 2005; Wahlster, 2005) are part of a natural evolution.

CONCLUSION

For mobile applications to continue to provide a high quality-of-service (QoS) to their user community, their information architecture must be evolvable. The incorporation of Semantic Webtechnologies can be much more helpful in that regard. The adoption of these technologies does not have to be an "all or nothing" proposition: the evolution of a mobile application to a

semantic mobile application could be gradual, transcending from one layer to another in the aforementioned framework. In the long term, the benefits of transition outweigh the costs.

Ontologies form one of the most important layers in semantic mobile applications, and ontological representations have certain distinct advantages over other means of representing knowledge. However, engineering an ontology is a resource-intensive process, and an ontology is only as useful as the inferences (conclusions) that can be drawn from it.

To be successful, semantic mobile applications must align themselves to the Semantic Webvision of inclusiveness for all. For that, the semiotic quality of representations, particularly that of ontologies, must be systematically assured and evaluated.

REFERENCES

Alesso, H. P., & Smith, C. F. (2002). *The intelligent wireless Web.* Boston: Addison-Wesley.

Baader, F., McGuinness, D., Nardi, D., & Schneider, P. P. (2003). *The description logic handbook: Theory, implementation and applications.* Cambridge University Press.

Bertini, E., Catarci, T., Kimani, S., & Dix, A. (2005). A review of standard usability principles in the context of mobile computing. *Studies in Communication Sciences, 1*(5), 111-126.

Gandon, F. L., & Sadeh, N. M. (2004, June 1-3). Context-awareness, privacy and mobile access: A Web semantic and multiagent approach. *Proceedings of the 1st French-Speaking Conference on Mobility and Ubiquity Computing* (pp. 123-130), Nice, France.

Gruber, T.R. (1993). Toward principles for the design of ontologies used for knowledge sharing. *Formal ontology in conceptual analysis and knowledge representation*. Kluwer Academic.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The semantic Web. *Scientific American, 284*(5), 34-43.

Kamthan, P. (2001, March 20-22). Markup languages and mobile commerce: Towards business data omnipresence. *Proceedings of the WEB@ TEK 2001 Conference,* Québec City, Canada.

Kamthan, P., & Pai, H.-I. (2006, May 21-24). Human-centric challenges in ontology engineering for the Semantic Web: A perspective from patterns ontology. *Proceedings of the 17th Annual Information Resources Management Association International Conference (IRMA 2006),* Washington, DC.

Khushraj, D., & Lassila, O. (2004, November 7). CALI: Context Awareness via Logical Inference. *Proceedings of the Workshop on Semantic Web Technology for Mobile and Ubiquitous Applications,* Hiroshima, Japan.

Lassila, O. (2005, August 25-27). Using the Semantic Web in ubiquitous and mobile computing. *Proceedings of the 1st International IFIP/WG 12.5 Working Conference on Industrial Applications of the Semantic Web (IASW 2005),* Jyväskylä, Finland.

Sadeh, N. M., Chan, T.-C., Van, L., Kwon, O., & Takizawa, K. (2003, June 9-12). A Semantic Web environment for context-aware m-commerce. *Proceedings of the 4th ACM Conference on Electronic Commerce* (pp. 268-269), San Diego, CA.

Stamper, R. (1992, October 5-8). Signs, organizations, norms and information systems. *Proceedings of the 3rd Australian Conference on Information Systems* (pp. 21-55), Wollongong, Australia.

Tsounis, A., Anagnostopoulos, C., & Hadjiefthymiades, S. (2004, September 13). The role of Semantic Web and ontologies in pervasive computing environments. *Proceedings of the Workshop on Mobile and Ubiquitous Information*

Access (MUIA 2004), Glasgow, Scotland.

Wagner, M., & Paolucci, M. (2005, June 9-10). Enabling personal mobile applications through Semantic Web services. *Proceedings of the W3C Workshop on Frameworks for Semantics in Web Services,* Innsbruck, Austria.

Wahlster, W. (2005, June 3). Mobile interfaces to intelligent information services: Two converging megatrends. *Proceedings of the MINDS Symposium,* Berlin, Germany.

Wilson, M., Russell, A., Smith, D. A., Owens, A., & Schraefel, M. C. (2005, November 7). mSpace mobile: A mobile application for the Semantic Web. *Proceedings of the 2nd International Workshop on Interaction Design and the Semantic Web,* Galway, Ireland.

KEY TERMS

Delivery Context: A set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user, and other aspects of the context into which a resource is to be delivered.

Knowledge Representation: The study of how knowledge about the world can be represented and the kinds of reasoning can be carried out with that knowledge.

Ontology: An explicit formal specification of a conceptualization that consists of a set of terms in a domain and relations among them.

Personalization: A strategy that enables delivery that is customized to the user and user's environment.

Semantic Web: An extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

Semiotics: The field of study of signs and their representations.

User Profile: An information container describing user needs, goals, and preferences.

This work was previously published inEncyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 375-380,

Chapter 3.3
Mobile Portal Technologies and Business Models

David Parsons
Massey University, New Zealand

INTRODUCTION

Mobile portals have become a common entry point to the mobile Internet, and take a number of forms. They may be service provider portals, such as Vodafone's Live! portal (Vodafone, 2006), offering access to both in-house and brokered external services. Alternatively, they may be public pure play sites that provide some kind of managed access to resources using a yellow-pages approach. Good examples of this kind of mobile portal are WordDial (WordDial, 2006) and graBBit (Grabbit, 2006), though they have very different approaches to the way that they provide targeted access to resources, with WordDial using a keyword approach and graBBit modeled on more traditional search engines. As well as mobile and pure play operators, mobile portals are also provided by device manufacturers (e.g., Palm (Palm, 2006)), software companies (e.g., MSN (Microsoft, 2006)) existing Web portal providers (e.g., Yahoo (Yahoo, 2006)), mass media companies (e.g., AOL (AOL, 2006)) and transaction providers (m-commerce sites).

MOBILE PORTAL ADVANTAGES

The advantages that mobile portals have over standard Web portals are in ubiquity, convenience, localization, and personalization. Ubiquity means that the portal can be accessed anywhere, regardless of location. With ever widening coverage by mobile network providers, mobile portals have an increasingly ubiquitous presence. Availability at all times, via mobile devices, provides for convenience, with the ability for users to access portals at the point of need, for example to get up to date information on flight times or traffic conditions. Wireless connectivity is integrated into the mobile phone, whereas alternative ways of connecting to the Internet while traveling, such as accessing

wireless or fixed networks, or using publicly available computers, can be difficult and/or expensive to access in many locations. Localization is a specific strength of mobile portals, since they can use location awareness to provide services that are targeted to the user's current locality (e.g., local weather). Location awareness can be supported by a number of technologies, including triangulation from a mobile phone network or the satellite based global positioning system (GPS). Finally, personalization is a key component of mobile portals for two reasons. First, the difficulty of navigation and the small screen size of mobile devices means that it is important to target Web-based material as much as possible. Second, such targeting is easier for subscription type services that are common with mobile phone contracts, where the carrier is likely to be able to gather considerable information about users and construct accurate profiles of their activities and requirements. All of these characteristics are important features in the potential for mobile commerce, which relies on giving the best value-for-time service. Portals that are easily customizable, technically flexible, and contain relevant content are those that are most likely to be successful tools for mobile commerce (Clarke, Flaherty, & Madison, 2003).

MOBILE PORTAL TECHNOLOGIES

The technology of mobile portals is evolving as mobile devices become more sophisticated. Early portals were based on the wireless access protocol (WAP) version 1.0, using the Wireless Markup Language (WML) with very limited user interface features and severe limits on the type of content that could be accessed. In many cases, content was based on a transformation from HyperText Markup Language (HTML) pages, designed for standard Web browsers, into WML pages. These conversions, performed by WAP gateways that linked the mobile device network to the wider Internet, were slow and the content was not optimized for mobile users. Current WAP-based portals take advantage of the improvements in WAP technology that were introduced with version 2.0 (e.g., WAP push and end-to-end security) and more powerful handsets to provide richer interaction and media types. In addition, content is more likely to be tailored especially for mobile devices rather than being converted from HTML, developed either directly in WML or in XHTML-MP (eXtensible HyperText Markup Language – Mobile Profile) which is the evolutionary pathway from WML and is now the recommended markup language for mobile Internet domains (Cremin & Rabin, 2006).

Portals that were developed in the context of second generation (2G) mobile phone networks suffered from slow connection speeds, limiting the range of contents that could be provided. Portals running over third generation (3G) networks benefit from much faster data transfer speeds, so they can deliver rich multimedia content, such as TV and movie feeds and MP3 downloads. However, despite the market dominance of entertainment content, with the huge popularity of ring tones and screen savers, mobile portal services are not limited to entertainment alone. Some portals also host location based services, for example the provision of MapPoint access via the Vodafone portal in certain territories, and portal-hosted M-Payment services are increasingly popular.

DESIGN ASPECTS OF MOBILE PORTALS

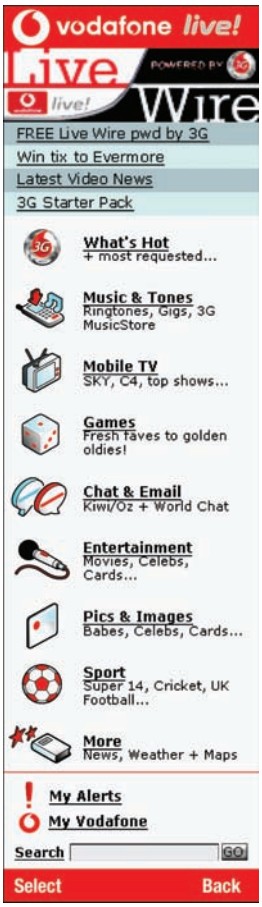
Mobile portals have had to be designed to provide the easiest access to services within the usual constraints of mobile devices, such as limited screen space, varying navigation button layouts on phones from different manufacturers, and lack of a consistent programming platform. Unlike portals designed for the desktop that are usually based around table-like structures containing separate portlets, mobile portals are structured

around nested menu lists, often with images, that provide quick scrolling access to services. The initial WAP portal pioneered by Vodafone Live! exemplified the typical style for mobile portals, with a brand header followed by a list of headlines that lead to other pages. Figure 1 shows a top level menu page from Vodafone Live! Although this type of mobile portal design has become a little more sophisticated over time with the move towards larger screens and XHTML-MP markup, the basic principles of using brief headline links and/or small images still apply.

Typical top-level mobile portal menus contain links to services such as news, weather, TV, downloads (games, ring tones, screen savers) and search engines. Mobile portals are not, however, only designed for one way services. One of the more unique features of a mobile portal is the ability to register for alerts, sent via SMS or using push technologies.

Because of the difficulties of configuring connections to the mobile Internet and managing page navigation with limited control keys, mobile carriers have worked with handset manufacturers to provide branded phones that include single key access to the carrier's mobile portal. This makes it easier to access the carrier's own portal but harder to access other portals.

Figure 1. The Vodafone Live! mobile portal (image courtesy of Vodafone New Zealand Ltd.)



BUSINESS MODELS FOR MOBILE PORTALS

There are three basic business models for mobile portals, which may be used in combination. Either they are based on subscription, payment for individual services or advertising. The role of the mobile network operator in the m-commerce value chain will vary between contexts, but at the most active level the operator will provide the network, the WAP gateway, the mobile portal and also act as an intermediary and trusted third party between the customer and other content and service providers (Tsalgatidou & Veijalainen, 2000).

The first generation of mobile portals, introduced in the late 1990s, had limited success due to factors including cost, limited browser capability and slow transmission speed. However in Japan, NTT DoCoMo's subscription-based I-mode portal showed that it was possible to achieve success in the mobile portal market by developing a large customer base built using youth targeted branding, low costs and suitable technology (CNET News, 2001). A key aspect of success in Japan, as opposed to early failure in Europe, was that DoCoMo successfully integrated the three value chains that comprise mobile telecommunications, the devices, the infrastructure, and the services (Sigurdson, 2001). More recent success outside Japan has been based on integrating these three components, via Web portals, that link devices to carriers by building portal access into their menus and brokering services from other providers.

Mobile portals have been an important revenue generator for mobile phone network providers because they have been the main driver for use of data services by personal, as opposed to corporate, users. For example, UK figures provided on a regular basis by the Mobile Data Association show that WAP page impressions (i.e., requests for one or more WML files that construct a single page) have increased hugely since 2002, when the first UK mobile portals were introduced, from about 200 million per month to nearly 2 billion by the end of 2005 (Mobile Data Association, 2006).

Portals provided by network providers sometimes use a walled garden approach to browsable content, which integrates third party content. In many cases, this content has to be paid for. Access to the portal is built into phones provided by the carrier, making access easy, but locking the user into one point of access to the mobile internet. From the user's perspective, the walled garden is useful in that the control of content means that all content will be appropriate to the mobile device. However, it limits the user's ability to browse the internet more widely. On many devices, although it is possible to do so it is much more difficult to set up than using the built in portal. As an alternative approach, some carriers simply provide direct access to the Web via a specific home page, such as T-Mobile's use of the Google home page (Mobile Pipeline, 2005).

FUTURE TRENDS IN MOBILE PORTALS

Beyond the current WAP generation, future mobile portals will take advantage of smart phone and Java Micro Edition devices to deliver more sophisticated content and interactivity, using dynamically loaded applications and leveraging XHTML-MP markup as the common evolution path from WAP, cHTML and XML. To enable two way interaction between users and portal providers, many portals include push elements, enabling alerts to be sent to users based on their user profiles, and increasingly, Podcasts will be integrated into mobile portals to enable more sophisticated push content (Lewin, 2005). As mobile devices evolve from WML based markup to XHTML-MP, and screen size and resolution increases, there will be less distinction between pages designed for the Web in general and those designed specifically for mobile devices. The

distinction between mobile and Web portals will blur, and eventually the distinction between them may well fade way almost altogether. In the interim, with the increasing number of portals available, and the increasing flexibility of devices, it is unlikely that providers will be able to sustain purely walled garden approaches. Rather, they will need to use their branded sites to provide unique content through their partners, and leverage the usability advantages of customized handsets, in order to retain users.

As devices and networks evolve, portal providers will have to adapt to changing technologies and markets. There will, however, still be significant differences in content provision between mobile portals and the rest of the Internet, because of the value added services that are possible through localization and personalization. Because of this, even when the mobile portal ceases to exist as a separate entity, Web portals will still include some elements that are unique to the mobile user.

CONCLUSION

Mobile portals have been an important component of the mobile Internet, providing mobile users with easier access to Web-based resources and enabling service providers to provide targeted content. Partnerships between network carriers and mobile device manufacturers are an important part of the business strategy of many mobile portals, enabling a walled garden approach that manages the user's Internet access. Early mobile portals had to be developed in the context of the limited form factor of WAP phones and restrictions on connection availability and speed. With the development of mobile phones with bigger, better screens (full color, high resolution, etc.) and high speed data connections, mobile portals have become both more sophisticated in the user interface and able to deliver a wider range of content.

REFERENCES

AOL. (2006). *AOL Mobile Portal*. Retrieved January 31, 2006, from http://aolmobile.aol.com/portal/

Clarke, I., Flaherty, T., & Madison, J. (2003). Mobile portals: The development of m-commerce. In B. Mennecke & T. Strader (Eds.), *Mobile commerce: Technology, theory and applications* (pp. 185-201). Hershey, PA: IRM Press.

CNET News. (2001). *Wireless Web portals duke it out*. Retrieved March 2006, from http://news.com.com/Wireless+Web+portals+duke+it+out/2009-1033_3-255977.html?tag=st.num

Cremin, R., & Rabin, J. (2006). *dotmobi switch on! Web browsing guide*. Retrieved March 2006, from http://pc.mtld.mobi/documents/dotmobi-Switch-On!-Web-Browsing-Guide.html

Grabbit. (2006). *Grabbit*. Retrieved January 31, 2006, from http://www.grabbit.co.nz/

Lewin, J. (2005). *Podcasting emerges as an ebusiness tool*. Retrieved January 31, 2006, from http://smallbusiness.itworld.com/4427/nls_ecommercepod050601/page_1.html

Microsoft. (2006). *MSN Mobile*. Retrieved January 31, 2006, from http://mobile.msn.com/

Mobile Data Association. (2006). *Mobile Data Association home page*. Retrieved January 26, 2006, from http://www.mda-mobiledata.org/mda/

Mobile Pipeline. (2005). *T-Mobile to use Google as mobile portal, dumps 'walled garden.'* Retrieved January 26, 2006, from http://informationweek.com/story/showArticle.jhtml?articleID=164903968

Palm. (2006). *Palm Mobile portal*. Retrieved January 31, 2006, from http://mobile.palmone.com/

Sigurdson, J. (2001). *WAP OFF—Origin, failure and future*. Retrieved January 26, 2006, from

http://www.telecomvisions.com/articles/pdf/wap-off.pdf

Tsalgatidou, A., & Veijalainen, J. (2000, September 4-6). Mobile electronic commerce: Emerging issues. In *EC-WEB 2000, 1st International Conference on E-Commerce and Web Technologies*, Greenwhich, UK (LNCS 1875, pp. 477-486). London: Springer.

Vodafone. (2006). *Vodafone Live!* Retrieved January 31, 2006, from http://www.vodafone.co.nz/vlive/vlive.jsp

WordDial. (2006). *WordDial home page*. Retrieved January 31, 2006, from http://www.worddial.com/

Yahoo. (2006). *Yahoo Mobile*. Retrieved January 31, 2006, from http://mobile.yahoo.com/

KEY TERMS

Global Positioning System (GPS): A network of satellites that enables ground based devices to acquire their latitude, longitude and altitude. Since line of sight is required to four satellites for accurate positioning, availability and accuracy will vary depending on the device context. For example, GPS location finding cannot be used indoors.

Localization: The delivery of services to the user that are aware of the user's current location and therefore tailored to that context.

Mobile Portal: Access point to the mobile Internet that provides a gateway to mobile applications.

Personalization: Providing content to the user that is based on their user profile.

Ubiquity: The availability of a service in most, if not all, locations.

Vodafone Live!: The original WAP portal, launched by Vodafone in 2002.

WAP Gateway: Part of the infrastructure of the mobile internet, providing a gateway between the World Wide Web and mobile telephone infrastructure.

WAP Push: Technology that allows a server to push content to WAP phone without requiring the phone's browser to make a client request.

Wireless Access Protocol (WAP): A communications protocol developed specifically for mobile phones, which supports page markup using the Wireless Markup Language (WML).

Wireless Markup Language (WML): XML compliant markup syntax, developed by the WAP forum, for creating pages for display on mobile phones.

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 583-586,

Chapter 3.4
Mobile Learning Technologies

Diane M. Gayeski
Ithaca College, USA

INTRODUCTION

While educational and corporate training environments have made large investments in getting wired to high-speed Internet connections, our work and social environments are rapidly becoming more mobile and flexible. The Internet and organizationally based intranets are powerful learning and performance tools, as long as users have a high-speed connection and up-to-date computing equipment. Online learning and information is not nearly as convenient or reliable when learners need to access sites from their homes, hotel rooms, client locations, or while on the road. In corporate settings, large numbers of critical employees such as factory engineers, health care professionals, builders, and maintenance workers often do not even have offices in which to use a computer.

Beyond the need for fast and flexible access to interactive learning, the field of online learning is being reshaped by several important economic and professional trends:

1. **Many corporate training departments are moving towards a more holistic professional model of "performance consulting" in which they provide solutions beyond instructional interventions; the trend is to provide short performance support tools and job aids that are meant to be used in the process of actually performing work.** Thus, the need for small and wireless media devices is even more pressing. Even the smallest laptop is a clumsy reference device when one is trying to troubleshoot a telephone system in a cramped equipment closet or when attempting to develop specs for a new manufacturing facility while walking around a construction site. Moreover, constantly updated information is often needed in tasks such as order-tak-

ing or quality control; therefore, some type of wireless connectivity, sometimes called "persistent computing," is needed.

2. **On the education front, many colleges and universities are attempting to broaden their student base and income by offering courses and noncredit educational experiences to adult learners who are already in the workplace.** Because of this, distance technologies often need to be able to be accessed in nontraditional locations, such as during commuting time. INSEAD, a large international business school based in France and Singapore, is already using applications that run on cell phones to allow their students to collaborate and check on course updates and grades. This is especially useful because many of their students are busy executives who can use their very limited time (such as while commuting or waiting to be served in a restaurant) to engage in coursework. These applications run in coordination with traditional classroom learning as well as WBT-based course materials.
3. **While interactive learning and Web browsing have become second nature even to very young children, it is expensive and cumbersome to equip elementary and high-school classrooms with traditional computers.** Thus, some other more mobile and less expensive devices that still offer the immediacy and interactivity of the Web are desirable. Many schools have purchased or received grants to equip every student with a small personal digital assistant.
4. **As education and business are becoming more global, there is a need to provide information and instruction to regions that are not well served by high-speed Internet connections or even good phone service.** In many developing countries, it is much easier and cheaper to use wireless connectivity than to attempt to install conventional wired access. The prevalence of cellular technologies in areas such as Japan, Singapore, and Scandinavia make it attractive to employ wireless systems because of the large installed base of users.
5. **Finally, millions of personal digital assistants and other handheld technologies such as MP3 audio players are now in the hands of users worldwide.** Many businesspeople already use palm-type computers as date books, and they would like to leverage their investments to use these devices for more applications. Digital audio players and other multifunction devices, including cell phones that connect to the Internet and share pictures, are becoming popular among young people; again, this rapidly growing installed base makes it inexpensive to offer instructional programs on them.

THE RAPID MOVE TOWARD WIRELESS COMPUTING

The communications landscape is becoming wireless even more quickly than it became wired. For example:

- By 2009, there will be more than 61 mobile workers in the United States alone.
- Almost 2 billion people now have cell phones worldwide.
- Of the 9 million handheld computer units purchased in the last few years, 80% are synchronized to some corporate user's computer at work.
- Laptops and personal digital assistants now outnumber conventional desktop computers.
- Over two-thirds of the telephone numbers issued worldwide are for cellular phones.

Today, it is common for office workers to have both cell phones and some type of mobile computing device. For less than the price of lunch for a corporate project team, one can buy wristwatch audio players, a credit card-sized organizer that will store thousands of phone contacts and appointments, or a digital camera that can directly send its pictures via wireless e-mail and browse the Web. New software systems can now turn a home or office PC into a mini-server which you can access from anywhere by using a variety of devices, so

you can be sitting in a traffic jam and download an MP3 audio file from your home PC and listen to it on your cell phone, or use your personal digital assistant to retrieve a user's manual from your office server while you are at a client site. College campuses are quickly adopting wireless networks to make instruction, collaboration, and scheduling available at a lower cost per delivery medium and networking than conventional desktop computers and wired campuses (Loh, 2001). Finally, there are literally thousands of "hot spots," which are wireless Internet connections in public locations such as hotels, libraries, airports, and parks (Wi-Fi Freespot Directory, 2004).

NEW PHILOSOPHIES OF LEARNING

Being tethered to an Internet line or being saddled with a laptop is not always effective or even possible, but this is not the biggest limitation of typical Web-based training. Beyond the technical constraints, there is a philosophical mismatch between what typically passes as good online training and what modern organizations need to improve performance. Most Web-based training is conceptualized and managed as courses which are designed, produced, and deployed by trainers, and which are in turn taken and passed by learners. These courses mimic college curricula more than workplace performance improvement efforts. This idea of both the course and the learner ever being "done" is a fundamental mismatch with the kind of continuous improvement systems in place in most organizations. For example, is a course on leadership or negotiation ever really complete and up-to-date—and is a learner ever completely through adding to his or her knowledge in this area?

Instead of courses, many training experts feel that we need "learning bytes": little packages of content and job aids that people can access at the peak moment of "teachability" and performance enhancement. Ideally, learning would occur when and where people need it most, and where it can most readily impact the success of their work. The information would be completely up-to-date, and would be interactive in its style and content.

In order for communication, documentation, and training professionals to align their efforts more closely with the emerging landscape of the modern workplace, they need to move:

- From producing courses to developing learning and performance systems that "grow themselves";
- From managing certification to encouraging continuous learning and collaboration;
- From developing curricula to building structures to capture and store knowledge bytes;
- From teaching and telling to aiding performance; and
- From creating schedules and managing facilities to adapting to the schedules and locations of the workers.

TYPES OF MOBILE DEVICES

Many types of mobile devices are potential platforms for information and collaboration that support learning and performance improvement in the workplace.

So, envision this: instructional designers now have a mobile delivery system for instruction, communication, collaboration, and performance aids that combines the following:

- Phone calls
- Music or audio playback
- Video and still-frame picture capture
- The ability to browse Web sites online
- E-mail send and receive
- Text and voice paging
- Global positioning and interactive map directions

Table 1.

PDAs	Personal digital assistants are handheld computers whose built-in software includes an address book, calendar, to-do list, and which also may include miniature versions of word processors, spreadsheets, and email clients. There are two major operating system platforms, Palm OS and Windows Mobile for Pocket PCs. Short training courses, documentation, and expert systems can easily be deployed on these.
MP3 players	These are the next generation Walkmans, but instead of playing cassettes or CDs, they play back digital audio files. The files are typically downloaded from the Web or "ripped" from your own CDs, using a conventional computer. Then they are downloaded into the MP3 player via a simple cable. You listen to files through a headset or adapter that lets you play it through your car radio. Beyond listening to music, trainers can record auditory instructional materials, and major website services provide audio versions of popular management books and magazines.
Digital text, audio, still picture, and video capture devices	A plethora of devices let you capture text, pictures, audio, and motion video. Besides the familiar digital cameras, small hand scanners, the size and shape of pens, are now available. Some of these are stand-alone devices while some (including audio capture devices and small digital still cameras) are built into PDAs. These allow you to create materials for training or documentation, and they also allow end-users to capture pictures or sounds on the job so that they can consult with colleagues or experts.
Tablet computers	These are basically laptops without the keyboard (or with a detachable keyboard) and they are very useful for performance support and data entry on the job. The user draws on the screen itself, or uses a stylus to tap on menu items. They can be strapped onto vehicles such as a truck dashboard or forklift truck, or carried around a construction site or storeroom.
Smart cell phones and pagers	"Smart" cell phones, widespread in Europe, are becoming more popular in other parts of the world; they allow you to wirelessly browse miniature versions of web sites, type short messages to other users, or receive automatic notifications of information such as flight delays or stock prices. Some universities are already employing these in graduate and continuing education courses.
GPS	Global positioning devices use a system of satellites to pinpoint your location—down to a couple of feet—using a small device that periodically sends out signals. These are used, naturally, for locating yourself on a map and creating routes to a destination. Companies with a mobile workforce use these to track performance and as security devices to locate employees who may need help.
Wearable devices	Remember Dick Tracy? Futurists say we'll soon be wearing wrist watches that are cell phones and web browsers, earrings that play music, and necklaces that carry our ID and medical information. They are being employed as performance aids especially within factory and equipment maintenance applications.

- Multimedia, hypertext training, and documentation
- Expert systems and smart job aids
- Programs that monitor performance or take measurements or perform calculations for workers

APPLICATIONS

Many corporations are already creating innovative and highly effective applications for mobile computing devices. For example, Cisco has created a small documentation and job aid program on one of its routers that can be downloaded from its Web site onto a PDA. Intel uses video clips on a Pocket PC to disseminate best practices to technicians in their chip factories. Maines Paper and Food service uses GPS systems and Palm computers to aid in the accuracy and efficiency of its delivery system to restaurants. Location scouts in the film industry use wirelessly connected video cameras to shoot and e-mail pictures of potential shooting sites to producers, saving days in their schedules. Cypress Semiconductor's Vice President of Marketing creates MP3 audio news programs that his staff can download on their computers or on portable MP3 players to listen to as they jog or commute. Telecommunications repair fleets for SBC Corporation are using ruggedized laptops that combine computer-based training, expert systems, and actual testing devices. And BOC Gases deploys wearable computers with interactive documentation that allows technicians in factories to quickly inspect and repair refrigeration systems.

In educational applications, Drexel University now uses a Web-based infrastructure that allows students to access their e-mail, pay bills, check on grades, and work with course materials either using conventional computers or wireless devices such as cell phones and personal digital assistants. One math teacher in Irvine, California has a class set of 40 (numbered) PDAs that she uses to give quizzes; the devices have become very popular with the students. The students are given the quiz ID of the assessment for that day, and they enter their student ID and quiz ID into the Classroom Wizard program. After answering the questions on the PDA, a student beams his or her answers to the teacher's computer, which scores it right away, and the student's score appears on his or her PDA. Medical residents in anesthesia in Great Britain use a program developed by their professors to access information about prescribing the right form and dosage of medications. And even elementary school students are using quizzes and games that run on handheld computers, thanks to simple and inexpensive authoring programs that any teacher can learn.

The challenge will not be in learning how to use and create programming for these devices: this is actually quite simple and versions of popular authoring tools are being introduced. Rather, this next revolution in technology will bring about discontinuities that will require a change in the fundamental approach to training, communication, collaboration, and performance support in the workplace. These devices are already being used widely in elementary schools, high schools, and colleges; the new workforce will not only accept this technology, but will demand it.

ACKNOWLEDGMENT

This chapter is derived from the author's latest book, *Learning Unplugged*.

REFERENCES

Gayeski, D. (2002). *Learning unplugged: Using mobile devices for organizational learning and performance improvement.* New York: AMACOM.

Loh, C. S. (2001, November-December). Learning tools for knowledge nomads: Using personal digital assistants (PDAs) in Web-based learning environments. *Educational Technology,* 5-14.

Wi-Fi Freespot Directory. (n.d.). Retrieved July 29, 2004, from http://www.wififree spot.com/

This work was previously published in Flexible Learning in an Information Society, edited by B. Khan, pp. 146-152 , copyright 2007 by Information Science Publishing (an imprint of IGI Global).

Chapter 3.5
Enhancing Learning Through Mobile Computing

Marsha Berry
RMIT University, Australia

Margaret Hamilton
RMIT University, Australia

Naomi Herzog
RMIT University, Australia

Lin Padgham
RMIT University, Australia

Ron Van Schyndel
RMIT University, Australia

ABSTRACT

The mission of this chapter is to explore ways in which mobile computing via the employment of Tablet PCs can assist the human computer interaction in the design and project development process and thereby enhance learning. We follow the process of ethnographic action research and report on the learning, observations, and communications of students in a multimedia program who were given the use of a Tablet PC for their second year of their degree. We discuss the educational design and customized agent software developed for this project and draw conclusions for wireless networks, and benefits and issues involved in enabling mobile computing and encouraging group dynamics among students.

INTRODUCTION

It is interesting to consider how much people learn when mobile. When on the train traveling from home to university all manner of observations might influence the way a person thinks and reinforce some learning experience. It is also

often a good time to revise notes before an exam or interview. Similarly, when walking from one lecture to another or over to the cafeteria, students may exchange information that contributes greatly to their learning. In the study discussed in this book chapter, we observe and analyze the learning experiences of students who were each given a Tablet PC for a semester of their course.

Ethnographic action research is a methodology for investigating the impact of technology on a community. It was first devised in 2002 to explore the use of computers on communities in India (Tacchi, Slater, & Hearn, 2004). Its principles are that one change rarely impacts on only one individual, and changing one aspect may affect many other aspects of a student's life within his or her community.

BACKGROUND

In this chapter we report our research into mobile computing and the design process. This research has been supported by HP Mobile Technology for Teaching Grant Initiative—2004 Higher Education, and we have undertaken exploratory ethnographic action research to explore and analyze to what extent Tablet PCs enhance learning within the context of students learning the design development process.

Formal RMIT student surveys (the top 10 student concerns are available through RMIT University) indicate that students would like to engage more fully with the University and fellow students in a manner that meets both their social and academic needs. Observations and conclusions drawn from this research indicate that students:

- undertake more hours of paid employment to support their study costs, resulting in increased pressure to maximize time and resources in academic hours;
- want to interact with the University in ways that best suit their personal circumstances and preferred learning practices; and
- have limited amounts of quality contact time with fellow students on campus.

Tablet PCs have the potential to facilitate students' engagement with the University and fellow students in a manner that does meet their social and academic needs, and our research explores the extent to which this may occur through the use of mobile computing devices.

We chose students from a multimedia design degree (Bachelor of Design [Multimedia Systems]) for this study because they spend considerable time engaged in group projects. It is also a challenging use of mobile technology as students spend time generating, analyzing, and collaborating around images. The students are diverse, with academic interests ranging from creative media design to software development. The aim of this study trial is to explore new methods for applying mobile technologies within both formal and ad hoc study groups. Students from the multimedia design program are generally expected to work in their groups both in and outside of the classroom on design projects. Interaction between students is not moderated; rather it is supported by a learning program that emphasizes team skills. Students enrolled in this program are typically local school leavers and have completed Year 12 Mathematics and English. A very small proportion is international students, mainly drawn from China and India. The students are enrolled in two design courses that require teamwork and collaboration for one semester.

Applications used in multimedia design are typically central processing unit (cpu) intensive and require a large display screen with keyboard, mouse, and WACOM (registered brand name) Tablet as input devices. The Tablet PCs with digitized screen, and pen and ink technology present an opportunity to explore the extent to which

Tablet PCs may become an enabling technology for students learning design processes. A detailed description of Tablet PCs follows in Table 1.

The features we believe to be enabling are the digitized screen and ink technology, wireless capability, which means we can examine the use of enabling Web technologies such as blogs, and mobility, which means that students may capture inspiration as and when it occurs and store it locally on the hard drive or in a blog when in a wireless zone.

THE DESIGN PROCESS

The design process can be described as cyclical with iterative loops whereby an initial idea is developed into a concept. Feedback is sought and then the concept is expanded further into a proof of concept (in this case electronic) and finally manufactured into a product (in this case a Web site) that is ready for consumption with feedback sought and integrated at key junctures in the production process.

The participants in the study reported in this chapter were drawn from the students enrolled in a core second year design course of a Bachelor of Design (multimedia systems) degree program. In

Table 1. Work environment summary

Number and kind of Tablet PC's	16 Hewlett Packard TC1100 Tablet PCs
Hardware Configuration	• Tablet PC with builtin microphone • Detachable Keyboard • Pen/stylus • Earphones
Software Configuration (under University Licenses)	• Microsoft Windows XP Tablet Edition • Customised Agent Software • Microsoft Office (Word, Excel, Powerpoint, Frontpage); • Adobe Acrobat; • Macromedia products (Flash, Studio-MX2004); • Appropriate WiFi VPN network access software, including customised security and virus-checking software. This enables limited internet and university website access, but also allows wireless communication between Tablets; • Remote access to a personalised blog server for use in delayed project-related intercommunication and archiving. WordPress is installed on a server which is rendered visible to the WiFi network for student use.
Accessories for group use	3 docking stations, 2 digital cameras
Blog Server Hardware	Standard PC (Macintosh)
Blog software	WordPress
Number of students in whole class	80
Students involved in project	16 for individual work in semester 1, and team work for semester 2

this particular course students engage in a group project where they design and build a Web site for a client so as to gain direct experience of working for a client with specific team role responsibilities. The student learning is scaffolded through a design and production process modeled on a simulated student-centered work-based learning approach that is described in the following section.

Workplace-Based Learning

Industry practice is often explored and tested in an educational institution via simulated projects that seek to meet specific and relevant learning outcomes. However, a simulated project can be less than effective as it often may lack critical detail and commercial imperatives to solve specific problems and challenges.

Central to the strategy of situated learning in the workplace is the direct experience and subsequent knowledge gained in the process. However, guided learning in the workplace can be limited due to commercial constraints and lack of mentoring skills and processes. In the absence of a structured pathway of learning, students are required to integrate the formal theory gained in a university and practical knowledge gained in the workplace.

Leaving learners, particularly novices, to piece together a picture of the complex workplace environment without guidance is more likely to result in incorrect and fragmented understandings. (Cornford & Beven, 1999)

The issue then is how to integrate the hands-on learning that occurs in industry and then bring that back into a formal learning environment that will assist in contextualizing their experiences and skills gained.

Guided learning can augment many of the strengths of learning through everyday activity, and also be able to address some of its weaknesses. (Billet, 2001)

As Billet proposes, a combination of strategies is required to achieve a meaningful outcome for students. An integration of the strengths of the workplace-based model, coupled with the benefits of a face-to-face learning environment (such as a tutorial seminar program) and underpinned by a lecture program would provide an effective structure in which to enhance learning outcomes.

In the absence of situated learning where students are located in the workplace, the structure of the curriculum focuses on the development of commercial projects over the duration of the semester within the educational institution. Students form small working groups within their tutorial classes and are provided with a commercial project and a client.

The design of this problem-based learning environment supports constructivist teaching and learning practices. Students are guided through a process that encourages, through the design process, a construction of knowledge and understanding based on direct experience. Savoie and Hughes (1994), outline several actions to put this into effect:

- Identify a problem suitable for the students;
- Connect the problem with the context of the students' world so that it presents authentic opportunities;
- Organize the subject matter around the problem, not the discipline;
- Give students responsibility for defining their learning experience and planning to solve the problem;
- Encourage collaboration by creating learning teams;
- Expect all students to demonstrate the results of their learning through a product or performance.

The forming strategies presented to the students are drawn from the Savoie and Hughes (1994) outline of actions. Students form work teams comprising of three to five individuals based on friendship groups.

MOBILE COMPUTING AND THE DESIGN PROCESS

Like those in the design industry, students often work together in concentrated patches, and then continue development largely on their own. They frequently juggle multiple projects and other concerns. Communication can be patchy and often is not centralized. Retrieval of communication items and files can also be unreliable. The result may take up valuable resources and development time and manifest in poor outcomes, frustration, and lack of momentum.

Development of a communication strategy and document retrieval process has been implemented to facilitate the design and development process. This is made possible by the use of mobile computing in the form of streamlining processes. The group members are asked to maintain a personal blog (Weblog) that is intended to provide a record of their research into their specific area of concern, and at the same time, also make this research transparent to the rest of the group. A Rich Site Summary (RSS) aggregate has been put in place to alert other team members that the new posts are up on the blog and can be viewed at their convenience. RSS is described in wikipedia as:

RSS is a family of Web feed formats, specified in eXtensible Markup Language (XML) and used for Web syndication. RSS is used by (among other things) news Web sites, Weblogs, and podcasting. The abbreviation is variously used to refer to the following standards:

- Rich Site Summary (RSS 0.91)
- Resource Description Framework (RDF) Site Summary (RSS 0.9 and 1.0)
- Really Simple Syndication (RSS 2.0)

Web feeds provide Web content or summaries of Web content together with links to the full versions of the content and other metadata. RSS in particular delivers this information as an XML file called an RSS feed, Web feed, RSS stream, or RSS channel. In addition to facilitating syndication, Web feeds allow a Web site's frequent readers to track updates on the site using an aggregator. (http://en.wikipedia.org/wiki/RSS_%28file_format%29)

RSS feeds allow people to remain up to date with changes that have been made to a blog or Web site that they visit regularly. An aggregator is a way of getting a customized and consolidated view of all the sites one regularly visits.

A group blog has also been set up to facilitate group communication. The groups use these in different ways:

- introduction of a new item of interest to the group overall;
- meeting minutes and actions to be taken;
- posting up documents, drafts, or approved designs;
- general considerations or concerns to the group;
- miscellaneous items such as notification of next meetings, introduction of new group members, and communication with tutor/exec producer.

Additionally, a client blog (now largely standard industry practice) has been put into place, with clients being given read and write access to monitor development and design of their project. This has been particularly useful in confirming outcomes and minimizing unnecessary client contact (usually in place to prevent clients from being uniformed).

Customized agent software has also been developed to facilitate document retrieval and assist in the process of design and programming.

The agent is required to locate files on the server via a server agent and report back to the Tablet holder using a Tablet agent to notify updates to files and filing systems.

The overall intentions of these items are to:

- maintain communication across the groups and their clients;
- provide a map of next steps and actions;
- maintain momentum and focus throughout the design process;
- help members locate themselves within the framework of the group (particularly useful when they are involved in other projects or other commitments);
- aid group dynamics as lack of communication can contribute heavily to unsuccessful outcomes;
- enable contact maintenance with the client.

Tablet Personal Computers and the Wireless Network

Sixteen Hewlett Packard TC1100 Tablet PCs are given to the students, as well as three docking stations and two digital cameras for shared use. This creates an excellent opportunity to test two different aspects of mobility: wireless networking and the differences between the Tablet PCs and desktop computers.

The first aspect of mobility is the wireless ad-hoc networking made possible by the hardware. This enables two different kinds of interactions: person-to-person (via WiFi or Bluetooth—a short-range networking protocol chiefly used for wireless peripheral control or file transfer) and person-to-blog/server via WiFi—a longer-range networking protocol that is principally used for wireless network access.

As shown in the student comments later in this chapter, significant use is made of the wireless components, due partly to the communicative nature of the projects, and partly to the particular team assignment work.

One problem we encounter in implementing the system is the use of a specialized network protocol for the communication between agent and server (the agent system is described in more detail in the next section). A previously unallocated TCP/IP port (or channel number) is used as the communication port by default. Although the software allows any port to be chosen, it needs to be the same port for all tablets, as they must all use the same channel. In addition, the network needs to know that this port has been allocated for specific use, and all appropriate permissions need to be obtained. This effectively means that although the project can exist on a wireless network, that network has to be preconfigured to accept it.

While wireless networks exist that will accept any port (and thus need no such pre-configuration), they are generally insecure because of this blanket acceptance policy. The alternative is to allow the network packets of information to be encrypted on the Tablet PC and at the server end. This is effectively a virtual private network (VPN). The use within our University of a VPN as the wireless medium initially causes some concern because of the necessity to customize the VPN settings for using this port. In addition, our security policies do not directly allow access to the Internet (via port 80—the HTTP "channel") and prevent direct access to the internal wired network.

Students are required to access their blogs, which are installed on a server on the wired intranet. For this project, special provision and permissions are required to enable access to this server from the wireless network. This exposes the server to the wireless network, exchanging its internal security status for the wireless security arrangements. So the safety of a private intranet needs to be exchanged for that for a VPN. Because of the differing security models for these two configurations, we recommend that it would be preferable in future to have a wirelessly ac-

cessed server dedicated to this task alone, and any unrelated server activity be moved to a different separate internally wired server.

The second aspect of mobility that the students are exposed to is the user interface differences to a desktop computer. As seen in Table 1, the Tablet PCs have a keyboard, a screen that can be separated from the keyboard, a pen or stylus, and appropriate software to accept pen gestures, commands, and handwriting storage and recognition. For this Tablet PC, the keyboard is normally intended to be used only for extended text entry, and the pen used for normal Tablet functionality.

The tasks the students undertake often involve drawing and sketching (see the included software in Table 1), so it is encouraging to see some students using the pen in all activities, while others are content to use it only when appropriate.

The principal difference between a pen and a mouse depends on the dialog mode—input or selection mode. Neither pen nor mouse works in isolation as a positional input device. In both cases, these move a cursor, which is the positional input coordinate identifier. Herein lies the difference. A mouse moves the cursor using relative coordinates, which implies that the cursor's initial position is known to the user. So to press a button on the screen, the user must home in the cursor to the button, and then press the mouse button to select/activate. Moving the cursor is a necessarily interactive process, and any person who has experienced slow mouse response lag can appreciate the crucial dependence on interactivity that the mouse requires. Contrast this with a pen, where the user identifies the target button and using his/her hand positions the pen to the target. No intermediate interactivity is required to enable accurate positioning.

In input mode, the difference between pen and mouse is smaller, and more subject to personal taste. The benefit of the mouse is consistent hand positioning, support for the wrist and elbow, and having the hands away from the desktop so that it is always visible (note that these are all desktop-based advantages). While pen has direct positioning and given fast interaction can easily be used for drawing, the arm is usually not in a comfortable position for long duration detailed work. However, quick sketches in a mobile context are easily facilitated. Lastly, the direct positioning of the cursor facilitates handwriting as easily as with a real pen, allowing text to be entered via handwriting recognition, a tool that some students have mentioned using in our study.

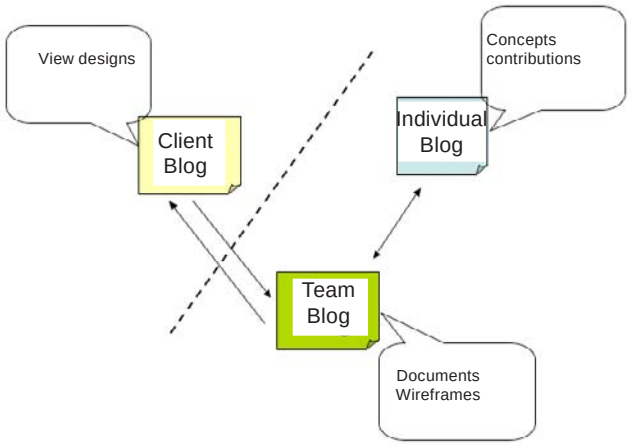
Customized Agent Software

Customized agent software has been developed, primarily for content synchronization between Tablets within each project group, and between Tablets and blog, or server-based data storage. Major check-points in content developments are stored on a communal server and could be accessed by participants and also the project coordinator(s) for progress verification and eventual project assessment.

We include an overview of the software model or architecture employed and the role each component plays in it. As will be explained in the next section, actual student usage of the software did not precisely follow the usage patterns envisaged by the previous design. However, this provides insight into the students' view of the environment, and how they adapt it for their own use.

Part of the process of designing a supportive environment includes the specification of the agent-based software. It must be able to provide automated support for a range of tasks such as proactive notification to students if the group files have been updated, to ensure that they check and access any relevant changes. This removes the tedium of repetitive manual checking. Intelligent agents are a popular technology for open environments such as the Internet (Singh & Huhns, 2005). Small autonomous pieces of software (agents) can reside on different machines, and communicate with each other to achieve tasks. In this project we have an agent on each Tablet, which communicates

Figure 1. Team communication using three blogs

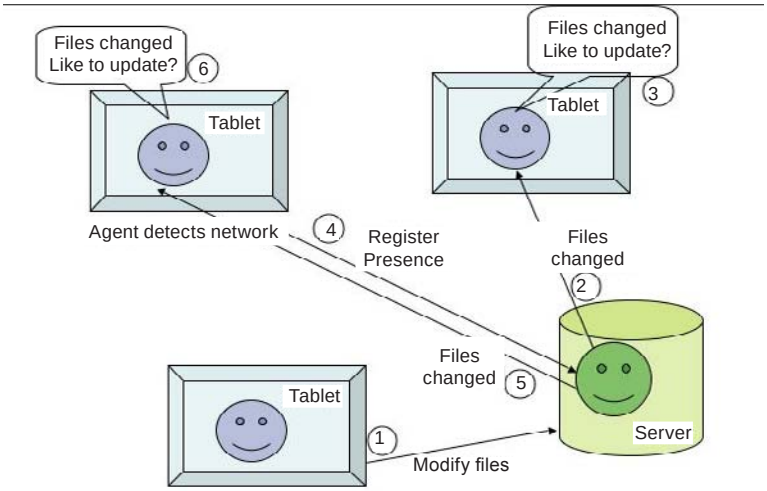


with an agent on the central file server where files are located. Figure 1 shows the interaction between the agents that results in the user being notified if there are any updates.

While this initial task is quite straightforward, it is envisaged that there will be potentially a range of automated tasks that can be performed by the agent software on behalf of, or in support of the user. The intention is to interact with students to develop the particular support they will find useful. Observation of student behavior during the trial period indicates that the availability of the wireless network to the Tablets led to far less use of the central server and the file structure contained there than had been the case previously. Consequently we consider that it would be advantageous to adapt the agent software to better support the communication mechanisms the students actually use. For example they e-mail versions of their files to each other frequently. While this appears to have worked for this small project, this mechanism does not offer the same level of backup and support as does storing files on the central file server. We could, for example, adapt the agent functionality to better support the way the participants end up doing things, by having the Tablet agent recognize when new files are sent to other team members, and automatically store these into the central server.

The agent approach used is what is referred to as *belief desire intention* (BDI) agents. This approach develops agents in terms of these mental concepts, as developed in philosophy, to explain how focused practical reasoning happens in humans (Bratman, 1987). Beliefs are the information that the agent has (or believes it has, as it may not always be accurate) about the world, other agents, or even itself. Desires are the goals the agent wishes to accomplish—which may arise based on information from the environment: for example, if a file is updated in the central repository, an agent managing that repository may then have a goal to ensure that all group members are notified of the change as soon as possible, potentially using various communication means such as e-mail or even sms, if the user is not accessible via the network. Intentions are the plans the agents has regarding how it *intends* to accomplish its goals. If, during execution of the plan, there is some problem, the agent will adapt its intentions (or plans) to attempt to still achieve the goal. One advantage of this approach is the ease with which it can be adapted and evolved to manage an extremely large number of complex

Figure 2. Specification of the customized agent-based software



interactions regarding which decisions should be made. This enables the kind of support to be easily tailored to both very specific situations, and also to individual groups and/or users. For example the agent residing on a particular user's machine can build up beliefs regarding the way that user likes to work. Agent technology also facilitates pro-active behavior on the part of the system (to achieve *desires*, or goals). The agent software can persist in trying to achieve a particular goal (of which user notification of change is a simple example) using alternative approaches if attempts are unsuccessful. This makes it possible to build flexible and robust systems in dynamic environments.

Research Questions

In this research, we consider the question of how Tablet PCs help human computer interaction in the design and project development process. Our focus is on:

- The immediate circle of students in terms of how they are organized, how they carry out their work, and how the Tablet PC fits into their lives as students of multimedia design;
- The everyday lives of the participants, their ways of doing things, and the impact mobile technology in the form of wireless Tablet PCs has on this (if any);
- The construction of knowledge and meaning through the use of blogs and the impact mobile technology in the form of wireless Tablet PCs has on this (if any);
- The wider social context in terms of access to a wireless intranet, access to mobile computing, access to a personal blog.

Discourse analysis and ethnography provide the base for qualitative research into the contents of the blogs, structured interviews, and focus groups. Ethnography, in its literal sense, means "to write or represent a culture" (Tacchi et al., 2004). Ethnography is an approach that may encompass several methods rather than a specific methodology. The methods are integrated to provide a holistic account of a culture, in this case, students enrolled in the Bachelor of Design

(multimedia systems) and their use of Tablet PCs in their everyday design studies. The structured interviews, informal feedback conversations with participants recorded as field notes by the researchers, questionnaires, blogs, and focus group conversations will be integrated so that the knowledge and experience gained through one method informs the other methods. This is in keeping with ethnographic and action research principles. This is in contrast to other studies where students used Tablet PCs only during their classroom learning, such as Tutty (2005; Tutty, White, & Pascoe, 2005, 2006). Our focus is on mapping the experience of the student participants in detail and to enable the use of the Tablet PC to become part of everyday life as ubiquitous computing. Therefore we did not use a comparison group of students who were not given Tablet PCs for their learning and personal use.

Our intervention is the introduction of the Tablet PCs, the wireless network, and the use of blogs. Students are provided with Tablet PCs for the whole semester and are able to carry them around with themselves wherever they go and to take them home. This allows the Tablet PCs to become a part of the students' normal daily living and learning routines.

WHAT THE STUDENTS SAID

In his discussion of the computer for the 21st century, Mark Weiser introduces his concept by suggesting that:

The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it. (Weiser, 1991)

In this study, we are interested in how the students have woven the use of the mobility afforded by the Tablet PC into their lives. To this end, we administer an initial survey to measure their initial use and experience of computers generally. We find that the majority of students consider them-

Table 2. Baseline survey of students' skill levels

Responses – Skill Level Questions	5 Expert	4 inter	3 novice	2 none	1 unsure
Drawing with a mouse	14%	43%	43%	0%	0%
Drawing with a pen on a WACOM tablet	10%	19%	43%	24%	5%
Writing with a pen on a WACOM	5%	33%	38%	14%	5%
Using wireless networking	19%	43%	38%	0%	0%
Using the Internet	81%	19%	0%	0%	0%
Using a computer	71%	29%	0%	0%	0%
Installing software	57%	38%	5%	0%	0%
Downloading software from the Internet	62%	29%	0%	5%	0%
Using handwriting recognition software	10%	33%	33%	24%	0%
Using a PDA	10%	19%	29%	43%	0%
Using blogs	0%	10%	43%	43%	5%
Putting together my own blog	5%	10%	29%	43%	14%
Developing my own website	14%	38%	33%	5%	10%
Using chat software such as ICQ	52%	19%	24%	5%	0%
Using agent software	5%	10%	19%	43%	19%
Finding information I need using the Internet	62%	33%	5%	0%	0%

selves as novices and have no experience using pens with their computers. However, all indicate they believe they have average or expert Internet skills. All have computers at home, and all save one has Internet at home—most have broadband. The details of this questionnaire and the student responses have been published in Berry and Hamilton (2006). These results bear out our initial assumptions about the communicative ecology of the students. Our initial assumptions were that students would not have had extensive experience using pens and digitized screens.

The second part of the survey is designed to collect baseline information about the participants perceived skill levels in human computer interaction. The survey items fall into several categories: drawing skills, ability to use pens and Tablets or digitized screens, ability to use wireless networks, experience with handwriting recognition software, ability to use and maintain blogs, general use of the Internet to locate information, download and/or install software, ability to use mobile devices (specifically PDAs), and Web design and development skills. The snapshot that emerges from the survey response data indicates that a significant number felt themselves to be either novices or average users across most items except those dealing with use of the Internet to locate information, and downloading and installing software.

Students were asked to read each statement and circle the number that best suits their thoughts about their skill level at this time:

1 I am unsure of my skill level
2 I have no knowledge about this
3 I am a beginner/novice at this stage with a basic introductory skill level
4 I have intermediate skills
5 I am an expert user with highly proficient skills

Next, we conduct focus group meetings so we can meet the students face to face and discuss how they initially use the Tablet PCs, what they intend to do, and any issues they currently have. The questions for these focus groups may be viewed at http://raws.adc.rmit.edu.au/~e01913/blog, and a discussion of the results is also in our earlier paper (Berry & Hamilton, 2006), as well as selected segments posted on a University Web site, http://www.lts.rmit.edu.au/mobcommec/index.htm.

For the first semester, the students mostly work individually; however, for the second semester, they are required to work in groups on an industry-based project as discussed previously. They are placed in groups regardless of whether they have a Tablet PC or not, as the majority of students in the class do not have a Tablet PC. Originally we had wanted all participants in a particular group to have a Tablet PC; however, in practice, this was not possible, as groups are composed of friendships, which are formed irrespective of how the Tablet PCs have been allocated. Each group of students has four or five members, and all have Tablets in the majority of groups; however, there are two groups where only one student has a Tablet PC and one group where all except one member has a Tablet PC.

We analyzed the focus group tapes using discourse analysis and have selected the comments that best reflect the general trends. Discourse analysis is a technique that has its origins in linguistics and is also used in ethnography, which examines lived experience as its subject. We allow the students, our research collaborators, to speak for themselves. They become the experts in the uses and usefulness of the Tablet PCs. We adopted and maintained a nonjudgmental position and treated everything as valid. Sharp, Woodman, and Robson (2000) also used discourse analysis as part of their methodology to study software engineering practices. They follow the principle that "all views should be attended to, and given equal weight." We also adhere to this principle in our presentation of the student views. Ethnography is not empirical in its approach to phenomena.

In this section we present the student responses

reflecting their lived experience and the impact possession of a Tablet PC has had on their learning and group work practices. We asked a very open-ended question that led to conversations about how the Tablet PCs fit into their work practices in their work groups. We chose to allow their voices to speak instead of overlaying their responses with an empirical set of categories. This is in keeping with postmodern anthropology and ethnography. Clifford (1986) foregrounds the dilemma of writing about lived experience and suggests possible ways of structuring and articulating the text:

Whatever else an ethnography does, it translates experience into text. There are various ways of effecting this translation, ways that have significant ethical and political consequences. One can "write up" the results of an individual experience of research. This may generate a realistic account of the unwritten experience of another group or person. One can present this textualization as the outcome of observation, of interpretation, of dialogue. One can construct an ethnography composed of dialogues. One can feature multiple voices, or a single voice. One can portray the other as a stable, essential whole, or one can show it to be the product of a narrative of discovery, in specific historical circumstances. (Clifford, 1986, p. 115)

Tyler (1986) discusses the multivocal or polyphonic nature of post-modern writing that aims to set down lived experience:

A post-modern ethnography is a cooperatively evolved text consisting of fragments of discourse intended to evoke in the minds of both reader and writer an emergent fantasy of a possible world of commonsense reality, and thus to provoke an aesthetic integration that will have a therapeutic effect. (Tyler, 1986, p. 125)

We asked the students the following question:

How are you Using the Tablets so far This Semester?

Their responses range from:

I have found the Tablet extremely useful in this scenario [work group]. I have used it in a way, so as a group, we can have lunch in the cafeteria and also discuss/browse relevant Internet sites—this makes it convenient and efficient instead of having lunch first then going over to the labs to discuss Web sites (since we are very busy students). Sometimes, when our lab is too packed, or fellow students use two computers at once to render images, and so forth, I take out my Tablet and use it instead. As a group, we did our proposal presentation to the client today; we had to use the Tablet to set up with the classroom projector to run our PowerPoints and Flash files. Without this, we would have had to do it in our client's office, which would have not been a terribly good experience.

And:

I have used the Tablet to create the Flash, and showing other team members and students the work. I have also used the Tablet to do other course homework.

From students who have the only Tablet PC in their group, to:

…really good, been using it to do the wireframes, site maps, and some other stuff. Also working on my blog banner with it, which I will be animating in Flash.

And:

I have been using the Tablet for all of my project work this semester. Some examples are using Flash to do some test animations, using Dreamweaver to put together some html docs, and we have all been taking notes using Windows Journal. We

have been attempting to create a wireless network between our Tablets for easy transfer of files; I'm still working on that...

Both of these comments are from students who all had Tablet PCs in their group. Hence it would appear that while the majority of students are adopting the use of the Tablet in their immediate circle of friends, if their friends do not have one, they are becoming the note-taker, keeper of resources, and demonstrator of group work for the client.

Other responses include:

The Tablets are a great resource. When developing our proposal for our client, we took in mind that using a multimedia enriched presentation, especially for a children's writer, allows for the group to typically show where they are heading with the project. There is nothing easier than images and animation to describe the design flow of the group. The Tablets have allowed us to work with this content and to have it displayed on big projectors easily, knowing that it will work the way we want it to. We have been able to do designs on the spot for our client as well as take notes and even voice record our whole meeting so we know exactly what the client likes and dislikes about the project.

I am using my Tablet in every design class and also out of class for this project. It is very useful as when we have meetings we can view items on this Tablet and the work we have produced on this Tablet.

The use of Tablet PCs is becoming part of the everyday lives of the participants and their ways of doing things:

This semester I used my Tablet PC mainly for updating the blog and maintaining design research in a folder stored on my laptop that can be easily taken from home to school.

The students are adopting it not just for their coursework, but also for their own construction of knowledge:

I have also been using the Microsoft Journal program to consolidate design ideas and do rough sketches (as I find Firework's FreeHand far too awkward because of its vectors), so essentially I just use it as good old 'butchers paper' in terms of sketching and making notes as design ideas come to me.

I use the Tablet to do a lot of writing as I'm a writer; I write on the train a lot thanks to the Tablet, and this writing often includes rationale for my designs to post on my blog. I use the laptop to connect to the RMIT wireless to send e-mails in the caf [student union cafeteria], and our group often meets in the caf so we can access the Internet and do any kind of research, e-mailing, or other Internet related tasks together whilst discussing our design project.

People have been using them for virtually every aspect of their studies, from note taking to sketching out designs.

This semester I have decided to use the Tablet PC for everything. This means any notes taken, and design ideas, Web page development, assignments, and so forth I am doing with the Tablet. Thus far it has been pretty good. We have tried to establish a wireless link between the Tablets, but have been so far unsuccessful.

Unfortunately access to the wireless network has been patchy:

I was having a great run with the wireless access, and was looking past the constant connectivity

failures. Unfortunately what this meant was that I am now getting the message that I have logged on too many times. Other than that, the Tablet has been a success and is most definitely a very good tool in the learning environment. In regards to using the imaging and graphics software on it, the small screen has been my major issue, as well as the processor not being able to handle images with a large file size.

And there have been issues with screen size. Some students have found the processing power slower, especially when they are using a large package like Photoshop:

I have decided to use the Tablet PC for everything: all note taking, all assignments, and most of my Internet research via the wireless network. This has been relatively successful, although I have had many complications with wireless connectivity and now with me getting the message saying that I have exceeded the maximum amount of logins. I have used the Tablet for a lot of the design development, and have basically transferred all of the usual things I would do on paper onto the Tablet. All of my note taking and assignment thoughts are now done on the journal software. More recently the PC has been running exceptionally slow. I am not sure, but I think it may be a virus. I have noticed a file labeled MediaGateway.exe, which should not have been there. I must have accessed it to view something on the Web. And it has rendered the Tablet almost unusable at the moment. I can still operate the PC in safe mode and have uninstalled that file following the Windows instructions. Anyway, I have backup plans in the interim, and will push on regardless. Other than this, the Tablet has been running exceptionally well and has been a brilliant asset to my development this year, and much more so this semester.

The majority of students have found using the Tablet PC enhanced their social context in terms of their access to the wireless intranet from the cafeteria, to their interactions with their clients:

So far this semester I have been using the Tablet PC in a number of ways. I used to the Tablet to construct the Home Hardware Web site proposal presentation. It was useful because I was able to carry it around anywhere, and my group members and I were all able to work on the presentation outside of class (e.g., the Cafe).

- I also used it to and from uni in the train to construct mock-ups for the Web site.
- The Tablet PC was also used after our client meetings to type up any notes, which were then uploaded onto my blog.
- I have also been using it to 'sketch' in ideas using the journal.

Some students are experimenting with the construction of knowledge beyond their immediate classrooms, to experimenting with various different new aspects provided by the mobility and wireless access to the Internet:

…sketching and listening to music. I bring it to meetings and use my handwriting to jot down notes. We haven't managed to set up MSN Meeting or use the agent, so it is like a laptop, to type up documentation.

I'm actually utilizing my Tablet quite a lot this semester. I'm currently using it for taking notes in lectures and classes, brainstorming, doing work on public transport, in the city, and at RMIT. I'm mainly using the programs Word, Photoshop, Illustrator, Journal, and Notepad on the Tablet.

Some light Flash work and using it as more of a tool to take down notes in class/lectures. More handy then a normal book that we write in, as we don't have to search through annoying papers, and so forth. Also use it sometimes in the cafeteria to

research for our projects whenever other classes are taken. Very handy.

Other students are aware of these different options, but are busy enough adapting the technologies to their everyday needs as students:

So far we are using our Tablets like normal laptops, bringing them to class and working on them with Flash, Photoshop, and Illustrator. We haven't actually made use of the Tablet function in any way, as our project design is quite basic and comprised without the Tablet. It is good for us all to have one to bring and log on to wireless.

The Tablet goes everywhere; it's my test environment at home and at Uni. I keep the most up to date information on my Tablet, and I access it over a network to modify files stored on its Web server. I don't think I've rebooted it for about two weeks; it's great to turn on and off very quickly. I'm amused on the train with it, and wherever I go I can get work done. It's the Tablet's portability that's most valuable to me. It's very flexible in how I can use it (standing up/sitting down), and it's powerful enough to do my HTML/PHP/CSS work for the project.

During client meetings I am using the Tablet to take notes and track the meeting. I am also using it to check my e-mails at Uni and to transfer data from home to Uni.

Tablets this semester are really helpful! They have been great when working on my ideas for design and being able to show the other members in my group my work by simply turning on my computer. It almost works as a folio for me, in that I can transfer my work onto the Tablets (or even the work I have produced on the Tablets), bring it to Uni, and connect it up to the net to do simple things like update the blog and share my ideas with my teammates. I have been using the wireless connect a lot more this semester—once again really helpful, whereas now we can have meetings in the caf and not have to rely on computer rooms to be able to connect to the Internet to discuss our work and ideas during the meetings.

Of the 16 allocated to students in this design course, only one student found he was not using his Tablet PC, and his issues were:

To be totally honest, I haven't been using the Tablet at all this semester. I'm finding that the Tablets are extremely slow loading and laggy. Every time when you open up a browser, e-mail, or any application at all the system just hangs and you cannot do anything else until the application has opened. On top of that, because my role consists of using the keyboard a lot, the small keyboard makes things difficult to work with; therefore, I haven't used the Tablet at all.

We have the distinct impression that the Tablet PCs with their digitized screens and capacity for wireless networking have largely disappeared in to the fabric of the students' everyday life (Weiser, 1991). To check this impression we administer the final survey that contains items identical to the baseline survey. The results indicate that while their general abilities with the use of the Internet remained unchanged, the responses in the other categories at the beginning of this section are all now at the average and expert skill level, thus indicating that the participants feel their skills have definitely improved over the study period, especially with regard to using digitized Tablets with a pen and wireless networking.

Students were again asked to read each statement and circle the number that best suits their thoughts about their skill level at this time:

1 I am unsure of my skill level
2 I have no knowledge about this
3 I am a beginner/novice at this stage with a basic introductory skill level

Table 3. Exit survey of students' skill levels

Responses – Skill Level Questions	5 Expert	4 inter	3 novice	2 none	1 unsure
Drawing with a mouse	18%	55%	18%	9%	0%
Drawing with a pen on a WACOM tablet	27%	45%	18%	9%	0%
Writing with a pen on a WACOM	36%	45%	9%	9%	0%
Using wireless networking	18%	55%	18%	0%	0%
Using the Internet	73%	18%	9%	0%	0%
Using a computer	64%	36%	0%	0%	0%
Installing software	45%	45%	9%	0%	0%
Downloading software from the Internet	36%	45%	9%	9%	0%
Using handwriting recognition software	9%	73%	9%	9%	0%
Using a PDA	0%	18%	55%	27%	0%
Using blogs	27%	55%	18%	0%	0%
Putting together my own blog	18%	64%	18%	0%	0%
Developing my own website	45%	45%	9%	0%	0%
Using chat software such as ICQ	64%	18%	18%	0%	0%
Using agent software	0%	9%	45%	36%	9%
Finding information I need using the Internet	91%	0%	9%	0%	0%

4 I have intermediate skills
5 I am an expert user with highly proficient skills

CONCLUSION

Implementation of mobile computing effectively facilitates the establishment of a learning community among the students. Many of the structures initially set up seem to have broader application than originally anticipated. Mobile computing assists project groups by extending their communication beyond traditional methods.

Students are able to communicate with other groups effectively and solve a wide range of problems. Instant messaging and e-mail are a key component of contact over this semester and appear to be more prominent than previously observed. Additionally, alternate forms of record management such as document posting on secure sites to be retrieved seem to also be utilized. Weblogs are a key component in record keeping and are often linked to other student Weblogs recording a series of interesting or relevant information.

Many of the strategies set up to facilitate the mobile computing in fact benefit all projects. Teaching staff are able to monitor project development with ease and often recognize issues that come up early in the piece, as the process is far more effectively documented. This has enabled early intervention and in many circumstances provided the group with clear direction to move forward and work through the challenges faced in each project.

The students are able to monitor their own work in relationship to their team members. Problems encountered in previous semesters do not appear to be the status quo with the presence of the mobile computing. There are often typical issues in the

teaching and learning process such as:

- students being unprepared for meetings;
- leaving items at home and not available for discussion;
- technical difficulties of transporting files and data across platforms and via e-mail;
- lack of opportunity for meetings with each other out of class times due to work or study constraints;
- difficulties of interpretation or communication between group members and the client;
- progress slowed by weekly meetings.

The capability of the Tablet PCs and mobile computing allows students to communicate, review, and update their information and development strategies with fewer delays.

FUTURE TRENDS

So far our findings have been encouraging. However, our investigations have opened up further avenues to explore and have posed future research questions relating to allowing all students access to Tablet PCs for all their courses.

We note that the Tablet PCs provide a platform for the development of ideas and enable the centralizing of notes, sketches, various media, and drafts. This means that projects can move along faster, as more information and record keeping can be provided at meetings, more groundwork is covered in a more sophisticated way, and the progression of ideas can be seen more easily. It also results in less information being misplaced or lost, as might be the case with scraps of paper. However, memory storage can become an issue, and many students purchased memory sticks, both for the memory storage and faster transfer of files.

Tablet PCs can help students by providing a communication strategy to expand and support development processes and provide an immediacy to what students are doing, bypassing some pen and paper steps (as an aside: we noticed 40% of students in the cafeteria in Building 8, on Thursday July 28, at 10:30am were using notebook computers). In providing the means to show clients the current status of work, they also enable a clearer understanding of where the project is currently positioned. One student's final comments were:

The tablet has been very convenient for the year. It has assisted me with my design client work and everyday tasks. Thanks!

ACKNOWLEDGMENTS

The authors acknowledge the help and support of associate professor Jim McGovern, associate professor Vic Ciesielski, professors Mark Shortis, Evan Smith, Laurie Davies, and Matt Maddocks, and all members of the HP Mobility Grant team within our University. Also, this research is partially funded by HP by means of their HP Mobile Technology for Teaching Grant Initiative—2004 Higher Education.

REFERENCES

Berry, M., & Hamilton, M. (2006). Mobile computing, visual diaries, learning and communication: Changes to the communicative ecology of design students through mobile computing. In the *Eighth Australasian Computing Education Conference (ACE2006)*, Hobart, Tasmania, Australian Computer Society, Inc.

Billet, S. (2001). *Learning in the workplace: Strategies for effective learning.* Sydney: Allen & Unwin.

Bratman, M. (1987). *Intentions, plans, and practical reason.* Harvard University Press.

Clifford, J. (1986). On ethnographic allegory. In J. A. M. Clifford (Ed.), *Writing culture: The politics of ethnography.* G.E. University of California Press.

Cornford, I. R., & Beven, F. A. (1999). Workplace learning: Differential learning needs of novice and more experienced workers. *Australian and New Zealand Journal of Vocational Education Research,* 28.

Savoie, J. M., & Hughes, A. S. (1994). Problem-based learning as classroom solution. *Educational Leadership,* 54-57.

Sharp, H., Woodman, M., & Robson, H. (2000). Using ethnography and discourse analysis to study software engineering practices. *IEEE Software, 17*(1).

Singh, M. P., & Huhns, M. N. (2005). *Service-oriented computing: Semantics, processes, agents.* John Wiley & Sons, Ltd.

Tacchi, J., Slater, D., & Hearn, G. (2004). *Ethnographic action research.*

Tutty, J., White, B., & Pascoe, R. (2005a). Experiences from a wireless-enabled tablet classroom. In *Australasian Computing Education Conference*. Newcastle, Australia: CRPIT.

Tutty, J., White, B., & Pascoe, R. (2005b). Experiences from a wireless-enabled tablet classroom. In *Australasian Computing Education Conference* (pp. 165-172). Newcastle, Australia: CRPIT. 42.

Tutty, J., White, B., & Pascoe, R. (2006). Experiences from a wireless-enabled tablet classroom. In *The 8th Australasian Computing Education Conference*. Hobart, Australia: CRPIT.

Tyler, S. (1986). Post-modern ethnography: From document of the occult to occult document. In J. A. M. Clifford (Ed.), *Writing culture: The politics of ethnography.* G.E. University of California Press.

Weiser, M. (1991). The computer for the 21st century. *Scientific American,* 94-110.

This work was previously published in Enhancing Learning Through Human Computer Interaction, edited by E. McKay, pp. 57-74, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.6
Mobile Technology and its Applications in Instructional Conversation

Jason Caudill
Independent Consultant, USA

INTRODUCTION

Mobile learning (m-learning) is the most recently developed category of electronic learning (e-learning), both of which are valuable tools in instructional conversation. What makes m-learning unique, and thus deserving of its study as an independent concept, is mobility; learners have the opportunity to go not just beyond the classroom, but beyond the limits of desktop and even laptop computers to engage in instructional environments. As an independent concept, m-learning has its own hardware and network technology, as well as a relationship with and difference from e-learning. As a component of instructional conversation, m-learning provides learners with opportunities to engage in discussion from almost any location at any time, making the conversations much more natural and beneficial to the group.

M-learning technology is, to support the uniqueness of the discipline, mobile. Devices that people carry on a regular basis and can access at almost any location are what drive m-learning practice. Working in concert with these devices is mobile networking technology, which provides the mobile learner with access to instructional material from a wide variety of locations and frees them from being tied to a cabled network connection at a static location.

Given that m-learning is using modern technology to achieve its goals, it is reasonable to associate m-learning with e-learning, and this is entirely correct. M-learning can in many ways be viewed either as an extension of e-learning or as a specific component of e-learning. While a discussion of detailed definitions will come later, it is important to recognize that entering into m-learning is not a departure from e-learning; the practitioner is simply adding new tools to their box.

There are four main objectives of this chapter, all of which relate to preparing instructional conversation practitioners to integrate m-learning

into their teaching. The first of these objectives is to gain knowledge of m-learning hardware. As with any technical application, learning the available tools is a critical first step in applying them. Second, readers should gain knowledge of m-learning networking. As will be explored, hardware mobility is of little use without a mobility of information, and information mobility is enabled by mobile networking. Third, readers will gain an understanding of the relationship of m-learning to e-learning which will provide them with a base from which to launch their own m-learning applications. Fourth, readers will review current applications of m-learning technology in the field of instructional conversation to provide examples of how to apply their new knowledge of m-learning to their own instructional conversation environments.

BACKGROUND

To begin an exploration of m-learning, it is first necessary to build a vocabulary of terms for the field and how those terms will be used here.

Defining m-learning is somewhat complicated, in part because m-learning has not been in use for long enough to firmly establish what it is and what it is not. Some of the definitions currently in the literature include: Kambourakis, Kontoni, and Sapounas (2004) define m-learning as being, "The point at which mobile computing and e-learning intersect to produce an anytime, anywhere learning experience." Colazzo, Ronchetti, Trifonova, and Molinari (2003) state that, "A mobile learning educational process can be considered as any learning and teaching activity that is possible through mobile tools or in settings where mobile equipment is available."

With this definition in mind, it is important to define exactly what is meant by the term mobile device. By most definitions personal digital assistants (PDAs), mobile phones, and MP3 players can be considered mobile devices (Mellow, 2005; Andronico, Carbonaro, Casadei, Colazzo, Molinari, & Ronchetti, 2003). Outside of these definitions, however, there is a certain amount of disagreement over exactly what constitutes mobile technology. The biggest question is whether or not laptop computers are mobile devices. In one sense they are; laptops can operate on battery power and access wireless networks for communication. In another sense, however, laptops are not truly mobile because they are not handheld devices, they require a user to be seated or standing beside a table or shelf to use, and they also need to be carried either independently or in a large case. For the purposes of the chapter, mobile devices will be defined as those devices that are small enough to fit in a shirt or jacket pocket and can be used in a variety of different environments, so laptops are not included. That said, it is important to remember that laptop and even desktop computers have the capability to access m-learning media, they are just not m-learning devices.

Now that the hardware being used in m-learning has been defined, the actual media formats of m-learning will be explored. As will become obvious, some media formats are platform-specific, while others will work across a wide variety of different mobile devices.

Probably the most basic and also the most proprietary media being used in m-learning is the short message service (SMS) via mobile phone. SMS is, in essence, a small format mobile e-mail application. Users can send short messages, usually around 150-200 characters, over mobile networks where they are received by users' mobile phone handsets. What defines this technology as basic in the field of m-learning is that instead of being user-requested the messages are sent to users at the discretion of the provider. Because SMS applications lack the on-demand features seen with other m-learning technology, it is classified here as a lower-level application. This is not in any way to say that SMS is not important or useful, quite the contrary, and later sections will highlight just how useful SMS ap-

plications can be to certain environments. The second factor that somewhat limits SMS is the proprietary platform required to support it. For now, SMS only works with mobile phones, which necessitates anybody in a system using SMS to have a handset and service. Many, if not most, students already have access to this technology, but there is no guarantee of universal access and, additionally, service contracts vary as to the cost, if any, of each text message received, which can add to the expense of the student.

Personal digital assistants (PDAs) are the next category of mobile learning technology to define. These handheld devices have changed dramatically from their early days of being just address books and calendars with a calculator. The best of these devices today are much closer to being palmtop computers than anything else, integrating wireless networking technology, e-mail, and even Internet access along with the traditional date book functions. Also, the most fully featured models today have the option of storing and playing media files of multiple formats and utilizing removable storage media to provide the storage capacity for those files. What this utility means for the m-learning environment designer is that any number of m-learning media formats can be accessed by the learner with just one device. This accessibility lends considerable flexibility to the instructional designer, and also to the learner, who has the option of using their choice of media at any given time with their PDA.

Probably the fastest growing and most universally accepted m-learning media format is the podcast. Podcast, at their most basic level are simply digital audio files, most often formatted as MP3 files. While the term Podcast was derived from the Apple Computer company's MP3 player, the iPod, it is not necessary to have an iPod to listen to podcasts. Any electronic device capable of playing MP3 files can play a podcast. As MP3 technology spreads, these devices include not only computers and dedicated MP3 players, but also PDAs, mobile phones, and even such unexpected items as sunglasses and wristwatches. All podcasts contain audio information that can be listened to on a mobile device, but they are not limited to just audio information. Podcasts can include still images that change in time with the audio playback, giving the user the opportunity to view slides or other visual aides during the presentation, much like watching an instructor's PowerPoint presentation during a live lecture. The next step beyond the inclusion of still images actually steps beyond being a podcast and turns into something different.

Video podcasts, sometimes referred to as vodcasts, are self-contained audio and video files that are compiled in one of a variety of digital video file formats. Vodcasts can be distributed as instructional video or can be a full video file of a lecture, as opposed to the audio only or audio plus still images that are possible with a regular podcast. While vodcasts can obviously convey larger amounts of information, they are not as accessible to users as the audio podcast. To play a vodcast, a device must have a screen and video playback capability, requirements that eliminate many MP3 players and other devices. Also, vodcasts require the viewer to watch the presentation, as opposed to just listening to information, which eliminates vodcasts as an option for use while driving or doing other activities that are compatible with listening to audio but not viewing video.

There are two sides to mobile technology, the first being hardware and the second being networking. Currently, there are two main wireless networking standards being used for two primary purposes.

The IEEE 802.11 communication protocol, commonly referred to as Wi-Fi, is the most prevalent wireless networking technology currently in use. Wi-Fi is commonly available in two speeds, B, which communicates at a speed of 11 kilobits per second (kps) and G, which communicates at a speed of 54 kps. There are other standards in existence, but B and G are what most networks operate on and both B and G wireless network

devices can operate on both B and G wireless networks. What Wi-Fi technology provides to the mobile computing user is the opportunity to access networked resources while away from wired connections, and also to quickly and easily collaborate with other mobile technology equipped teammates. Details of exactly how the technology is employed will be explored following the other primary wireless technology in use, Bluetooth.

The IEEE 802.15.1 wireless communication standard is what is commonly referred to as Bluetooth. While Wi-Fi is traditionally a wireless network connection to the Internet, Bluetooth functions much more frequently as a device-to-device data transfer mechanism. Using Bluetooth, it is possible, without cables or external storage media, to transfer data between two Bluetooth-equipped hardware devices quickly and easily. This technology gives mobile technology users the capability of sharing important information between, for example, their laptop, PDA, and mobile phone, all without having to carry cables or physically connect the devices to each other. The advantage to this technology is that it provides the mobile learner with an easy way to utilize information on multiple devices, thereby freeing them from having to rely on any single piece of hardware for their mobile computing.

With these two widespread wireless protocols defined, how are they used? There are two different applications of Wi-Fi technology that will be defined here, hotspots and ad-hoc networks.

Hotspots are simply areas where Internet access is available via Wi-Fi technology. Originally seen in coffee shops and similar urban locations as a draw for customers to be able to stay connected to e-mail and Internet resources while spending time in the business, the hotspot phenomenon has expanded rapidly. Currently, there are many businesses and hotels that offer free Wi-Fi access as a service to their customers, and in some cities entire downtown districts have been equipped with broad coverage Wi-Fi so that anywhere in the area, people can access the Internet from their wireless devices. Wi-Fi compliments a wide variety of different m-learning applications. Podcasts can be downloaded over a wireless network, course management systems can be accessed, and e-mail can be used as a direct communication between learners or between learners and instructors.

The other capability of Wi-Fi is to create what are called ad-hoc networks between devices. This process does not require a router or other hardware, only compatible devices that include Wi-Fi functionality. What this ad-hoc networking capability allows users to do is to gather at any location and quickly, easily, without cables, transfer files and information between hardware devices while working collaboratively.

The ad-hoc networking capabilities of Wi-Fi are somewhat similar to what Bluetooth technology gives to users. While it is technically possible to offer Internet access via a Bluetooth network, it is very rarely used as such. Bluetooth's primary use is as a device-to-device networking service. File transfers, data backups, and even the dialing of mobile phones can be accomplished through Bluetooth communication between the correct types of devices. In the context of mobile learning, this connectivity between devices enables users to easily share information between different pieces of mobile hardware so that they can access that information on whatever type of device is most convenient.

To finish the discussion of mobile networking technology, the newest and potentially most useful addition to the field will be examined. Third-generation technology, marketed as 3G, is broadband wireless networking via cellular phone networks. The speed of 3G is 384 kilobits per second. The key advantage to 3G technology is that unlike Wi-Fi and Bluetooth, both of which are local area networks (LANs), 3G is a wide area network (WAN). WAN technology covers a much broader area and has the capability of providing wireless Internet services to 3G-equipped hardware anywhere in a city or even beyond city limits; the scope of coverage depends only on

the presence of compatible transmission towers. What this means to users of 3G technology is that they have the ability to connect in a much wider variety of locations than the users of Wi-Fi technology, thus giving them even more opportunities to participate in a learning activity on their own time at the location of their own choosing. The one serious drawback to this technology is that at the moment it is still relatively expensive, around $60 a month, versus the usually free access to hotspots afforded by Wi-Fi technologies.

Discussions of the Topic

Having defined the technology that is involved in m-learning, how does m-learning contribute to instructional conversation? To answer this question it is first necessary to determine just what instructional conversation is and how it is being used in today's classrooms.

Putnam and Borko (2000) define instructional conversation as, "…a mode of instruction that emphasizes active student involvement in goal and meaning-oriented discussions." Instructional conversation as a teaching tool has been found to be effective in working with students who are at risk educationally (Watson, 2000). Doherty, Hilberg, Epaloose, and Tharp (2002) make the point that "fully inclusive, small-group ICs are stressed because they increase the participation of all students, including the more passive learners." Indeed, the increased accessibility to learners who may not be likely to join a live conversation but are comfortable conversing via discussion board or other computer-mediated format where they have the time to collect their thoughts, draft, and edit their responses, has long been a benefit of online education, and this advantage carries through to the mobile applications of instructional conversation.

From a technical standpoint, "the increasing availability of computer-based tools and resources and the growing emphasis on using these in subject teaching and learning has a potentially significant impact upon established patterns of classroom interaction" (Hennessy, Deaney, and Ruthven, 2005). In relation to multimedia, Mayer (2003) discusses the personalization effect, defined as the fact that, "…students learn more deeply from a multimedia explanation when the words are presented in conversational style rather than formal style."

What is it about m-learning that makes it a valuable component in instructional conversation? There are considerable similarities between instructional conversation and contextual life-long learning, whose features are that:

- "learning is not confined to pre-specified times or places, but happens whenever there is a break in the flow of routine daily performance and a person reflects on the current situation, resolves to address a problem, to share an idea, or to gain an understanding
- formal education can not provide people with all the knowledge and skills they need to prosper throughout a lifetime. Therefore, people will need to continually enhance their abilities, in order to address immediate problems and to participate in a process of continuing vocational and professional development."

Sharples (2000) discusses the connection of m-learning technology to lifetime learning and outlines the attributes of technology devices that facilitate this goal:

- **Highly portable:** So they can be available wherever the user needs to learn
- **Individual:** Adapting to the learner's abilities, knowledge, and learning styles and designed to support personal learning, rather than general office work
- **Unobtrusive:** So that the learner can capture situations and retrieve knowledge without the technology obtruding on the situation
- **Available anywhere:** To enable communi-

cation with teachers, experts and peers
- **Adaptable:** To the learner's evolving skills and knowledge
- **Persistent:** To manage learning throughout their lifetime, so that the learner's personal accumulation of resources and knowledge will be immediately accessible despite changes in technology
- **Useful:** Suited to everyday needs for communication, reference, work, and learning
- **Intuitive:**To use by people with no previous experience with technology"

It is apparent from Sharples' technology attributes that accessibility and usability are keys to the success of learning technology, particularly m-learning technology whose purpose is to be a part of a learner's daily life. Combined with the purpose of instructional conversation, that being to engage learners in educational discussions of a topic, m-learning technology can serve to greatly enhance the learning experience. Having precisely defined the components, how this contribution is made by the technology can be explored.

MAIN THRUST OF CHAPTER

Issues

M-learning as a technology has some very important issues in its application and use. User access is a major consideration when considering the implementation of a m-learning program, as is the choice of what kind of technology, which directly affects pedagogy options, will be utilized. Also important are questions of instructional design and training for both instructors and learners.

The number one consideration when embarking on an m-learning program is that of learner access to the technology. In practice, the best design, the best content, and the best intentions for a program are all useless if the target population does not possess the technological resources to access the material. It is absolutely critical to any m-learning application that there is market saturation of the technology in the targeted group of learners (Viteli, 2000). The challenge presented by this issue varies by situation, with some environments being much easier to manage than others. A professional graduate-level program, for instance, can easily place a technology requirement on all incoming students, specifying hardware, software, network, and operating system requirements as a prerequisite for participating in the program. Other situations, however, do not have a population with the resources to buy into new technology. Most K-12 public schools, for instance, have neither their own resources nor the universal student resources to levy technology requirements on learners. What this means for the instructional designer, for the practitioner of instructional conversation, is that the resources of the learner have to be considered and accounted for before employing a new technology. To work, m-learning requires population access to the technology.

Assuming user accessibility, the choice of what kind of m-learning technology to use is critical to the design of an m-learning system. While some technologies are directly involved in instructional conversation, others are facilitators of conversations, serving to make the conversations that do occur more effective, or even just making them accessible, and therefore possible. Because of this, multiple technologies may be chosen in any given instructional design. Some of the questions that can be asked before making this decision are: does communication need to be one-way or two-way? Can the information be conveyed with text only, audio only, with the inclusion of pictures, or is video required? Can messages be sent out at the discretion of the instructor or do learners need to be able to access information at will? Do messages need to be confidential or shared among all participants? All of these things factor into making decisions about the components of an m-learning system. Table 1 provides a list of

Table 1. M-learning technologies and their applications in instructional conversation

M-learning Technology	Application in Instructional Conversation
e-mail	distribution tool for information or media, conversation tool via listserv or direct learner/instructor discussion
SMS	quick, reliable distribution of notes and information to learners
PDA	mobile network access for e-mail, discussion boards, mobile computing applications to prepare for assigned conversation topics
Podcasts	deliver media to learners in advance of discussions, archive discussions for later reference
Vodcasts	deliver video examples of activities, distribute instructional video files of the topic for discussion
Wi-Fi	mobile network access on a variety of mobile devices providing access to online learning materials
Bluetooth	ad-hoc networking for group discussions and local file transfer of learning materials

m-learning technologies and their related application in instructional conversation.

Controvercies

The biggest controversy facing m-learning technology, and especially podcasting and vodcasting is one of intellectual property rights. Not only is there the very real concern of violating an existing copyright by distributing material that an instructor does not have the rights to, but the rights to material created by an instructor are a matter of serious debate as well. It will be largely left to the legal system to define who owns specific content in specific situations, but considerable litigation can be expected in the future to determine exactly how intellectual property rights will be determined for instructors and professors using digital media to record and distribute their course content.

Problems

Currently the question of how to provide technology access to learners is a serious problem when considering a move to m-learning technology. In a corporate training environment, this problem is simply one of whether or not the company has the resources, or the desire, to invest in the technology for employees to use. While a definite problem, this is not something insurmountable. For public universities and preparatory schools, however, the issue is much more serious.

As technology continues to play a more and more important role in education there are serious effects being seen from what is called the digital divide. Basically, the digital divide is the performance gap between students from families that do have computers and other technology in the home and those who do not. A full discussion of all the related issues is beyond the scope of this chapter, but the fact that the digital divide exists does have real implications on m-learning integration in public schools at all levels.

Where private schools have the luxury, more often than not, of mandating the purchase of specific hardware and/or software as a school supply, public schools do not. This puts public schools in the uncomfortable position of having to either provide students with technology or not pursue m-learning. Given the funding difficulties being faced by most public schools, preparatory and university alike, the purchase of multiple PDAs,

much less mobile phones with service, is highly unlikely. This being the case, many, if not most m-learning applications remain beyond the reach of public institutions of learning.

What these problems all have in common is that they are problems of access. How does an organization provide its learners with access to m-learning materials? The access is key, the best instructional design and the fastest servers are unable to improve the learning environment if learners do not have the means to access the instructional resources on that server.

Solutions

Unfortunately for practitioners, solutions to the access problems are not yet here. It can be predicted that in the near future there will be hardware solutions inexpensive enough to provide wide access to mobile technology but when that time is, can not be known.

The most encouraging current work involves the MIT Laptop Project, which has designed a Wi-Fi capable, hand-crank-powered laptop running open source software that will be available when in production for $100 per unit. The goal of this project is to bridge the digital divide for developing countries and communities. While by most definitions laptops are not considered to be mobile technology, it is encouraging that efforts are being made to provide affordable, wirelessly networked hardware and software to those people who are least able to afford it. It is impossible to project when this type of work will extend to truly mobile devices, but as with everything in technology, it is reasonable to forecast drastic drops in the price of mobile devices compared to the performance that they offer.

For solutions to come, m-learning likely needs to mature as a discipline and be well established enough to demand the attention of development efforts. Currently, the technology and the pedagogy of mobile learning are so new that very little outside the field of instructional technology is being done.

Recommendations

The best recommendation for current or future practitioners of m-learning is to establish a good knowledge base of not only technical expertise, but of e-learning pedagogy. While hardware, software, and networking technologies change, the basic purpose of m-learning, to convey information to learners, does not. If a practitioner has a clear picture of what they are trying to convey and how they want learners to interact with the system, then virtually any technology can be adapted to serve that purpose.

M-learning in Instructional Conversation Practice

M-learning technology, as an emerging field of study, is still new to the practice of instructional conversation but there are already some excellent examples of its usefulness. Mobile technology as a conversational tool needs to meet several criteria:

…the minimum equipment needed to hold conversations that promote effective learning consists of the following: a shared language in which to express commands, questions, instructions, agreements and disagreements; minds capable of giving rise to conversation about some shared phenomenon; and an external representation of the phenomenon that can provide a common framework for exploring differences of conception. Relating this to the design of learning technologies, we require more than transparent channels of communication and a means for transmitting knowledge, we also need a shared language (among learners, and between learners and computer systems), a means to capture and share phenomena, and a method of expressing

and conversing about abstract representations of the phenomena (Sharples, 2002).

To facilitate these conversations, mobile technology is being used to enable impromptu collaboration, facilitate discussion in language learning classes, and provide practical learning experiences for groups of students.

Mobile devices using Wi-Fi or Bluetooth networking technology are very well-suited to participate in quickly assembled, wireless networks between multiple devices that are being used close to each other. The general term for this type of connection between machines is an ad-hoc network, a network that comes together as the users of the devices come together and then ceases to exist when the users leave the area. Using these ad-hoc networks, it is possible for users to collaborate and exchange information via mobile devices at virtually any location quickly and easily. The attributes of ad-hoc mobile networking that support its use in instructional conversation are that it is opportunistic, spontaneous, proximity-based, and transient (Kortuem, Schneider, Preuitt, Thompson, Fickas, & Segall, 2001). By capitalizing on these attributes, mobile learners have the freedom to effectively engage in collaborative activities in many more locations and at many more times than they would with more traditional file storage and transfer technologies.

Related to ad-hoc networking applications is the use of mobile technology in language learning environments. Liang, Liu, Wang, and Chan (2005) discusses the use of modified pocket electronic dictionaries in foreign language courses, citing one characteristic exhibited as being, "…interactive-based, connecting it with the student's ELMD to perform individualized, group-based, or whole-class learning activities expanding the degree of student participation." By utilizing the technical capabilities of the mobile device, in this case a pocket electronic dictionary, discussion among students and instructors can be facilitated. This application is very similar to what can be accomplished with student PDA devices, enabling participants to share files and other information via mobile device to enhance the learning experience of everyone involved in the conversation.

The value of these experiences, as delivered by mobile technology, is explained by Lowand O'Connell (2002):

The data connectivity and communication aspects of many mobile devices support social interaction, collaboration and construction of learning—for example:

- *The ability of mobile phones to exchange media via MMS, e-mail, and SMS*
- *The ability of PDAs to connect to the Internet and interface with both synchronous and asynchronous communication tools*
- *The exchange of compatible memory cards between learners to copy reviewed and recommended audio or video resources across media players, PDAs or mobile phones.*

What these communication abilities address in the m-learning arena is the need for the learner to relate, one of the four R's of mobile learning, the other three being record, recall, and reinterpret (Low and O'Connell, 2002). The practical application of mobile technology to relating to others in a learning environment involves:

- *The learner may use a portable device to communicate with other people (for example, with other learners), or with a teacher (that is, in a learning relationship).*
- *The learner can use the device to communicate directly and synchronously (for example, mobile phone conversation), or access asynchronous communication services (for example, Web discussion board or Web log).*

- *They can also recommend and share resources, for example, linking mobile devices (usually wirelessly) and sending a file from one to the other.*
- *Communicative and collaborative: underpinned by a social constructivist or connectivist theory of learning (Low and O'Connell, 2002).*

Mobile applications of instructional conversation as a whole are growing rapidly, as is the field m-learning. Patten, Sanchez, and Tangney (2006) describes several new applications that have the goal of facilitating learner interaction. Four of the applications are described here:

- **TxtIT:** An SMS messaging application "to support interactivity in the classroom"
- **Mapping Challenge:** A treasure hunt game using mobile phones
- **SortIT:** A collaborative problem-solving application for handheld devices

The final m-learning collaboration tool to be highlighted here is an automated tutoring system called Alykko. Silander and Rytkonen (2005) introduce Alykko with the explanation that, "Communication, collaborative knowledge building, observations, and finding new innovations describe the student's learning activity in the authentic learning. When using mobile devices, students are able to construct useful knowledge in a real situation." The Alykko system works with students to construct an instructional conversation by providing "…focusing and deepening questions to a student based on the student's progress and activity" (Silander and Rytkonen, 2005). As such, the system actually gives learners the opportunity to engage in an environment of instructional conversation while working independently; the electronic device prompts the learner to think and analyze the material in question, much the same as a conversation with peers would do.

It is clear from these examples that a wide range of things can be accomplished in the realm of instructional conversation using mobile technology. As was said earlier, m-learning, and the technology behind it, are still new innovations and their full use in instructional conversation has yet to be determined. By starting now, and building a solid understanding of the basic components, it will be possible to find new applications of the technology as capacities and capabilities change.

FUTURE TRENDS

Emerging Trends

One of the most important emerging trends in m-learning is the shifting undergraduate student population. With more and more students being what are classified as "nontraditional," that is, older students with full-time jobs and possibly families, demands for distance and mobile learning opportunities will likely increase.

To best serve these nontraditional students, it is often important to engage them in conversation and encourage them to share their professional experiences from the workplace. Given the breadth of experiences that the average nontraditional student has versus the traditional undergraduate, instructional conversation becomes even more important. Combining the need for instructional conversation with the demanding schedules of this new category of learners lends itself very well to the incorporation of m-learning technologies to facilitate the conversations.

Future Research Opportunities

Because m-learning is such a new discipline, the field is open to a wide variety of future research projects. In relation to the use of m-learning technologies with nontraditional learners, it would be particularly interesting to do a study on the use of

m-learning with traditional versus nontraditional students to see if there is any performance gap based on familiarity with the technology.

For any group of students, it would be useful to study the actual impact of m-learning technology on instructional conversation, perhaps by running a study where one section of a course uses the technology and another does not. As with many educational studies, the design of the assessment instrument would be very important, but the results could help to better design instructional conversation via m-learning in the future.

CONCLUSION

Overall Coverage

There is a broad and expanding selection of mobile learning technologies available. Just during the writing of this chapter, the competition for video-capable MP3 players has exploded, with multiple large firms competing with the iPod. What matters far more than the technology, however, is the application of the technology and its relationship to instructional conversation. Regardless of the medium, discussions among learners and instructors are the key component to instructional conversation and an important contribution to any learning environment. In order to capitalize on these advantages the practitioner must have a good understanding of both pedagogy and technology.

Concluding Remarks

Technical advancements have given instructors a wide variety of tools to use in their goal of conveying information to learners. As practitioners, it is important to stay current in technical trends, but also to remain grounded in good design and good practice. Think of the learner first, and tailor any new innovation to serve them. Not every new idea will be successful, but if they are student-focused every new idea will be worth pursuing.

FUTURE RESEARCH DIRECTIONS

Because m-learning is so new and the research into the field is at such an early stage, there are many possibilities for mobile technology and instructional conversation research. Perhaps the most important is practical research in the effects of mobile technology when applied to instructional conversation environments. Closely linked to this study of effectiveness would be a study of the different effectiveness of different mobile technologies.

Before studies could be conducted it is necessary to construct an assessment instrument to determine the volume and quality of activity in the instructional conversation. This instrument would need to define specific points that researchers could use to analyze conversation participation. These points might classify types of posts to a conversation, what constitutes a significant contribution, and how contributions can be categorized in regard to their level of significance to the conversation. Before research studies could be conducted it would be necessary to validate this instrument.

The application of the instrument would be to implement mobile technology as a component of instructional conversation into a course design that already uses instructional conversation with some other technology. Ideally, this experimentation would occur in two different sections of a single course taught during the same semester by the same instructor to minimize differences between instructional environments so that observed differences in student activities in the instructional conversations can best be attributed to the difference in technology. By analyzing the results of these two courses with the assessment instrument, it should be possible to establish the differences

in student activity with or without the inclusion of mobile technology. Ideally, this study would be run repeatedly, comparing each of the different mobile technologies with an environment not using mobile technology to discover which mobile technologies are effective and which are not.

As a second step of this process, subsequent studies of the same design could be run to determine the effectiveness of different types of mobile technology. By using two different types of mobile technology in a study instead of one class using mobile technology and one not, it should be possible to identify the most effective mobile technologies in instructional conversation. It is highly likely that a long-term, intensive study of this type would result in pairing different mobile technologies to different instructional conversation environments and thus some technologies would be useful for some environments but not necessarily for others.

For instructional designers and other technology researchers, there is much work to be done in testing the use of mobile technology in instructional conversation applications. As the technology continues to change research opportunities will expand.

REFERENCES

Andronico, A., Carbonaro, A., Casadei, G., Colazzo, L., Molinari, A., & Ronchetti, M. (2003). *Integrating a multi-agent recommendation system into a mobile learning management system*.

Doherty, R.W., Hilberg, R. S., Epaloose, G., & Tharp, R. (2002). Standards performance continuum: Development and validation of a measure of effective pedagogy. *The Journal of Education Research, 96*(2), 78-89.

Hennessy, S., Deaney, R., & Ruthven, K. (2005). Emerging teacher strategies for mediating 'Technology-integrated Instructional Conversations': a socio-cultural perspective. *The Curriculum Journal, 16*(3), 265-292.

Kambourakis, G., Kontoni, D-P. N., & Sapounas, I. (2004). Introducing attribute certificates to secure distributed e-learning or m-learning services. In *Proceedings of the IASTED International Conference* (pp. 436-440). Innsbruck, Australia:.

Kortuem, G., Schneider, J., Preuitt, D., Thompson, T. G. C., Fickas, S., & Segall, Z. (2001). When peer-to-peer comes face-to-face: Collaborative peer-to-peer computing in mobile ad hoc networks. In *Proceedings of the First International Conference on Peer-to-Peer Computing.*

Liang, J., Liu, T., Wang, H., & Chan, T. (2005). Integrating wireless technology in pocket electronic dictionary to enhance language learning. In *Proceedings of the Fifth IEEE Conference on Advanced Learning Technologies* (pp. 495-497).

Low, L., & O'Connell, M. (2002). Learner-centric design of mobile learning. In *Proceedings for ASCILITE 2002.*

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction, 13,* 125-139.

Mellow, P. (2005). The media generation: Maximize learning by getting mobile. In *Proceedings for Ascilite 2005: Balance, Fidelity, Mobility: Maintaining the Momentum?* (pp. 469-476).

Patten, B., Sanchez, I., & Tangney, B. (2006). Designing collaborative, constructionist and contextual applications for handheld devices. *Computers & Education, 46,* 294-308.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4-15.

Sharples, M. (2000). The design of personal mobile technologies for lifelong learning. *Computers & Education, 34*, 177-193.

Sharples, M. (2002). Disruptive devices: mobile technology for conversational learning. *International Journal of Continuing Engineering Education and Life Long Learning, 12*(5-6), 504-520.

Silander, P., & Rytkonen, A. (2005, October 25-28). An intelligent mobile tutoring tool enabling individualisation of students' learning processes. In *Proceedings M-learn 2005 4th World Conference on M-learning* Cape Town, South Africa.

Viteli, J. (2000). Finnish future: From e-learning to m-learning? In *Proceedings for the ASCILITE Conference*.

Watson, J. (2000). Constructive instruction and learning difficulties. *Support for Learning, 15*(3), 134-140.

ADDITIONAL READINGS

Bannasch, S. (2001). Educational innovations in portable technologies. In R. Tinker & J. Krajcik (Eds.), *Portable technologies: Science learning in context* (pp.121-145). New York: Kluwer Academic/Plenum Publishers.

Bull, G., Bull, G., Garofalo, J., & Harris, J. (2002). Grand challenges: Preparing for the technological tipping point. *Learning and Leading with Technology, 29*(8).

Chen, Y. S., Kao, T. C., & Sheu, J. P. (2003). A mobile learning system for scaffolding bird watching learning. *Journal of Computer Assisted Learning, 19*(3), 347-359.

Churchill, E., Snowdon, D., & Munro, A. (2001). *Collaborative virtual learning environments, digital places and spaces for interaction.* London: Springer-Verlag.

Cole, H., & Stanton, D. (2003). Designing mobile technologies to support co-present collaboration. *Pers Ubiquit Comput* (7).

Curtis, M., Luchini, K., Bobrowski, W., Quintana, Ch., & Soloway, E. (2002, August). Handheld use in K-12: A descriptive account. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)* (pp. 23-30). Växjö, Sweden:

Danesh, A., Inkpen, K. M., Lau, F., Shu, K., & Booth, K. S. (2001). Geney: Designing a collaborative activity for the palm handheld computer. In *Proceedings of the Conference on Human Factors in Computing Systems.*

Davis, S. M.(2002, August). Research to industry four years of observations in classrooms using a network of handheld devices. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)* (pp. 31-38). Växjö, Sweden:

Divitini, M., Haugalokken, O. K., & Norevik, P.(2002, August). Improving communication through mobile technologies: Which possibilities?" In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)* (pp. 86-90). Växjö, Sweden:

Dvorak, J. D., & Burchanan, K.(2002, June). Using technology to create and enhance collaborative learning. In *Proceedings of 14th World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA 2002)* Denver, CO, USA:

Garner, I., Francis, J., & Wales, K.(2002, June). An evaluation of the implementation of a short messaging system (SMS) to support undergraduate students. In *Proceedings of the European Workshop on Mobile and Contextual Learning* (pp. 15-18). Birmingham, UK:

Gay, G., Rieger, R., & Bennington, T. (2002). Using mobile computing to enhance field study. In T. Koschmann, R. Hall, & N. Miyake (Eds.), *CSCL 2: Carrying forward the conversation*. London: Erlbaum.

Hakkarainen, K., Lipponen, L., & Järvelä, S. (2001). Epistemology of inquiry and computer-supported collaborative learning. In T. Koschmann et al. (Eds), *CSCL2: Carrying forward the conversation* (pp. 128-156). Mahwah, NJ: Erlbaum.

Kirner, T. G., Kirner, C., Kawamoto, A. L. S., Cantao, J., Pinto, A., & Wazlawick, R. S. (2001). Development of a collaborative virtual environment for educational applications. In *Proceedings WEB3D 2001* (pp. 61-68).

Klopfer, E., Squire, K., & Jenkins, H.(2002, August). Environmental detectives: PDAs as a window into a virtual simulated world. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)* (pp. 95-98). Växjö, Sweden:

Liu, T. C., Wang, H. Y., Liang, J. K., Chan, T. W., Ko, H. W., & Yang, J. C. (2003). Wireless and mobile technologies to enhance teaching and learning. *Journal of Computer Assisted Learning, 19*(3), 371-382.

Liu, T., Wang, H., Liang, J., Chan, T., & Yang, J.(2002, August). Applying wireless technologies to build a highly interactive learning environment. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)* (pp. 63-70). Växjö, Sweden:

Markett, C., Arnedillo Sa´nchez, I., Weber, S., & Tangney, B. (2004). Pls turn ur mobile on: Short message service (SMS) supporting interactivity in the classroom. In Kinshuk, D. G. Sampson, & P. Isaias (Eds.), *Cognition and exploratory learning in digital age* (pp. 475-478). Lisbon: International Association for Development of the Information Society Press.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: when presenting more material results in less understanding. *Journal of Educational Psychology, 93*, 187-198.

McGreen, N., & Arnedillo Sa´nchez, I. (2005a). Mapping challenge: A case study in the use of mobile phones in collaborative, contextual learning. In P. Isaias, C. Borg, P. Kommers, & P. Bonanno (Eds.), *Mobile learning 2005* (pp. 213–217). Malta: International Association for Development of the Information Society Press.

Mifsud, L.(2002, August). Alternative learning arenas—pedagogical challenges to mobile learning technology in education. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)* (pp. 112-116). Växjö, Sweden:

Milrad, M., Perez, J., Hoppe, U. (2002, August). C-notes: Designing a mobile and wireless application to support collaborative knowledge building. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)* (pp. 117- 120). Växjö, Sweden:

Moreno, R., & Mayer, R. E. (2000). Engaging students in active learning: the case for personalized multimedia messages. *Journal of Educational Psychology, 92*, 724-733.

Silander, P., Sutinen, E., & Tarhio, J. (2004). Mobile collaborative concept mapping—combining classroom activity with simultaneous field exploration. In *Proceedings of The 2nd IEEE International Workshop on Wireless and Mobile Technologies In Education (WMTE 2004)* (pp. 114-118).

Snowdon, D., Churchill, E. F., & Munro, A. J. (2001). Collaborative virtual environments: Digital spaces and places for CSCW: An introduction. In E. F. Churchill, D. N. Snowdon, & A. J. Munro (Eds), *Collaborative virtual environments: Digi-*

tal places and spaces for interaction (pp. 3-17). London: Springer-Verlag.

Stone, A., Briggs, J., & Smith, C.(2002, August). SMS and interactivity—some results from the field, and its implications on effective uses of mobile technologies in education. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)* (pp. 147-151). Växjö, Sweden:

Wang, H. Y., Liu, T. C., Chou, C. Y., Liang, J. K., Chan, T. W., & Yang, S. (2004). A framework of three learning activity levels for enhancing the usability andfeasibility of wireless learning environment. *Journal of Educational Computing Research, 30*(4), 331-35.

Waycott, J., Scanlon, E., & Jones, A.(2002, June). Evaluating the use of PDAs as learning and workplace tools: An activity theory perspective. In *Proceedings of theEuropean Workshop on Mobile and Contextual Learning* (pp. 34-35). Birmingham, UK:

Zurita, G., & Nussbaum, M. (2004). A constructivist mobile learning environment supported by a wireless handheld network. *Journal of Computer Assisted Learning, 20*(4).

Zurita, G., Nussbaum, M., & Sharples, M. (2003). Encouraging face-to-face collaborative learning through the use of handheld computers in the classroom. In *Proceedings of the Mobile HCI 2003.* Berlin, Heidelberg: Springer-Verlag.

This work was previously published in Handbook of Conversation Design for Instructional Applications, edited by R. Luppicini, pp. 388-402, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.7
Embedded Agents for Mobile Services

John F. Bradley
University College Dublin, Ireland

Conor Muldoon
University College Dublin, Ireland

Gregory M. P. O'Hare
University College Dublin, Ireland

Michael J. O'Grady
University College Dublin, Ireland

INTRODUCTION

A significant rise in the use of mobile computing technologies has been witnessed in recent years. Various interpretations of the mobile computing paradigm, for example, ubiquitous and pervasive computing (Weiser, 1991) and more recently, ambient intelligence (Aarts & Marzano, 2003)—have been the subject of much research. The vision of mobile computing is often held as one of "smart" devices operating seamlessly and dynamically, forming ad-hoc networks with other related devices, and presenting the user with a truly ubiquitous intelligent environment. This vision offers many similarities with the concept of distributed artificial intelligence where autonomous entities, known as agents, interact with one another forming ad-hoc alliances, and working both reactively and proactively to achieve individual and common objectives.

This article will focus on the current state of the art in the deployment of multi-agent systems on mobile devices and smart phones. A number of platforms will be described, along with some practical issues concerning the deployment of agents in mobile applications.

BACKGROUND

In the most general terms, an agent is one entity that acts, or has the authority to act, on behalf of another. In terms of information technology, an agent is a computational entity that acts on behalf of a human user, software entity, or another agent. Agents have a number of attributes that distinguish them from other software (Bradshaw, 1997; Etzioni & Weld, 1995; Franklin & Graesser, 1996; Wooldridge & Jennings, 1995):

- **Autonomy:** The ability to operate without the direct intervention from any entity, and possess control over their own actions and internal state.
- **Reactivity:** The ability to perceive their environment and react to changes in an appropriate fashion.
- **Proactivity:** The ability to exhibit goal-directed behavior by taking the initiative.
- **Inferential Capability:** The ability to make decisions based on current knowledge of self, environment, and general goals.
- **Social Ability:** The ability to collaborate and communicate with other entities.
- **Temporal Persistence:** The ability to have attributes like identity and internal state to continue over time.
- **Personality:** The ability to demonstrate the attributes of a believable character.
- **Mobility:** The ability to migrate self, either proactively or reactively, from one host device to another.
- **Adaptivity:** The ability to change based on experience.

An agent requires some space where it can exist and function, and this is provided for by an agent platform (AP). An AP comprises "the machine(s), operating system, agent support software,…agent management components…and agents" (FIPA, 2000, p. 6). The AP allows for agent creation, execution, and communication.

The majority of computer systems currently in operation use algorithms that are based on the concept of perfect information. The problem is that in the real world, businesses often require software functionality that is much more complex than this (Georgeff, Pell, Pollack, Tambe, & Wooldridge, 1999). Typically, computational entities within these systems should have an innate ability to deal with partial information and uncertainty within their environment. These types of systems are highly complex and are intractable using traditional approaches to software development. The rate at which business systems must change, due to market pressures and new information coming to light, requires software architectures and languages that more efficiently manage the complexity that results from alterations being made to the code and the specifications.

Agent architectures, and in particular belief-desire-intention (BDI) (Rao & Geogeff, 1995) agent architectures, are specifically designed to deal with these types of issues and thus contain mechanisms for dealing with uncertainty and change. A problem with traditional systems is that they assume that they exist within a static or constant world that contains perfect information. The types of mobile systems that we are concerned with are dynamic and perhaps even chaotic, embedded with agents that have a partial view of the world and which are resource bounded.

Agents rarely exist in isolation, but usually form a coalition of agents in what is termed a multi-agent system (MAS). Though endowed with particular responsibilities, each individual agent collaborates with other agents to fulfill the objectives of the MAS. Fundamental to this collaboration is the existence of an Agent Communications Language (ACL), which is shared and understood by all agents. The necessity to support inter-agent communication has led to the development of an international ACL standard, which has been ratified by the Foundation for Intelligent Physical Agents (FIPA).

JAVA 2 MICRO EDITION (J2ME)

Most agent platforms developed for mobile devices have been written in the Java programming language—on mobile devices that usually means Java 2 Micro Edition (J2ME). This edition of Java contains a cut down API, a reduced footprint Java Virtual Machine, and a slightly different syntax (e.g., parameterized classes in Java 5). Java applications that contain dependencies on the idiosyncrasies of the different editions cannot be ported to a different range of devices without making alterations to the code. Their performance, however, is improved because the code is no longer developed to the lowest common denominator. Different algorithms and coding styles are now used for desktop machines and embedded devices rather than adopting comprised or overarching approaches that do not maximize the performance or maintainability of either.

A NUMBER OF AGENT PLATFORMS EXISTS FOR MOBILE DEVICES

3APL-M

3APL-M (Koch, 2005) is a platform that enables the fabrication of agents using the Artificial Autonomous Agents Programming Language (3APL) (Dastani, Riemsdijk, Dignum, & Meye, 2003) for internal knowledge representation. Its binary version is distributed in J2ME and J2SE compilations. 3APL provides programming constructs for implementing agents' beliefs, goals, basic capabilities, and a set of practical reasoning rules. The framework comprises an API that allows a Java application to call 3APL logic and deliberation structures.

Agent Factory Micro Edition

Agent Factory Micro Edition (AFME) (Muldoon, O'Hare, Collier, & O'Grady, 2006) is an agent platform developed for the construction of lightweight intelligent agents for cellular digital mobile phones and other compatible mobile devices. AFME is broadly based on Agent Factory (Collier, 2001), a pre-existing J2SE framework for the fabrication and deployment of agents. AFME differs from the original version of the system in that it has been designed to operate on top of the Constrained Limited Device Configuration (CLDC) Java platform augmented with the Mobile Information Device Profile (MIDP). CLDC and MIDP form a subset of the J2ME platform specifications. Though sharing the same broad objectives of the other projects mentioned in this section, AFME differs in a number of ways. With a jar size of 85k, it is probably the smallest footprint FIPA-compliant deliberative agent platform in the world. The platform supports the development of a type of software agent that is: autonomous, situated, socially able, intentional, rational, and mobile. An agent-oriented programming language and interpreter facilitate the expression of an agent's behavior through the formal notions of belief and commitment. This approach is consistent with a BDI agent model.

LEAP

Probably the most widely known agent platform for resource-constrained devices is the Light Extensible Agent Platform (LEAP) (Berger, Rusitschka, Toropov, Watzke, & Schichte, 2002). LEAP is a FIPA-compliant agent platform developed to be capable of operating on both fixed and mobile devices with various operating systems in wired or wireless networks. Since version 3.0, LEAP extends the Java Agent DEvelopment Framework (JADE) (Bellifemine, Caire, Poggi,

& Rimassa, 2003) by using a set of profiles that allow it to be configured for various Java Virtual Machines (JVMs). The architecture of the platform is modular and contains components for managing the lifecycle of the agents and controlling the array of communication protocols. The platform is split into several agent containers, one for every device used. These containers are responsible for passing messages between agents and choosing the appropriate communication protocol. One of these containers, known as the main container, includes agents that fulfill the white and yellow pages services as necessitated by the FIPA specification.

MAE

The MAE (mobile agent environment) (Mihailescu, Binder, & Kendall, 2002) agent platform has been designed to be independent of device and language implementations. To accomplish this, the platform is divided into a reference API specification, reference implementation, and non-standard implementation additions. The reference API specification is not dependent on programming language or hardware, and it contains the core platform components. The Reference Implementation contains all the device-dependent code required by the reference API specification. The third part, non-standard implementation additions, is used for application-specific components.

While this approach gives a high degree of platform independence, unless it is being deployed in an environment of homogeneous devices, it means a lot of work as each platform may require its own implementation.

MicroFIPA-OS

MicroFIPA-OS is an agent toolkit based on the standard FIPA-OS but optimized for resource-constrained mobile devices (Tarkoma & Laukkanen, 2002). It targets the personal Java platform and thus operates on personal data assistants. The system can run in minimal mode whereby agents do not use task and conversation managers. Yellow and white page services are provided in compliance with the FIPA specification. The platform is entirely embedded, however it is recommended that only one agent operate on low-specification devices.

NON-EMBEDDED AGENTS FOR MOBILE SERVICES

There are other types of agent platforms suitable for mobile services that do not embed the agents in the mobile device:

- platforms that use the mobile device as just an interface while the agents are executed on more capable hosts, for example MobiAgent; and
- platforms that do part of the execution on the mobile device, while simultaneously executing the remainder the task on other hosts such as KSACI (Hübner, 2000a, 2000b).

MobiAgent

A MobiAgent (Mahmoud, 2001) platform comprises a handheld mobile wireless device and an agent gateway, which are networked and communicate through hypertext transfer protocol (HTTP). The agent gateway executes the agent and its associated apparatus. The user interacts with the agent through an interface on the mobile device, which connects to the agent gateway and configures the agent. After the agent carries out a task, it reports back through the interface. This approach requires the minimum amount of processing and memory resources on the mobile device, but it makes the connectivity essential.

KSACI

Simple agent communication infrastructure (SACI) is a framework for creating agents that communicate using the Knowledge, Query, and Manipulation Language (KQML) (Finin, 1997). Each SACI agent has a mailbox to communicate with other agents. Infrastructure support is provided for white and yellow pages, but the platform is not FIPA compliant. KSACI is a smaller version of SACI suitable for running on the kVM (Albuquerque, Hübner, de Paula, Sichman, & Ramalho, 2001). The platform is not entirely situated on the constrained device and only supports the running of a single agent, which communicates via HTTP with a proxy running on a desktop machine.

DISCUSSION

A mobile computing environment is typified by resource constraints. Issues like processing power, memory, battery life, connectivity, and input/ouput (I/O) all require careful consideration.

It is often reported that intelligent agent platforms are unsuited for mobile applications because of their excessive computational overhead. This problem is usually due to particular agent platform implementations rather than an innate problem with the agent paradigm itself. Improving the efficiency of the reasoning algorithms within these systems can often lead to significant gains in efficiency. Additionally, the programming style adopted by the developer can have a considerable impact on performance. Developing in a style that conforms to the Law of Demeter (Lieberherr, Holland, & Riel, 1988) can reduce the footprint of the software by minimizing duplicated code while also improving maintainability in that internal implementation details of the object model are hidden. Further performance gains may be obtained through the use of autonomic procedures. An example of such a procedure, termed Collaborative Agent Tuning, may be found in Muldoon, O'Hare, and O'Grady (2005). Tuning enables agents collectively to alter their response times and computational overhead so as to maximize system performance.

The communication infrastructure is another fundamental resource that must be managed astutely when developing multi-agent systems. It is particularly important when working with lightweight devices that have limited battery power since sending messages consumes significantly more battery resources than normal processing. Mobile devices often have limited bandwidth and must make intelligent decisions as to what information to download and when to download it.

Additionally there is the issue of human-agent communication. Consideration must be made for the I/O capabilities of the devices. Most would have some form of keyboard input in the form of a touch screen or keypad. How the agent would convey information would be a bigger modality issue—is there a screen, does it allow for graphics or just text, how big is the screen, and how much of it is available to the agent?

FUTURE TRENDS

In the future, agents will emerge that are endowed with autonomy, mobility, and human-computer interaction facilities (Bradley, Duffy, O'Hare, Martin, & Schön, 2004). Such agents will opportunistically migrate, based on their tasks at hand, to different platforms (each offering varying capabilities and prospects), which would usually be for the benefit of an associated user. The presence of the agent moving through cyberspace as the user moves through physical space allows the associated user to be contactable at anytime through the agent.

A clear application of such nomadic agents is that of an autonomous "intelligent" digital assistant that is independent of any one physical device. These entities will effectively give any

user his or her own personal assistant that will help with the information overload in daily life, assisting with personal communications and offer a generic interface to any number of devices. These devices will have the ability to react to the current needs of their user, and beyond this, grow and learn to anticipate future needs and requirements. Perhaps our vision can be best summed up by Luc Steels' metaphor for what the robots of the future will be like:

[It] is related to the age-old mythological concept of angels. Almost every culture has imagined persistent beings which help humans through their life. These beings are ascribed cognitive powers, often beyond those of humans, and are supposed to be able to perceive and act in the real world by materialising themselves in a bodily form at will. (Steels, 1999)

He goes on to detail how angels may "project the idea of someone protecting you, preventing you from making bad decisions or actions, empowering you, and defending you in places of influence."

CONCLUSION

Agents encapsulate a number of features that make them an attractive and viable option for realizing mobile services. At a basic level, their autonomous nature, ability to react to external events, as well as an inherent capability to be proactive in fulfilling their objectives make them particularly suitable for operating in complex and dynamic environments. Should an agent be endowed with a mobility capability, its ability to adapt and respond to unexpected events is further enhanced. However, there are a few negative aspects to using agents. These systems can be more complex and require more device resources than the equivalent application-specific programs. Having no native support, agents require their own agent platforms for creation, execution, and communication. These problems will be reduced with advancements in mobile computing technologies, however in order to optimize system performance, agents will still have to manage their resources in a prudent and intelligent manner.

REFERENCES

Aarts, E., & Marzano, S. (Eds.). (2003). *The new everyday: Views on ambient intelligence*. Rotterdam, The Netherlands: 010.

Albuquerque, R. L., Hübner, J. F., de Paula, G. E., Sichman, J. S., & Ramalho, G. L. (2001, August 1-3). KSACI: A handheld device infrastructure for agents communication. *Pre-proceedings of the 8th International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, Seattle, WA.

Bellifemine, F., Caire, G., Poggi, A., & Rimassa, G. (2003, September). *JADE*. White Paper.

Berger, M., Rusitschka, S., Toropov, D., Watzke, M., & Schichte, M. (2002). Porting distributed agent-middleware to small mobile devices. *Proceedings of the Workshop on Ubiquitous Agents on Embedded, Wearable, and Mobile Devices held in conjunction with the Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Bologna, Italy.

Bradley, J. F., Duffy, B. R., O'Hare, G. M. P., Martin, A. N., & Schön, B. (2004, September 7-8). Virtual personal assistants in a pervasive computing world. *Proceedings of IEEE Systems, Man and Cybernetics, the UK-RI 3rd Workshop on Intelligent Cybernetic Systems (ICS'04)*, Derry, Northern Ireland.

Bradshaw, J. M. (1997). An introduction to software agents. In J. M. Bradshaw (Ed.), *Software agents* (pp. 3-46). Boston: MIT Press.

Collier, R. W. (2001, March). *Agent factory: A framework for the engineering of agent-oriented applications.* PhD thesis, Department of Computer Science, University College Dublin, National University of Ireland.

Dastani, M., Riemsdijk, B., Dignum, F., & Meye, J. J. (2003). A programming language for cognitive agents: Goal directed 3APL. *Proceedings of the 1st Workshop on Programming Multiagent Systems: Languages, Frameworks, Techniques, and Tools* (ProMAS), Melbourne.

Etzioni, O., & Weld, D. S. (1995). Intelligent agents on the Internet: Fact, fiction, and forecast. *IEEE Expert, 10*(4), 44-49.

Finin, T., & Labrou, Y. (1997). KQML as an agent communication language. In J. M. Bradshaw (Ed.), *Software agents* (pp. 291-316). Boston: The MIT Press.

FIPA (Foundation for Intelligent Physical Agents). (2000). *FIPA agent management specification.* Retrieved from http://www.fipa.org

Franklin, S., & Graesser, A. (1996). Is it an agent or just a program? A taxonomy for autonomous agents. *Proceedings of the 3rd International Workshop on Agent Theories, Architectures, and Languages.* New York: Springer-Verlag.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., & Wooldridge, M. (1999). The belief-desire-intention model of agency. *Proceedings of the 5th International Workshop on Intelligent Agents V: Agent Theories, Architectures, and Languages (ATAL-98),* Paris, France.

Hübner, J. F., & Sichman, J. S. (2000a). SACI: Uma ferramenta para implementação e monitoração da comunicação entre agents. *Proceedings of IBERAMIA.*

Hübner, J. F., & Sichman, J. S. (2000b). *SACI programming guide.*

Koch, F. (2005, July 25-29). 3APL-M platform for deliberative agents in mobile devices. *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS''05)* (pp. 153-154), The Netherlands. New York: ACM Press.

Lieberherr, K. J., Holland, I., & Riel, A. J. (1988). Object-oriented programming: An objective sense of style. Object oriented programming systems, languages and applications conference. *SIGPLAN Notices (Special Issue),* (11), 323-334.

Mahmoud, Q. H. (2001). MobiAgent: An agent-based approach to wireless information systems. *Proceedings of the 3rd International Bi-Conference Workshop on Agent-Oriented Information Systems*, Montreal, Canada.

Mihailescu, P., Binder, W., & Kendall, E. (2002). MAE: A mobile agent platform for building wireless m-commerce applications. *Proceedings of the 8th ECOOP Workshop on Mobile Object Systems: Agent Applications and New Frontiers*, Málaga, Spain.

Muldoon, C., O'Hare, G. M. P., Collier, R. W., & O'Grady, M. J. (2006, May 28-31). Agent factory micro edition: A framework for ambient applications. *Proceedings of Intelligent Agents in Computing Systems, a Workshop of the International Conference on Computational Science (ICCS 2006),* Reading.

Muldoon, C., O'Hare, G. M. P., & O'Grady, M. J. (2005). Collaborative agent tuning. *Proceedings of the 6th International Workshop on Engineering Societies in the Agents' World (ESAW 2005),* Kusadasi, Turkey.

Rao, A. S., & Georgeff, M. P. (1995, June). BDI agents: From theory to practice. *Proceedings of the 1st International Conference on Multi-Agent Systems (ICMAS'95)* (pp. 312-319), San Francisco.

Steels, L. (1999). *Digital angels.* Retrieved from http://arti.vub.ac.be/steels/sued-deutsche.pdf

Tarkoma, S., & Laukkanen, M. (2002). Supporting software agents on small devices. *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS),* Bologna, Italy.

Weiser, M. (1991). The computer for the twenty-first century. *Scientific American,* (September), 94-100.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review, 10*(2), 115-152.

KEY TERMS

Agent: A computational entity that acts or has the authority to act on behalf of a human user, software entity, or another agent.

Agent Communication Language: A formal language used for communication between agents.

Agent Platform: Provides the necessary infrastructure on which an agent operates.

Ambient Intelligence: Computing and networking technology that is unobtrusively embedded in the environment.

Embedded Agent: An agent that is contained wholly, along with its platform, on a particular device.

Mobile Service: One of several services provided through devices in a mobile computing environment (i.e., mobile phones, personal data assistants, wearable computers, etc.).

Multi-Agent System: A system comprising several agents—on the same platform or across multiple platforms—with a common goal.

Pervasive Computing: Computing involving computers, usually mobile devices, in all aspects of daily life.

Ubiquitous Computing: Computing in which the computers are embedded in everyday objects and all computing is done in the background.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 243-248, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.8
A Database Service Discovery Model for Mobile Agents

Lei Song
University of Guelph, Guelph, Canada

Xining Li
University of Guelph, Guelph, Canada

Jingbo Ni
University of Guelph, Guelph, Canada

INTRODUCTION

The number of services that will become available in distributed networks (in particular, on the Internet) is expected to grow enormously. Besides classical services such as those offered by printers, scanners, fax machines, and so on, more and more services will be available nowadays. Examples are information access via the Internet, music on demand, Web services, and services that use computational infrastructure that has been deployed within the network. Moreover, the concept of service in mobile agent systems, which will be described in this article, has come into prominence recently.

The mobile agent model is a new distributed software development paradigm as compared to the traditional client-server model. Instead of calling operations on servers with some form of synchronization, the user passes on his or her goal to an agent that can migrate within the computational environment and knows how to handle it without being controlled. In brief, mobile agents are active, autonomous, intelligent objects that are able to move between locations in a so-called agent system. Mobile agents must interact with their hosts in order to use their services or to negotiate services with other agents (Song & Li, 2004). Discovering services for mobile agents comes from two considerations. First, the agents possess local knowledge of the network and have a limited functionality, since only agents of limited size and complexity can migrate efficiently in a network and have little overhead. Hence, specific services are required that aim at deploying mobile agents efficiently in the system and the network.

Second, mobile agents are subject to strong security restrictions, which are enforced by the security manager. Thus, mobile agents should find services that help to complete security-critical tasks other than execute code that might jeopardize remote servers. Following this trend, it becomes increasingly important to give agents the ability to find and make use of services that are available in a network (Bettstetter & Renner, 2000).

Some of the mobile agent systems developed in the last few years are Aglets (Lange & Ishima, 1998), Voyager (Recursion Software Inc, 2005), Grasshopper (Baumer et al., 1999), Concordia (Mitsubishi Electric, 1998), and D'Agents (Gray et al., 2000). Research in the area of mobile agents looked at languages that are suitable for mobile agent programming, and languages for agent communication. Much effort was put into security issues, control issues, and design issues. Some state-of-the-art mobile agent systems focus on different aspects of the above issues (e.g., Aglets on security, D'Agents on multi-language support, Grosshopper on the implementation of the FIPA [FIPA, 2002], and MASIF [Milojicic et al., 1998] standard). However, few research groups have paid attention to offering an environment to combine the concept of service discovery and mobile agent paradigm. Most existing mobile agent systems require their programmers to specify agent migration itinerary explicitly. This makes mobile agents weak in their ability to sense their execution environment and to react autonomously to dynamic distributed systems.

In this article, we propose a new service discovery model DSSEM (discovery service via search engine model) for mobile agents. DSSEM is based on a search engine, a global Web search tool with centralized index and fuzzy retrieval. This model especially aims at solving the database service location problem and is integrated with our IMAGO (intelligent mobile agent gliding online) system. The IMAGO system is an infrastructure for mobile agent applications. It includes code for the IMAGO server—a multi-threading logic virtual machine, the IMAGO-Prolog—a Prolog-like programming language extended with a rich API for implementing mobile agent applications, and the IMAGO IDE, a Java-GUI-based program from which users can perform editing, compiling, and invoking an agent application. In our system, mobile agents are used to support applications, and service agents are used to wrap database services. Service providers manually register their services in a service discovery server. A mobile agent locates a specific service by submitting requests to the service discovery server with the description of required services. Web pages are used to advertise services. The design goal of DSSEM is to provide a flexible and efficient service discovery protocol in a mobile agent environment.

The rest of the article is organized as follows. The next section presents a brief background related to this article and discusses the problem of service discovery in mobile agent systems. The section Disvoery Services Via Search Engine Model (DSSEM) introduces DSSEM and compares it with several service discovery protocols (SDPs) currently under development. The comparison criteria include functionality, dependency on operating systems, and platforms. The section titled Service Discovery in the IMAGO system gives an overview of the design of service discovery module and integration with the IMAGO system. Finally, the last section provides some discussion and concluding remarks as well as future work.

BACKGROUND AND MOTIVATION

The general idea of distributed services is that an application may be separated from the resources needed to fulfill a task. These resources are modeled as services, which are independent of the application. Services do not denote software services alone but any entity that can be used by a person, a program, or even another service (Hashman & Knudsen, 2001). Service discov-

ery is a new research area that focuses not just on offering plug-and-play solutions but aims to simplify the use of mobile devices in a network, allowing them to discover services and also to be discovered (Ravi, 2001).

In general, the service usage model is role-based. An entity providing a service that can be utilized by other requesting entities acts as a provider. Conversely, the entity requesting the provision of a service is called a requester. To provide its service, a provider, in turn, can act as a requester, making use of other services. To form a distributed system, requesters and providers live on physically separate hosting devices. Providers from time to time should advertise services by broadcasting to requesters or by registering themselves on third-party servers. From requests' point of view, it must do the following:

- Search and browse for services
- Choose the right service
- Utilize the service

Before a service can be discovered, it should make itself public. This process is called *service advertisement.* The work can be done when services are initialized or every time they change their states via broadcasting to anyone who is listening. A service advertisement should consist of the service identifier plus a simple string saying what the service is or a set of strings for specifications and attributes. An example is given in Table 1.

There are several ways that a client looks up services that it requires. If the client knows the direct address of services, it can make direct requests, or it can listen to broadcasting advertisements and select those that it needs. The common method, however, is that the client forms a description of the desired service and asks a known discovery server if there is any service matching the request.

Table 1. A typical advertisement of service

Identifier: office-printer-4
Type: printer/postscript/HP20
Speed: 24ppm
Color: yes

A variety of service discovery protocols (SDPs) are currently under development by some companies and research groups. The most well-known schemes are Sun's Java-based Jini™ (Sun, 1999), Salutation (Salutation Consortium, 1998), Microsoft's UPP (Universal Plug and Play, 2000), IETF's draft Service Location Protocol (SLP) (Guttman et al., 1999), and OASIS UDDI (OASIS, 2005). Some of these SDPs are extended and applied by several mobile agent systems to solve the service discovery problem. For example, GTA/Agent (Rubinstein & Carlos, 1998) addresses the service location issue by extending SLP, a simple, lightweight protocol for automatic maintenance and advertisement of intranet services. Though SLP provides a flexible and scalable framework for enabling users to access service information about existence, location, and configuration, it only possesses a local function for service discovery and is not scalable up to global Internet domain, because it uses DHCP and multicast instead of a central lookup mechanism. AETHER (Chen, 2000) makes an improvement to Jini by building a dynamic distributed intelligent agent framework. Jini provides a flexible and robust infrastructure for distributed components to find each other on the Internet. However, it relies on the use of standard Java-based interfaces implemented by both clients and servers in their work. This requires existing systems to be modified for use with Jini; however, a significant amount of the production software currently available around the world is unlikely to be modified. After a study of different SDPs and mobile agent systems that are adopting

these methods, we found that several problems cannot be solved easily by the existing protocols due to their limitations.

First of all, most existing works support an attribute-based discovery as well as a simple name lookup to locate a service. Usually, there is only a set of primitive attribute types in the service description, such as string and integer, to characterize a service. Thus, the service discovery process is achieved primarily by type matching, string comparison, or integer comparison. Here, we define a service description as a sequence of flags or codes that can be multicast to service requesters or registered on a third-party server for advertisement purposes. Generally speaking, a service description is composed of two parts: property and access. The property of a service description describes the type, characteristics, constraints, and so forth of a service, which will be published in the service discovery server for advertising purposes. The access of a service is more complicated. It may contain multiple access method tags, as there could be multiple ways to invoke a service (e.g., using the interface of services, downloading the client-proxy code, locating a database, RPC, RMI, or URL location).

For example, Table 2 shows a service description in SLP, where the value of type tag (i.e., service:printer) indicates the property of the service. It also contains some other property tags to describe this resource in detail, such as paper per minute or color support. In the searching phase, much of the power of SLP derives from its ability to allow exact selection of an appropriate service from among many other advertised services with the same tags. This is done by requesting only the service or services that match the required keywords and attribute values specified by requesters. These keywords and attribute values can be combined into boolean expressions via "AND" and "OR" or common comparison operators "<=", ">" or substring matching. Considering the previous example again, the search request from a requester could be "< service:printer, bmw, ((name = lj4050) (page per min.>8)) >".

Table 2. Example of SLP service description

type = service: printer
scope = administrator, bmw
name = lj4050
paper per min. = 9
Color-support = yes
usage = //li4050.com: 1020/queue1

A further step in SDP's development is using eXtensible Markup Language (XML) to describe services. In fact, Web service discovery protocol UDDI, its description language WSDL, as well as the communication protocol SOAP are all based on XML. In addition, an XML description can be converted to a Java document object model (DOM) so that it can be merged into a service registry system. The example in Table 2 can be described in XML as follows:

<description ID="0198">
<type> service: printer </type>
<scope> administration, bmw </scope>
<name> lj4050 </name> .....
<usage> //li4050.com: 1020/queue1 </usage>
</description>

No matter what kind of description format is applied, the lack of rich representation for services has not been changed. The problem arising directly in our project is that these protocols are not adequate to advertise some special services such as database services. A database system already has a well-defined interface, and all a mobile agent requires is a way of finding the location of specific databases and deciding where to move. In this situation, the only way we can accomplish this is by registering the database's name and attributes for future discovery. However, for a database service, people care more about the content of the database than its name or structure. Considering an example of a bookstore, before placing an order

to the bookstore, customers would like to know if the books they require are available in the store by checking the summary of all books with some keywords or a fuzzy search criterion. From this point of view, a simple string identifier or XML identifier cannot meet the requirement.

The second problem is ranking. After requesters have searched all services that may be required, they still need to select the right one for utilization. Just imagine that over the entire Internet, tens of thousands of providers could publish their services by their own will. We should be able to know which ones provide the most realistic and highest quality services that users want. Obviously, moving to the hosts one by one to find out the required information is not a wise choice. Therefore, generating a service rank is essential. However, none of the existing SDPs offers such a mechanism for ranking discovered services. They are satisfied only with finding a service without considering whether the service would be able to serve the requester.

The most significant contribution of our research is that we enrich the service description by using Web page's URL (later the search engine will index the content referenced by this URL) to replace the traditional string-set service description in mobile agent systems. Because of their specific characteristics, such as containing rich media information (text, sound, image, etc.), working with the standard HTTP protocol, and being able to reference each other, Web pages may play a key role as the template of the service description. On the other hand, since the search engine is a mature technology and offers an automated indexing tool that can provide a highly efficient ranking mechanism for the collected information, it is also useful for acting as the directory server in our model. Of course, DSSEM also benefits from previous service discovery research in selected areas but is endowed with a new concept by combining some special features of mobile agents as well as integrating service discovery tools with agent servers.

DISCOVERY SERVICES VIA SEARCH ENGINE MODEL (DSSEM)

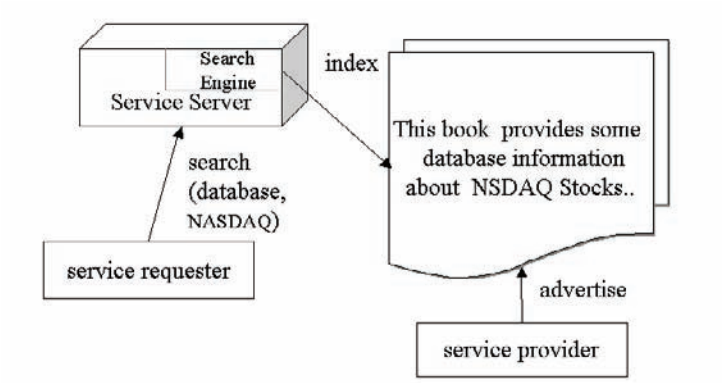
As the most important media type on the Internet today, hypertext Web pages are posted to advertise the information by industries and individuals. Though millions of them are published on the Internet, these Web pages still increase rapidly every day for a variety of reasons. They are spidered and indexed by commercial search engines such as Google, Yahoo!, AltaVista, and so forth. Users easily can find Web pages' locations by submitting the search request to those search engines.

In principle, if we install a lightweight search engine on the service discovery server that can retrieve Web pages posted by service providers and design a Web search interface for the incoming mobile agents, then the problems described previously could be solved in an easy way. In this situation, service providers don't need to register the service description on the service discovery server. Instead, they register the URLs of their Web sites that advertise all the information concerning services. As a middleware on the service discovery server, the search engine periodically will retrieve the document indicated in the URLs and all their referencing documents, parse all the tags and words in the documents, and set up the relationship between the keywords and the host address of these service providers.

On the other hand, mobile agents can utilize the system interface by accessing the search engine's database and obtain a destination itinerary that includes a list of ranked host addresses of the service providers. Based on the previous discussion, Figure 1 shows the service discovery process of DSSEM.

The current version of DSSEM concentrates on the database service discovery. The database service advertisement information can be converted easily to Web page representation. The specific characteristic of a Web page is that it contains rich media information and flexible layout and can reference each other. As an example in Figure 2,

Figure 1. Service discovery process of DSSEM



we find that a two-dimensional database table can be converted into a one-dimensional Web page. Moreover, some binary data stored in the database, such as image, can be extracted from higher-dimensional space to a lower-dimensional space as the text representation in the Web page.

To use Web pages as a medium to advertise services for service providers, we should modify the template in the service description of SLP. The remarkable change is that some properties once represented by strings or XML language now are represented as a Web site's home URL. Table 3 illustrates a service description template of a bookstore example.

The proposed model is similar to SLP and Jini with respect to the service discovery process; however, it extends those protocols by setting up a centralized, seamless, scalable framework on the Internet. Unlike some multicasting services protocols, the centralized service discovery server makes DSSEM service discovery available on the Internet worldwide. The process of registration is similar to UDDI, and the process of discovery is similar to the lookup service in Jini. Besides that, features of mobile agents bring DSSEM other incomparable advantages. First, code mobility is almost impossible in most distributed systems. Therefore, a client must download the resource

Figure 2. Web representation of database

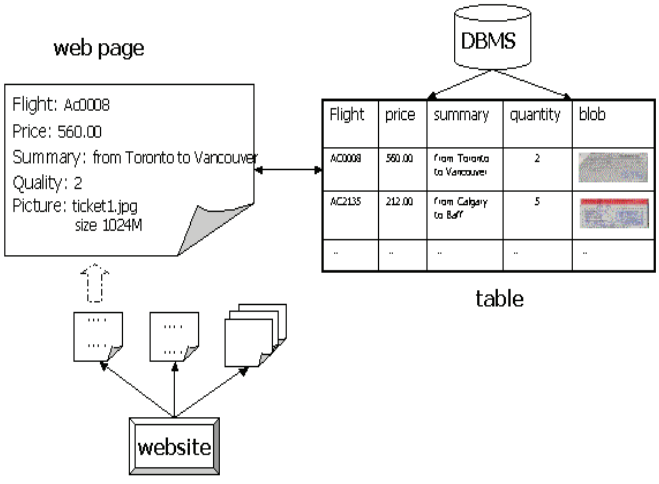


Table 3. An example of service description

type = service: database
name = bookstore
URL = //www.cis.uoguelph.ca/
location(URL)= www.uoguelph.ca
interface = dp_get_set(Handler, 'SQL statement', Result_handler)

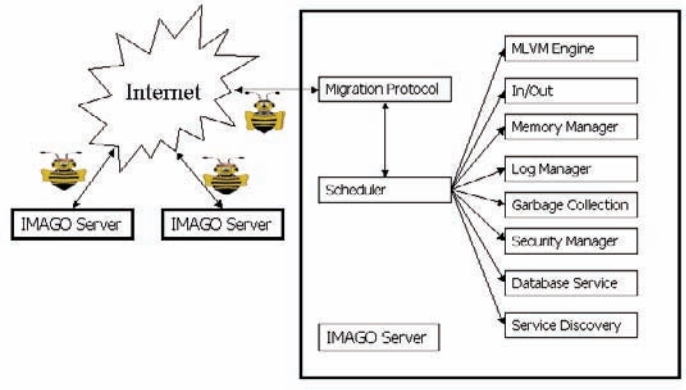
drivers to invoke services. Although RPC or RMI mechanism can help to call services remotely, it might consume tremendous network bandwidth when dealing with services involving a huge amount of data, such as database services. DSSEM goes one step further. It makes agents migrate to the destination hosts and utilize services locally. Second, the security issue is seldom considered in current service discovery protocols. However, a mobile agent server requires a strict security concern for authorization and authentication when it accepts the incoming agents and provides them services for utilization.

SERVICE DISCOVERY IN THE IMAGO SYSTEM

IMAGO is a mobile agent system in which agents are programs written in IMAGO Prolog that can move from one host on the Internet to another. Briefly speaking, an agent is characterized as an entity that is mature (autonomous and self-contained), mobile, and bears the mental model of the programmer (intelligent) (Li, 2001). From an application point of view, the IMAGO system consists of two kinds of agent servers: stationary server and remote server. The stationary server of an application is the home server where the application is invoked. On the other hand, agents of an application are able to migrate to remote servers. Like a Web server, a remote server must have either a well-known name or a name searchable through the service discovery mechanism. Remote servers should provide services for network computing, resource sharing, or interfaces to other Internet servers, such as Web servers, database servers, and so forth.

In fact, an IMAGO server, no matter if it is stationary or remote, is a multithreading logical virtual machine to host agents and provides a protected agent execution environment. The IMAGO system is portable in the sense that its servers run on virtually all Linux boxes with Gnu C compiler and Posix package. Tasks of an IMAGO server include accepting agents, creating a secure run time environment, and supervising agent execution. It also must organize agent migration from or to other hosts, manage communications among agents, authenticate and control access for agent operations, recover agents and the information

Figure 3. The infrastructure of IMAGO system



carried by them in case of network and computer failures, and provide some basic services for the agents, such as database service and discovery service.

The architecture of the IMAGO server is shown in Figure 3. In this architecture, the system modules are configured to deal with different tasks. The core module of the IMAGO server is the scheduler. It maintains an agent list, where each entry on the list matches different stages of the life cycle of an agent, such as creation, execution, memory-related processing (i.e., expansion, contraction, or garbage collection), termination, and migration. The agent list is sorted with respect to system-defined priorities. For example, the highest priority is given to agent migration, followed by agent creation and execution, memory manipulation, and, finally, database service and service discovery.

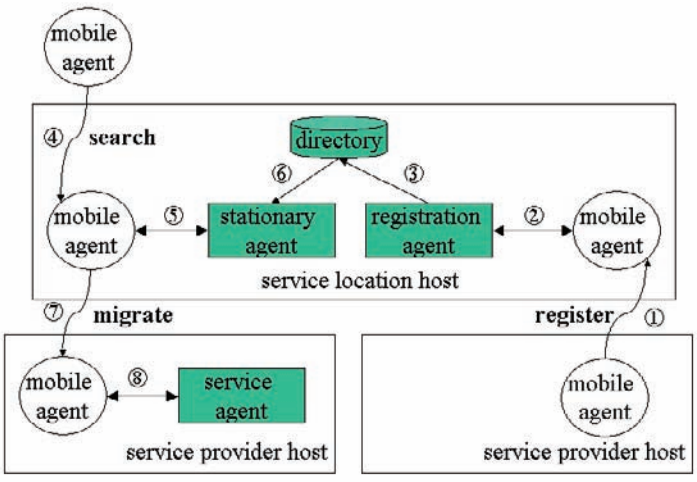
In the early phase of system design, database operation becomes the major service to applications in the IMAGO system. Thus, the problem of service discovery focuses on how to find such services effectively. Once a database server has been found, agents may migrate to that remote server and invoke database access locally through built-in primitives. As an example, the following code shows a typical database access issued by an IMAGO agent:

```
dp_connect(URL, DatabaseName, UserName,
     Password, AccessMode), //connection
dp_get_set(Handler, 'select …..', ResultHandler),
          //data access
dp_disconnect(Handler).        //disconnection
```

Before a database service is advertised, the service provider should fill out a registration form and submit the form to an IMAGO service discovery server. The contents of the form include service type, database name, URL of the service provider host, access mode, HTTP URL of the service Web site, interface function, and the verification information. We chose URL as the host address, since it is compatible with most commonly used Web browsers and independent of address families (i.e., IP, IPv6, and IPX).

To illustrate how DSSEM works, Figure 4 shows the steps involved in the service registration and discovery process in our IMAGO system. A service discovery server is called the service location host. In order to gather useful information, the search engine, IMAGOSearch, should be installed

Figure 4. The processes of Web search module



independently on the service location host. This search engine maintains a special database system designed to index Internet addresses (i.e., URLs, Usenet, Ftp, image locations, etc.). Like traditional search engines, IMAGOSearch consists of three main components: spider, indexer and searcher. They are grouped into two modules, where one module includes spider and indexer, running in the background of a service location host, and the other module is the searcher, running in the foreground to provide discovery services. First, the spider gets the URLs from a URL list that contains initial Web site URLs registered by service providers. The spider then traverses along these URLs in the breadth-first manner and loads the referred hypertext documents into the service discovery server. The salient words of these documents are extracted by the indexer. Some related information such as text also is saved into the database for user retrieval. In addition, the indexer looks for URL anchors that reference other documents and adds them to the URL list. Besides considering the weight of each word in the documents (e.g., a word occurrence in the title should be assigned a higher weight than that which occurs in the body), IMAGOSearch also pays attention to positions of each word and its relative distance during ranking. The ranking algorithm we use is called the shortest-substring ranking (Charles et al., 2000), which offers a simple way to weight each Web page based on a search criteria and total them up to form Web site ranking. The searcher behaves as a bridge between the IMAGO server and the search engine. It is responsible for accepting the search requests from mobile agents, querying the database, ranking search results, and, finally, returning a destination itinerary.

The application programming interface for mobile agents is a built-in predicate; namely, Web_*search(query, number, Result)*, where *query* is a compound term, such as *locate("tsx", "stock transaction", "imago server"), number* is an integer indicating the maximum number of results expected, and *Result* is a variable to hold the returned values. When an agent issues a Web_search(…) predicate, the agent is blocked, and control is transferred to the service discovery module of the hosting IMAGO server. This module will communicate with the searcher, wait for search results, and resume the execution of the blocked agent. Search results will be delivered to the agent in the form of a list, where list entries are ranked in terms of priorities from high to low.

Table 4. Comparison of different SDPs

Feature	SLP	Jini	Salutation	UPnP	DSSEM
Network transport	TCP/IP	Independent	Independent	TCP/IP	SITP
Programming language	Independent	Java	Independent	Independent	Independent
OS and platform	Dependent	Independent	Dependent	Dependent	Dependant
Code mobility	No	On demand	No	No	Yes
Srv attributes searchable	Yes	Yes	Yes	No	Yes
Leasing concept	Yes	Yes	No	Yes	Yes
Event notification	No	Remote event	Periodic and automatic	Publish events	No
Security	No	Java-based	Authentication	No	Strict
Service Description and Scope	Service type and attribute matching	Interface type and attribute matching	Functional units and attributes within it	Description in XML	Web page description and fuzzy matching

DISCUSSION AND CONCLUSION

In this article, we have discussed the design of a service discovery protocol—DSSEM—and its implementation in the IMAGO system. Table 4 summarizes the main features of selected protocols compared with DSSEM. From an implementation point of view, the most critical issue about the performance of a search engine is the quality of search results. However, we cannot make a comparison with other major commercial search engines, since they are operating at different levels. Thus, user evaluation is beyond the scope of this article. In order to show that our search engine does return useful information, Table 5 gives the experimental results for a query using the keywords *imago lab*. The results show that all server URLs have come from reasonably high-quality Web sites, and, at last check, none were broken links. An R_w value is calculated according to word occurrence, weight, and a factor value measuring the distance of keywords by a ranking algorithm (Charles et al., 2000). We define the result that has the highest R_w value as the highest priority and assign it a 100% rate; therefore, the percentage of other results are rated relatively. Of course, a true test of the quality of a search engine would involve extensive experiments, analysis, and user evaluation, which is part of our future work.

Aside from the search quality, IMAGOSearch is designed to scale up cost effectively, as the sizes of Web pages grow. Because IMAGOSearch only indexes Web servers registered by IMAGO Server users, we do not need to worry about indexed pages exceeding the maximum size of the database. One endeavor that we are undertaking is to reduce the table redundancy and to use the storage efficiently. Our experiment shows that indexing 22,000 different documents consumes only 140Mb disk space. The search time is dominated mostly by the performance of CPU, disk I/O, and the underlying database system.

When a mobile agent wants to locate certain services, it first must move to the service discovery server and then make a local query and migrate to the destination hosts after obtaining the itinerary. This brings us to the problem that, as a central unit, the service discovery server might become a bottleneck, especially when it is handling thousands of Web pages every day and simultaneously hosting as many incoming mobile agents as possible. A possible solution is to duplicate service discovery servers. Replicas not only make the service discovery mechanism very efficient but also increase the ability of fault tolerance.

The results of our work are encouraging, and further studies in this field are being planned. First, the current implementation of search engine deals with only the AND logical relationship between search strings; it could be enhanced to parse some complex search criteria that combine keywords with boolean expressions (AND, OR) and conditional expressions (<=, >=, substring match, etc.). Second, since a database contains multidimensional information, how to reflect dimensional relationship by a flat Web page is a big challenge. A possible way to address this issue is to use XML metadata to describe the database dimension.

Table 5. Search results for imago lab keyword

draco.cis.uoguelph.ca	R_w= 13.8 100%
www.cis.uoguelph.ca	R_w = 10.65 77%
www.uoguelph.ca	R_w= 4.6 33%
www.cas.mcmaster.ca	R_w= 4.23 30.6%

ACKNOWLEDGMENT

We would like to express our appreciation to the Natural Science and Engineering Council of Canada for supporting this research.

REFERENCES

Baumer, C., Breugst, M., & Choy, S. (1999). Grasshopper—A universal agent platform based on OMG MASIF and FIPA standards. *Proceedings of the First International Workshop on Mobile Agents for Telecommunication Applications (MATA'99)* (pp. 1-18).

Bettstetter, C., & Renner, C. (2000). A comparison of service discovery protocols and implementation of the service location protocol. *Proceedings of the EUNICE 2000, Sixth EUNICE Open European Summer School*, The Netherlands.

Charles L., Clarke, A., & Gordon V. (2000). Shortest substring retrieval and ranking. *ACM Transactions on Information Systems* (pp. 44-78).

Chen, H. (2000). *Developing a dynamic distributed intelligent agent framework based on Jini architecture* . Master's thesis, MD: University of Maryland.

FIPA. (2002). Agent management specification. Retrieved from http://www.fipa.org

Gray, R., Cybenko, G., & Kotz, D. (2002). D'agents: Applications and performance of a mobile-agent system. *Software—Practice and Experience, 32*(6), 543-573

Guttman, E., Perkins, C., & Veizades, J. (1999). *Service location protocol, version 2* (white paper). IETF, RFC 2608.

Hashman, S., & Knudsen, S. (2001). *The application of Jini technology to enhance the delivery of mobile services* [white paper]. Retrieved , from http://wwws.sun.com/

John, R. (1999). *UPnP, Jini and salutaion—A look at some popular coordination framework for future network devices* [technical report]. California Software Labs.

Lange, D., & Ishima, M. (1998). *Programming and deploying Java, mobile agents with aglets.* Addison-Wesley.

Li, X. (2001). IMAGO: A prolog-based system for intelligent mobile agents. *Proceedings of the Mobile Agents for Telecommunication Applications (MATA'01)*, Lectures Notes in Computer Science, 21-30. Springer Verlag

Li, X. (2003). *IMAGO prolog user's Mannual, version 1.0* [technical report]. University of Guelph.

Milojicic, D., Breugst, M., & Busse, I. (1998). MASIF: The OMG mobile agent system interoperability facility. *Proceedings of the Second International Workshop on Mobile Agents* (pp. 50-67).

Mitsubishi Electric ITA. (1998). *Mobile agent computing* (white paper).

OASIS UDDI. (2005). *UDDI* [white paper]. Retrieved from http://www.uddi.org

Ravi, N. (2001). *Service discovery in mobile environments* [technical report]. Arlington, TX: University of Texas, Arlington.

Recursion Software Inc. (2005). *Voyager product documentation.* Retrieved from http://www.recursionsw.com/voyager_Documentation.html

Rubinstein, M., & Carlos, O. (1998). Service location for mobile agent system. *Proceedings of the IEEE/SBT International Telecommunications Symposium (ITS'98)* (pp. 623-626).

Salutation Consortium. (1998). *Salutation architecture overview* [white paper]. Retrieved from http://www.salutation.org/whitepaper/originalwp.pdf

Sun Technical. (1999). *Jini architectural overview* [white paper]. Retrieved from http://www.sun.com/jini/

Universal Plug and Play Forum. (2000). *Universal plug and play device architecture, version 0.91* (White Paper).

This work was previously published in Intelligent Information Technologies and Applications, edited by V. Sugumaran, pp. 173-189, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 3.9
Databases for Mobile Applications

Indranil Bose
University of Hong Kong, Hong Kong

Yip Yee Shing
University of Hong Kong, Hong Kong

Wang Ping
University of Hong Kong, Hong Kong

Chan Lit Tin
University of Hong Kong, Hong Kong

Mok Wai Shan
University of Hong Kong, Hong Kong

Shiu Ka Wai
University of Hong Kong, Hong Kong

Wong Ka Shing
University of Hong Kong, Hong Kong

INTRODUCTION

Owing to the rapid development of mobile technology over the past few decades, there have been many different kinds of mobile devices emerging in the market, and most of them work with databases seamlessly. Mobile phone gaming, downloading of ringtones, and e-calendaring are some of the prominent examples of mobile applications that require the close integration of mobile devices with databases. Mobile devices take various forms and configurations. The packaging, form factors, hardware platforms, operating system support, and functional capabilities vary across these devices. There are, however, many common attributes shared by the devices, such as notebook computers, pen-based computers, handheld computers, and the like, all of which are used in mobile computing. These devices can be categorized into the following categories according to their functionalities and features, as detailed in Dhawan (1997). They are:

- notebook computers
- personal digital assistants
- tablet computers
- hybrid mobile devices
- mobile phones

In this article, we focus on personal digital assistants (PDA) and mobile phones as they are the most popular and commonly used mobile devices in the industry.

Mobile Computing Applications

Mobile applications include basic applications like datebook, address book, to-do list, and memos and also horizontal and vertical industry applications that mainly fall into the following three categories (Dhawan, 1997):

- Shrink-wrapped horizontal industry mobile computing applications that can be used in broad segments of various industries, e.g., electronic mail, electronic messaging via paging, and sales force automation.
- Generic horizontal industry applications requiring extensive customization, and these include database access from an information server, computer-aided dispatch (CAD), and intrasite and intersite mobility applications among others.
- Vertical industry applications include the applications that are specific to industries like insurance, banking, airlines, government, utilities, and transportation, e.g., finance industry insurance and financial planning, and stock trading.

The diverse variety of the types of mobile applications demonstrates the reach of mobile computing into almost every facet of personal and business life. One of the applications that is gaining popularity is mobile e-commerce. Mobile e-commerce refers to commercial activities performed electronically. An example of this is an online shopping mall (via the mobile devices to the Internet). Mobile commerce is one of the most popular applications these days in addition to obtaining stock quotes, directions, weather forecasts, and airline flight schedules from mobile devices (Munusamy & Hiew, 2004).

Comparison Between Mobile Devices and Desktop Computers

Compared to desktop computers, mobile devices have small memory, low computing capabilities, limited interaction facilities, and limited display and network processing capabilities. With recent technological advancements, hybrid devices combining the functionality of mobile phones together with PDAs have been developed. The differences are mainly attributed to their hardware design and system configurations. Table 1 compares desktop computers, PDAs, and mobile phones with respect to their processing power, memory, storage capacity, connection speed, and display. The data presented is current as of June 22, 2004. Data related to specifications for the desktop, PDA, and mobile phone have been downloaded from the Web sites of Dell Inc. and Nokia Inc. and relate to the Dell Dimension 8400 Desktop, the Dell Axim X3 Pocket PC 400 MHz WiFi, and the Nokia 7610, respectively.

From Table 1, it can be observed that mobile devices have smaller memory size and storage capacity as well as display size than desktop computers. So, the amount of data that can be transferred and displayed at a time is less than that of desktop computers. Furthermore, the processing power of mobile devices is usually limited when compared with desktop computers. The amount of data that can be processed at a given time is also small. Also, mobile devices have lower connection speeds and less stable network connections. They must have ways to overcome these deficiencies in order to ensure good performance in retrieving data from remote databases.

Challenges for Mobile Devices

Some of the challenges faced by mobile devices when connecting to remote databases include challenges in network connectivity, data transmission, security, and data consistency.

Table 1. Comparison between a desktop computer and mobile devices

	Desktop	PDA	Mobile Phone
Processing Power	2.8–3.4 GHz	400 MHz	Unknown
Memory	512 MB–2 MB	64–1024 MB	8 MB (internal)
Storage Capacity	80–400 GB	N/A	N/A
Connection Speed	56 Kbps–100 Mbps	56 Kbps–11 Mbps	Up to 40.3 Kbps
Display	15–19 inch	3.5 inch	1.3 inch

- **Network connectivity:** Mobile devices usually work in an unstable network environment. The network stability is affected by many factors, such as weak signal and strong interference. Without physical network connections, mobile devices often lose connection with the network.
- **Data transmission:** Wireless networks have limited bandwidth compared to traditional cable networks. The slow transmission speed imposes problems in uploading and downloading data. Large network latency constraints also result in long response time.
- **Security:** Any message between the database system and mobile devices is sent over the air, and it is possible for hackers to sniff the message and perform eavesdropping. Advanced encryption and user authentication technology is needed to prevent any such types of hacking activities.
- **Data consistency:** Database applications apply extensive caching and replication to boost performance, which can lead to possible data inconsistency. Mobile devices with little memory storage and slow connection speed cannot obtain all the information from the central database system instantly. The narrow bandwidth of the devices also affects immediate updates from the mobile devices to the database server. It is thus quite difficult to keep data consistent between mobile devices and the database server.

The objective of this article is to give a brief overview on the design of databases for mobile applications and to describe how the database design is currently being done for a successful mobile application called mBroker that is operational in Hong Kong. The article provides a description of the functionalities of the mBroker system and highlights the database design being used by the mBroker solution at the present time.

BACKGROUND

The database design for mobile applications is different from normal database applications running on personal computers. Due to the limited hardware configurations and network settings of mobile devices, database vendors usually provide special database systems and APIs (application programming interfaces) tailored for mobile devices.

Database manufacturers like Oracle, IBM, and Sybase usually follow a similar architecture to build mobile database applications. The main components include a micro database engine, synchronization middleware, and wireless networking. The difference is often the naming of different components. Figure 1 depicts the typical database architecture for mobile applications.

Micro Database Engine

Mobile devices have limited hardware in terms of processing power, memory size, and battery life. A normal database engine requires a minimum 20 MB memory to operate, which is unavailable for mobile devices. Therefore, database vendors develop robust micro database engines for most mobile devices.

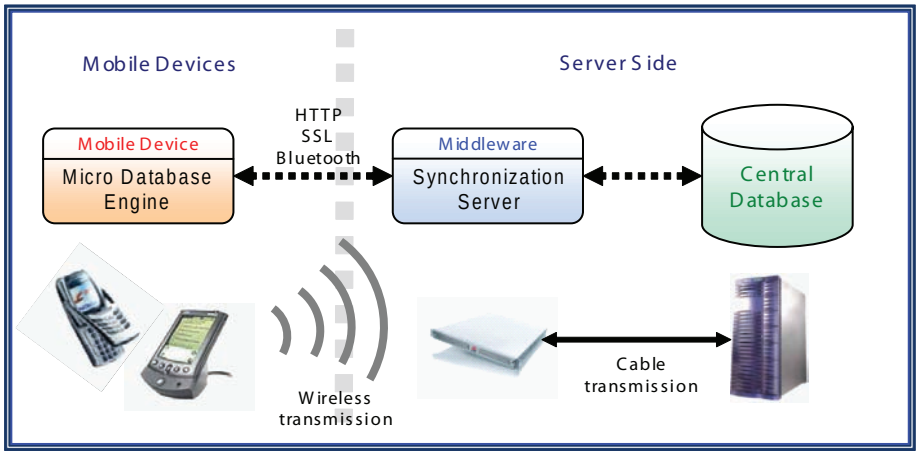
Micro database engines only focus on the most frequently used functionalities which are relevant to mobile applications. These include basic SQL statements, Join, Group By, Order By, scrollable cursors, and simple primary key and foreign key operations. For high-end mobile devices, advanced indexing features are sometimes included to improve performance. Database operations which are rare and less useful are removed from the micro engine. For example, view creation, subqueries, stored procedures, triggers, and user-defined functions are not provided in micro database engines. The DB2 Everyplace and Oracle Lite (Viellard, 2001) are examples of micro database engines.

Synchronization Middleware

The network capabilities of mobile devices are limited. Real-time online database access requires a large and stable network bandwidth, which is expensive and not always required for mobile applications. Thus, database vendors employ extensive caching and synchronization techniques in mobile database applications. Caching is to preload data to the devices for offline browsing. Users can then browse data seamlessly even if the wireless network is unstable or even unavailable. If instant update to the central server is not necessary, data modification operations are further optimized. Updates are made to the data on the devices first, and the modifications are not reflected on the server database immediately. Data are updated to the remote central database later by synchronization.

Synchronization is done by middleware which is sometimes called a synchronization server or mobile server. Advanced synchronization methods keep data in both mobile devices and the cen-

Figure 1. A typical database architecture for mobile applications



tral database consistent. The middleware acts as a middleman between the mobile devices and the central database. It collects changes in the mobile devices and executes appropriate SQL statements to update the central database. At the same time, the middleware also propagates updates in the central database to the mobile devices.

Wireless Networking

Different types of mobile devices communicate using different network protocols. Database vendors usually include a number of network protocol supports in their products, for example, GSM, IEEE 802.11, and Bluetooth.

Security

To overcome the security problems of mobile database applications such as eavesdropping, the mobile database architecture needs to support both username/password authentication and encrypted communication based on the Secure Sockets Layer (SSL) protocol and other popular encryption algorithms.

In spite of the well designed architectural model, there still exist a number of limitations in the integration of the databases with mobile devices.

Literature Review

Previous research related to database design for mobile applications has involved various mechanisms for avoiding compromise in the performance of the database due to the use of the wireless network. Some of the techniques used involve reducing the number of data exchanged over the wireless network and providing a data cache on the mobile host (Chan, Si, & Leong, 1998). To address the challenges related to maintenance of data consistency for mobile data access, several techniques have been suggested in the literature. These have ranged from transaction management (Mazumdar & Chrysanthis, 1999) and concurrency control for mobile databases (Prabhu, Ray, & Yang, 2004) to replication of mobile databases which are allocated on the fixed network (Budiarto, Harumoto, Tsukamoto, Nishio, & Takine, 1998). It is argued in Budiarto, Nishio, and Tsukamoto (2002) that without replication, mobile databases have a very low availability, and it is shown using simulation that the performance of replication strategies depends on various factors such as network size, mobility, access ratio, and access concentration. Another important issue for mobile databases is how to provide consistent results for location-specific continuous queries, the likes of which may be encountered when navigating road maps using mobile devices. In Gok and Ulusoy (2000), several approaches are compared in terms of relative performance for providing answers to location-dependent queries from mobile users. An analytical model based on the idle replacement policy is described by Hung, Lin, Peng, and Yang (2001) to solve the problem of overflows of visitor location registers for mobile databases. For a detailed discussion on the various issues related to database design for mobile databases with respect to factors such as mobile location data management, transaction processing and broadcast, cache management and replication, query processing, and mobile Web services, the interested reader may refer to Barbara (1999), Madria, Mohania, Bhowmick, and Bhargava (2002), and Yang, Bouguettaya, Medjahed, Long, and He (2003).

MAIN THRUST

Sixteen Hong Kong brokerage firms are currently improving their productivity and customer satisfaction levels by using the "mBroker" solution offered jointly by Heracle Technologies Limited and Hutchison Telecom, built on the Palm OS platform for Palm handhelds (Lai, Tam, & Lemaitre, 2004). The mBroker solution provides a

secure trading platform to remotely access stock information in real time and to conduct stock trading activities.

Problem Description

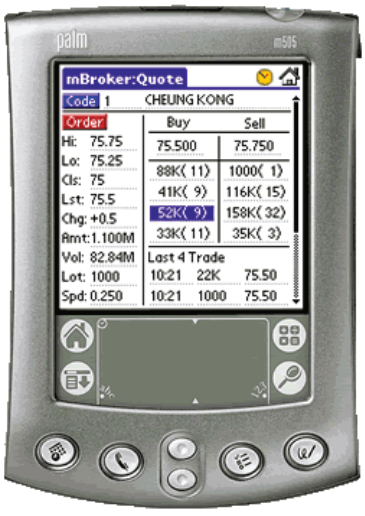
It is time-consuming and labor-intensive for investors to rely on desktop computers or consult brokerage agents to obtain the latest stock quotes or the trading history of any particular stock index, as well as to place an order. Mistakes such as overlooked orders, data entry errors, or delays due to congested telephone lines often occur when dealing with an agent.

It is critical to guarantee high speed and accuracy for this type of situation. Speeding up the order processing and accessing real-time information without compromising accuracy are goals that all brokers strive for to retain their competitive edge.

The mBroker Solution

mBroker—an innovative, secure, and cost-effective wireless stock trading system with specific design for PDA stock trading—allows PDA users to get stock quotes and trade stocks at any time in any place in the world. More importantly, it is totally secure as it uses the "Hongkong Post Mobile e-Cert" and also Oracle8i. With mBroker, investors can place an order remotely via an intuitive touch screen interface of Palm handhelds, without the need to contact the agent.

Figure 2. Application interface of mBroker on a PDA device



Functionalities of mBroker

There are a diverse range of services related to stock trading activities provided by mBroker on Palm. These include:

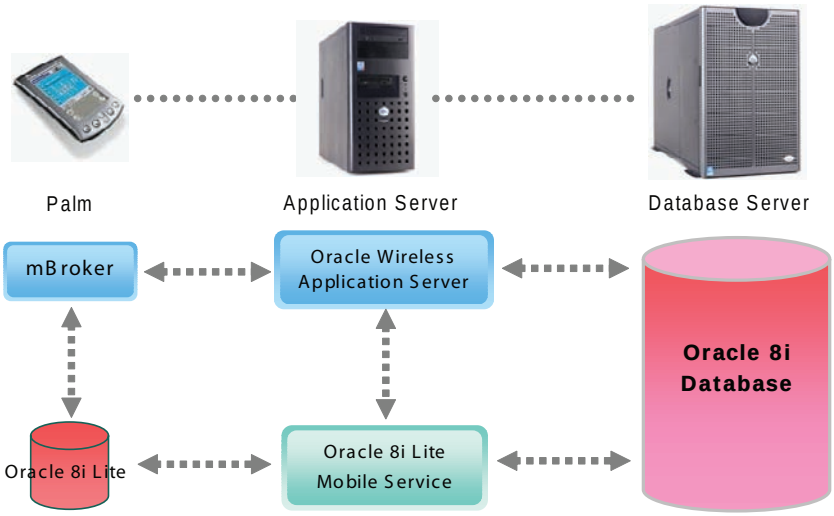
- Place, modify, and cancel orders.
- Access real-time stock quotes.
- Keep track of the Hang Seng Index.
- View order status and transaction history.

The mBroker application can be run on Palm OS 3.5 or above and only occupies approximately 190 Kb of memory space. What makes this solution attractive is the minimal deployment cost. Since there is no special software or hardware requirements for participating brokerage firms, their customers can enjoy the service simply by subscribing to the mBroker service. The user interface is very user-friendly since it has a similar look and feel to any other Palm application. Both English and Chinese versions of this software are available to the users.

Architecture of mBroker

Figure 3 shows a simplified view of the mBroker application and the database system architecture. The Oracle database stores stock information and statistics. The stock data includes stock codes, stock names, and stock prices. In order to help the investors in making investment decisions, the database also stores information such as past selling prices of the stock, the high/low price information, and transaction volume statistics.

Figure 3. A simplified view of the architecture of mBroker



A subset of this information is loaded onto the Palm for fast offline browsing.

Transaction details are also required for order tracking and enquiry. Therefore, all the transaction details are logged in the Palm and the central database. These include the broker ID, status, stock type, price, quantity, transaction timestamp, etc.

The mBroker application follows an architecture that is similar to that discussed earlier. The stock and transaction information is originally stored in a central Oracle 8i database server. When a mobile device with mBroker wants to retrieve stock quotes from the central database server, it will send requests to a middleware named Oracle Lite Mobile Service in order to subscribe to the information (Viellard, 2001). This middleware is installed in a server machine and acts as a middleman between different mobile devices and the central database server. It obtains the required information from the database server and sends it back to the mobile devices. The middleware is responsible for synchronization, keeping the mobile devices informed of any changes on the database server. When the user buys or sells stock via mBroker, requests are sent from the mobile devices to the middleware. The middleware then updates the central database accordingly. The middleware is capable of receiving thousands of requests at the same time and decides the sequence of processing the transaction requests following predefined business rules written in PL/SQL.

Choice of Database

Heracle Technologies Limited chose to use Oracle8i as the central component because of mobile service provided by Oracle8i Lite. Oracle8i Lite is accessed by users through the application server and provides the necessary workspace for end users to request context switching between online and offline modes. It also automatically initiates the necessary two-way replication of data and applications between server and client, depending on changes of mode. When it comes to stock order transactions, it is imperative to ensure a maximum level of security. Not all proprietary databases, e.g., DB2, can support a public key infrastructure (PKI; Browder, 2002); Oracle8i supports PKI and therefore this was the obvious choice for mBroker.

Security of mBroker

To ensure the security of the stock order information (OI), the OI is encrypted using the PIN input and sent to Hongkong Post's Certification Authority server for authentication. All transaction details are secured by Hongkong Post Mobile e-Cert, secured by PKI technology. The Mobile e-Cert can be obtained from Hutchison Telecom, the world's first certified Registration Authority for the issuance of Mobile e-Cert. A high security solution ECC163 cryptography is adopted. Upon verification, the order is allowed to pass through the Order Routing System gateway of Hong Kong Exchanges and Clearing Limited (HKEx) to the designated brokerage firm's system for subsequent processing. Once the transaction is completed, a confirmation note with a transaction reference number is sent back to the user's mBroker interface.

Disaster Recovery in mBroker

Oracle8i database has a component called Data Guard, which offers the most complete and robust disaster recovery solution and high availability through the use of a transactionally consistent standby database. The Data Guard automates the complex tasks of disaster recovery and provides monitoring, alerting, and control mechanisms to maintain a standby operation. Moreover, Data Guard reduces planned downtime by utilizing the standby server for maintenance and routine operations in addition to reporting.

Current Performance of mBroker

mBroker takes less than six seconds on average to complete an order via a handheld, whereas the user takes approximately one minute to complete the same using a phone call. The issue of wireless security is addressed by the introduction of user/server authentication and digital signatures. As a result, an automated, speedy order placement in a highly secured wireless environment is ensured.

FUTURE TRENDS

In view of the issues discussed so far it is obvious that there are many areas which need to be improved. In the future, the goal is to provide the following functionalities for mBroker (Huntsman, 2003):

- Providing intelligent roaming capabilities to enable users to work without interruption, even when network connections are disrupted.
- Exploiting multiple network interfaces in a single device or being able to select the fastest or least costly connection.
- Successfully synchronizing databases by caching contents to local devices through asynchronous connections.
- Allowing portability to a range of devices.
- Conserving power at the operating system level and maximizing performance.

CONCLUSION

In this article we have provided a brief background on the use of databases for mobile applications. This is a growing area and is facing a number of challenges at this time. The database design for mobile applications is also discussed in this article . We have also discussed a successful mobile application called mBroker which is currently in use in Hong Kong. Another similar example is the new mobile workforce effectiveness solution called SMARTselling, developed by Eleven Technology and powered by SQL Anywhere Studio from iAnywhere Solutions. This technology is currently being used by Pepsi Bottling Group and by Proctor and Gamble. This software helps in automated order entry and runs on a small

handheld device that communicates wirelessly with back-office systems. With the help of this application, tedious, error-prone, and costly paper-based processes can be eliminated, and the time spent on checking inventories and shelf displays can be significantly reduced. It is hoped that in the future more mobile applications like mBroker and SMARTselling will be developed, which will affect the various facets of everyday life for people around the globe.

REFERENCES

Barbara, D. (1999). Mobile computing and databases: A survey. *IEEE Transactions on Knowledge and Data Engineering, 11*(1), 108-117.

Browder, K. (2002). *Technical comparison of Oracle database and IBM DB2 UDB: Focus on security.* Retrieved December 2, 2003, from *http://www.oracle.com/ip/se/o9idb_db2_techcompar.pdf*

Budiarto, K., Harumoto, M., Tsukamoto, M., Nishio, S., & Takine, T. (1998). Replica allocation strategies for mobile databases. *IEICE Transactions on Information and Systems, E81-D1,* (pp. 112-121).

Budiarto, K., Nishio, S., & Tsukamoto, M. (2002). Data management issues in mobile and peer-to-peer environments. *Data and Knowledge Engineering, 41,* 183-204.

Chan, B. Y. L., Si, A., & Leong, H. V. (1998). Cache management for mobile databases: Design and evaluation. *Fourteenth International Conference on Data Engineering, ICDE-98,* Orlando, Florida (Vol. 2, No. 7, pp. 54-63).

Dhawan, C. (1997). *Mobile computing: A systems integrator's handbook*. New York: McGraw-Hill.

Gok, H. G., & Ulusoy, O. (2000). Transmission of continuous query results in mobile computing systems. *Information Sciences, 125,* 37-63.

Hung, H.-N., Lin, Y.-B., Peng, N.-F., & Yang, S.-R. (2001). Resolving mobile database overflow with most idle replacement. *IEEE Journal on Selected Areas in Communication, 19*(10), 1953-1961.

Huntsman, J. B. (2003). *Introducing the Intel mobile application architecture guide*. Retrieved December 2, 2003, from http://www.intel.com/update/contents/sw12031.htm

Lai, A., Tam, A., & Lemaitre, S. (2004). *Hong Kong stock trading industry embarks on a new era with mBroker capabilities on Palm OS platform*. Retrieved December 2, 2003, from *http://www.dv net.com/pdf/casestudies/palm.pdf*

Madria, S. K., Mohania, M., Bhowmick, S. S., & Bhargava, B. (2002). Mobile data and transaction management. *Information Sciences, 141,* 279-309.

Mazumdar, S., & Chrysanthis, P. K. (1999). Achieving consistency in mobile databases through localization in PRO-MOTION. *Second International Workshop on Mobility in Databases and Distributed Systems (MDDS99),* Florence, Italy (pp. 82-89).

Munusamy, M., & Hiew, P. L. (2004). *Characteristics of mobile devices and an integrated m-commerce infrastructure for m-commerce deployment.* Retrieved December 2, 2003, from *http://www.wayneyeung.com/files/papers/FP-102.pdf*

Prabhu, N., Kumar, V., Ray, I., & Yang, G.-C. (2004, March). Concurrency control in mobile database systems. *18th International Conference on Advanced Information Networking and Application (AINA04), 2,* 83-86.

Viellard, E. (2001). *Oracle9i Lite Business White Paper.* Retrieved December 2, 2003, from *http://*

otn.oracle.com/products/lite/pdf/o9ilite_bwp.pdf

Yang, X., Bouguettaya, A., Medjahed, B., Long, H., & He, W. (2003). Organizing and accessing Web services on air. *IEEE Transactions on Systems, Man, Cybernetics, Part A: Systems and Humans, 33*(6), 742-757.

KEY TERMS

Bluetooth: A wireless technology developed by Ericsson, Intel, Nokia, and Toshiba that specifies how mobile phones, computers, and PDAs interconnect with each other, with computers, and with office or home phones. The technology enables data connections between electronic devices in the 2.4 GHz range. Bluetooth can replace cable or infrared connections for such devices.

Caching: The technique of copying data from a server machine (the central storage place) to a client machine's local disk or memory; users then access the copy locally. Caching reduces network load because the data does not have to be fetched across the network more than once (unless the central copy changes).

Database Synchronization: When a database is being synchronized, no new update transactions are allowed, and all open update transactions are finished. After that, all updated blocks are written to disk.

Horizontal Industry Applications: A horizontal industry is one that aims to produce a wide range of goods and services. Horizontal industry applications are utilized across many different industries. While the core part of the application does not require changes, an organization needs customization at the front end or at the back end. Database access and service representative dispatch are typical examples of these applications.

IEEE 802.11: 802.11 refers to a family of specifications developed by the IEEE for wireless LAN technology. 802.11 specifies an over-the-air interface between a wireless client and a base station or between two wireless clients. The IEEE accepted the specification in 1997.

Load Balancing: It is the method of distributing system load evenly across server machines by placing identical copies of frequently accessed information among available server machines.

Middleware: This software manages the communication between a client program and a database. For example, a Web server connected to a database can be considered middleware as the Web server sits between the client program (a Web browser) and a database. The middleware allows the database to be changed without necessarily affecting the client and vice versa.

Mobile Application: A mobile application is any application that can be used on the move. It may or may not be wireless. It must be tailored to the characteristics of the device that it runs on. Limited resources, low network bandwidth, and intermittent connectivity are all important factors that affect the design of these applications.

Mobile Device: A mobile device is anything that can be used on the move, ranging from laptops to mobile phones. As long as the location is not fixed, it is considered mobile. Areas that are not included in the definition of mobile include remote offices, home offices, or home appliances.

Public Key Infrastructure (PKI): A system that enables users of a public network to exchange data securely and privately through the use of a public and private cryptographic key pair, which is obtained and shared through a trusted authority. It provides for a digital certificate that can identify an individual or an organization and director services that can store and, when necessary, revoke the certificates. The comprehensive architecture includes key management,

the registration authority, certificate authority, and various administrative tool sets.

Replication: The process of creating read-only copies of any data. Replication is supported by the security, directory, and file services in a distributed computing environment. Replication can improve availability and load balancing.

Secure Sockets Layer (SSL): SSL is a transaction security standard developed by Netscape Communications to enable commercial transactions to take place over the Internet. It's one of a few competing security standards.

Vertical Industry Applications: A vertical industry is one that is focused on a relatively narrow range of goods and services. Vertical industry applications are specific to certain industries. Usually there are some characteristics of the business processes unique to a particular industry that make certain applications very specific for that particular industry. As a result, some vendors develop turnkey software solutions for their own use.

This work was previously published in Encyclopedia of Database Technologies and Applications, edited by L. Rivero, J. Doorn, and V. Ferraggine, pp. 162-169, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.10
A Virtual Community for Mobile Agents

Sheng-Uei Guan
Brunel University, UK

Fangming Zhu
National University of Singapore, Singapore

INTRODUCTION

Electronic commerce (e-commerce) is booming with the increasing accessibility of the Internet. E-commerce is revolutionizing the concept of carrying out business functions. By using a Web browser, buyers are able to access numerous e-commerce Web sites, where they can make purchases within a reasonable price range. Suppliers realize that e-commerce is essential to the success and competitiveness of their businesses. The benefits of conducting business online include reduction of the cost for many transactions and streamlining of operations.

However, there are also some obstacles to the success of e-commerce. Firstly, buyers may be lost in the ocean of the items available. Secondly, it is a tedious task to search for a specific product through the Internet and it is difficult to bargain within the current infrastructure. Thirdly, some transactions are so complicated that they are too difficult to be dealt with. For instance, merchants often negotiate transactions with multiple issues of concern such as price, quantity, and method of delivery. Many strategies are adopted to accomplish these tasks, and both the negotiating counterparts and the environment can affect the choice of the strategies. However, in many existing auction Web sites, price is the main focus for both bidders and sellers. Bidders and sellers are seldom given a chance to negotiate the other issues, and many commercial opportunities are neglected.

This chapter discusses SAFER for e-commerce (secure agent fabrication, evolution & roaming for e-commerce), which uses secure agents to alleviate problems in e-commerce

BACKGROUND

Software agents have demonstrated potential in conducting transactional tasks in e-commerce through the Internet. It acts on behalf of an entity to carry out a delegated task. One of the earliest agents in e-commerce is the shopping agent, which carries out automatic comparative price shopping on the Web. A client can assign one or many shopping agents to carry out the shopping task. Agents can gather price information and present it to the client for a decision. Certainly, the task of a software agent involves more than online data gathering and filtering. For example, software agents are also used in negotiation (Guttman & Maes, 1998; Krishna & Ramesh, 1998). Negotiation agents are instructed with expected prices, quantities, delivery modes, and/or negotiation strategies (Oliver, 1996; Kang, 1998). Besides, software agents can also undertake other tasks, such as payment (Guan & Hua, 2003; Guan et al., 2004), mediation, distribution, interaction and sales promotion in e-commerce.

Software agents (Bradshaw, 1997; Poh & Guan, 2000; Wang et al., 2002; Guan & Zhu, 2002; 2004; Guan et al., 2004) can be endowed with attributes such as mobility, intelligence and autonomy. To alleviate concerns such as authorization, traceability, integrity, and security in e-commerce and the Internet, constructing appropriate architecture for agent systems in e-commerce is a fundamental consideration in facilitating agent-based transactions (Lee, 1997; Guan & Yang, 2004). As software agents become more common, there is a need for skilled programmers and even ordinary e-commerce clients to manipulate them. A practical way is to provide sites with methods to fabricate various agents according to the requirements of the clients. Agents should have an evolutional ability to enhance its intelligence and survivability. Roaming is one of the basic capabilities for agents so that they can fully utilize the power of network computing. They can achieve timesaving and cost cutting in completing its task without compromising security by roaming from one host to another (Yang & Guan, 2000; Guan & Yang, 2002).

THE SAFER ARCHITECTURE

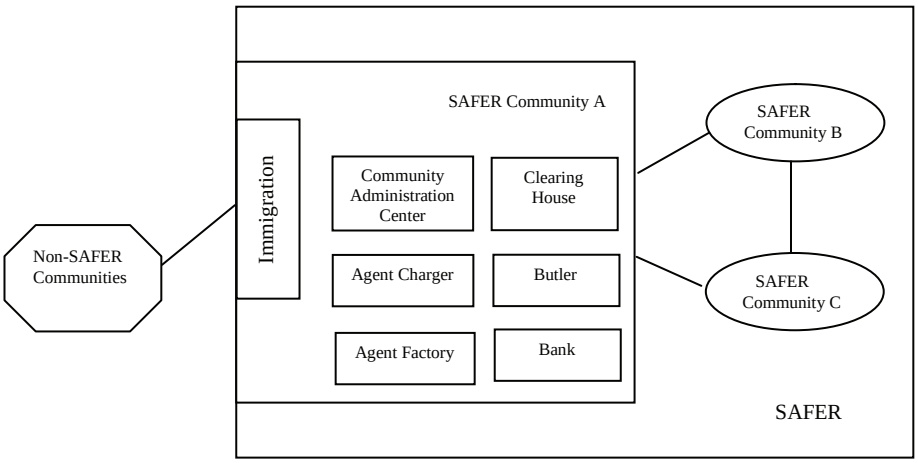
SAFER is an infrastructure to serve agents in e-commerce and establish the necessary mechanisms to manipulate them. The goal of SAFER is to recommend standard, dynamic and evolutionary agent systems for e-commerce. The SAFER architecture comprises different communities as described in Figure 1. Each community consists of the following components: owner, butler, agent, agent factory, community administration center, agent charger, agent immigration, clearing house and bank. Each component will be elaborated in the following subsections.

Community

Agents can be grouped into many communities based on certain criteria. In order to distinguish agents in the SAFER architecture from those that are not, we divide them into SAFER communities and non-SAFER communities as shown in Figure 1. We shall only discuss the SAFER community. Each SAFER community can possess a set of the facilities and individuals as described in Figure 2. Figure 2 only lists the necessary components in one community. Some community may have more entities than those depicted in the figure. For instance, there can be two agent chargers in a large community.

In order to become a SAFER community member, an applicant should apply to his local community administration center. The center will issue a certification to the applicant whenever it accepts the application. A digital certificate will be issued to prove the certified status of the applicant. To decide whether a facility or individual belongs to a community, one can look up the roster in the community administration center. A

Figure 1. SAFER architecture (1)



registered agent in one community may migrate into another community. In addition to permanent residence in a community, an agent can carry out its tasks in a foreign community. For agents to visit a foreign community, they must register in the foreign community administration center as guests. When an agent roams from one SAFER community to another, it will be checked by a trusted machine—agent immigration (Guan et al., 2003) with regard to its authorization and security before it can perform any action in this community.

Agent community is the basic unit in SAFER e-commerce. It offers factories and evolution vehicles to streamline e-commerce agents. Under these organized communities, agents can be regulated in a tidy order and perform their tasks more efficiently. The tighter structure also provides a solid base for enhancing the security of agents, which is one of the most important concerns in agent-based e-commerce systems.

Owner

Agent owners stand at the top of the SAFER architecture's hierarchy. They are the real users during the transactions and agents are acting on behalf of them. An owner has the priority and responsibility for all his agents. An owner controls his agents from creation to termination. An owner can request an agent factory (Guan et al., 2004) to fabricate agents responsible for specific e-commerce activities. Sometimes an owner needs to initiate important decisions of a transaction so that his agent can complete its tasks. For instance, negotiation agents need to request the final agreement from its owner before it can sign the contract. Each owner should register in the community administration center before he can have access to the facilities in the community. To relieve his burden, an owner can authorize a butler to handle most of his tasks.

Butler

An agent butler assists its agent owner in coordinating agents for him. In the absence of the agent owner, an agent butler will, depending on the authorization given, make decisions on behalf of the agent owner.

As agents are dispatched for certain missions, the agent owner will issue authorization to them. These authorizations may include the amount of credit an agent is allowed to spend, the range of host this agent is allowed to roam and others.

One function of the agent butler is to make payments when an agent is involved in any transaction with external parties. For example, suppose the agent owner authorizes the agent butler to handle transactions involving less than one hundred dollars. When one of the agents reaches an agreement to buy a book from Amazon.com and requests the payment from the agent butler, the agent butler can immediately issue the payment without further consulting the agent owner. The agents require the presence of agent butler in any transaction because in the SAFER architecture it is not given any capability to make payments. If the agent is allowed to make payments without consulting its owner, it will have to carry certain cash credit with it. The cash credit may be compromised in an event of agent abduction.

In addition, an agent butler also keeps track of the agent's activities and its location. For example, an information-gathering agent will send information like sites visited and information collected back to the agent butler. With this information, the agent butler may ensure that the other agents do not visit the sites again.

Another function of agent butlers is to act as receptionists in agent roaming (Yang & Guan, 2000). It services both the incoming and outgoing agents as well as coordinating agent transport.

Agent

Agent plays an active role in SAFER e-commerce. It is agent that brings to life the promising aspects in the next generation of e-commerce. All the facilities in the SAFER architecture serve agents in one way or another. Each agent has a unique identification and belongs to one specified owner. To identify itself, an agent should have a digital certificate issued by its creator. According to the tasks assigned by the owner, we can classify agents into many categories, such as negotiation agents, payment agents, mediation agents, and so on.

As an agent acts on behalf of its owner, it should have certain degree of intelligence. For example, it accepts the owner's assignment and carries out the task delegated to it. Agents have the mobility to travel through the Internet. Mobile agents can carry important information when they are roaming through the network to complete transactions. It should be immune from attacks from hackers or malicious agents. Security is thus the most crucial issue in agent roaming as well as other activities in e-commerce.

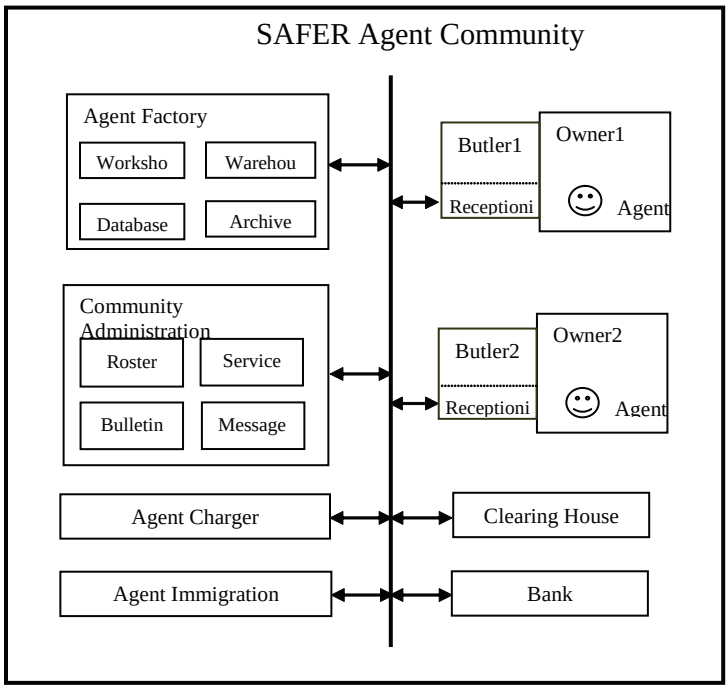
Agent Factory

Agent factory is the kernel of SAFER, as it undertakes the primary task of "creating" agents. In addition, agent factory has the responsibility to fix and check agents, which is an indispensable function in agent evolution and security.

As illustrated in Figure 2, an agent factory consists of four components, namely, workshop, warehouse, database and archive. Workshop is the site where an agent is fabricated, fixed, and checked. Database includes various ontology structures and standard modules to assemble different agents in the workshop. Archive is the set of the factory logs and information of agents that have undergone certain processes in the agent factory.

An agent factory provides an interface, such as choices of fabrication or repair instructions for owners (or butlers), to customize agents with desirable functionality. After the owner specifies the parameters and information, the agent factory starts to assemble or update the agent using the ontology structures and standard modules in the database. The agent factory also undertakes the liability to check the fitness and integrity of the agents in the community, which is essential to the evolution procedures and security protection of the agents. Any work done by the agent factory is recorded in the archive for later references.

Figure 2. SAFER architecture (2)



Community Administration Center

The community administration center is responsible for administrative matters in the community. It has the privileges of coordinating and facilitating the activities of all the entities in the community. Its aims are to assure the smooth running of routine operations and security of the whole community.

The administration center has a roster of the community, which includes basic data on the registered facilities, owners, agents and guest agents from other communities. This roster is updated periodically. When an agent factory has fabricated a new agent, it will inform the center to add the new item. When an agent is terminated, it should also be reflected in the roster. In addition, the center does a thorough routine examination of all components in the community and updates the roster periodically. If a foreign agent wants to enter this community for some purposes, it should request a "visa" from the agent immigration and the agent immigration will forward the registration to the administration center. For these reasons, the center is well aware of any incoming agents and events happening in it so that it is well guarded from any intruder.

Agent Charger

The agent charger is a trusted machine deployed to ensure agent integrity in SAFER. Since SAFER is designed for e-commerce, roaming agents must be protected from malicious attacks. One important aspect of agent protection is agent integrity. Roaming agents must not be interfered with during roaming or its execution at the remote host. The protection of agent integrity (both code integrity and data integrity) during roaming operation has been addressed in Yang and Guan (2000). In order to protect agent integrity during an agent's execution at a remote host, the concept of "agent battery" is now introduced.

Before explaining agent battery, the concept of "agent action" will first be defined. Agent action is any activity involving an agent with one or more third parties that may cause dispute or damage to another party. Typical agent actions include agent roaming, agent negotiation, and execution of transactions. Without the ability to perform "actions," an agent will be disabled like an electric toy car out of battery. Agent battery refers to a battery carried by each agent that specifies the number of "actions" it can perform. The agent battery decreases its energy level (i.e., number of "actions") by one each time before an action is executed. If the level reaches zero, the agent is not allowed to perform any more actions. In order to restore its energy level, the agent approaches an agent charger to regain its energy. Agent charger is located in the SAFER community. In large communities, there may be more than one agent charger to facilitate the charge-up operations. Before restoring an agent battery, the agent charger should inspect the agent for its fitness and integrity. The inspection should include both agent code or agent data. If an agent is found to be intact (code integrity and data integrity is not compromised), the charger can increase the battery quota in the agent. The amount of quota to be increased should be specified by the agent owner/butler. If no value is specified, the default amount will be used. In case the agent has been found interfered with during charging, the agent charger will detain the agent and informs either the agent's owner or its butler.

With this battery system, an agent is forced to go through a "medical check-up" periodically. Moreover, each agent when arriving at a machine, its battery will be checked to see if it runs out of battery and whether any compromise has been made.

Agent Immigration

In an effort to promote open architecture, SAFER is designed to allow interaction with agents and hosts from within its family and other non-SAFER architectures. This leads to another problem: how do agents roam from one community to another community?

Agent immigration is introduced to provide a mechanism in the administration of agents across community boundaries. If an agent needs to roam outside the community, it has to obtain a "visa" from the visiting community's immigration (Guan et al., 2003). A host will ensure that only foreign agents with valid visas are allowed to execute in its premise. Therefore, if an illegal agent sneaks into a community without going through agent immigration, it will not be able to perform any action.

The policy of issuing visas may differ from immigration to immigration. It is important to identify where an agent comes from and updates the information in the immigration log. If the agent causes any damage in the community, a trace is available to identify the malicious agent's owner. An agent is forbidden for entry if the immigration detects any problem. On the other hand, if the immigration exercises trusts on incoming agents, the above check can be waived. Different policy may be applied to agents from SAFER communities and non-SAFER communities since agents from non-SAFER communities are more likely to be malicious.

Clearing House and Bank

In order to facilitate financial transactions and clearance, clearing house and bank are included as separate entities in each SAFER community.

Each agent and host should open an account with the bank that resides at its originating community. The personal particulars of the agents will only be known by the local bank and are not disclosed. If a transaction takes place within a community that does not involve any party from other communities, they can make an appropriate request directly to a local bank for immediate settlement. However, if a transaction involves

parties from other communities, clearing house must be used as the medium for settlement with different banks. Different from banks, clearing house does not contain any account information as the bank does. It is merely a medium through which inter-bank settlement can be facilitated. Anonymity across different communities can be provided through the use of clearing house. The detailed payment scheme has been elaborated by Guan and Hua (2003).

AGENT FABRICATION

In SAFER, agents are fabricated by an agent factory in its community. There are many supporting arguments to adopt this mode of fabrication:

- Although some users may design agents by themselves, most users do not have the ability to do so. Also software agents in e-commerce have many types. It will be more convenient if an agent can be customized according to its own specification by using the agent factory.
- E-commerce agents implemented individually can lead to lack of standardization among owners. This may result in communication break-down.
- Adopting this mode of agent fabrication will enhance the security of SAFER e-commerce. Since information of all fabricated agents is stored in agent factories, agents can be administered more efficiently and safely.

Under SAFER, the fabrication of agents obeys prescribed routine procedures. An owner customizes new agents through the interface provided by an agent factory. When an agent factory fabricates a new agent, it chooses the corresponding ontology structure from the database according to the requests from the owners. The agent factory then assembles the agents according to both the ontology structure and the owner's specification.

Each ontology structure defines the components and construction of a specific type of agents. Different types of agents are defined with different ontology structures depending on their prototypes in agent factory. Examples are ontology structures for negotiation agents and information collection agents. The same type of agents may have different structures because of different requirements. The ontology structures are stored in the database of the agent factory.

AGENT EVOLUTION

One of the most prominent aspects of SAFER is agent evolution. As numerous agents are distributed throughout the Internet and act on behalf of different owners in different communities, collaborations and competitions exist among them. For example, agents with the same goal of finding the pricing for a certain type of computer can collaborate with one another. One agent can inform the others the Web sites it has visited and the information gathered so that other agents do not need to visit the same Web sites again. In another scenario, some agents may cooperate to negotiate with several sellers. They can share information and adjust their strategies accordingly during the negotiation process in order to reach satisfactory deals. On the other hand, competitions are inevitable when resource is limited. For example, if only a limited number of computers are available, agents will have to compete against each other to get them. In the end, some of them will succeed, while the others will fail. The successful agent may become more powerful, and the failed agent may lose some fitness. This is similar to collaborations and competitions in natural ecosystems.

The fitness of an agent is an indicator of an agent's ability to survive and adapt to the environment. The higher the fitness of an agent, the stronger it is. The evaluation of fitness is performed in agent factories and agent chargers using the following criteria:

- **Integrity of agents:** The integrity of agents may be compromised during the process of evolution, roaming or communicating. It is caused by intentional damages from malicious agents or accidental errors during legal formalities.
- **History of agents:** History of agents includes the number of tasks carried out and the quality of completed tasks. Every time an agent completes a task and reports to its owner, the owner will assess the quality and give a corresponding mark. Through analyzing the trend of agent performance by combining every task and its mark, the fitness of agent can be evaluated.
- **Evolution record:** Evolution record can be an auxiliary tool in the evaluation of agent fitness. Every result of fitness evaluation is stored in the community administration center. It shows whether an agent's growth is healthy.

If an agent charger finds the fitness of an agent too low, it can reject to recharge the agent and adopt some other measures. Furthermore, the agent charger checks the evolution record of the agent. If the fitness of an agent is decreasing rapidly, it can be suspected that the agent might have been attacked or something may have gone wrong in the working experience of the agent. The agent charger can then send it back to an agent factory for a thorough examination. The agent factory has the right to detain or terminate an agent. It will inform the community administration center as well as the owner about the measures taken.

AGENT ROAMING

A set of agent transport protocols has been designed for SAFER in Yang and Guan (2000) to allow intelligent agents to roam from host to host. The transport protocols designed provide a secure mechanism for agents in e-commerce to roam across different hosts and communities in SAFER.

Supervised agent transport protocol allows controls to agent owner/butler during an agent's roaming operation. The agent has to obtain an approval from its owner/butler before roaming to a new host. The owner/butler can thus control the agent's roaming destination and prevent the agent from moving to certain undesirable hosts. The drawback of this protocol is the lack of efficiency since each agent's movement involves the agent owner/butler. This involvement will inevitably delay the transport process and incur additional network traffics. If the agent owner/butler happens to be using a low bandwidth connection to the Internet, the situation may worsen as agents roaming in high bandwidth networks suffer from the bottleneck in low bandwidth communication with the agent owner/butler.

On the other hand, unsupervised agent transport protocol does not involve the agent owner/butler directly in the transport process. Agents do not need to request for a permission before roaming. Instead, an indirect notification of the roaming operation is sent to the agent owner/butler for recording purpose. The agent owner/butler is unable to control the agent's roaming destination directly. The advantage is the increased efficiency of agent roaming since fewer parties are directly involved in the transport process, thus leading to shorter turnaround time.

Based on different concerns on efficiency, roaming control as well as level of security, individual SAFER agents can choose to use the most appropriate transport protocol or even a combination of different transport protocols in their roaming operations.

FUTURE TRENDS

We have planned the following steps to further improve the architecture and its functions. Firstly, we have started to implement the architecture

using Java. We are trying to provide the basic modules and facilities first, as we regard SAFER as our infrastructure for further research on e-commerce. Secondly, as evolution is proposed in the SAFER architecture, its mechanism and theory presentation is being constructed. Thirdly, we are developing the payment mechanism in SAFER, as it is the essential part in e-commerce.

CONCLUSION

In this article, we have proposed the SAFER architecture for agent-based e-commerce. SAFER provides the facilities to serve agents in e-commerce and establishes the necessary mechanisms to manage and control their activities. SAFER covers the whole lifecycle of an agent from its fabrication to its termination. We elaborate the functions of the components in SAFER, and the three aspects of SAFER: fabrication, evolution & roaming. Tree structure is employed to present the agent and ontology structure in agent fabrication and evolution.

SAFER e-commerce provides an opportunity in standardization for dynamic, secure and evolutionary agent architecture. With the SAFER architecture, conducting transactions in the Internet will be more convenient, secure and efficient.

REFERENCES

Bradshaw, J.M. (1997). *Software agent*. Cambridge, MA: MIT Press.

Guan, S.-U., & Hua, F. (2003). A multi-agent architecture for electronic payment. *International Journal of Information Technology and Decision Making (IJITDM), 2*(3), 497-522

Guan, S.-U., Tan, S.L., & Hua, F. (2004). A modularized electronic payment system for agent-based e-commerce. *Journal of Research and Practice in Information Technology, 36*(2), 67-87.

Guan, S.-U., Wang, T., & Ong, S.-H. (2003). Migration control for mobile agents based on passport and visa. *Future Generation Computer Systems, 19*(2), 173-186.

Guan, S.-U., & Yang, Y. (2002). SAFE: Secure agent roaming for e-commerce. *Computer & Industrial Engineering Journal, 42*, 481-493.

Guan, S.-U., & Yang, Y. (2004). Secure agent data integrity shield. *Electronic Commerce and Research Applications, 3*(3), 311-326.

Guan, S.-U., & Zhu, F. (2002). Agent fabrication and its implementation for agent-based electronic commerce. *International Journal of Information Technology and Decision Making (IJITDM), 1*(3), 473-489.

Guan, S.-U., & Zhu, F. (2004). Ontology acquisition and exchange of evolutionary product-brokering agent. *Journal of Research and Practice in Information Technology, 36*(1), 35-46.

Guan, S.-U., Zhu, F., & Maung, M.T. (2004). A factory-based approach to support e-commerce agent fabrication. *Electronic Commerce and Research Applications, 3*(1), 39-53.

Guttman, R.H., & Maes, P. (1998). Agent-mediated negotiation for retail electronic commerce. In Selected Papers from the First International Workshop on Agent Mediated Electronic Trading on Agent Mediated Electronic (pp. 70-90). Minneapolis, MN. London: Springer-Verlag.

Kang, J.Y., & Lee, E.S. (1998). A negotiation model in electronic commerce to reflect multiple transaction factor and learning. In *Proceedings of the 12th International Conference on Information Networking.*

Krishna, V., & Ramesh, V.C. (1998). Intelligent agents for negotiation in market games, Part 1: Model. *IEEE Transactions on Power Systems, 13*(3).

Lee, J.G., Kang, J.Y., & Lee, E.S. (1997). ICOMA: An open infrastructure for agent-based intelligent electronic commerce on the Internet. In *Proceedings of the 1997 International Conference on Parallel and Distributed Systems* (pp. 648-655). Seoul, South Korea.

Oliver, J.R. (1996). On artificial agents for negotiation in electronic commerce. In *Proceedings of the 29th Annual Hawaii International Conference on System Sciences* (Vol. 4, pp. 337-346).Wailea.

Poh, T.K., & Guan, S.U. (2000). Internet-enabled smart card agent environment and applications. In S.M. Rahman & M. Raisinghani (Eds.), *Electronic commerce: Opportunities and challenges.* Hershey, PA: Idea Group Publishing.

Wang, T., Guan, S.U., & Chan, T.K. (2002). Integrity protection for code-on-demand mobile agents in e-commerce. *Journal of Systems and Software*, *60*(3), 211-221.

Yang, Y., & Guan, S.U. (2000). Intelligent mobile agents for e-commerce: Security issues and agent transport. In S.M. Rahman & M. Raisinghani (Eds.), *Electronic commerce: Opportunities and challenges.* Hershey, PA: Idea Group Publishing.

KEY TERMS

Agents: A piece of software, which acts to accomplish tasks on behalf of its user.

Anonymity: The degree to which a software system or component allows for or supports anonymous transactions.

Authentication: The process of ensuring that an individual is one who he or she claims to be.

Client: In this work, it refers to customers who pay for good and services.

Digital Certificate: A certificate that uses a digital signature to bind together a public key with an identity—information such as the name of a person or an organization, the address, and so forth. The certificate can be used to verify an agent's identity, for example.

E-Commerce: The act of conducting business transactions over networks and through computers.

Integrity: Regarding the protection of data or program code from being modified by unauthorized parties.

Security: The effort to create a secure computing platform, designed so that agents (users or programs) can only perform actions that have been allowed.

This work was previously published in Encyclopedia of Networked and Virtual Organizations, edited by G. Putnik and M. Cunha, pp. 1764-1771, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.11
Concepts and Operations of Two Research Projects on Web Services and Mobile Web Services

Zakaria Maamar
Zayed University, UAE

ABSTRACT

Today, Internet technologies are enabling a wave of innovations that have an important impact on the way businesses deal with their partners and customers. Most businesses are moving their operations to the Web for more automation, efficient business processes, and global visibility. Web services are one of the promising technologies that help businesses in achieving these operations and being more Web-oriented. Besides the new role of the Internet as a vehicle of delivering Web services, a major growth in the field of wireless and mobile technologies is witnessed. Because users are heavily relying on mobile devices to conduct their operations, enacting Web services from mobile devices and possibly downloading these Web services for execution on mobile devices are avenues that academia and industry communities are pursuing. M-services denote the Web services in the wireless world. In this chapter, two research initiatives carried out at Zayed University are presented and referred to as SAMOS, standing for Software Agents for MObile Services, and SASC, standing for Software Agents for Service Composition.

OVERVIEW

Today, several businesses are adopting Web-based solutions for their operation, aiming for more process automation and more worldwide visibility. Thanks to the Web technology, users from all over the world can satisfy their needs by browsing and triggering the services of these businesses. Such services are usually referred to as Web services (Boualem, Zeng & Dumas, 2003). The advantages of Web services have already been demonstrated in various projects and highlight their capacity to

be composed into high-level business processes. For example, a vacation business process calls for the collaboration of at least four Web services: flight reservation, accommodation booking, attraction search, and user notification. These Web services have to be connected with respect to a certain flow of control (first, flight reservation, then accommodation booking and attraction search). Multiple technologies are associated with the success of Web services, namely, WSDL (Web Services Definition Language), UDDI (Universal Description, Discovery, and Integration), and SOAP (Simple Object Access Protocol) (Curbera, Duftler, Khalaf, Nagy, Mukhi & Weerawarana, 2002). These technologies support the definition, advertisement, and binding of Web services.

Besides the Web expansion, we witness the tremendous progress in the field of wireless technologies. Telecom companies are deploying new services for mobile devices. Reading e-mails and sending messages between cell phones are becoming natural. Surfing the Web, thanks to the Wireless Application Protocol (WAP), is another evidence of the wireless technology development. The next stage (if we are not already in it) for telecom and IT businesses is to allow users to enact Web services from mobile devices and, possibly, to make these Web services runnable on mobile devices. M-services (M for mobile) denote these new type of Web services (Maamar & Mansoor, 2003).

It is accepted that composing multiple services (whether Web services or M-services) rather than accessing a single service is essential. Berardi et al. (2003) report that composition addresses the situation of a client's request that cannot be satisfied by any available service, whereas a composite service obtained by combining a set of available services might be used. Searching for the relevant services, integrating these services into a composite service, triggering the composite service, and monitoring its execution are among the operations that users will be in charge of. Most of these operations are complex, although repetitive, with a large segment suitable for computer aids and automation. Therefore, software agents are deemed appropriate candidates to assist users in their operations (Jennings, Sycara & Wooldridge, 1998).

Throughout this chapter, two research initiatives that our research group is conducting at Zayed University are presented. These initiatives are respectively **SAMOS**, standing for **Software Agents for MObile Services**, and **SASC,** standing for **Software Agents for Service Composition**. Both initiatives deal with the composition of services using software agent-oriented approaches. This chapter is structured as follows. The Background section outlines the concepts that are used in our research work, such as mobile computing and software agents. The next section overviews some research projects related to mobile computing. The SAMOS Research Initiative and SASC Research Initiative sections present SAMOS and SASC in terms of architecture, types of agents, and operation. In the last section, we draw our conclusions.

BACKGROUND

Mobile Computing

Mobile computing refers to systems in which computational components, either hardware or software, change locations in a physical environment. The ability to move from one location to another is because of the progress in several technologies: component miniaturization, wireless networks, and mobile-code programming languages. Categories of mobility include (Wand & Chunnian, 2001): hardware mobility, software mobility, and combined mobility. A code that is downloaded from a server to a mobile phone combines both hardware and software mobility. The Overview of Some Research Projects Related to Mobile Computing section provides more details on mobile computing using research projects as examples.

Web Services and M-Services

A Web service is an accessible application that can be automatically discovered and invoked by other applications and humans. An application is a Web service if it is (Benatallah et al., 2003): (i) independent as much as possible from specific platforms and computing paradigms; (ii) developed mainly for interorganizational situations rather than for intraorganizational situations; and (iii) easily composable so its composition with other Web services does not require the development of complex adapters.

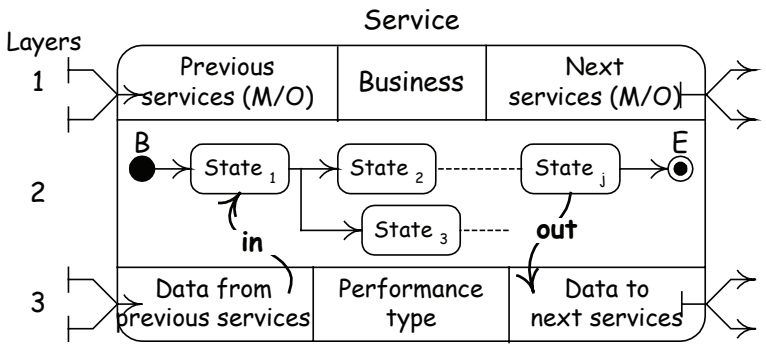
Two definitions are associated with an M-service (Maamar & Mansoor, 2003). The weak definition is to remotely trigger a Web service for execution from a mobile device. In that case, the Web service acts as an M-service. The strong definition is to wirelessly transfer a Web service from its hosting site to a mobile device where its execution happens. In that case, the Web service acts as an M-service that is: (i) transportable through wireless networks; (ii) composable with other M-services; (iii) adaptable with regard to the computing features of mobile devices; and (iv) runnable on mobile devices. In both SAMOS and SASC initiatives, only the M-services that comply with the strong definition are considered.

The differences between Web services and M-services are depicted at two levels. The first level concerns the communication medium (wired channel for Web services versus wireless channel for M-services). And the second level concerns the location of where the processing of the service occurs (server side for Web services versus user side for M-services).

In Maamar, Benatallah, and Mansoor (2003), we introduced the concept of service chart diagram as a technique for modeling and specifying the component services that participate in composite services. A service chart diagram enhances the state chart diagram of UML. In fact, the emphasis this time is on the context surrounding the execution of a service rather than only on the states that a service takes (Figure 1).

A service chart diagram wraps the states of a service into five perspectives, each perspective has a set of parameters. The state perspective corresponds to the state chart diagram of the service. The flow perspective corresponds to the execution chronology of the composite service in which the service participates (Previous services/Next services parameters; M/O respectively stands for Mandatory and Optional). The business perspective identifies the organizations (that is, providers) that make the service available (Business parameter). The information perspective identifies the data that are exchanged between the services of the composite service (Data from previous services/Data for next services parameters). Because the services participating in a composition can be either mandatory or optional, the information per-

Figure 1. Service chart diagram of a component service



spective is tightly coupled to the flow perspective with regard to mandatory data and optional data. Finally, the performance perspective illustrates the ways the service can be invoked for execution (Performance type parameter).

Software Agents

A software agent is a piece of software that autonomously acts to undertake tasks on behalf of users (Jennings et al., 1998). The design of many software agents is based on the approach that the user only needs to specify a high-level goal instead of issuing explicit instructions, leaving the how and when decisions to the agent. A software agent exhibits a number of features that make it different from other traditional components including autonomy, goal orientation, collaboration, flexibility, self-starting, temporal continuity, character, communication, adaptation, and mobility. It is noted that not all of these characteristics have to embody an agent.

Besides the availability of several approaches and technologies related to the deployment of Web services (for example, SOAP, UDDI, Salutation), they are all tailored to a context of type wired. In a similar context, all the computing resources are fixed and connected through a permanent and reliable communication infrastructure. The application of these approaches and technologies to a context of type mobile computing is not straightforward. Indeed, major adjustments are required because of multiple obstacles ranging from potential disconnections of mobile devices and unrestricted mobility of persons to power scarcity of mobile devices and possibility of capturing the radio signals while in the air. These obstacles highlight the suitability of software agents as potential candidates to overcome them. First, an agent is autonomous. Thus, it can make decisions on the user's behalf while this one is disconnected. Second, an agent can be mobile. Thus, it can move from one host to another. Continuous network connectivity is not needed. Third, an agent is collaborative. Thus, it can work with other agents that identify, for example, the providers of Web services. Last but not least, an agent is reactive. Thus, it can monitor the events that occur in the user's environment, so relevant actions can be promptly taken.

OVERVIEW OF SOME PROJECTS RELATED TO MOBILE COMPUTING

There exist several research projects that have studied how mobile devices can change the way of doing business and undertaking operations. In HP Laboratories, the authors in Milojicic et al. (2001) worked on delivering Internet services to mobile users. This work was conducted under the project Ψ for Pervasive Services Infrastructure (PSI). The Ψ vision is "any service to any client (anytime, anywhere)". The project investigated how offloading parts of applications to midpoint servers can enable and enhance service execution on a resource-constrained device.

The Odyssey project aimed at providing system support for mobile and adaptive applications (Noble et al., 1997). Odyssey defined a platform for adaptive mobile data access on which different applications, such as Web browser, video player, and speech recognition, can run on top. The Odyssey approach is to adjust the quality of accessed data to match available resources.

Ninja aimed at suggesting new types of robust and scalable distributed Internet services (Ninja, 2001). The objective in Ninja is to meet the requirements of an emerging class of extremely heterogeneous devices that would access these services in a transparent way. In Ninja, the architecture considered four elements: bases, units, active proxies, and paths. Proxies are transformational intermediaries that are deployed between devices and services to shield them from each other. A service discovery service is also suggested in Ninja for two reasons: (i) enable services to announce their presence and (ii) enable users and programs to locate the announced services.

SAMOS RESEARCH INITIATIVE

In addition to the role of the Internet as a vehicle of provisioning Web services, it is noticed that more Web services will be delivered to people who use mobile devices and, particularly, to those who are on the move most of the time (for example, sales representatives). It is also noticed that mobile devices are being enhanced with extra computing resources and advanced functionalities (Yunos, Gao & Shim, 2003). Unfortunately, the growth in the development and use of mobile devices is subject to multiple challenges. For instance, mobile devices are still bound to their batteries for operation, which leads to limit, to a certain extent, their computation performance.

It occurs that mobile users have to postpone their operations because they lack appropriate facilities running on their mobile devices (for example, an application that converts a drawing file into a format that the user's mobile device can display). In SAMOS, we support such users by allowing them: (i) to search for additional facilities, when needed; (ii) to fetch these facilities to their mobile devices; and (iii) to conduct these two operations in a transparent way. Various solutions are put forward to handle these points and are discussed throughout this part of the chapter. A solution to point (i) consists of devising brokering mechanisms. A solution to point (ii) consists of using wireless communication channels. Finally, a solution to point (iii) consists of using Software Agents (SAs) to make the search for and fetch the facilities transparent to users.

Architecture and Software Agents of SAMOS

Brokering mechanisms and SAs are considered in the design and development of SAMOS. The salient features of the architecture of SAMOS are:

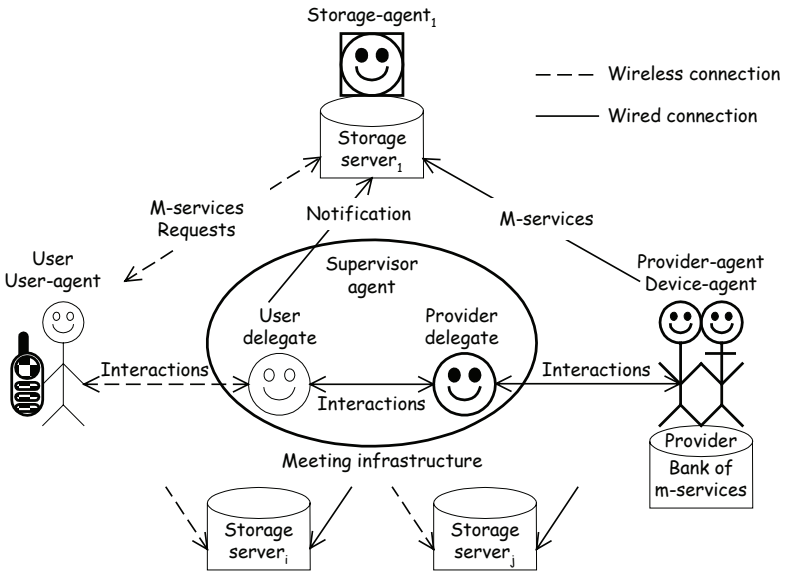
- Three types of SAs: user-agent, provider-agent, and device-agent. The first type is associated with users of M-services, whereas the second and third types are associated with providers of M-services.
- A software platform, called Meeting Infrastructure (MI), is headed by a supervisor-agent. This MI has a brokering role (Maamar, Dorion & Daigle, 2001).
- Two types of delegates, namely, provider-delegate and user-delegate. Delegates respectively interact on behalf of user-agents and provider-agents in the MI.
- Storage servers that save the sequence of M-services to be submitted to mobile devices for execution. Storage servers are spread across networks, and storage-agents are responsible for managing these servers. In SAMOS, a sequence corresponds to a composite service that has M-services as primitive components.

Figure 2 illustrates the architecture of SAMOS. It consists of four parts: user, provider, MI, and storage. The MI and storage parts are wirelessly linked to the user component, whereas the MI and storage parts are linked to the provider component with wires.

The user part consists of users and user-agents. User-agents accept users' needs, convert them into requests, and submit them to user-delegates. The supervisor of the MI creates user-delegates on requests from user-agents. To satisfy users' requests, user-delegates interact with provider-delegates.

The provider part consists of providers, provider-agents, and device-agents. Provider-agents act on behalf of providers by advertising their M-services to user-delegates through provider-delegates. Plus, provider-agents monitor the behavior of providers when new M-services are offered and, thus, need to be announced. In Figure 2, M-services are gathered into a bank on

Figure 2. Architecture of SAMOS



which provider-agents and device-agents reside. Provider-agents create provider-delegates. In the provider part, device-agents support the work of provider-agents, whereas the role of device-agents is to wrap the M-services before they are sent to mobile devices for execution. The rationale of device-agents is to consider the differences that exist between mobile devices (for example, screen size, processor power).

The MI part is a software platform in which user-delegates and provider-delegates interact in a local and secure environment (Maamar et al., 2001). In an open environment, most of the interactions occurring between requesters of services and providers of services are conducted through third parties (referred to as brokers). Despite its important role, a third party can easily become a bottleneck. To overcome this problem, requesters and providers need a common environment in which they meet and interact directly. The MI corresponds to this common environment. In SAMOS, the supervisor-agent of the MI has several responsibilities including monitoring the interactions that occur within the MI and making the MI a safe environment.

The storage part receives the sequence of M-services that will be submitted to mobile devices for performance. In SAMOS, one of the operation principles is to submit the M-services to mobile devices for execution one at a time. This restriction is due to the limited resources of these devices. However, the restriction can be handled (that is, adjust the number of M-services to be submitted) based on the computing resources of a mobile device and the bandwidth of the wireless communication channels. Several advantages are obtained from the use of storage servers. For instance, a user-agent does not have to deal with several providers. Its unique point of contact for getting the M-services is the storage-agent. The same thing applies to device-agents that will only be interacting with few storage-agents instead of multiple user-agents. Security is increased for both users and providers. Indeed, storage servers are independent platforms where security controls are carried out.

User-Oriented Components

A user-agent resides in a mobile device. First, the user interacts with the user-agent to arrange

requests. After submitting those requests to the user-delegate, the user-agent takes a standby state and waits for notifications from its user-delegate. Notifications concern the sequence of M-services that satisfies the user's requests. Before executing them on the user's device, the M-services are put in a storage server. The MI supervisor-agent suggests to the user-delegate the storage server to be used based, for example, on the server's location. To download the M-services one at a time from the storage server to the user's device, the user-agent communicates with the storage-agent. The user-agent keeps track of the execution of the M-services before it asks the storage-agent to submit further M-services. When an M-service is received, the executed M-service is deleted from the mobile device. Finally, the user-agent informs the user about the completed requests.

A user-delegate resides in the MI, acting on behalf of the user-agent. The user-delegate receives the user's requests from the user-agent. Afterward, it interacts with provider-delegates. The purpose of these interactions is to match the requests of users to the M-services of providers that are announced. In case there is a match (we assume that there is always a match), the user-delegate designs the sequence of M-services that satisfies the user's requests. Information about this sequence is sent afterward to the storage-agent. The objective is to make the storage-agent ready for receiving the M-services from device-agents. Furthermore, the user-delegate notifies the user-agent about the sequence of M-services it has prepared for its user. To set up a sequence, the storage-agent knows the M-service that comes before and after the M-services to be submitted by a device-agent (flow perspective of a service chart diagram, Figure 1). Instead of creating a user-delegate on a mobile platform and shipping that delegate to the MI, we suggested to perform this operation in the MI for two main reasons: (i) even if we expect a major improvement in the resources of mobile devices, those resources have to be used in a *rationale* way and (ii) the wireless connection that transfers the user-delegate to the MI is avoided.

Provider-Oriented Components

A Provider-agent resides in a provider site running on top of its resources such as M-services. Provider-delegates broadcast the M-services to user-agents through user-delegates. The provider-agent is in constant interaction with its provider-delegate. For instance, it notifies the provider-delegate about the negotiation strategy it has to follow with user-delegates.

A device-agent resides in a provider site. Its responsibility is to wrap the M-services according to the features of the devices to which these M-services will be submitted for performance. Initially, the M-services are sent to storage servers. The provider-agent has already submitted the contact details of the storage server to the device agent. We recall that the user-delegate informs the storage-agent of the storage server about the M-services it will receive. Double checking the information that user-delegates and provider-delegates submit to a storage-agent offers more security to the agents of SAMOS.

A provider-delegate resides in the MI, acting on behalf of a provider-agent. In SAMOS, the provider-delegate is responsible for interacting with user-delegates regarding the M-services it offers. In addition, the provider-delegate interacts with its provider-agent for notification purposes. Notifications are then forwarded to device-agents for action. We recall that the provider-agent is responsible for creating the provider-delegate and its transfer to the MI.

MI-Oriented Components

The supervisor-agent resides in the MI and has several responsibilities: it supervises the operations that occur in the MI; it mediates in case of conflicts between user-delegates and provider-delegates; it sets user-delegates and assigns them

to user-agents; it checks the identity of provider-delegates when they arrive from provider sites; and finally, it suggests to user-delegates the storage server to be used.

Storage-Oriented Components

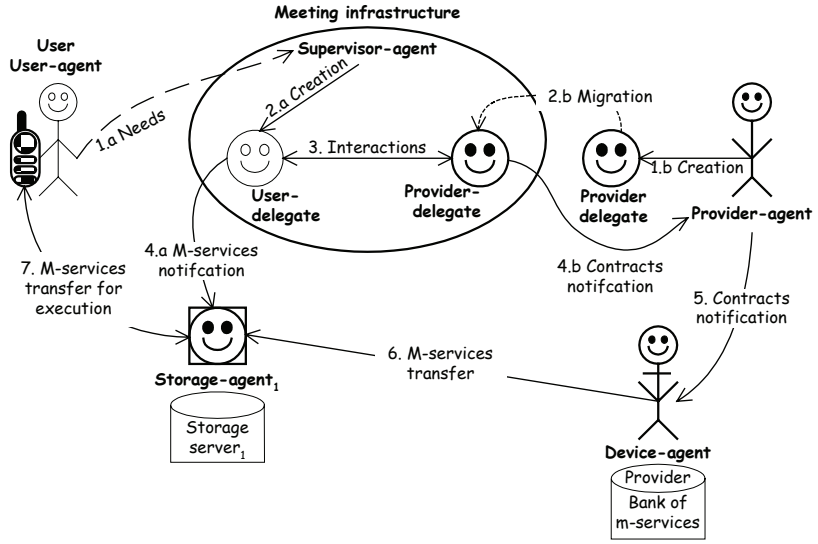
A Storage-agent runs on top of a storage server. This server saves the M-services to be sent to mobile devices for performance. According to the information on the sequence of M-services it receives from the user-delegate, the storage-agent arranges the sequence as the M-services start arriving from providers. As soon as this sequence is completed, it notifies the user-agent in order to get ready for receiving the M-services. Based on the requests it receives from the user-agent, the storage-agent submits the M-services one at a time. These M-services are ready for execution. The deletion of M-services from the storage servers and mobile devices follows certain reliability rules. These rules ensure that the M-services to be sent to a mobile device for execution are successfully received and executed.

OPERATION OF SAMOS

The operation of SAMOS consists of five stages (Figure 3): agentification, identification, correspondence, notification, and realization. The purpose of the agentification stage is to set up the different infrastructures and agents that constitute SAMOS. User-agents are established at the user level. Provider-agents and device-agents are established at the provider level, too. Finally, the meeting infrastructure and storage servers, including their storage-agent, are deployed. In Figure 3, operations (1.a) and (1.b) illustrate the agentification stage.

The purpose of the identification stage is to inform the supervisor-agent of the MI about the existence of users and providers who are interested in using SAMOS. At the agentification stage, user-agents and provider-agents are respectively installed on top of mobile devices of users and resources of providers. The outcome of the identification stage is the creation of user-delegates and the reception of provider-delegates arriving from provider-sites. Creation and reception operations occur in the MI. Provider-agents notify the supervisor-agent about their readiness to submit the

Figure 3. Operation of SAMOS



provider-delegates to the MI. User-agents inform the supervisor-agent about the users' requests they would like to submit. In Figure 3, operations (2.a) and (2.b) illustrate the identification stage.

The purpose of the correspondence stage is to enable user-delegates and provider-delegates to get together. In Figure 3, operation (3) illustrates this stage. User-delegates have requests to satisfy, and provider-delegates have services to offer. First, the user-delegate searches for the provider-delegates that have the M-services it needs. Two approaches are offered (Maamar & Mansoor, 2003):

a) The user-delegate asks the supervisor-agent to suggest a list of provider-delegates that have the services it needs.
b) The user-delegate requests from the supervisor-agent the contact details of all the provider-delegates that exist in the MI.

Independently of the approach that is adopted, the user-delegate submits its needs of services to a shortlist of selected provider-delegates. Based on different parameters, such as workload and commitments, provider-delegates answer the user-delegate. At this time of our research in SAMOS, it is assumed that providers do not have services in common. Consequently, there is no need for a user-delegate to look for the best service. Once the user-delegate and provider-delegates agree on the M-services to use, notifications are sent to different recipients as it is discussed in the next stage.

The purpose of the notification stage is to inform different agents about the agreements between user-delegates and provider-delegates. In Figure 3, operations (4.a), (4.b), (5), and (6) illustrate this stage. Regarding the user-delegate, it is in charge of informing (i) the user-agent about the sequence of M-services it has established to satisfy its user's request and (ii) the storage-agent about the sequence of M-services it will receive from different device-agents. Regarding the provider-delegate, it notifies the provider-agent about its agreements with a user-delegate. Based on the information it receives from its provider-delegate, the provider-agent forwards this information to the device-agent. This information is about the M-services that are involved and the storage server that is used. Among the actions the device-agent takes is to submit the M-services to the storage-agent of the storage server.

The purpose of the realization stage is to execute the sequence of M-services that the user-delegate has designed. User-agent and storage-agent participate in this stage. We recall that the user-delegate has already informed the storage-agent about the M-services it will receive from device-agents. Before the user-agent starts asking the storage-agent for the M-services it has, it waits for a notification message from the storage-agent mentioning that the sequence is ready for submission and, thus, for execution. In Figure 3, operation (7) illustrates the realization stage. In the realization phase, reliability is one of the concerns that have been considered in SAMOS. We consider a storage server as a backup server for the M-services. When a storage-agent sends an M-service to a user-agent, the storage-agent keeps a copy of this service at its level. The storage-agent deletes that M-service when the user-agent asks for the M-service that follows the one it has received. For the last M-service of a sequence, the user-agent sends an acknowledgment message to the storage-agent, so this M-service can be deleted.

Summary on SAMOS

In this part of the chapter, we discussed the use of M-services in the context of SAMOS. M-services are seen as a logical extension to the widespread use of Web services in the wireless world. Considering mobile devices as computing platforms is becoming a reality as the networks that make them reachable are in constant progress, offering more bandwidth and ensuring more reliability and efficiency. For instance, third-generation com-

munication systems are providing high quality streamed Internet content (Chisalita & Shahmehri, 2001). In addition to higher data rates, these systems back the provision of new value-added services to users, such as geographical positioning and mobile payment.

SASC RESEARCH INITIATIVE

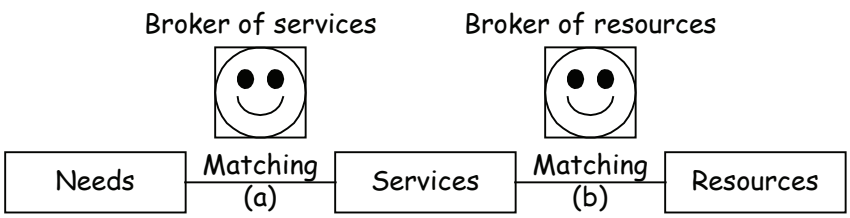
Despite that provisioning Web services is a very active area of research and development, very little has been done to date regarding their integration with M-services. Several obstacles still exist including throughput and connectivity of wireless networks, limited computing resources of mobile devices, and risks of communication channel disconnections.

A framework that composes services, whether Web services or M-services, should offer more opportunities to users to conduct operations regardless of (i) the type of services, (ii) the location of users, and (iii) the computing resources on which services will be performed. This situation is challenging due to the gap existing between wired and wireless. First, Web services are associated with fixed devices. However, M-services are associated with mobile devices. Second, the execution of Web services occurs in the server side, whereas the execution of M-services occurs in the client side (according to the strong definition of what an M-service is). Third, fixed devices are not resource-constrained which is not the case for mobile devices. Despite the multiple opportunities that could be offered to users, few research efforts are being dedicated to the composition of Web services and M-services.

Because the information space is already full of several providers of services, a broker that matches services to needs of users is one step in the design of the SASC framework (Figure 4a). On the other side, because services require resources on which they can be computed, there is a need for another broker as a second step in the design of the SASC framework (Figure 4b). This broker matches services (those that satisfy users' needs) to the resources of providers.

In the previous paragraph, it is shown that two types of providers are involved: provider of services and provider of resources (a provider can play both roles). Due to this distinction of providers, a user with a fixed or mobile device is also seen in the SASC framework as a provider of resources in the composition framework (that is, users' devices are advertised to the broker of resources). Considering users as providers of resources enables them to play an active role instead of always being limited to their traditional passive role of consumers. The rationale is to take advantage of the spare resources that are available on devices. It is observed that many of the systems are often underutilized due to geography factor. Busy hours in one time zone tend to be idle hours in another zone. Therefore, demands for computational resources can be met with hosts that have idle resources. For the needs of the SASC initiative, the term composite service denotes the set of component services (whether

Figure 4. Needs versus services and services versus resources



composite services, Web services, or M-services) that take part in a composition.

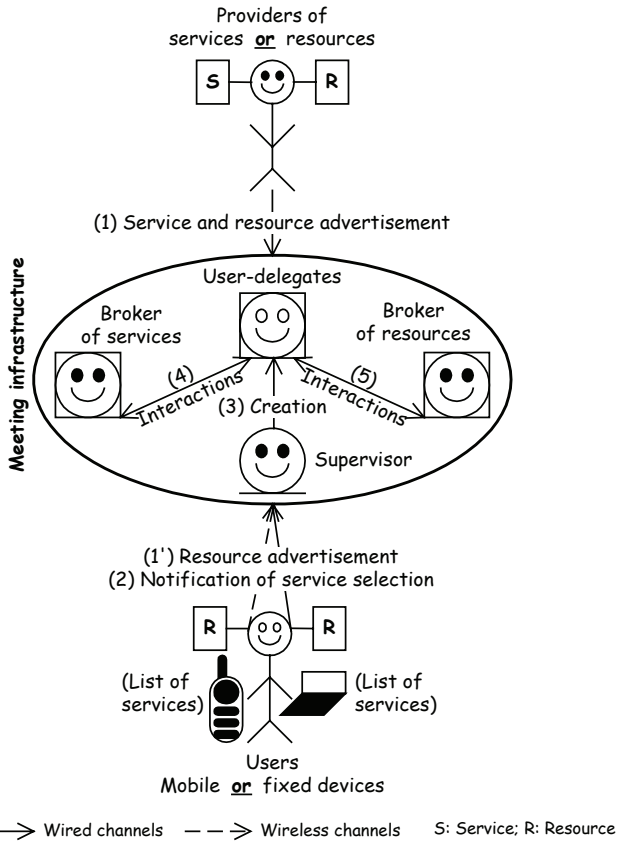
Architecture and Software Agents of SASC

Figure 5 is the agent-based architecture upon which the SASC framework is deployed. The architecture has three parts. The first part corresponds to providers of services (S) or resources (R). The second part corresponds to consumers of services (that is, users) with their fixed or mobile devices. Finally, the third part corresponds to the meeting infrastructure (similar to the one that is used in SAMOS) on which brokers carry out the matching operations between needs of users and services of providers and, later on, between services of providers and resources of providers (Figure 4). The meeting infrastructure connects the provider and consumer parts. To keep Figure 5 clear, the different agents that populate the architecture are not represented. The core agents of the framework are briefly described below.

Provider-agents are specialized into two types: resource-provider-agents and service-provider-agents. Resource-provider-agents handle the execution of the services of service-provider-agents. In the MI, resource-delegates and service-delegates, respectively, represent resource providers and service providers (delegates are agents but are given a different name to avoid confusion).

User-agents reside in the devices of users and are specialized into two types: fixed-user-agents (for users of fixed devices) and mobile-user-agents

Figure 5. Architecture of SASC



(for users with mobile devices). In the meeting infrastructure, user-delegates represent users to whom their needs are submitted.

Broker-agents are specialized into two types (Figure 4): service-broker-agent and resource-broker-agent. A service-broker-agent receives (i) notifications from service-delegates about their offers of services and (ii) requests from user-delegates about their needs of services. Whereas a resource-broker-agent receives (i) notifications from resource delegates regarding their respective offers of resources and (ii) requests from service-delegates regarding their needs of resources.

The supervisor-agent is in charge of the MI. For instance, it creates user-delegates and checks the security credentials of service-delegates and resource-delegates once both arrive from their original host. It should be noted that the security of delegates is beyond the scope of this chapter. However, the security of the services that run on computing resources is discussed in Maamar, Hamdi, Mansoor, and Bhati (2003).

Rationale of User/Resource/Service-Delegates

- Because mobile devices are resource-constrained, several authors, such as Jailani, Othman, and Latih (2002) and Messer, Greeberg, Bernadat, and Milojicic (2002), observe that it is appropriate to offload computing from mobile devices to fixed ones. In SASC, once the service-broker-agent matches users' needs to providers' services, the next step for a user-delegate is to integrate the component services into a composite service. This integration requires resources that have to be used in a *rationale* way when it comes to mobile devices. Therefore, it is preferable to undertake the development of composite services in the MI rather than in mobile devices. In addition, it may happen that the user-delegate needs further information from a broker to complete its work on a composite service. Since the user-delegate already resides in the MI, it locally interacts with the broker. This constitutes another argument in favor of using user-delegates. Because of the advantages that local interactions offer, even users of fixed devices are encouraged to develop their composite services in the meeting infrastructure.
- When there is a match between the needs of a user and the offers of services, the service-broker-agent locally notifies the user-delegate and remotely notifies the relevant service-provider-agents. Since remote exchanges are subject to obstacles (for example, network reliability, transfer safety), providers of services are associated with service-delegates. Service-delegates are transferred from the sites of their respective provider-service-agents to the MI. After the first match is over, the service-delegate informs the resource-broker-agent about its needs of resources; certain services have been selected and, thus, need to be executed. Once the resources are identified, the service-delegate remotely interacts with the resource-provider-agents about the modalities of using their resource. Similarly to service-delegates, it is more convenient if the interactions between service-delegates and resource-provider-agents occur locally. Therefore, resource-provider-agents have resource-delegates to act on their behalf in the meeting infrastructure.

Operation of SASC

The operation of SASC consists of six stages: initialization, advertisement, search for services, search for resources, refinement, and completion. Below is a summary of the main actions that occur in each stage.

Initialization stage:

- Agentify users and providers.

- Create supervisor and brokers and deploy them in the meeting infrastructure.
- Embody agents with operation mechanisms.

Advertisement stage:
- Create service-delegates and resource-delegates.
- Transfer delegates to the meeting infrastructure.
- Check delegates before they enter the meeting infrastructure.
- Advertise services and resources to brokers.

Search for services stage:
- Create user-delegates in the meeting infrastructure.
- Submit users' needs to user-delegates.
- Interact with service-broker-agent.
- Match user's needs with providers' services.
- If positive match, return list of service-delegates to user-delegate.

Search for resources stage:
- Interact with resource-broker-agent.
- Match selected service with providers' resources.
- If positive match, return list of resource-delegates to service-delegates.
- Select a specific resource-delegate for a service.
- Transfer service for execution to resource-delegate site.

Refinement stage:
- Combine outcomes of search for services and search for resources stages.
- Submit new details (version and processing type) on service to user-delegate.
- Finalize selection of service-delegate by user-delegate.

Completion stage:
- Work on next service based on details of previous service.
- Select a specific resource-delegate for a service.
- Transfer service for execution to resource-delegate site.
- Submit new details (version and processing type) on service to user-delegate.
- Finalize selection of service-delegate by user-delegate.Keep running completion stage until all services are processed.

The purpose of the initialization stage is to perform the agentification of the components of SASC (that is, provider and user). Each provider/user is associated with an agent that exhibits a behavior in terms of resources to have, services to offer, and needs to satisfy. User-agents and provider-agents are respectively installed on top of users' devices and providers' resources/services. Moreover, further agents (supervisor, service-broker, and resource-broker) are created in the MI. Afterwards, information on "services versus needs" is loaded into the knowledge base of the service-broker-agent (operation done by the administrator of SASC). Likewise, information on "resources versus services" is loaded into the knowledge base of the resource-broker-agent. Finally, the supervisor-agent is embodied with the mechanisms of creating user-delegates as well as verifying and installing service-delegates and resource-delegates.

The purpose of the advertisement stage is to notify the brokers about the available services and resources that are made available to the user community. As a first step, service/resource-provider-agents create service/resource-delegates and transfer them to the MI. Because mobile devices are resource-constrained, the supervisor-agent creates the resource-delegates on behalf of the users of these devices. Mechanisms that embody a service/resource-delegate are several, including how to announce itself to the supervisor-agent,

how to register at the service/resource-broker-agent, and how to notify its respective service/resource-provider-agent. When service/resource-delegates access the meeting infrastructure, they register at the appropriate broker to submit their offers of services/resources. It should be noted that service-delegates have a dual role (Figure 4): (i) as a provider of services when they interact with the service-broker-agent and (ii) as a consumer of resources when they interact with the resource-broker-agent. The purpose of the search for services stage is to look for the services that satisfy a user's needs. On reception of the needs, the supervisor-agent creates a user-delegate to be in charge of user satisfaction. First of all, the user-delegate interacts with the service-broker-agent. The purpose is to identify the services of service-delegates that satisfy the user's needs. In case certain services are identified, the service-broker-agent notifies the user-delegate and the service-delegate of these services. Because service-delegates may have services in common, the user-delegate has to select a particular service-delegate. However, the user-delegate delays its selection until further details on services are provisioned. These details concern the cost, version, and processing type of each service.

The purpose of the search for resources stage is to identify the resources that support the execution of the services (that is, those that have been identified in the search for services stage). In the MI, service-delegates trigger the matching between services and resources. The identification of the resources is conducted service per service. As it will be described, the selection of a resource for any service depends on the version and type of processing (that is, remote processing or local processing) of the direct predecessor service of this service. On receiving the service-delegates' requests, the resource-broker-delegate identifies the appropriate resource-delegates. Since several resource-delegates can support the execution of the same service, a service-delegate has to select a resource-delegate. In SASC, the selection strategy consists of minimizing the cost of running a service on a resource considering the version and type of processing of this service. At this time of the operation of SASC, each service-delegate knows exactly for its service the version and type of processing to offer to the user-delegate.

The purpose of the refinement stage is to improve the outcome of the search for services stage. Since a service-delegate is aware of the version and type of processing of the service it will offer to the user-delegate, the service-delegate prepares a cost for that service. In its offer, the service-delegate includes the cost of running the service on a resource. After it receives all the offers from service-delegates, the user-delegate selects for a service a particular service-delegate. The user-delegate minimizes the cost of getting the service from all the service-delegates. When the user-delegate selects a service-delegate, this service-delegate submits to the resource-provider-agent the following details: (i) the service this resource-provider-agent will receive for processing; (ii) the version of this service; (iii) the user-delegate that will trigger the processing of this service; and (iv) the way this service will be invoked for processing. Completion is the final stage in the operation of SASC. Here, the selection of any service directly depends on the version and type of processing of its direct predecessor service. In addition to the cost criterion that was used in the previous stages, another selection criterion is now included, namely, location. Location criterion aims at gathering the maximum number of services for execution in the same computing site[1]. By computing site, it is meant: location of resource-provider-agents and current location of the user-delegate. By gathering services in the same computing site, the following advantages are obtained: (i) extra moves of the user-delegate to distant sites of resource providers are avoided and (ii) extra remote communication and data exchange messages between user-delegates and resource-provider-agents are avoided, too. Therefore, the location criterion is privileged over the

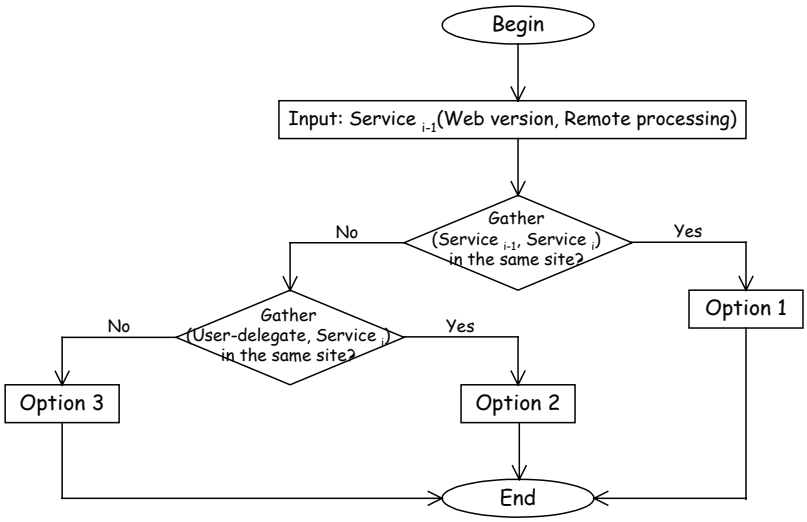
cost criterion. When the details on a service are known, the user-delegate requests from the service-delegates to identify the resource-delegates for the next service. The work on service$_{(i)}$ is decomposed into three cases: Web version and remote processing of service$_{(i-1)}$, Web version and local processing of service$_{(i-1)}$, and M-version and local processing of service$_{(i-1)}$. To keep the chapter self-contained, only the first case is presented.

Web Version and Remote Processing of Service$_{(i-1)}$ Case

Since the processing of the Web version of service$_{(i-1)}$ has been remotely conducted, this means that the user-delegate is in a different site to the execution site of service$_{(i-1)}$. Three exclusive options are offered to the user-delegate to make a decision on service$_{(i)}$ (Figure 6).

- Option 1: the processing of service$_{(i)}$ takes place in the site of service$_{(i-1)}$ in order to comply with the location criterion. Therefore, the user-delegate requests from the service-delegates of service$_{(i)}$ to check with the resource-broker-agent what follows: does resource-delegate$_{(i-1)}$ support the remote processing of the Web version of service$_{(i)}$? If yes, then the service-delegates have to select resource-delegate$_{(i-1)}$. Afterwards, the user-delegate selects a service-delegate based on the cost criterion. As a result, the Web versions of service$_{(i)}$ and service$_{(i-1)}$ will be both installed in site$_{(i-1)}$ of resource-delegate$_{(i-1)}$. The user-delegate will remotely process them.
- Option 2: the processing of service$_{(i)}$ takes place in the site of the user-delegate in order to comply with the location criterion. Option 2 exists because resource-delegate$_{(i-1)}$ does not support the remote processing of the Web version of service$_{(i)}$. Therefore, the user-delegate requests from the service-delegates of service$_{(i)}$ to check with the resource-broker-agent what follows: does the resource-delegate of the current site of the user-delegate support the local processing of the Web version of service$_{(i)}$? If yes, then the service-delegates have to select resource-delegate$_{(i-1)}$. Afterwards, the user-delegate selects a service-delegate

Figure 6. Application of location criterion to service selection



based on the cost criterion. As a result, the Web version of service$_{(i-1)}$ and the Web version of service$_{(i)}$ will be located in different sites. However, the user-delegate will locally process service$_{(i)}$.

- Option 3: the processing of service$_{(i)}$ takes place in any site (different from the site of the user-delegate and the site of service$_{(i-1)}$). Option 3 happens because the site of the user-delegate does not support the local processing of the Web version of service$_{(i)}$. In that case, the location criterion does not hold. Search for services and search for resources stages as previously described are carried out in order to define the version and type of processing of service$_{(i)}$ and the respective resource-delegate.

Summary on SASC

Future computing environments will involve a variety of devices with different capacities in terms of processing power, screen display, input facilities, and network connectivity. Furthermore, a variety of services will be offered to users making the use of these devices important in their performance. In this second part of the chapter, we presented SASC that aims at composing services whether Web services or M-services. The backbone of SASC is a software agent-based architecture that integrates several agents such as user, provider, service, and resource. SASC also aims at provisioning services independently of the location of users and the resources they may be using. Service provisioning has relied on two selection criteria (execution cost and resource location) to identify which resources should be assigned to which services.

CONCLUSION

In this chapter, we presented the research initiatives that are carried out @ Zayed University on Web services and Mobile Web services. Among these initiatives, we cited SAMOS, standing for Software Agents for MObile Services, and SASC, standing for Software Agents for Service Composition. New issues that are related to mobile Web services and their integrations with traditional Web services are raised, varying from low bandwidth and high latency of wireless networks to screen sizes of mobile devices. To deal with these issues, software agents are considered due to their various features. For instance, a software agent is autonomous. Thus, it can make decisions on the user's behalf while this one is disconnected. Second, a software agent can be mobile. Thus, it can move from one host to another. Continuous network connectivity is not needed. The major progress happening in the wireless field will be offering the right mechanisms to users to conduct their daily activities over a variety of mobile devices. Three major factors should boost the penetration and expansion of mobile Web services, namely: personalization, time-sensitivity, and context-awareness.

ACKNOWLEDGMENTS

The author would like to thank the referees for their valuable comments and suggestions of improvements. The author also acknowledges the contributions of Q. H. Mahmoud (Guelph University, Canada) and W. Mansoor (Zayed University, U.A.E.) to SAMOS, and B. Benatallah (University of New South Wales, Australia) and Q. Z. Sheng (University of New South Wales, Australia) to SASC.

REFERENCES

Benatallah, B., Sheng, Q. Z., & Dumas, M. (2003, January/February). The SELF-SERV environment for Web services composition. *IEEE Internet Computing, 7*(1).

Berardi, D., Calvanese, D., De Giacomo, G., Lenzerini, M., & Mecella, M. A. (2003). *Foundational vision for e-services.* Proceedings of the Workshop on Web Services, e-Business, and the Semantic Web (WES'2003) in conjunction with The 15th Conference On Advanced Information Systems Engineering (CAiSE'2003), Klagenfurt/Velden, Austria.

Chakraborty, D., Perich, F., Joshi, A., Finin, T., & Yesha, Y. (2002). *A reactive service composition architecture for pervasive computing environments.* Proceedings of the 7th Personal Wireless Communications Conference (PWC'2002), Singapore.

Chisalita, I., & Shahmehri, N. (2001). *Issues in image utilization with mobile e-services.* Proceedings of the 10th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE'2001), Boston, Massachusetts.

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., & Weerawarana, S. (2002, March/April). Unraveling the Web services web: An introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing, 6*(2).

Jailani, N., Othman, M., & Latih, R. (2002). *Secure agent-based marketplace model for resource and supplier broker.* Proceedings of the 2nd Asian International Mobile Computing Conference (AMOC'2002), Langkawi, Malaysia.

Jennings, N., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, Kluwer Academic Publishers, *1*(1).

Maamar, Z., Dorion, E., & Daigle, C. (2001, December). Towards virtual marketplaces for e-commerce. *Communications of the ACM, 44*(12).

Maamar, Z., Benatallah, B., & Mansoor, W. (2003). *Service chart diagrams - Description & application.* Proceedings of the 12th International World Wide Web Conference (WWW'2003), Budapest, Hungary.

Maamar, Z., Yahyaoui, H., Mansoor, W., & Bhati, A. (2003). *Towards an environment of mobile services: Architecture and security.* Proceedings of the 2003 International Conference on Information Systems and Engineering (ISE'2003), Montreal, Canada.

Maamar, Z., & Mansoor, W. (2003). Design and development of a software agent-based and mobile service-oriented environment. *e-Service Journal, Indiana University Press, 2*(3).

Messer, A., Greeberg, I., Bernadat, P., & Milojicic, D. (2002). *Towards a distributed platform for resource-constrained devices.* Proceedings of the IEEE 22nd International Conference on Distributed Computing Systems (ICDCS'2002), Vienna, Austria.

Milojicic, D., Messer, A., Bernadat, P., Greenberg, I., Fu, G., Spinczyk, O., et al. (2001). *Pervasive services infrastructure* (Tech. Rep. No. HPL-2001-87). HP Laboratories, Palo Alto, CA.

Ninja. (2001). The Ninja project. Retrieved August 15, 2004, from *http://ninja.cs.berkeley.edu*

Noble, B. D., Satyanarayanan, M., Narayanan, D., Tilton, J. E., Flinn, J., & Walker, K. R. (1997). *Agile application-aware adaptation or mobility.* Proceedings of the 16th ACM Symposium on Operating Systems Principles, France.

Wand, A. I., & Chunnian, L. (2001). Process support for mobile work across heterogeneous systems (Tech. Rep.). Norwegian University of Science and Technology, Department of Information Sciences.

Yunos, H. M., Gao, J. Z., & Shim, S. (2003, May). Wireless advertising's challenges and opportunities. *IEEE Computer.*layers: network, service discovery, service composition, service execution, and application.

ENDNOTE

[1] The use of the location criterion is backed by the work of Chakraborty, Perich, Joshi, Finin, and Yesha (2002). In this work, a reactive service composition architecture for pervasive computing environments has been designed. The architecture consists of five layers: network, service discovery, service composition, service execution, and application. We focus on the service execution layer. During the execution of services, this layer might want to optimize the bandwidth required to transfer data over the wireless links between services and, hence, execute the services in an order that minimizes the bandwidth utilization. This optimization is similar to the location criterion. With that criterion, the cross-network traffic between the resources can be reduced, which avoids extra data exchanges between distant resources.

This work was previously published in Service-Oriented Software System Engineering: Challenges and Practices, edited by Z. Stojanovic and A. Dahanayake, pp. 225-246, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 3.12
Handheld Computing and J2ME Programming for Mobile Handheld Devices

Wen-Chen Hu
University of North Dakota, USA

Jyh-haw Yeh
Boise State University, USA

I-Lung Kao
IBM, USA

Yapin Zhong
Shandong Institute of Physical Education and Sport, China

INTRODUCTION

Mobile commerce or m-commerce is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of Internet-enabled mobile handheld devices (Hu, Lee, & Yeh, 2004). It is expected to be the next milestone after electronic commerce blossoming in the late-1990s. Internet-enabled mobile handheld devices are one of the core components of a mobile commerce system, making it possible for mobile users to directly interact with mobile commerce applications. Much of a mobile user's first impression of the application will be formed by his or her interaction with the device, therefore the success of mobile commerce applications is greatly dependent on how easy they are to use. However, programming for handheld devices is never an easy task not only because the programming languages and environments are significantly different from the traditional ones, but also because various languages and operating systems are used by handheld devices and none of them dominates.

This article gives a study of handheld computing, especially J2ME (Java 2 Platform, Micro Edition) programming, for mobile commerce. Various environments/languages are available

for client-side handheld programming. Five of the most popular are (1) BREW, (2) J2ME, (3) Palm OS, (4) Symbian OS, and (v) Windows Mobile. They apply different approaches to accomplishing the development of mobile applications. Three themes of this article are:

1. Introduction of handheld computing, which includes server- and client- side computing.
2. Brief introductions of four kinds of client-side computing.
3. Detailed discussion of J2ME and J2ME programming.

Other important issues such as a handheld computing development cycle will also be discussed.

BACKGROUND

Handheld computing is a fairly new computing area and a formal definition of it is not found yet. Nevertheless, the authors define it as follows: Handheld computing is the programming for handheld devices such as smart cellular phones and PDAs (personal digital assistants). It consists of two kinds of programming: client- and server-side programming.

The definitions of client- and server- side computing are given as follows:

- **Client-Side Handheld Computing:** It is the programming for handheld devices and it does not need the support from server-side programs. Typical applications created by it include (1) address books, (2) video games, (3) note pads, and (4) to-do list.
- **Server-Side Handheld Computing:** It is the programming for wireless mobile handheld devices and it needs the support from server-side programs. Typical applications created by it include (1) instant messages, (2) mobile Web contents, (3) online video games, and (4) wireless telephony.

This article will focus on the client-side computing. The server-side computing is briefly given next.

Server-Side Handheld Computing

Most applications created by this kind of programming, such as instant messaging, require network programming such as TCP/IP programming, which will not be covered in this chapter. The most popular application of server-side handheld computing is database-driven mobile Web sites, whose structure is shown in Figure 1. A database-driven mobile Web site is often implemented by using a `three-tiered client/server architecture` consisting of three layers:

1. **User Interface:** It runs on a handheld device (the client) and uses a standard graphical user interface (GUI).
2. **Functional Module:** This level actually processes data. It may consist of one or more separate modules running on a workstation or application server. This tier may be multi-tiered itself.
3. **Database Management System (DBMS):** A DBMS on a host computer stores the data required by the middle tier.

The three-tier design has many advantages over traditional two- or single- tier design, the chief one being: the added modularity makes it easier to modify or replace one tier without affecting the other tiers.

CLIENT-SIDE HANDHELD COMPUTING

Various environments/languages are available for client-side handheld programming. Five of the most popular are (1) BREW, (2) J2ME, (3) Palm OS, (4) Symbian OS, and (5) Windows Mobile. They apply different approaches to accomplishing

the development of mobile applications. Figure 2 shows a generalized development cycle applied by them and Table 1 gives the comparison among the five languages/environments. The second half of this article is devoted to J2ME details and brief introductions of the other four are given in this section.

BREW (Binary Runtime Environment for Wireless)

BREW is an application development platform created by Qualcomm Inc. for CDMA-based mobile phones (Qualcomm Inc., 2003). CDMA is a digital wireless telephony transmission technique and its standards used for 2G mobile telephony are the IS-95 standards championed by Qualcomm. BREW is a complete, end-to-end solution for wireless applications development, device configuration, application distribution, and billing and payment. The complete BREW solution includes

- BREW SDK (software development kit) for application developers,
- BREW client software and porting tools for device manufacturers, and
- BREW distribution system (BDS) that is controlled and managed by operators—enabling them to easily get applications from developers to market and coordinate the billing and payment process.

Table 1. A comparison among five handheld-computing languages/environments

	BREW	J2ME	Palm OS	Symbian OS	Windows Mobile
Creator	Qualcomm Inc.	Sun Microsystems Inc.	PalmSource Inc.	Symbian Ltd.	Microsoft Corp.
Language/ Environment	Environment	Language	Environment	Environment	Environment
Market Share (PDA) as of 2004	N/A	N/A	2nd	N/A	1st
Market Share (Smartphone) as of 2005	?	N/A	3rd	1st	2nd
Primary Host Language	C/C++	Java	C/C++	C/C++	C/C++
Target Devices	Phones	PDAs & phones	PDAs	Phones	PDAs & phones

Figure 1. A generalized system structure of a database-driven mobile Web site

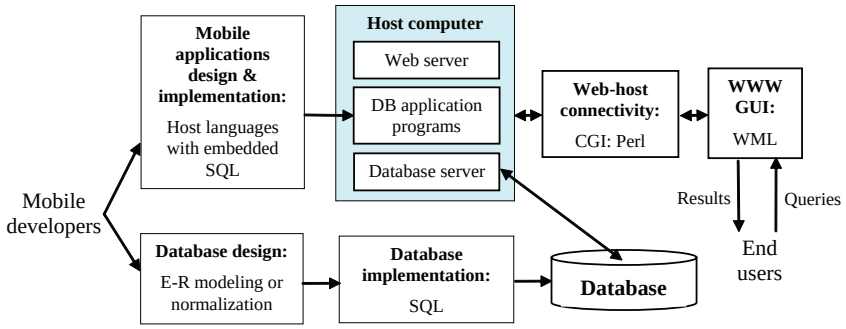
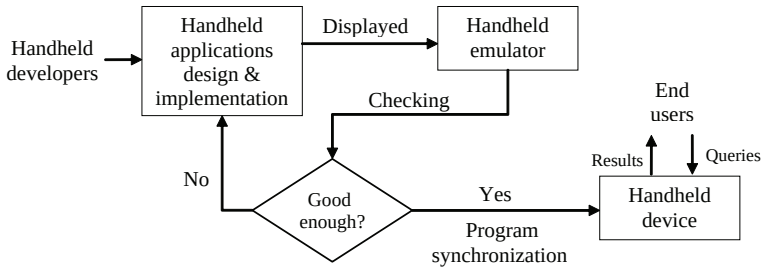


Figure 2. A generalized client-side handheld computing development cycle



Palm OS

Palm OS, developed by Palm Source Inc., is a fully ARM-native, 32-bit operating system running on handheld devices (PalmSource Inc., 2002). Palm OS runs on almost two out of every three PDAs. Its popularity can be attributed to its many advantages, such as its long battery life, support for a wide variety of wireless standards, and the abundant software available. The plain design of the Palm OS has resulted in a long battery life, approximately twice that of its rivals. It supports many important wireless standards, including Bluetooth and 802.11b local wireless and GSM, Mo-bitex, and CDMA wide-area wireless networks. Two major versions of Palm OS are currently under development:

- **Palm OS Garnet:** It is an enhanced version of Palm OS 5 and provides features such as dynamic input area, improved network communication, and support for a broad range of screen resolutions including QVGA.
- **Palm OS Cobalt:** It is Palm OS 6, which focuses on enabling faster and more efficient development of smartphones and integrated wireless (WiFi/Bluetooth) handhelds.

As of August 2005, no hardware products run Palm OS Cobalt and all devices use Palm OS Garnet. Likely as a result of Palm OS Cobalt's lack of adoption, PalmSource has shifted to developing Palm OS Cobalt's APIs on top of a Linux kernel.

Symbian OS

Symbian Ltd. is a software licensing company that develops and supplies the advanced, open, standard operating system—Symbian OS—for data-enabled mobile phones (Symbian Ltd., 2005). It is an independent, for-profit company whose mission is to establish Symbian OS as the world standard for mobile digital data systems, primarily for use in cellular telecoms. Symbian OS includes a multi-tasking multithreaded core, a user interface framework, data services enablers, application engines, integrated PIM functionality, and wireless communications. It is a descendant of EPOC, which is a range of operating systems developed by Psion for handheld devices.

Windows Mobile

Windows Mobile is a compact operating system for mobile devices based on the Microsoft Win32 API (Microsoft Corp., 2005). It is designed to be similar to desktop versions of Windows. In 1996, Microsoft launched Windows CE, a version of the Microsoft Windows operating system designed specially for a variety of embedded products, including handheld devices. However, it was not well received primarily because of battery-hungry hardware and limited functionality, possibly due to the way that Windows CE was adapted for handheld devices from other Microsoft 32-bit desktop operating systems. Windows Mobile includes three major kinds of software:

Figure 3. A screenshot of KToolbar after launching

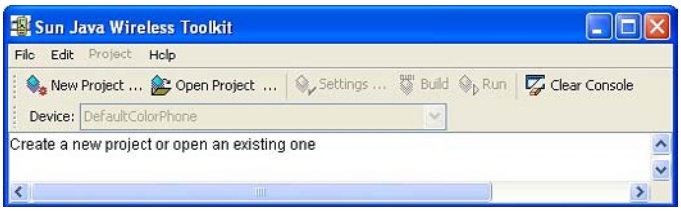


Figure 4. A screenshot of a pop-up window after clicking on the button New Project of KToolbar

- **Pocket PCs:** Pocket PC enables you to store and retrieve e-mail, contacts, appointments, games, exchange text messages with MSN Messenger, browse the Web, and so on.
- **Smartphones:** Smartphone supplies functions of a mobile phone, but also integrates PDA-type functionality, such as e-mails, instant messages, music, and Web surfing, into a voice-centric handset.
- **Portable Media Centers:** Portable media centers let users take recorded TV programs, movies, home videos, music, and photos transferred from Microsoft Windows XP-based PC anywhere.

Windows Mobile-Based Pocket PCs

Pocket PCs were designed with better service for mobile users in mind and offers far more computing power than Windows CE. It provides scaled-down versions of many popular desktop applications, including Microsoft Outlook, Internet Explorer, Word, Excel, Windows Media Player, and others. It also includes three major kinds of software:

- **Pocket PC:** It puts the power of Windows software into a Pocket PC, giving you time to do more with the people and things that matter.
- **Pocket PC Phone Edition:** It combines all the standard functionality of a Windows Mobile-based Pocket PC with that of a feature-rich mobile phone.
- **Ruggedized Pocket PC:** It lets you do more of what matters to you even in the toughest user environments.

Windows Mobile-Based Smartphones

Windows Mobile-based smartphone integrates PDA-type functionality into a voice-centric handset comparable in size to today's mobile phones. It is designed for one-handed operation with keypad access to both voice and data features. The Smartphone is a Windows CE-based cellular phone. Like the Pocket PC, all Smartphones regardless of manufacturer share the same configuration of Windows CE. Also, Smartphones come bundled with a set of applications such as an address book, calendar, and e-mail program.

J2ME (JAVA 2 PLATFORM, MICRO EDITION)

J2ME provides an environment for applications running on consumer devices, such as mobile phones, PDAs, and TV set-top boxes, as well as a broad range of embedded devices (Sun Microsystem Inc., 2002a). Like its counterparts for the enterprise (J2EE), desktop (J2SE) and smart card (Java Card) environments, J2ME includes Java virtual machines and a set of standard Java APIs defined through the Java Community Process, by expert groups whose members include device manufacturers, software vendors, and service providers.

J2ME Architecture

The J2ME architecture comprises a variety of configurations, profiles, and optional packages that implementers and developers can choose from, and combine to construct a complete Java runtime environment that closely fits the requirements of a particular range of devices and a target market. There are two sets of J2ME packages, which target different devices:

- **High-End Devices:** They include connected device configuration (CDC), foundation and personal profile.
- **Entry-Level Devices and Smart Phones:** They include connected limited device configuration (CLDC) and mobile information device profile (MIDP).

Configurations comprise a virtual machine and a minimal set of class libraries and they provide the base functionality for a particular range of devices that share similar characteristics, such as network connectivity and memory footprint. Profiles provide a complete runtime environment for a specific device category.

J2ME Programming

This sub-section gives an example of J2ME programming (Sun Microsystem Inc., 2004). Other client-side handheld programming is similar to this. Figure 3 shows the Sun Java Wireless Toolkit©, which is a toolbox for developing wireless applications that are based on J2ME's CLDC and MIDP. The toolkit includes the emulation environments, performance optimization and tuning features, documentation, and examples that developers need to bring efficient and successful wireless applications to market quickly. The following steps show how to develop an MIDP application, a simple "Hello, World!" program, under Microsoft Windows XP:

1. Download Sun Java Wireless Toolkit 2.3 Beta, which includes a set of tools and utilities and an emulator for creating Java applications that run on handheld devices, at http://java.sun.com/products/sjwtoolkit/download-2_3.html .
2. Run MIDlet, an MIDP application, development environment `KToolbar` as shown in Figure 3 by selecting the following Windows commands:

   `Start ▶ All Programs ▶ Sun Java Wireless Toolkit 2.3 Beta ▶ KToolbar`

3. Create a new project by giving a project name such as `HelloSuite` and a class name such as `HelloMIDlet` as shown in Figure 4. After the project HelloSuite is created, the KToolbar will display the message shown in Figure 5, which tells where to put the Java source files, application resource files, and application library files.
4. Create a J2ME source program and put it in the directory `C:\WTK23\apps\HelloSu-`

`ite\src\`. Figure 6 gives a J2ME example, which displays the text "Hello, World!" and a ticker with a message "Greeting, world."

5. Build the project by clicking on the `Build` button. The `Build` includes compilation and pre-verifying.
6. Run the project by clicking on the `Run` button. An emulator will be popped up and displays the execution results of the built project. For example, Figure 7 shows an emulator displays the execution results of HelloSuite.
7. Upload the application to handheld devices by using USB cables, infrared ports, or Bluetooth wireless technology.

Figure 5. A screenshot of KToolbar after a project HelloSuite created

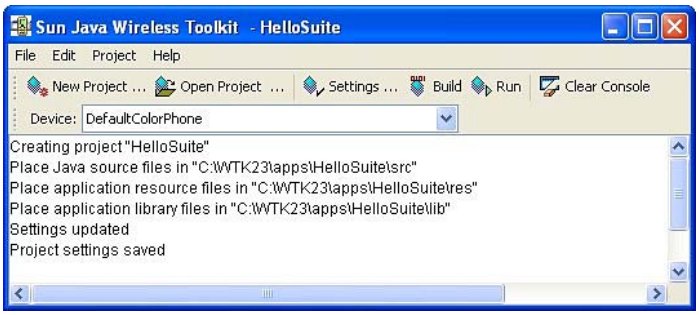


Figure 6. An example of an MIDlet program HelloMIDlet.java

C:\WTK23\apps\HelloSuite\src\HelloMIDlet.java

```
// This package defines MIDP applications and the interactions between
// the application and the environment in which the application runs.
import  javax.microedition.midlet.*;

// This package provides a set of features for user interfaces.
import  javax.microedition.lcdui.*;

public class HelloMIDlet  extends MIDlet  implements CommandListener {

  public void  startApp( ) {
    Display  display     = Display.getDisplay( this );
    Form     mainForm    = new  Form    ( "HelloMIDlet" );
    Ticker   ticker      = new  Ticker  ( "Greeting, World" );
    Command  exitCommand = new  Command( "Exit", Command.EXIT, 0 );

    mainForm.append            ( "\n\n          Hello, World!" );
    mainForm.setTicker         ( ticker );
    mainForm.addCommand        ( exitCommand );
    mainForm.setCommandListener( this );
    display.setCurrent         ( mainForm );
  }

  public void  pauseApp ( ) { }

  public void  destroyApp( boolean unconditional ) {
    notifyDestroyed( );
  }

  public void  commandAction( Command c, Displayable s ) {
    if ( c.getCommandType( ) == Command.EXIT )
      notifyDestroyed( );
  }
}
```

Figure 7. A screenshot of an emulator displaying the execution results of HelloSuite

Mobile Information Device Profile (MIDP) Packages

Table 2 shows the packages provided by the MIDP (Sun Microsystem Inc., 2002b). The packages javax.* are the extensions to standard Java packages. They are not included in the JDK or JRE. They must be downloaded separately.

FUTURE TRENDS

A number of mobile operating systems with small footprints and reduced storage capacity have emerged to support the computing-related functions of mobile devices. For example, Rearch In Motion Ltd.'s BlackBerry 8700 smartphone uses RIM OS and provides Web access, as well as wireless voice, address book, and appointment applications (Research In Motion Ltd., 2005). Because the handheld device is small and has limited power and memory, the mobile OSs' requirements are significantly less than those of desktop OSs.

Table 2. Mobile Information Device Profile (MIDP) package list

Package	Classes and Descriptions
User Interface	**`javax.microedition.lcdui`**: The UI API provides a set of features for implementation of user interfaces for MIDP applications.
	`javax.microedition.lcdui.game`: The Game API package provides a series of classes that enable the development of rich gaming content for wireless devices.
Persistence	**`javax.microedition.rms`**: It provides a mechanism for MIDlets to persistently store data and later retrieve it.
Application Lifecycle	**`javax.microedition.midlet`**: The MIDlet package defines MIDP applications and the interactions between the application and the environment in which the application runs.
Networking	**`javax.microedition.io`**: The MID Profile includes networking support based on the `Generic Connection` framework from the *Connected, Limited Device Configuration*.
Audio	**`javax.microedition.media`**: The MIDP 2.0 Media API is a directly compatible building block of the Mobile Media API (JSR-135) specification.
	`javax.microedition.media.control`: This package defines the specific Control types that can be used with a `Player`.
Public Key	**`javax.microedition.pki`**: Certificates are used to authenticate information for secure Connections.
Core	**`java.io`**: Provides classes for input and output through data streams.
	`java.lang`: MID Profile Language Classes included from Java 2 Standard Edition.
	`java.util`: MID Profile Utility Classes included from Java 2 Standard Edition.

Although a wide range of mobile handheld devices are available in the market, the operating systems, the hub of the devices, are dominated by just few major organizations. The following two lists show the operating systems used in the top brands of smart cellular phones and PDAs in descending order of market share:

- **Smart Cellular Phones:** Symbian OS, Microsoft Smartphone, Palm OS, Linux, and RIM OS (Symbian Ltd., n.d.).
- **PDAs:** Microsoft Pocket PC, Palm OS, RIM OS, and Linux (WindowsForDevices, 2004).

The market share is changing frequently and claims concerning the share vary enormously. It is almost impossible to predict which will be the ultimate winner in the battle of mobile operating systems.

CONCLUSION

Mobile commerce is a coming milestone after electronic commerce blossoming in the late-1990s. The success of mobile commerce applications is greatly dependent on handheld devices, by which mobile users perform the mobile transactions. Handheld computing is defined as the programming for handheld devices such as smart cellular phones and PDAs. It consists of two kinds of programming: client- and server- side programming. Various environments/languages are available for client-side handheld programming. Five of the most popular are

1. **BREW:** It is created by Qualcomm Inc. for CDMA-based smartphones.
2. **J2ME:** J2ME is an edition of the Java platform that is targeted at small, standalone or connectable consumer and embedded devices.
3. **Palm OS:** It is a fully ARM-native, 32-bit operating system running on handheld devices.
4. **Symbian OS:** Symbian OS is an industry standard operating system for smartphones, a joint venture originally set up by Ericsson, Nokia, and Psion.
5. **Windows Mobile:** Windows Mobile is a compact operating system for handheld devices based on the Microsoft Win32 API. It is a small version of Windows, and features ma ny "pocket" versions of popular Microsoft applications, such as Pocket Word, Excel, Access, PowerPoint, and Internet Explorer.

They apply different approaches to accomplishing the development of handheld applications and it is almost impossible to predict which approaches will dominate the client-side handheld computing in the future, as the Windows to desktop PCs.

REFERENCES

Hu, W.-C., Lee, C.-W., & Yeh, J.-H. (2004). Mobile commerce systems. In Shi Nansi (Ed.), *Mobile Commerce Applications* (pp. 1-23). Hershey, PA: Idea Group Publishing.

Microsoft Corp. (2005). *What's new for developers in Windows Mobile 5.0?* Retrieved August 29, 2005, from http://msdn.microsoft.com/mobility/windowsmobile/howto/documentation/default.aspx?pull=/library/en-us/dnppcgen/html/whatsnew_wm5.asp

PalmSource Inc. (2002). *Why PalmOS?* Retrieved June 23, 2005, from http://www.palmsource.com/palmos/Advantage/index_files/v3_document.htm

Qualcomm Inc. (2003). *BREW and J2ME: A complete wireless solution for operators commit-*

ted to Java. Retrieved February 12, 2005, from http://brew.qualcomm.com/brew/en/img/about/pdf/brew_j2me.pdf

Research In Motion Ltd. (2005). *BlackBerry application control: An overview for application developers*. Retrieved January 05, 2006, from http://www.blackberry.com/knowledgecenterpublic/livelink.exe/fetch/2000/7979/1181821/832210/BlackBerry_Application_Control_Overview_for_Developers.pdf?nodeid=1106734&vernum=0

Sun Microsystem Inc. (2002a). *Java 2 Platform, Micro Edition*. Retrieved January 12, 2006, from http://java.sun.com/j2me/docs/j2me-ds.pdf

Sun Microsystem Inc. (2002b). *Mobile information device profile specification 2.0*. Retrieved October 25, 2005, from http://jcp.org/aboutJava/communityprocess/final/jsr118/

Sun Microsystem Inc. (2004). *J2ME Wireless Toolkit 2.2: User's guide*. Retrieved October 21, 2005, from http://java.sun.com/j2me/docs/wtk2.2/docs/UserGuide.pdf

Symbian Ltd. (2005). *Symbain OS Version 9.2*. Retrieved December 20, 2005, from http://www.symbian.com/technology/symbianOSv9.2_ds_0905.pdf

Symbian Ltd. (n.d.). *Symbian fast facts*. Retrieved January 26, 2005, from http://www.symbian.com/about/fastfaqs.html

Wilson, J. (2005). *What's new for developers in Windows Mobile 5.0*. Retrieved January 14, 2006, from http://msdn.microsoft.com/smartclient/default.aspx?pull=/library/en-us/dnppcgen/html/whatsnew_wm5.asp&print=true

WindowsForDevices.com. (2004). *Windows CE zooms past Palm*. Retrieved August 23, 2005, form http://www.windowsfordevices.com/news/NS6887329036.html

KEY TERMS

Binary Runtime Environment for Wireless (BREW): BREW is an application development platform created by Qualcomm Inc. for CDMA-based mobile phones.

Client-Side Handheld Programming: It is the programming for handheld devices and it does not need the supports from server-side programs. Typical applications created by it include (1) address books, (2) video games, (3) note pads, and (4) to-do list.

Handheld Computing: It is the programming for handheld devices such as smart cellular phones and PDAs (Personal Digital Assistants). It consists of two kinds of programming: client- and server-side programming.

Java 2 Platform, Micro Edition (J2ME): J2ME provides an environment for applications running on consumer devices, such as mobile phones, PDAs, and TV set-top boxes, as well as a broad range of embedded devices.

Palm OS: Palm OS, developed by Palm Source Inc., is a fully ARM-native, 32-bit operating system running on handheld devices.

Server-Side Handheld Programming: It is the programming for wireless mobile handheld devices and it needs the supports from server-side programs. Typical applications created by it include (1) instant messages, (2) mobile Web contents, (3) online video games, and (4) wireless telephony.

Symbian OS: Symbian Ltd. is a software licensing company that develops and supplies the advanced, open, standard operating system—Symbian OS—for data-enabled mobile phones.

Windows Mobile: Windows Mobile is a compact operating system for mobile devices based on the Microsoft Win32 API. It is designed to be similar to desktop versions of Windows.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 302-309, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.13
Tools for Rapidly Prototyping Mobile Interactions

Yang Li
University of Washington, USA

Scott Klemmer
Stanford University, USA

James A. Landay
University of Washington & Intel Research Seattle, USA

ABSTRACT

We introduce informal prototyping tools as an important way to speed up the early-stage design of mobile interactions, by lowering the barrier to entry for designers and by reducing the cost of testing. We use two tools, SUEDE and Topiary, as proofs of concept for informal prototyping tools of mobile interactions. These tools address the early stage design of two important forms of mobile interactions: speech-based and location-enhanced interactions. In particular, we highlight storyboarding and Wizard of Oz (WOz) testing, two commonly used techniques, and discuss how they can be applied to address different domains. We also illustrate using a case study: the iterative design of a location-enhanced application called Place Finder using Topiary. In this chapter we hope to give the reader a sense of what should be considered as well as possible solutions for informal prototyping tools for mobile interactions.

INTRODUCTION

The iterative process of prototyping and testing has become an efficient way for successful user interface design. It is especially crucial to explore a design space in the early design stages before implementing an application (Gould et al., 1985). Informal prototyping tools can speed up an early-stage, iterative design process (Bailey et al., 2001; Klemmer et al., 2000; Landay et al., 2001; Li et al., 2004; Lin et al., 2000). These tools are aimed at lowering the barrier to entry for interaction designers who do not have technical backgrounds, and automatically generating early-stage prototypes that can be tested with end users.

The informal look and feel of these tools and their fluid input techniques, for example using pen sketching (Landay et al., 2001), encourage both designers and end users to focus on high level interaction ideas rather than on design or implementation details (e.g., visual layouts or colors). These details are often better addressed at a later stage. In this chapter, we focus on informal tool support for the early stage design of interactive mobile technologies. In particular, we describe informal prototyping tools that we developed for two increasingly important forms of mobile interaction: speech-based interactions (Klemmer et al., 2000) and location-enhanced interactions (Li et al., 2004).

The first of these two types of interactions, speech-based, works well on mobile phones, the major platform of mobile computing. These devices often have tiny screens and buttons to increase mobility, which makes speech interaction an important alternative. Although the accuracy of speech recognition is an important concern for a successful speech-based UI, the real bottleneck in speech interface design is the lack of basic knowledge about user "performance during computer-based spoken interaction" (Cohen et al., 1995). Many interaction designers who could contribute to this body of knowledge are excluded from speech design by the complexities of the core technologies, the formal representations used for specifying these technologies, and the lack of appropriate design tools to support iterative design (Klemmer et al., 2000). SUEDE (Klemmer et al., 2000) demonstrates how tool support can be used in the early stage design of speech-based user interfaces.

The second of these two types of interactions, location-enhanced, is important because of its implicit nature. While the explicit input channels (e.g., keyboarding or mouse pointing) available on mobile technology are more limited than on the desktop, the bandwidth of implicit input (using contextual information) is greatly expanded on mobile platforms. Mobile technology is more available in our context-rich, everyday lives than traditional desktop computing. One especially promising form of context-aware computing that has begun to see commercialization is location-enhanced computing, applications that leverage one's current location as well as the location of other people, places, and things (Li et al., 2004). For example, mobile phone services allow users to locate friends and family (LOC-AID), provide real-time navigation (InfoGation) and monitor and motivate users toward their fitness goals by using phone-based GPS to measure the user's speed, distance and elevation (BonesInMotion). E911 transmits a mobile phone user's current location when making emergency calls. However, location-enhanced applications are hard to prototype and evaluate. They employ sophisticated technologies such as location tracking and their target environment is mobile and in the field. Topiary (Li et al., 2004) demonstrates how high-level tool support can be provided for lowering the threshold and cost for designers to design and test location-enhanced applications.

Using SUEDE and Topiary as proofs of concept, we highlight two techniques commonly used in informal prototyping tools: storyboarding and Wizard of Oz (WOz) testing. To overcome the technical barrier for design, both SUEDE and Topiary employ a storyboarding-based approach for specifying interaction logic. To allow easy testing of prototypes, both tools employ WOz approaches where a human wizard simulates a sophisticated, nonexistent part of the prototype such as location tracking or speech recognition. To demonstrate how these types of tool can actually help prototype and test mobile technology, we introduce a case study using Topiary to design the Place Finder application.

BACKGROUND

User interface tools have been a central topic in HCI research. An extensive review of user

interface tools can be found in (Myers et al., 2001). A large number of research prototypes and commercial products have been developed for rapid prototyping of user interfaces (Apple, 1987; Bailey et al., 2001; Hartmann et al., 2006; Klemmer et al., 2000; Landay et al., 2001; Li et al., 2004; Lin et al., 2000; Macromedia; MacIntyre et al., 2004).

In particular, informal prototyping tools are aimed at the early stages of a design process, and are used to create early-stage prototypes for testing key design ideas rather than building full-fledged final systems (Landay et al., 2001). They often result in example-based interface mockups that are able to demonstrate exploratory interactive behaviors but ignore other non-exploratory aspects of a desired system. Informal tools have shown great potential to facilitate the early stages of a design process and have been developed for various domains. For example, SILK is a tool for designing graphical user interfaces (Landay et al., 2001) that allows designers to create GUI prototypes by sketching and storyboarding. DENIM (Lin et al., 2000), a tool for the early stage design of Web sites, has become one of the most popular informal prototyping tools (downloaded over 100,000 times since 2000). Informal prototyping tools are often grounded in current practices of designers, e.g., paper prototyping (Rettig, 1994; Snyder, 2003), and lower the barrier to entry by maintaining the affordance of an existing practice. At the same time, informal tools provide extra value by allowing the easy editing and maintenance of a design, and by generating testable prototypes.

MAIN FOCUS OF THE CHAPTER

In our research, two features have emerged as being particularly valuable for rapidly prototyping mobile interactions. The first is storyboarding, which is inspired by traditional paper prototyping where designers draw key interaction flows visually on paper. Storyboarding is enhanced by electronic tool support to create the states and transitions. Many systems have been influenced by Harel's Statecharts model (Harel, 1987). Storyboarding is employed by both SUEDE and

Figure 1. SUEDE allows designers to create example scripts of speech-based interactions (top) and speech UI designs (bottom) by storyboarding.

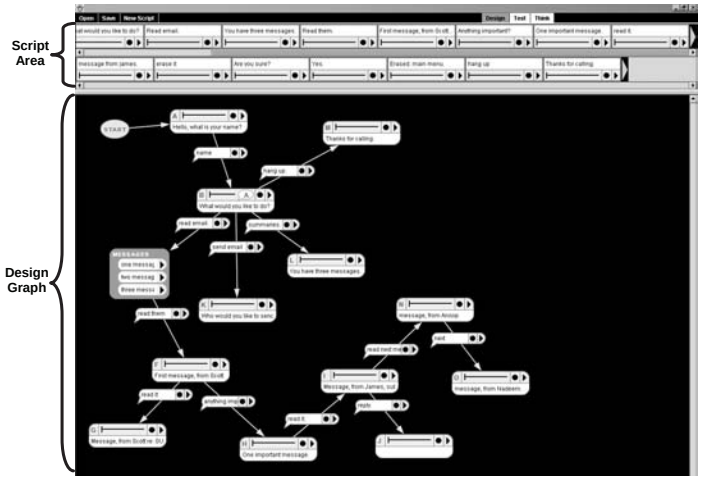


Figure 2. The active map workspace of Topiary is used to model location contexts of people, places and things and to demonstrate scenarios describing location contexts.

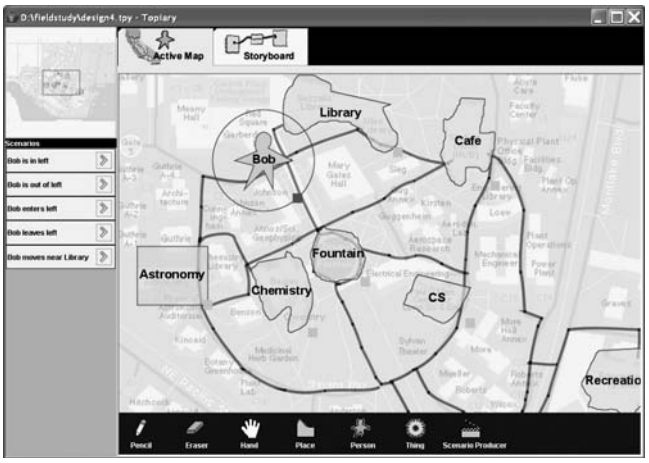
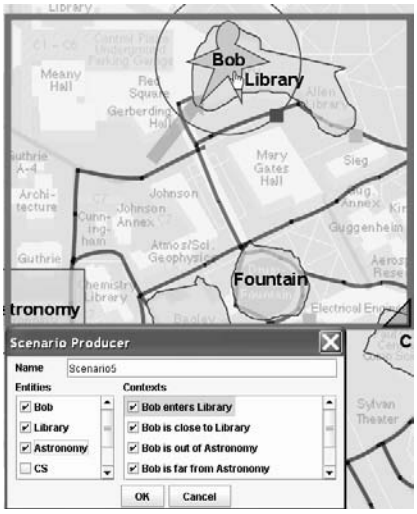


Figure 3. The designer drags Bob into the Library, with the context changing from "Bob is out of Library" to "Bob enters Library." As the entity CS (building) is unchecked, all related contexts to this place are filtered out.



Topiary to lower the technical barrier for creating early-stage prototypes.

The second valuable feature is Wizard of Oz (WOz) testing, where a designer simulates part or all of the application logic by manipulating the interface in response to user input. This significantly reduces the time and labor required to create a testable prototype. As both speech-based interfaces and location-enhanced computing involve a necessary but sophisticated component, that is speech recognition and location tracking, respectively, both SUEDE and Topiary employed a WOz approach to avoid the complexity of introducing these components. To give an example of how this type of tool can help design and evaluate mobile technology in practice, we describe a case study for the iterative design of a PDA-based mobile Place Finder application using the Topiary.

Prototyping with Storyboards

In the early stages of design, it is important that tools allow designers to focus on the high-level concerns of interaction design, rather than forcing designers to also specify how these interactions are implemented. Storyboarding is an efficient way for designers to describe how a user interface should behave by enumerating concrete interaction sequences including both user input and interface output. These sequences should cover the key interaction paths of a proposed system in a particular design space. The concerns of early-stage prototyping are distinct from those of constructing an actual system, which focus more on completeness than exploration of the design space.

SUEDE allows two kinds of storyboarding: linear (conversation examples) and non-linear (design graphs of an actual interface) storyboarding. Designers start a design by creating simple conversation examples (see the Script Area at the top of Figure 1). These examples then evolve into the more complex, graph structure representing the actual interface design (see the design graph at the bottom of Figure 1) (Klemmer et al., 2000). The process of creating linear examples first and then forming more general design graphs is based on the existing practices of speech UI designers:

Figure 4. Topiary's Storyboard workspace allows application prototypes to be created. The lower three links (in blue) are explicit links, representing the behavior of the OK button depending where "Bob" is. The top link (in green) is an implicit link, representing an automatic transition from the Map page to the Nearest Friends page when "Anyone moves near Bob"

we have found that often, designers begin the design process by writing linear dialog examples and then use those as a basis for creating a flowchart representation of the dialog flow on paper.

Designers lay out linear conversation examples horizontally as cards in the script area. *Prompts*, colored orange, represent the system's speech prompts. They are recorded by the designer for the phrases that the computer speaks. *Responses*, colored green, represent example responses of the end user. They are the phrases that participants make in response to prompts. System prompts alternate with user responses for accomplishing a task. A designer can record her own voice for the speech on both types of cards, as well as type in a corresponding label for each of the cards. By playing the recordings from left to right, the designer can both see and hear the example interaction. For example, in Figure 1, a designer has recorded a conversation example with the following alternating prompts and responses: "message from James," "erase it," "Are you sure," "Yes." After constructing example scripts, a designer can construct an actual design of a speech-based interface using the design graph (see Figure 1). A design graph represents a dialog flow based on the user's responses to the system's prompts. To create a design graph, designers can drag prompt or response cards from a script onto the design area, or create new cards on the design area, and link them into the dialog flow. SUEDE's storyboard mechanism embodies both the input and output of a speech interface in cards that can be directly manipulated (e.g., via drag & drop), and hides the complexity of using speech recognition and synthesis. This abstraction allows designers to focus on high-level design issues.

Topiary's storyboards also embed the specification of input and output interactions into a storyboard. Before introducing Topiary's storyboards, we first discuss Topiary's *Activity Map* workspace, a component designed for creating scenarios describing location contexts of people, places and things by demonstration (see Figures 2 and 3). The created scenarios can be used as input by Topiary storyboards when prototyping location-enhanced interactions (see Figure 4). Modeling implicit input, location context in this case, is a new challenge posed by mobile computing.

Topiary's Activity Map workspace employs an intuitive map metaphor for designers to demonstrate location contexts describing the spatial relationship of people, places and things. Designers can create graphical objects on the map to represent people, places and things (see Figure 2). Designers can move people and things on the map to demonstrate various spatial relationships. For example, in Figure 2, Bob is out of the library, the astronomy building and the café. However, Bob is close to the library because Bob's proximity region, indicated by the red circle around Bob, intersects with the library. The proximity region can be resized by dragging the rectangular handle. These spatial relationships can be captured via Topiary's *Scenario Producer*. Like a screen capture tool, a designer can position a *Scenario Producer* window over entities of interest to capture spatial relationships (see Figure 3). A dialog box is then brought up that allows designers to select contexts of interest. Designers can demonstrate dynamic contextual transitions such as "entering" or "leaving" by moving entities within the recording window. For example, dragging Bob into the Library changes the event "Bob is out of Library" into "Bob enters Library" (see Figure 3).

Based on the location scenarios captured in the Active Map workspace, designers can create application prototypes in the Storyboard workspace (see Figure 4). In Topiary, a storyboard page represents a screen of visual output and a link represents transitions between pages. The key innovation in Topiary's storyboards is that scenarios created in the *Active Map* workspace can be used as conditions or triggers on links (Li et al., 2004). Designers create pages and links by sketching. Topiary has two kinds of links (see

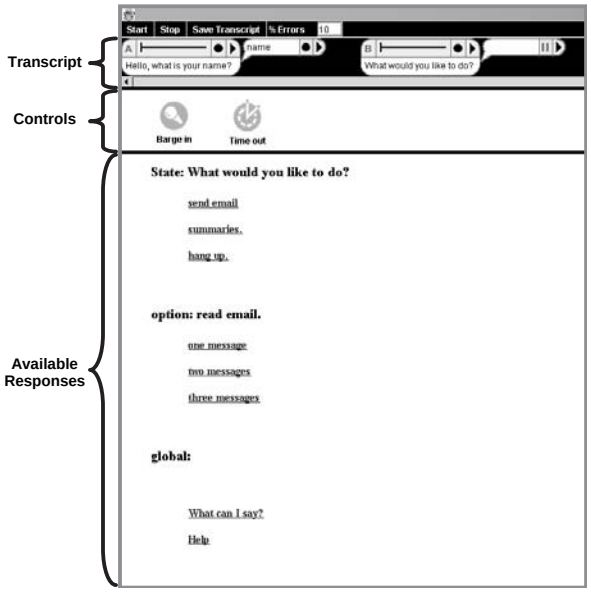
Figure 4). *Explicit links*, denoted in blue, start on ink within a page and they represent GUI elements that users have to click on, for example buttons or hyperlinks. *Implicit links*, denoted in green, start on an empty area in a page. They represent transitions that automatically take place when scenarios associated with that link occur. Explicit links model explicit interactions taken by end-users though they can be conditioned by sensed information, whereas implicit links model purely sensed data such as locations. One or more scenarios can be added to a link and multiple scenarios represent the logical AND of the scenarios. Multiple links starting from the same source represent the logical OR of transitions.

The Activity Map abstraction allows designers to focus on location contexts of interest rather than how these contexts can be sensed. Topiary's graphical storyboarding allows designers to specify rich interactions by drag & drop or sketching instead of specifying complex rules or Boolean logic expressions. From both SUEDE and Topiary, we conclude that the key to a successful informal tool is to devise an appropriate abstraction that matches designers' conceptual model for design and hides the less important aspects of exploring target interactions. Storyboards, as a meta-design metaphor, should be adapted and developed to fit within a specific domain when being applied.

Testing Using WOz Approaches

Speech-based or location-enhanced interactions resist rapid evaluation because the underlying technologies require high levels of technical expertise to understand and use, and a significant amount of effort to tune and integrate. For example, location-tracking infrastructures are not always available (e.g., GPS does not work well indoors) and they require a great deal of effort to deploy and configure. Incorporating these technologies too early in a design process may distract designers from fully exploring the design space. Consequently, we employed WOz approaches in these

Figure 5. SUEDE's Test mode is presented in a web browser and allows the wizard to focus on the current state of the UI (top) and the available responses for that state (bottom).



tools for testing early-stage prototypes. That is, a wizard (played by a designer or experimenter) simulates what these technologies would do in the final application.

Wizard of Oz (WOz) testing has been widely employed in user interface design. In a WOz test, a wizard (often played by a designer) fakes an incomplete (or nonexistent) system component to conduct early user evaluations (Dahlbäck et al., 1993). In its most basic form, a WOz test works by the wizard simulating the machine behavior entirely. There is no computation in the loop at all. Examples of this form include testing paper prototypes by having the wizard physically move around the paper-based windows and menus (Rettig, 1994) and testing potential speech interface interaction flows by having a human operator on the other side of the telephone, following a pre-specified interaction graph. When an interactive prototype has been created (at least partially), the wizard can simply use the implemented interface. As a variant of this approach, a programmer can implement a functionally complete but suboptimal interface, and have the wizard control this interface during testing as a means of eliciting users' conceptual models of the task for example (Akers, 2006).

Significant gains beyond these basic approaches can be achieved through tools designed explicitly to support a Wizard of Oz approach. The fundamental insight behind a WOz-enabled tool is that the wizard is provided with a distinct user interface from that of the end user, and that the primary goal for this interface is to enable the wizard to rapidly specify to the system what the user's input was. In SUEDE, the interaction flow and audio prompts are specified by the designer ahead of time, and the user's responses to the speech prompts are interpreted by the wizard and specified to the system using a graphical interface that is runtime-generated based on the user's current state within the interaction flow. During a test, a wizard works in front of a computer screen. The participant performs the test away from the wizard, in a space with speakers to hear the system prompts and a microphone hooked up to the computer to record his responses. During the course of the test session, a transcript sequence is generated containing the original system audio output and a recording of the participant's spoken audio input.

When the wizard starts a test session, SUEDE automatically plays the pre-recorded audio from the current prompt. The wizard interface in SUEDE displays hyperlinks that represent the set of possible options for the current state (see Figure 5); the wizard waits for the test participant to respond, and then clicks on the appropriate hyperlink based on the response. Here, the wizard is acting as the speech recognition engine. Additionally, effective wizard interfaces should provide a display of the interaction history (as well as capture this for subsequent analysis); global controls for options generally available in an interface genre but independent of a particular interface or interface state (these globals can be defined by the tool or specified by the designer); and support for simulated recognition errors. This set of functionality enables the wizard to customize the test as she sees fit, handle user input beyond what was originally designed, and test whether the application is designed in such a way that users can understand and recover from "recognition errors."

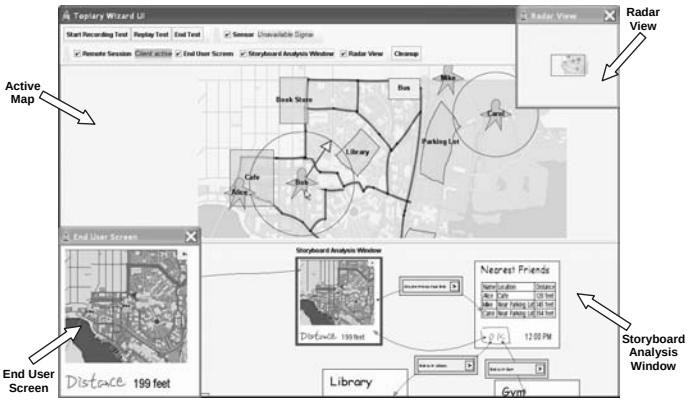
Location-enhanced interfaces introduce the additional challenge that, almost by definition, a test must be conducted while moving to be ecologically valid. To address this, Topiary's WOz interface was specifically designed for a wizard to interact with the interface while walking. Topiary automatically generates user interfaces for testing, including the Wizard UI and the End-User UI, based on the Active Map and the Storyboard workspace. The Wizard UI (see Figure 6) is where a wizard simulates location contexts, as well as observes and analyzes a test. The End-User UI is what an end user interacts with during a test and it is also shown in the End User Screen window

of the Wizard UI (see Figure 6) so that designers can monitor user interactions. The designer can also interact with the End-User Screen window for debugging purposes. The Wizard UI and End-User UI can be run on the same device (to let a designer try out a design) or on separate devices (one for the Wizard, the other for the user).

During a test, the wizard follows a user; each carries a mobile device, and these devices are connected over a wireless network. The wizard simulates location contexts by moving people and things around on the Active Map to dynamically update their location. The location changes of people and things on a map may trigger implicit transitions in the storyboard that will update the End-User UI. Topiary can also employ real location data if it is available, for more realistic testing at larger scales. A designer can choose to turn on a built-in location-tracking engine, based on Place Lab (LaMarca et al., 2005), which allows a WiFi-enabled or GSM-enabled device to passively listen for nearby access points to determine its location in a privacy-sensitive manner. In addition, a designer can analyze a design by recording a test and replaying it later. Topiary capture users' actions, like mouse movements and clicks, as well as physical paths traveled. The Storyboard Analysis window (see the bottom of Figure 6) highlights the current page and the last transition during a test or a replay session, which can help designers to figure out interaction flows.

Through our experience building SUEDE and Topiary, we have learned that effective tool support for Wizard of Oz testing comprises several key elements: the current state of the user interface (e.g., what is the current page in both tools), the current state of the user (e.g., the user's current location in Topiary) and the set of available actions (e.g., available responses in SUEDE). These elements should be provided to the wizard in an effective manner that allows the wizard to easily grasp and rapidly react. An effective Wizard interface should minimize the wizard's cognitive load by proactively maintaining a visible representation of state and having the displayed (and hence selectable) options for future action tailored to the state at hand.

Figure 6. The Wizard UI has four major parts: The Active Map (a clone of the Active Map workspace) for simulating location contexts, the End User Screen for monitoring a user's interaction or debugging a design, the Storyboard Analysis Window for analyzing interaction logic and the Radar View for easy navigation of the Activity Map.



A Case Study

To demonstrate how an informal prototyping tool can help at an early stage of the design process and how informal prototyping can inform the later design or development process, we report on our experience with the iterative design of a location-enhanced Place Finder using Topiary. A location-enhanced Place Finder embodies many features of location-enhanced, mobile applications. It allows users to find a place of interest more efficiently by leveraging the user's location (e.g., showing a path to the destination). With the help of Topiary, we were able to efficiently explore the usability issues of map-based navigation techniques on a PDA held by a user walking in the field. Map-based navigation is a key component of a Place Finder application. Based on two design iterations that involved creating five different designs and testing them with four participants in the field as well as an analysis of implementation issues, we built a high fidelity prototype of the Place Finder.

The first iteration included four different user interface designs that shared the same the underlying map of places and paths in the Active Map workspace (see Figure 2). At each iteration, a user test was conducted in the field on a college campus, using a Toshiba Tablet PC and an HP iPAQ™ Pocket PC. During each test, the wireless communication between the two devices was based on a peer-to-peer connection so that the connection was not affected by the availability of access points in the field.

Iteration #1

It took us only three hours in total to create four prototypes, each using a different navigation technique. The first design shows a map of the entire campus (see Figure 7a). The second design shows an area centered on the user and lets the user manually zoom in and out (see Figure 7b). The third design uses the user's current location to show different regions of the campus (see Figure 7c). The last design is similar to the second, except it automatically zooms in or out based on the user's current speed (see Figure 7d). This last design was based on the idea that people are reluctant to interact with a device while walking. All four designs showed the user's current location and shortest path (see the thick pink lines in Figure 7) to the target, both of which are updated dynamically by Topiary.

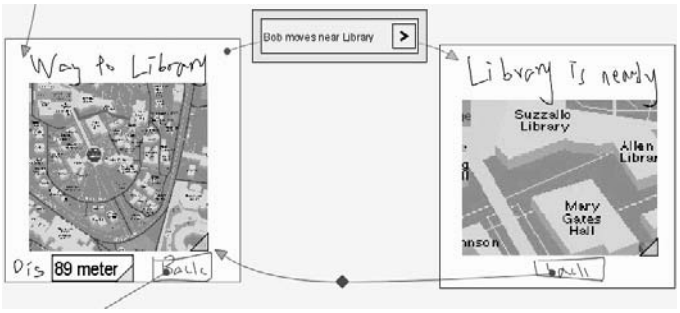
Four navigation segments were included in the test of Iteration #1, one segment for each of the four designs. These four segments were selected based on two principles. First, to smoothly connect the four experimental segments, the target of a segment should be the starting point of the following segment. Second, each segment should cover an area that requires a moderate walk, not too long or too short (e.g., an eight minute walk), and can produce a path with enough complexity to avoid simple paths (e.g., the entire path is a straight line.)

We had three participants try all four designs on a PDA in the field, with a wizard updating their location on a Tablet PC. Each experimental session lasted about one hour and each segment took about fifteen minutes to complete. During the test, we were able to make some minor changes to the design instantly in response to the participant's suggestions.

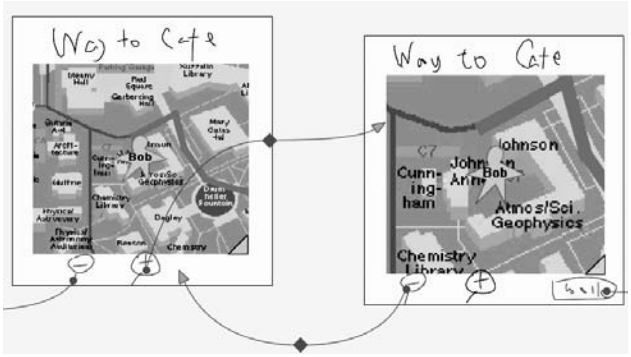
All three participants preferred the map centered on the user's current location (#2 and #4). The problem with the first design is that it shows the entire campus on a small PDA screen, which turned out to be hard to read. The third design does show more detail but it does not give a global view of the campus and the participants complained that they could not see the target until they were physically in that region, although they were still able to see the path.

Two participants preferred manual zooming to automatic zooming as they thought manual zooming gave them more control over the zoom levels. However, the other participant thought both kinds of zooming were good to use. All our participants

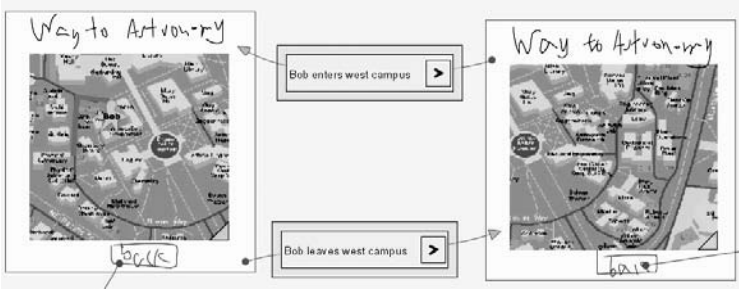
Figure 7. Storyboard fragments of the four designs in Iteration #1. A page, which holds maps and sketches, represents a screen of visual output of the user interface. Arrows (links) between pages represent transitions. The blue links represent GUI elements such as buttons for which scenarios can be used as conditions (not shown here). The green links represent transitions that can automatically take place when the associated scenarios occur.



(a) Design #1 shows the entire campus and a detailed map is automatically shown when a user gets close to a target. Here the scenario "Bob moves near library" triggers showing a detailed map around the library.

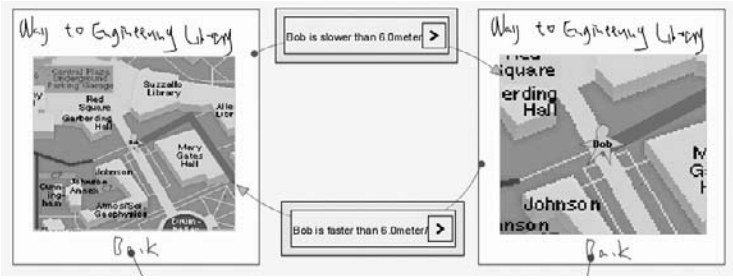


(b) Design #2 shows an area automatically centered on the user and lets the user manually zoom in or out by clicking on the sketched "+" or "–" buttons.



(c) Design #3 uses the user's current location to show different regions of the campus. Here the scenario "Bob enters (or leaves) west campus" triggers showing the west (or east) region of the campus.

continued on following page



(d) Design #4 is similar to Design #2 except it automatically zooms in or out based on the user's current speed. Here the scenario "Bob is slower (or faster) than 0.6 meter/s" triggers showing maps at different zoom levels.

thought the distance label from Design #1 was useful and they also suggested that we should flash the target when users get close to it.

One common problem with the four designs was that there was not enough orientation information provided. We originally thought users could figure out their orientation by referring to nearby buildings and the continuous change of their location on the map.

Iteration #2

Based on participant feedback and our observations during Iteration #1, we spent *one hour* creating a new design combining the best features of the four designs (see Figure 8). We added a page for users to choose automatic or manual zooming (see Figure 8a). We explored different ways of showing orientation on a map, including rotating the map, showing an orientation arrow, and showing trajectory arrows (see Figure 8b). These orientation representations are provided by Topiary. In addition, in response to the participants' request, we added the feature of flashing a target when it is nearby. We tested this new design again with three people[1]. Each test session lasted about half an hour in total. In the middle of the test, we turned on the sensor input that is built into Topiary to see how sensor accuracy affected our participants.

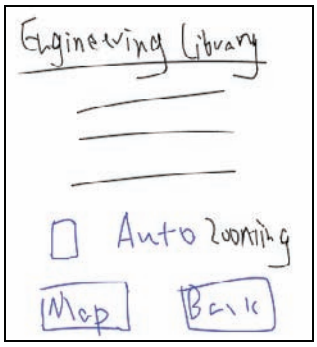
Our participants gave us many useful comments. For example, two of them suggested showing a movement trail to help to indicate orientation. Also, the inaccurate update of the user's location, either by the Wizard or by the sensor input (while it was turned on), did confuse the participants. As a result, one person suggested showing a region for the possible location instead of just a point. They also gave us some other suggestions, such as placing the distance label at the top of the screen rather than at the bottom.

Interestingly, some of our participants did not realize their location was being updated by a wizard rather than by real sensors. It was also observed that the prototype showed an optimal path to a participant who had spent three years on the campus but did not know the existence of this path. We did not know this path either and we simply drew a road network in Topiary by which this path was automatically constructed by the tool.

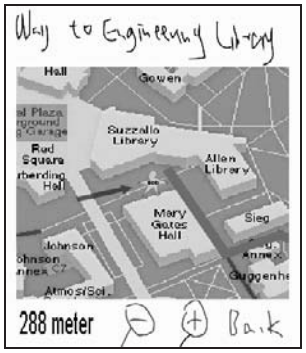
Building a High Fidelity Prototype

Through these two iterations of informal prototyping and testing, we got a rough view of what the Place Finder should be like. Then it was the time to consider implementation issues and to create a high fidelity prototype. Because we did

Figure 8. Two screens (pages) of the new design



(a) A user can select or deselect the checkbox to choose automatic or manual zooming.



(b) A map screen with zooming buttons and a trajectory arrow

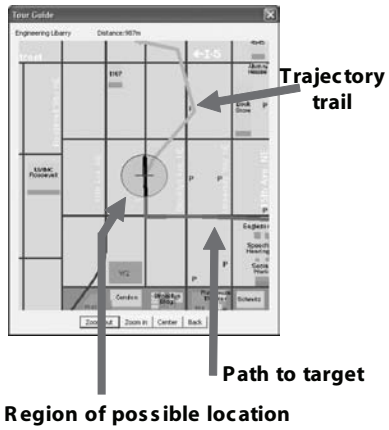
not want to add an extra device, like GPS, for the Place Finder PDA, we chose to use Place Lab for location sensing, since it requires only WiFi. However, Place Lab, like GPS, cannot provide precise orientation. As a result, we decided to show a movement trail (feedback from the earlier study) instead of showing potentially inaccurate directional arrows or employing map rotation. In addition, because the movement speed cannot be accurately measured, we cut the automatic zooming feature, although one participant showed interest in it. This also helped improve application performance on the PDA.

Based on the earlier tests and an analysis of the implementation issues, we built a high fidelity prototype in Java, using the IBM J9 SWT Java toolkit, in about *two weeks* (see Figure 9). We have used this prototype in the field for hours and it has helped us to find places that we had never been to before. We also got positive feedback from people to whom we demoed this prototype. However, performance on the PDA is still a major issue with this prototype and more profiling is necessary before widely testing it.

Lessons Learned

This study offers lessons in two areas. First, it identified usability issues as well as solutions for building map-based navigation techniques in a

Figure 9. The high fidelity prototype was built based on the informal prototypes and an analysis of implementation issues



location-enhanced Place Finder application. Second, the study gives an example of how early stage design iteration can be conducted using informal tools. The study showed obvious advantages over traditional paper prototyping since we were able to test our ideas in the field and leverage those results for later stage development.

Informal prototyping and testing in *hours* was much less expensive than directly building a high fidelity prototype over a period of *weeks* and then testing it with users. The tools allowed us to focus on interaction rather than implementation details. It turned out that little feedback from our participants was related to the informal look of the interface. *Focusing on key interactions* rather than specifying the behaviors of the entire application is important to efficiently conducting early stage design because prototyping tools often employ example-based approaches. In our study, only five places were modeled for testing the five low-fi designs. Once the early usability issues were solved, the design was scaled up to 35 places in the high fidelity prototype.

Carefully testing in the field is important for a successful early stage design because the field is where a mobile application design will be used. Testing in the field requires extra consideration when compared to controlled experiments in a lab setting. The *Wizard of Oz* technique was extremely useful in testing an early stage design since it can reasonably approximate realistic situations. On the other hand, using real sensor input, if not expensive, might help find more usability problems due to the inaccuracy of sensors in a test.

FUTURE TRENDS

Sensors such as accelerometers are becoming available on an increasing number of mobile devices to detect a user's context (e.g., movement, lighting or ambient noise) as well as other peripheral input (e.g., digital compass for the orientation of the device). With these sensors, richer interactions can be constructed. It is important for informal prototyping tools to support interaction design based on the available sensors of the platform. Multimodal interaction that combines multiple interaction modalities has shown promise. Speech-based interaction enhanced by location context is an extremely promising

research direction. By leveraging location context, a system can optimize speech recognition by focusing on phrases that have meaning in a particular context. This brings new research opportunities to the rapid prototyping of mobile technology. The two tools that we discussed in this chapter address speech-based interaction and location-enhanced computing separately. It would be interesting to combine the strengths of these types of tools for prototyping location-enhanced speech user interfaces.

CONCLUSION

Informal prototyping tools play an important role in the early stage design of interactive mobile technology. They lower the threshold for entry and reduce the cost for prototyping and testing. As a proof of concept of informal prototyping tools for mobile interaction, we discussed how SUEDE and Topiary address the design of speech-based interaction and location-enhanced interaction, respectively, the two representative types of interaction for mobile technology. We highlight two common features of these tools: graphical storyboarding and Wizard of Oz testing. To show how these tools can help an iterative design process, we reported on our experience in iteratively prototyping a location-enhanced Place Finder application, and testing its prototypes with real users in the field. The study indicated that this type of tool allowed a designer to effectively explore a design space in the early stages of design. As mobile computing becomes more powerful and prevalent, there will be more opportunities for research on informal prototyping tools for the design and evaluation of interactive mobile technology.

REFERENCES

Akers, D. (2006). CINCH: A cooperatively designed marking interface for 3D pathway selection. In *UIST'06* (pp. 33-42).

Apple (1987). *HyperCard User's Guide*. Apple Computer, Inc.

Bailey, B. P., Konstan, J. A., & Carlis, J. V. (2001). DEMAIS: designing multimedia applications with interactive storyboards. In *ACM Multimedia* (pp. 241-250).

BonesInMotion. BiM Active. Retrieved from http://bonesinmotion.com/corp/

Cohen, P. R., & Oviatt, S. L. (1995). The role of voice input for human-machine communication. In *Proceedings of the National Academy of Sciences, 92*(22), 9921-9927.

Dahlbäck, N., Jönsson, A. & Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. In *Intelligent User Interfaces '93* (pp. 193-200).

Gould, J. D., & Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM, 28*(3), 300-311.

Harel, D. (1987). Statecharts: A visual formalism for complex systems. *Science of Computer Programming, 8*(3), 231-274.

Hartmann, B., Klemmer, S. R., Bernstein, M., Abdulla, L., Burr, B., Robinson-Mosher, A., & Gee, J. (2006). Reflective physical prototyping through integrated design, test, and analysis. In *UIST'06* (pp. 299-308).

InfoGation. Odyssey Mobile. from http://www.infogation.com/

Klemmer, S. R., Sinha, A. K., Chen, J., Landay, J. A., Aboobaker, N., & Wang, A. (2000). SUEDE: A wizard of oz prototyping tool for speech user interfaces. In *CHI Letters: 2*(2), *UIST'00* (pp. 1-10).

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I. E., Scott, J., Sohn, T., Howard, J., Hughes, J. Potter, F., Tabert, J., Powledge, P, Borriello, G., & Schilit, B. N. (2005). Place lab: Device positioning using radio beacons in the wild. In *Proceedings of Pervasive'05* (pp. 116-133).

Landay, J. A., & Myers, B. A. (2001). Sketching interfaces: Toward more human interface design. *IEEE Computer, 34*(3), 56-64.

Li, Y., Hong, J. I., & Landay, J. A. (2004). Topiary: A tool for prototyping location-enhanced applications. In *CHI Letters: 6*(2), *UIST'04* (pp. 217-226).

Lin, J., Newman, M. W., Hong, J.I., & Landay, J. A. (2000). DENIM: Finding a tighter fit between tools and practice for Web site design. In *CHI Letters: 2*(1), *CHI'00* (pp. 510-517).

LOC-AID. LOC-AID People Service. from http://www.loc-aid.net/people_en.htm

MacIntyre, B., Gandy, M., Dow, S., & Bolter, J. D. (2004). DART: A Toolkit for Rapid Design Exploration of Augmented Reality Experiences. In *CHI Letters: 6*(2), *UIST'04* (pp. 197-206).

Macromedia. Director. from http://www.macromedia.com/software/director/

Myers, B., Hudson, S. E. & Pausch, R. (2001). Past, present and future of user interface software tools. In J. M. Carroll (Ed.), *The new millennium*, (pp. 213-233) New York: ACM Press, Addison-Wesley.

Rettig, M. (1994). Prototyping for tiny fingers. *Communications of the ACM, 37*(4), 21-27.

Snyder, C. (2003). *Paper prototyping: The fast and easy way to design and refine user interfaces*. Morgan Kaufmann.

KEY TERMS

Graphical Storyboarding: A technique that informal prototyping tools often employ for designers to describe how an interface should behave. Like a state transition diagram (STD), it has the concepts of states and transitions. However, in graphical storyboarding these states and transitions represent high level UI components or events rather than the computational elements found in a traditional STD.

Informal Prototyping: A type of user interface prototyping used in the early stages of design in which designers explore a design space by focusing on key interaction ideas rather than visual (e.g., color or alignment) or implementation details. These details are often better considered when creating **hi-fidelity** prototypes at a later stage. Paper prototyping is a representative form of informal prototyping in which designers draw interfaces as well as interaction flows on paper.

Informal UI Prototyping Tools: A type of UI prototyping tool that fluidly supports an informal UI prototyping practice. These tools maintain an "informal" look and feel, use fluid input techniques (e.g., sketching) and can automatically generate testable, interactive prototypes.

Location-Enhanced Applications: Computer applications that leverage the location of people, places and things to provide useful services to users. For example, based on the user's current location, show the nearby restaurants or friends. By using the location context, this type of application reduces explicit input required from a user (such as mouse clicks or typing).

Sketch-Based User Interfaces: A type of user interface in which users interact with a computer system by drawing with a pen. The drawings can be recognized and interpreted as commands, parameters or raw digital ink. This type of interface has shown promise in supporting domains such

as UI design, mechanical design, architectural design and note-taking.

Speech-Based Interfaces: A type of user interface in which the user input is submitted mainly via speech. A computer system responds based on either recognized words or vocal variation of the speech. The interface output is typically auditory (e.g., when it is on a phone) or visual.

User Interface Prototyping: A practice of creating user interface mockups to test some aspects of a target interactive system.

UI Prototyping Tools: Electronic tools supporting a user interface prototyping process.

Wizard of Oz Testing: A technique for testing an incomplete interface mockup, named after the movie the *Wizard of Oz*. In this technique, a wizard (often played by a designer) fakes an incomplete (or nonexistent) system component to conduct early user evaluations, (e.g., a wizard can simulate speech recognition when testing a speech-based interface or location tracking when testing a location-enhanced application).

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 330-345, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.14
Real–Time 3D Design Modelling of Outdoor Structures Using Mobile Augmented Reality Systems

Wayne Piekarski
University of South Australia, Australia

ABSTRACT

This chapter presents a series of new augmented reality user interaction techniques to support the capture and creation of 3D geometry of large outdoor structures. Named construction at a distance, these techniques are based on the action at a distance concepts employed by other virtual environments researchers. These techniques address the problem of AR systems traditionally being consumers of information, rather than being used to create new content. By using information about the user's physical presence along with hand and head gestures, AR systems can be used to capture and create the geometry of objects that are orders of magnitude larger than the user, with no prior information or assistance. While existing scanning techniques can only be used to capture existing physical objects, construction at a distance also allows the creation of new models that exist only in the mind of the user. Using a single AR interface, users can enter geometry and verify its accuracy in real-time. Construction at a distance is a collection of 3D modelling techniques based on the concept of AR working planes, landmark alignment, constructive solid geometry operations, and iterative refinement to form complex shapes. This chapter presents a number of different construction at a distance techniques, and are demonstrated with examples of real objects that have been modelled in the physical world.

INTRODUCTION

Current research in AR applications has focused mainly on obtaining adequate tracking and registration and then developing simple interfaces to present display information to the user (Azuma

et al., 2001). One important problem that has not been fully addressed is the authoring of the content that is displayed to the user. Since most AR systems are being used simply as a visualisation tool, the data is prepared offline with standard editing tools and then transferred to the AR system. While ourselves (Piekarski & Thomas, 2003) and others (Baillot, Brown, & Julier, 2001) have started to investigate outdoor AR modelling, this work is very preliminary and incomplete. Brooks states that one of the still unsolved problems in VR is the creation and capture of 3D geometry (Brooks, 1999), which is also relevant for AR models. To develop content for AR systems, we have developed a number of techniques collectively termed *construction at a distance* (CAAD). These techniques use the AR system itself to capture the 3D geometry of existing structures in the physical world, and create new 3D models of virtual objects that do not yet exist. CAAD makes use of the AR working planes and landmark alignment techniques presented in a previous paper (Piekarski & Thomas, 2004), and builds higher-level operations to perform the capture and creation of 3D models. While some of these CAAD techniques have been presented previously (Piekarski et al., Thomas, 2003), in this chapter, I describe new body-relative plane techniques and expand on previous work with a discussion of how AR working planes are used in the implementation.

The introduction section in this chapter describes the advantages of these modelling techniques over other existing methods. Next, the techniques are described over three sections and how they are implemented on a mobile outdoor AR system. An overview of the user interface that supports these techniques is discussed, followed by a discussion on the use of different viewpoints to support situational awareness. The chapter is then concluded with a discussion of possibilities for collaboration, and how the accuracy of the techniques are affected by various environmental factors.

Supplement Physical Capture L imitations

The purpose of these techniques is not to replace existing object capture methods, such as image-based reconstruction (Debevec, Taylor, & Malik, 1996) or laser scanning. These techniques are highly accurate and can produce excellent results given the proper conditions. However, there are a number of limitations and CAAD provides an alternative to existing techniques in the following ways:

- A human operator is capable of accurately estimating the geometry of planar shapes, even when partially occluded by other objects in the environment. When trees occlude the edges of a building, a human can estimate the layout based on incomplete visual information and a knowledge of the volumetric properties of buildings.
- The eye is a highly accurate input device capable of aligning along the walls of buildings (Cutting & Vishton, 1995; Piekarski et al., 2004). Accurate modelling is still possible when working from a distance and direct access to the object is not available.
- Existing capture techniques (Debevec et al., 1996) have a fixed operation time no matter what the complexity of the scene is, whereas in my methods the human can judge the most appropriate level of detail. In many cases, the user wants to create only simple shapes such as boxes to represent buildings, and so these techniques are ideal for quick operations.
- Existing techniques require the object to already exist so it can be captured, whereas my methods allow the human to specify any geometry desired. My techniques allow the creation of new shapes that do not physically exist and may be used to plan future construction work.

It is important to realise that there are limitations introduced by the resolution and accuracy of the tracking devices used to record the inputs. For example, when using a GPS accurate to 50 centimetres the object size that can be modelled is in the order of metres (such as a car), while using a 1 millimetre magnetic tracker allows much smaller objects (such as a drink can). This research does not attempt to address problems with registration or accuracy of tracking devices, but instead works within the limitations of current technology to provide the best solutions that can be achieved.

Working at a Fixed Scale

A number of VR techniques have been developed for use in modelling applications. These applications traditionally provide tools to create and manipulate objects in a virtual world, and to fly around and perform scaling operations to handle a variety of object sizes. While techniques for action at a distance such as spot lights, selection apertures, and image plane techniques (Pierce et al., 1997) have been developed, these only perform simple selections on existing objects and cannot be used to create new ones due to the lack of generating distance values. Techniques such as flying, worlds in miniature (Stoakley, Conway, & Pausch, 1995), and scaled world grab (Mine, Brooks, & Sequin 1997), can perform the creation of points by bringing the world within arm's reach, but accuracy is affected by the scale. Due to their non-exact freehand input methods, all of these systems are also limited to conceptual modelling tasks and not precision modelling. CAD systems use snapping functions or exact numerical entry to ensure accurate inputs, but require an existing reference to snap to or non-intuitive command-based entry.

Although AR environments share some similar functionality with VR, AR is unique in that it requires registration of the physical and virtual worlds. Flying and scaling operations require the breaking of AR registration and so cannot be used. Scaled world representations force the user to divert their attention from the physical world to perform miniature operations within the hands. Existing VR techniques cannot create models of objects the size of skyscraper buildings without breaking the 1:1 relationship between the user and the virtual world. With CAAD techniques, the scale of the world is fixed and only the user's head position controls the view. The virtual geometry is created using absolute world coordinates and is always registered and verifiable against the physical world in real-time. By using the physical presence of the user as an input device, the body can be directly used to quickly and intuitively control the view rather than relying on a separate input device.

Humans are much more capable of accurately estimating and specifying horizontal and vertical displacements compared to distances (Cutting et al., 1995). By using the AR working planes and landmark alignment techniques described previously (Piekarski et al., 2004), simple 2D input devices can be used to draw points in 3D. An AR working plane can be defined at any time from the body along the direction of view (maximising accuracy with landmark alignment) or relative to an existing object (maintaining the same accuracy as the source object), and the user can then move around to a different angle to draw against this surface. With AR working planes, the user is able to draw points that are at large distances and at locations that are not normally reachable, maintaining a 1:1 relationship between the virtual and physical worlds. The techniques in this chapter require any simple 2D input device with a cursor to draw against the AR working plane, with this particular implementation using a glove with fiducial-marker based tracking.

Iterative Model Refinement

CAAD relies on a set of fundamental operations that by themselves cannot generally model a physical world object. Combining a series of these fundamental operations by making iterative improvements can produce complex shapes how-

ever. As the modelling operation is taking place the user can see how well the virtual and physical objects compare, repeatedly making changes until the desired quality is gained. Constructive solid geometry (CSG) techniques used by CAD systems also rely on this principle to produce highly complicated shapes that would otherwise be difficult to specify. The ability to instantly verify the quality of models against the physical world helps to reduce errors and decrease the total modelling time. The process of iterative refinement for VR modelling is discussed by Brooks (1999), and he recommends that a breadth-first iterative refinement strategy is the most efficient. I use these VR guidelines for the proposed CAAD techniques, and take the refinement process one step further by using the unique ability of AR to compare virtual and physical worlds simultaneously.

Simplified Techniques

Some techniques have been developed previously for the interactive creation of data in virtual environments with no prior information. The CDS system by Bowman can create vertices by projecting a virtual laser beam against the ground plane (Bowman, 1996). By connecting these points together and extruding the 2D outline upwards, full 3D solid objects can be created although they are limited in complexity by being constant across the height axis. Baillot et al. performed the creation of vertices located at the intersection of two virtual laser beams drawn from different locations (Baillot et al., 2001). After defining vertices, these can then be connected together to form other shapes of arbitrary complexity, limited only by the time available to the user. Since these techniques both operate using vertex primitives that are then connected into edges, polygons, and objects, the complexity of this task increases as the number of facets on the object increases. Rather than treating objects as collections of vertices like the previously mentioned work, CAAD mainly operates using surfaces and solid objects, so an object with 10 facets can be modelled in 10 steps rather than as 20 vertices and 30 edges.

DIRECT OBJECT PLACEMENT TECHNIQUES

This section describes techniques involving the direct placement of objects within arm's reach. While not being truly CAAD, these techniques may be used as inputs for other operations. The simplest way to perform modelling is to use prefabricated objects and place them at the feet of the user as they stand in the environment, when commanded by the user. I have termed this technique street furniture, as it can be used to place down objects that commonly occur on the street (Piekarski et al., 2003). Furthermore, using the AR working planes techniques (Piekarski et al., 2003; Thomas 2004) the user is able to translate, scale, and rotate these objects in the AR environment. The street furniture method works well when objects to create are known in advance, and the user can avoid having to model the object each time. While this technique is not at a distance according to our requirements, it is the most basic and simplest operation that can be performed using a mobile outdoor AR computer. It is possible to use direct placement of markers at the feet to specify vertices and extrude the object upwards, but this is not always practical in the physical world because the user cannot stand on top of a building to mark its outline. Later techniques described in this chapter use direct placement for the creation of infinitely sized plane surfaces in the environment.

BODY-RELATIVE PLANE TECHNIQUES

This section describes a series of CAAD techniques based on the user's physical presence in the environment. Using simple head-based pointing,

the geometry of planes originating from the body can be specified, taking advantage of the user's sense of proprioception (Mine et al., 1997). Using CSG techniques, these planes can be used to easily define solid building shapes out of arm's reach. Since many buildings in the physical world can be modelled using planes, the process of modelling can be accelerated compared to the simplistic approach of creating each vertex and edge manually.

Orientation Infinite Planes

Buildings in the physical world tend to approximate collections of angled walls in many cases. A solid convex cube can be formed with the specification of six planes arranged perpendicular to each other and a CSG intersection operator. Instead of specifying these planes numerically, the user can create these same planes in an AR environment by projecting them along the line of sight. By looking along the plane of a wall of a building and aligning the two ends, the user can project an infinite plane along this wall in a similar way to AR working planes. Each plane defines a half space that when combined with a CSG intersect operation will form a finite solid shape.

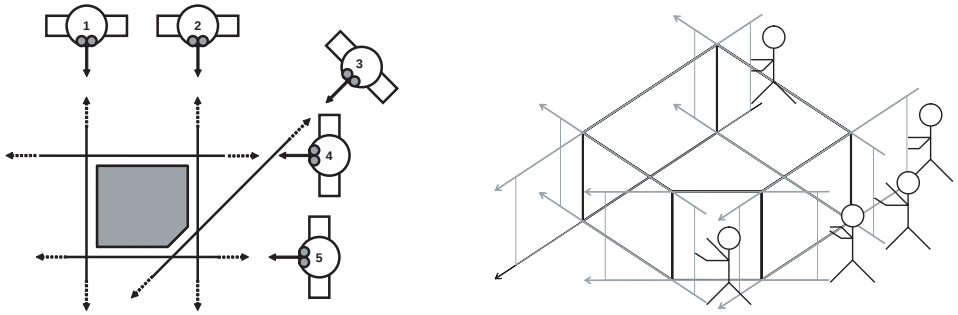
Figure 1 depicts a five-sided building and the location of the mobile AR user as they are sighting down each of the walls, showing the infinite volume being iteratively bound by the infinite planes. At the beginning of the operation, the AR system creates an (approximately) infinite solid volume that will be used for carving. When the user is aligned with a wall, they project an infinitely long vertical plane along the view direction into the world. This plane divides the previous infinite solid into two parts and the left or right portion (decided by the user) is carved away from the solid and removed. As the user sights along each wall, the solid will be reduced down to an object that is no longer infinite in the X and Y axes. At completion, a floor is automatically created at ground level, and the roof is left unbounded for carving using other techniques, since it is impractical to sight along the roof of a very tall building. The final 3D shape is stored using absolute world coordinates and reflects the geometry of the physical building.

With this technique, the object can be carved away iteratively and the user receives real-time feedback of the infinite volume being bounded, allowing immediate undo in case of a mistake. Compared to the direct methods described previously, this plane-based technique allows the capture of buildings from a distance without having to actually stand next to or on top of the building. Since the user is in direct control of the modelling process, the positions of occluded surfaces can be estimated using their knowledge of the environment. These features are useful because many existing physical capture methods require a full view of the object, GPS trackers do not work well near large buildings, and standing on top of a building may not be possible or too dangerous. This technique is also much more efficient than vertex and edge specification since each wall is handled with a single primitive that is easy to create accurately. A limitation of this technique is that using only planes and a CSG intersection to define objects restricts usage to convex buildings with no indentations, and this will be addressed further at the end of this section.

Position Infinite Planes

Another limitation of the orientation infinite planes technique is the dependence on an orientation sensing device for the head. While RTK GPS units may have reliable accuracies in the order of 2 cm, orientation sensors vary in accuracy due to problems with interference and limitations of the technology. These variations affect the placement of planes in the environment and as the distance from the user increases, angular errors cause increasing positional errors, but using techniques that can avoid the use of orientation

Figure 1. Infinite carving planes used to iteratively create a convex shape from an infinite solid



sensing should be able to produce much more accurate results.

In order to take advantage of the stability of position tracking, the orientation infinite planes technique described earlier can be modified to use two or more position points to specify orientation, making it invariant to errors in orientation tracking devices. Using the same landmark alignment concept discussed previously, the user can accurately sight along a wall and mark a position. To indicate direction, the user walks closer while maintaining their alignment and marks a second point. These two points can then be used to project an infinite carving plane. By increasing the spacing of the marker points or using a line of best fit amongst multiple points, the accuracy of this technique can be further improved.

The accuracy of this technique can be calculated based on the positional error of the GPS and the distance between the two marker points. To make this technique useful, it must have an accuracy that is better than is available using traditional orientation sensors. As an example, when a maximum allowable error of 1 degree is assumed, an RTK GPS unit with 2 cm accuracy will require a distance of 1.1 metres between the points. If 10 or more metres is used then the orientation accuracy will be orders of magnitude better than previously possible.

Fixed Infinite Planes

This technique is similar to the position infinite planes technique in that it is invariant to orientation sensing errors. The previous technique required the user to specify the orientation for each plane by using two points, but if the angles at each corner are known to be equal then only one orientation is needed and the others can be calculated automatically. The user creates the first plane using the same method described previously, but for each additional plane, only one position marker is recorded. Based on the number of positions marked, the system knows the number of walls to create and calculates the orientation for each position point based on the first plane. This technique uses nearly half the number of points and yet produces the same accuracy if the first plane is properly placed and the building meets the required properties.

CSG Operations

Many objects in the physical world are not the same shape as simple boxes, cylinders, spheres, and cones. While it may seem that many objects are too complicated to model, they may usually be described in terms of combinations of other objects. For example, the process of defining a cube with a hole using vertices is time consuming,

but can be easily specified with a CSG operation. CSG is a technique commonly used by CAD systems, supporting Boolean set operations such as inversion, union, intersection, and subtraction. The manufacture of objects in the physical world is also performed in a similar manner—a drill can be used to bore a hole very easily out of a solid cube. An example of CSG being used outdoors is applying the CSG difference operator to subtract cubes from a building shape. This could be used when the user needs to carve out indented windows. In Figure 2 part 1a and 2a, the user places a cube at a distance, and then drags it sideways until it enters the building shape. Alternatively, in Figure 2 part 1b and 2b, the cube is pushed into the surface of the building (similar to a cookie cutter), and requiring closer access to the building. As the cube is being positioned by the user, the CSG difference operator is interactively calculated and displayed to the user. Infinite planes are normally limited to producing only convex shapes, but using CSG techniques allows us to produce more complex concave shapes very intuitively.

AR WORKING PLANES TECHNIQUES

This section describes a series of CAAD techniques based on AR working planes (Piekarski & Thomas, 2004). The previous techniques are capable of placing prefabricated objects and capturing bounding boxes for large objects, but detailed modelling is not provided. Using AR working planes and a 2D input device, the user can specify much more intricate details to create realistic 3D models.

Projection Carving

The projection carving technique modifies existing objects by projecting points against surfaces and then cutting away extrusions to produce new highly concave shapes. This technique provides the ability to construct features such as zig-zag roofs and holes that are difficult or impossible to model using previously described techniques. Figure 3 depicts an example of how this technique can be used to carve two peaked roofs onto a building model. These building models may have been created using infinite planes and projection carving can be used to restrict the infinite roof to a finite volume. The AR working plane is created relative to a polygon that has been selected by the user. The object that contains the polygon is then used as the input for the upcoming carving operation. The user then creates vertices along the surface of the AR working plane and these are connected together to form a 2D concave outline. This outline is then extruded along the surface normal of the working plane and used as an input tool for a CSG difference carving operation.

The projection is performed using orthogonal extrusion from the AR working plane, and is position invariant so points can be entered from any location in front of the polygon. This enables the user to cut a flat roof on a 100 metre high building while standing at ground level and looking up. If the cursor was used to carve the object directly like a laser beam, the system would produce pyramid-shaped extrusions. For some buildings, the user may only desire to create a flat roof or a single slope, and by creating only one point the system will create a horizontal cutting plane, and with two points a diagonal cutting plane is created. More than two points implies the user wishes to cut with an outline and so it must be fully specified as in Figure 3. The CSG operation can be switched from difference to intersect if desired, with the effect being that the user can cut holes or split an object into separate parts instead of carving the outside. Used in this form, orthogonal extrusion is limited to carving operations that can be seen in a silhouette representation—other features such as indentations that are not visible from the side can not be captured with this technique. Some of these limitations can be overcome by limiting the depth of the extrusion used for carving. By using a small fixed value or controlling it by moving the

Figure 2. Box objects can be moved into a building surface to carve out windows

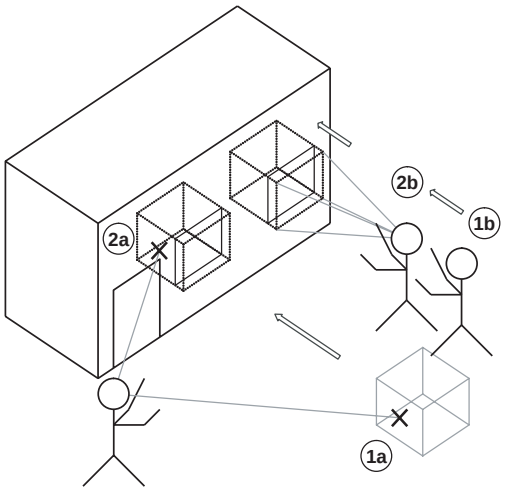
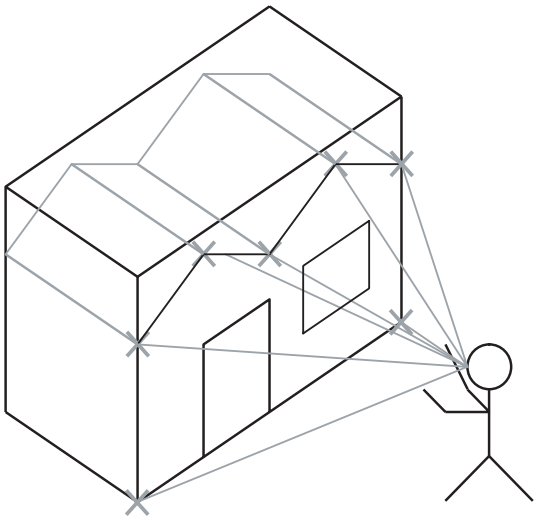


Figure 3. AR working planes are used to specify vertices and are projected along



body forward or backward, the extrusion can be controlled by the user and used for features such as windows or doors.

Figure 4 depicts the projection carving operation on a box that has been edited to match the shape of a building with a pitched roof. A second example demonstrating this technique is a small automobile being modelled outdoors in Figure 5. In both cases, a larger volume is placed down and the user then intersects points against the box surface to define the silhouette. Each frame in Figure 5 shows the process of specifying the solid region that approximates the car in the physical world. The object can then be carved along any of the other faces to further refine the model until it suits the user's requirements.

Projection Colouring

Once a building has been created, the user may desire to place windows, doors, and other extra details onto the model. While it may be possible to draw these details onto a texture map (which cannot be zoomed arbitrarily), or to place extra polygons outside the building to represent these (covering the original building), the building model itself remains untouched. If these new polygons are removed or manipulated, the original solid object remains since the changes are only superficial. A more desirable scenario is that polygons of a different colour are actually cut into the subdivided surface of an object, so that if they are deleted it is possible to see features inside the object that were previously concealed. I have named this technique projection colouring, and using similar steps as projection carving, vertices are projected against an AR working plane created relative to the surface and then connected into an outline. Instead of carving away the outline, the surface is subdivided and the colour of the outlined polygon is modified. The newly coloured polygons may then be deleted or manipulated freely by the user if desired. For example, a window and door can be cut out, with the door then openable using a rotation. Individual manipulation would not be possible with only the surface texture being modified.

Surface of Revolution

When working outdoors and modelling natural features such as trees and artificial features such as fountains, box-shaped objects are usually poor approximations to use. In an attempt to model these objects, surface of revolution techniques (as used in many desktop CAD systems) have been used to capture geometry that is rotated about an axis. The user starts by creating an AR working plane in the environment, with the most intuitive way being to sight toward the central trunk of the tree and project the AR working plane along the view direction. The user then projects vertices onto the AR working plane, defining one-half of the outline of the object. After specifying the vertices along the axis of rotation, the system generates a solid object by rotating the outline around the axis. Figure 6 shows an example where the vertices of a tree have been specified with a preview shape generated, along with the final shape from an external VR view. This technique generates good results when modelling natural objects such as pine trees that are highly symmetrical about the trunk. For trees that grow with deformities and other non-symmetrical features this technique may not generate suitable approximations. To improve the approximation, previously described carving techniques may be applied to refine the model until the user is satisfied with the object.

Texture Map Capture

When implementing live AR video overlay, the system can automatically match up images from the camera to polygons in the scene. Captured models are normally only presented using a single colour and texture maps increase the realism for users without having to add extra polygons for detail. To perform texture map capture, the user stands at a location where the texture for an object's polygon is clearly visible to the camera. The user selects the polygon to activate capture mode and the system projects the polygon vertices onto the AR video overlay to map the still image as a texture. The user repeats this operation for each polygon until the object is completely textured.

The best results for this technique are obtained when the object is fully visible and fills as much of the HMD as possible. Also, keeping the surface perpendicular to the user's viewing direction ensures that the texture is distorted as little as possible. Although techniques for capturing textures of 3D models have been described previously, this has not been performed in a mobile outdoor AR environment. Previously discussed work by Debevec et al. implemented

Figure 4. AR views of an infinite planes building with sloped roof being interactively carved

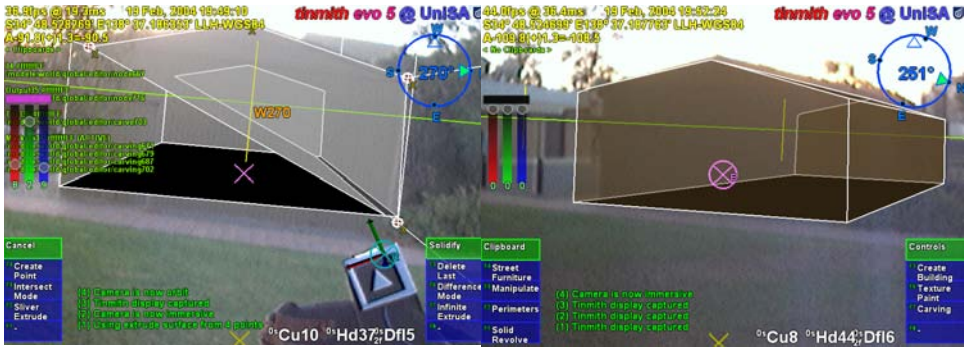
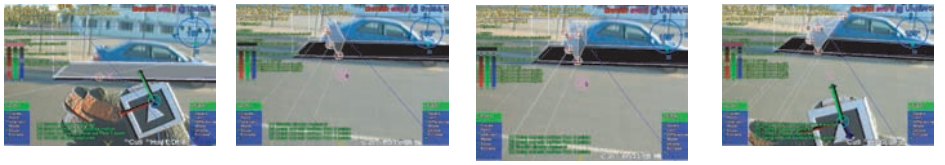


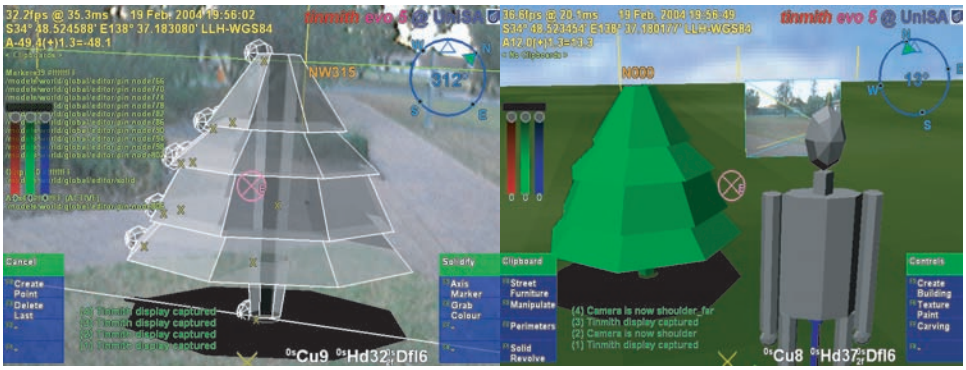
Figure 5. AR frames of an automobile being carving from a box, with markers placed at each corner indicating the silhouette of the object



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 6. (a) AR view of surface of revolution tree with markers on AR working plane (b) VR view of the final surface of revolution tree model as a solid shape



the capture of 3D models from photographs and extracted textures for each facet (Debevec et al., 1996). Lee, Hirota, and State also implemented the capture of textures in AR but with surfaces being modelled within arm's reach using a wand, with the system automatically capturing textures when video frames were deemed suitable (Lee et al., 2001). The video stream used with mobile AR suffers from problems with motion blur and tracker registration, and having the user choose the moment to capture the texture generates the highest quality models.

USER INTERFACE

The user interface is made up of three components: a pointer based on the tracking of the thumbs with a set of gloves worn by the user, a command entry system where the user's fingers interact with a menu for performing actions, and an AR display that presents information back to the user. The display for the interface is fixed to the HMD's screen and presents up to ten possible commands as menu options at any one time. Eight of these commands are mapped to the fingers as depicted in Figure 7, and the user activates a command by pressing the appropriate finger against the thumb. When an option is selected, the menu refreshes with the next set of options that are available. Ok and cancel operations are activated by pressing the fingers into the palm of the appropriate hand and are indicated in the topmost boxes of the menu. The interaction cursors are specified using fiducial markers placed on the tips of the thumbs, as shown in Figure 4 and Figure 5. With the use of vision tracking for cursor position, and metallic pads for finger press detection, it is possible to control the user interface in the harsh environmental conditions experienced outdoors. As discussed previously, a 2D input device is required to specify 3D points using the AR working planes techniques. This user interface provides the necessary 2D cursor to support this, as well as the command entry capability to control the various techniques described in this chapter.

EXTERNAL VIEWPOINTS

Our user interface is typically operated in an immersive mode where virtual objects are registered with the physical world. This view is intuitive because it is similar to how the user normally experiences the physical world. This view may cause problems in situations where very large objects such as buildings may exceed the field of view of the display, objects may be too distant to clearly view, or other objects may be occluding an object of interest. The immersive view restricts the user if it is impractical or impossible for the user to move to a new viewing position. In these cases, it is more useful to work in an external VR style view such as orbital view (Koller, Mine, & Hudson, 1996), where the user sees the virtual world from an external perspective. The advantages of external views are also discussed by Brooks, who mentions that users find a local map of the world useful to show where they are, what they are looking at, and to build a mental model for navigation (Brooks, 1988).

In the external views included in this chapter, the ground and sky are both rendered using texture maps so that the user understands they are looking at a fully virtual view and are no longer immersive. Since the external view is designed to be used while wearing the HMD outdoors, the body of the user is shown in the centre of the display and motion about the physical world is updated in real-time. Top down views, such as that shown in Figure 8(a), provide an aerial perspective of the area, with the display being fixed in north up mode or rotating freely according to the user's current view direction. In this example, an aerial photograph has been used instead of a grass texture to provide additional situational awareness. Orbital views such as that shown in Figure 6(b) and Figure 8(b) link all 3DOFs of head rotation

Figure 7. Each finger maps to a displayed menu option, the user selects one by pressing the appropriate finger against the thumb

to orbiting motions at a fixed distance around the user. These external views are generally only used while stationary because the physical world is completely blocked out and the user may not be able to safely move. The user is able to adjust the camera view using body position or head rotation (as discussed previously) but not hand gestures, and can freely switch between immersive and a number of pre-programmed external views using menu commands.

COLLABORATIVE MODELLING SCENARIO

While a number of techniques can perform the modelling of simple and useful shapes, the true power of CAAD is expressed when techniques are iteratively combined to produce more complicated real-world shapes. Furthermore, the usefulness of a system is enhanced when models can be collaboratively viewed by others at the same time. Customers, architects, and developers both onsite and at remote locations could work together to design buildings and landscapes. A user with a mobile AR system would walk outside to an empty piece of land to create a landscape to preview and perhaps construct in the future. The user creates the outline of the building using infinite planes and then carves out the roof of the building. Doors and windows are then added to the surface of the object. To finish off the model, the outline of a swimming pool can be added, and various street furniture accessories such as tables and chairs are added. Within 10-15 minutes, the user has created a simple model that they can iteratively adjust until they are happy with it.

Using the distributed nature of the Tinmith-evo5 software architecture (Piekarski & Thomas, 2003), indoor users can be connected so they can monitor the progress of the modelling operation on large fixed indoor displays. Using wireless 802.11 networks, the state of the remote system is sent indoors along with two-way voice data so the users can discuss the operation in progress. State information includes the full scene graph and tracking information, so the indoor display is able to reproduce any part of the outdoor system's state as required. The only current limitation is that live video is not streamed over the network due to bandwidth limitations. The indoor users could be remote experts such as architects or developers, observing the design that the customer wants and making comments in real-time. At the completion

Figure 8. (a) Top down view with aerial photography to improve situational awareness, (b) orbital view centred on the user showing building under construction

of the design, the indoor users can extract the model as VRML and then convert it into a proper set of building plans for construction.

OPERATIONAL PERFORMANCE

The CAAD techniques rely on the position and orientation sensors for all tracking, and so increasing the accuracy of these devices will produce improved results and affect the minimum model size that can be properly captured. Errors from each sensor have different effects on the captured models since one is measured as a distance and the other as an angle. Systems such as OSGAR (Coelho, MacIntyre, & Julier, 2004) attempt to model these errors for the registration of information, dynamically adjusting the display depending on the sensor errors. However, when rendering the AR display during 3D modelling, results are also affected not only by the errors in the current tracker data, but also those from the capture process. The position sensor used in these examples is a Trimble Ag132 GPS, with an accuracy of better than 50 centimetres and working reliably amongst small buildings and light tree cover. For orientation, an InterSense IntertiaCube2 hybrid magnetic and inertial sensor is used, although the tracking is unreliable when there are magnetic distortions present in the environment or when the user is moving quickly.

When modelling a new object, the accuracy of projection-based techniques is dependant on the user's current location and the direction they are looking. For the highest accuracy, it is desirable to be as close to the object as possible, minimising the distance the projection can stray from the desired direction caused by angular errors in the orientation sensor. When viewing an existing virtual object, the registration errors with the physical world caused by the GPS will be the most accurate when viewed from a distance due to perspective, while standing very close to an object will cause these errors to be more noticeable. For registration errors caused by the InertiaCube2, these remain constant on the display at all distances due to their angular nature.

CONCLUSION

This chapter has presented my novel CAAD techniques, designed to support the capture and creation of 3D models in outdoor environments using AR. CAAD takes advantage of the presence of the user's body, AR working planes, landmark alignment, CSG operations, and iterative refinement to perform modelling tasks with mobile

AR systems. When used in an AR environment, users can capture the geometry of objects that are orders of magnitude larger than themselves without breaking AR registration or having to touch the object directly. These modelling techniques are intuitive and support iterative refinement for detail in areas that require it with AR providing real-time feedback to the user. While existing techniques are available for the capture of physical world objects, these still have limitations and also cannot be used to create models that do not physically exist. The CAAD techniques were field tested using a number of examples to show how they may be applied to real world problems. By discussing insights gained from these examples, I have identified areas for improvement that currently cause accuracy problems.

REFERENCES

Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., & MacIntyre, B. (2001, November). Recent advances in augmented reality. *IEEE Computer Graphics and Applications, 21*(6), 34-47.

Baillot, Y., Brown, D., & Julier, S. (2001, October). Authoring of physical models using mobile computers. In *Proceedings of the 5th International Symposium on Wearable Computers*, Zurich, Switzerland (pp. 39-46).

Bowman, D. (1996). *Conceptual design space: Beyond walk-through to immersive design* (pp. 225-236). New York: John Wiley & Sons.

Brooks, F. P. (1988, May). Grasping reality through illusion: Interactive graphics serving science. In *Proceedings of the Conference on Human Factors in Computing Systems*, Washington, DC (pp. 1-11).

Brooks, F. P. (1999). What's real about virtual reality? *IEEE Computer Graphics and Applications, 19*(6), 16-27.

Coelho, E. M., MacIntyre, B., & Julier, S. J. (2004, October). OSGAR: A scene graph with uncertain transformations. In *Proceedings of the 3rd International Symposium on Mixed and Augmented Reality*, Arlington, VA.

Cutting, J. E., & Vishton, P. M. (1995). *Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth* (pp. 69-117). San Diego, CA: Academic Press.

Debevec, P. E., Taylor, C. J., & Malik, J. (1996, August). Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, New Orleans, LA (pp. 11-20).

Koller, D. R., Mine, M. R., & Hudson, S. E. (1996, November). Head-tracked orbital viewing: An interaction technique for immersive virtual environments. In *Proceedings of the 9th Annual Symposium on User Interface Software and Technology*, Seattle, WA (pp. 81-82).

Lee, J., Hirota, G., & State, A. (2001, March). Modeling real objects using video see-through augmented reality. In *Proceedings of the 2nd International Symposium on Mixed Reality*, Yokohama, Japan (pp. 19-26).

Mine, M., Brooks, F. P., & Sequin, C. H. (1997, August). Moving objects in space: Exploiting proprioception in virtual-environment interaction. In *Proceedings of the ACM SIGGRAPH 1997*, Los Angeles (pp. 19-26).

Piekarski, W., & Thomas, B. H. (2003, May). Interactive augmented reality techniques for construction at a distance of 3D geometry. In *Proceedings of the 7th International Workshop on Immersive Projection Technology / 9th Eurographics Workshop on Virtual Environments*, Zurich, Switzerland.

Piekarski, W., & Thomas, B. H. (2003, October). An object-oriented software architecture for 3D mixed reality applications. In *Proceedings of the 2nd International Symposium on Mixed and Augmented Reality,* Tokyo, Japan.

Piekarski, W., & Thomas, B. H. (2004, October). Augmented reality working planes: A foundation for action and construction at a distance. In *Proceedings of the 3rd International Symposium on Mixed and Augmented Reality,* Arlington, VA.

Pierce, J. S., Forsberg, A., Conway, M. J., Hong, S., Zeleznik, R., & Mine, M. R. (1997, April). Image plane interaction techniques in 3D immersive environments. In *Proceedings of the Symposium on Interactive 3D Graphics,* Providence, RI (pp. 39-43).

Stoakley, R., Conway, M. J., & Pausch, R. (1995, May). Virtual reality on a WIM: Interactive worlds in miniature. In *Proceedings of the Conference on Human Factors in Computing Systems,* Denver, CO (pp. 265-272).

This work was previously published in Emerging Technologies of Augmented Reality: Interfaces and Design, edited by M. Haller, B. Thomas, and M. Billinghurst, pp. 181-197, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 3.15
Mobile Ad Hoc Network

Subhankar Dhar
San Jose State University, USA

INTRODUCTION

A mobile ad hoc network (MANET) is a temporary, self-organizing network of wireless mobile nodes without the support of any existing infrastructure that may be readily available on conventional networks. It allows various devices to form a network in areas where no communication infrastructure exists. Although there are many problems and challenges that need to be solved before the large-scale deployment of an MANET, small and medium-sized MANETs can be easily deployed.

The motivation and development of MANET was mainly triggered by Department of Defense (DoD)-sponsored research work for military applications (Freebersyser and Leiner, 2002). In addition, ad hoc applications for mobile and dynamic environments are also driving the growth of these networks (Illyas, 2003; Perkins, 2002; Toh, 2002). As the number of applications of wireless ad hoc networks grows, the size of the network varies greatly from a network of several mobile computers in a classroom to a network of hundreds of mobile units deployed in a battlefield, for example. The variability in the network size is also true for a particular network over the course of time; a network of a thousand nodes may be split into a number of smaller networks of a few hundred nodes or vice versa as the nodes dynamically move around a deployed area.

Ad hoc networks not only have the traditional problems of wireless communications like power management, security, and bandwidth optimization, but also the lack of any fixed infrastructure, and their multihop nature poses new research problems. For example, routing, topology maintenance, location management, and device discovery, to name a few, are important problems and are still active areas of research (Wu & Stojmenovic, 2004).

Characteristics of MANET

- Mobile: The nodes may not be static in space and time, resulting in a dynamic network topology.
- Wireless: MANET uses a wireless medium to transmit and receive data.
- Distributed: MANET has no centralized control.
- Self-organizing: It is self-organizing in nature.
-

A message from the source node to destination node goes through multiple nodes because of the limited transmission radius.

- Scarce resources: Bandwidth and energy are scarce resources.
- Temporary: MANET is temporary in nature.
- Rapidly deployable: MANET has no base station and, thus, is rapidly deployable.
- Neighborhood awareness: Host connections in MANET are based on geographical distance.

SOME BUSINESS AND COMMERCIAL APPLICATIONS OF MANET

An ad hoc application is a self-organizing application consisting of mobile devices forming a peer-to-peer network where communications are possible because of the proximity of the devices within a physical distance. MANET can be used to form the basic infrastructure for ad hoc applications.

Some typical applications are as follows:

- **Personal-area and home networking:** Ad hoc networks are quite suitable for home as well as personal-area networking (PAN) applications. Mobile devices with Bluetooth or WLAN (wireless local-area network) cards can be easily configured to form an ad hoc network. With Internet connectivity at home, these devices can easily be connected to the Internet. Hence, the use of these kinds of ad hoc networks has practical applications and usability.
- **Emergency services:** When the existing network infrastructure ceases to operate or is damaged due to some kind of disaster, ad hoc networks enables one to build a network and they provide solutions to emergency services.
- **Military applications:** On the battlefield, MANET can be deployed for communications among the soldiers in the field. Different military units are expected to communicate and cooperate with each other within a specified area. In these kinds of low-mobility environments, MANET is used for communications where virtually no network infrastructure is available. For example, a mesh network is an ad hoc peer-to-peer, multihop network with no infrastructure. The important features are its low cost, and nodes that are mobile, self-organized, self-balancing, and self-healing. It is easy to scale. A good example is SLICE (soldier-level integrated communications environment), a research project sponsored by DARPA (Defense Advanced Research Projects Agency) in this area for this need. The idea is that every soldier is equipped with a mobile PC (personal computer) with a headset and a microphone. SLICE is supposed to create mesh networks that handle voice communications while mapping whereabouts of soldiers and their companions.
- **Ubiquitous and embedded computing applications:** With the emergence of new generations of intelligent, portable mobile devices, ubiquitous computing is becoming a reality. As predicted by some researchers (Weiser, 1993), ubiquitous computers will

be around us, always doing some tasks for us without our conscious effort. These machines will also react to changing environments and work accordingly. These mobile devices will form an ad hoc network and gather various localized information, sometimes informing the users automatically.

- **Location-based services:** MANET, when integrated with location-based information, provides useful services. GPS (Global Positioning System), a satellite-based radio navigation system, is a very effective tool to determine the physical location of a device. A mobile host in a MANET, when connected to a GPS receiver, will be able to determine its current physical location. A good example is that a group of tourists using PDAs (personal digital assistants) with wireless LAN cards installed in them along with GPS connectivity can form a MANET. These tourists can then exchange messages and locate each other using this MANET. Also, vehicles on a highway can form an ad hoc network to exchange traffic information.
- **Sensor network:** It is a special kind of hybrid ad hoc network. There is a growing number of practical applications of tiny sensors in various situations. These inexpensive devices, once deployed, can offer accurate information about temperature, detect chemicals and critical environment conditions (e.g., generate wild-fire alarms), monitor certain behavior patterns like the movements of some animals, and so forth. In addition, these devices can also be used for security applications. However, these sensors, once deployed, have limited battery power, and the lifetime of the battery may determine the sensor's lifetime. Recently, several government agencies (e.g., NSF [National Science Foundation]) have funded research projects on sensor networks.

MAC-LAYER PROTOCOLS FOR MANET

An ad hoc network can be implemented very easily using the IEEE 802.11 standard for WLAN. Since the mobile nodes in WLAN use a common transmission medium, the transmissions of the nodes have to be coordinated by the MAC (media-access control) protocol. Here we summarize the MAC-layer protocols.

- **Carrier-sense multiple access (CSMA):** Carrier-sense multiple-access protocols were proposed in the 1970s and have been used in a number of packet radio networks in the past. These protocols attempt to prevent a station from transmitting simultaneously with other stations within its transmitting range by requiring each station to listen to the channel before transmitting. Because of radio hardware characteristics, a station cannot transmit and listen to the channel simultaneously. This is why more improved protocols such as CSMA/CD (collision detection) cannot be used in single-channel radio networks. However, CSMA performs reasonably well except in some circumstances where multiple stations that are within range of the same receivers cannot detect one another's transmissions. This problem is generally called a hidden-terminal problem, which degrades the performance of CSMA significantly as collision cannot be avoided, in this case, making the protocol behave like the pure ALOHA protocol (Fullmer & Garcia-Luna-Aceves, 1995).
- **Multiple access with collision avoidance (MACA):** In 1990, Phil Karn proposed MACA to address the hidden-terminal problem (Karn, 1992). Most hidden-node problems are solved by this approach and collisions are avoided.
- **Multiple access with collision avoidance for wireless LANs (MACAW):** A group of

researchers, in 1994, proposed MACAW to improve the efficiency of MACA by adding a retransmission mechanism to the MAC layer (Bharghavan, Demers, Shenker, & Zhang, 1994).

- **Floor-acquisition multiple access (FAMA):** A general problem of MACA-based protocols was the collision of control packets at the beginning of each transmission as all terminals intending to transmit sends out RTS (request-to-transmit) signals. In 1995, another protocol called FAMA was proposed, which combined CSMA and MACA into one protocol where each terminal senses the channel for a given waiting period before transmitting control signals (Fullmer & Garcia-Luna-Aceves, 1995).
- **Dual-busy-tone multiple access (DBTMA):** Another significant cause of collision in MACA-based protocols is collision between control packets and data transmission. This problem can be solved by introducing separate channels for control messages, which was proposed in the DBTMA protocol published in 1998 (Haas & Deng, 1998).

ROUTING PROTOCOLS FOR MANET

Routing issues for ad hoc networks with different devices having variable parameters leads to many interesting problems, as evidenced in the literature (Das & Bharghavan, 1997; Dhar, Rieck, Pai, & Kim, 2004; Illyas, 2003; Iwata, Chiang, Pei, Gerla, & Chen, 1999; Liang & Haas, 2000; Perkins, Royer, & Das, 1999; Ramanathan & Streenstrup, 1998; Rieck, Pai, & Dhar, 2002; Toh, 2002; Wu & Li, 2001). This is also validated by industry as well as government efforts such as DoD-sponsored MANET work (Freebersyser & Leiner, 2002). A good network routing protocol may be one that yields the best throughput and response time. However, the very nature of ad hoc networks adds to the requirement for a good routing protocol a set of more, often conflicting, requirements. Accordingly, a good ad hoc routing protocol should also be scalable and reliable. Various routing algorithms and protocols have been introduced in recent years.

Wireless devices are often powered by batteries that have a finite amount of energy. In some ad hoc networks such as sensor networks deployed in a hostile zone, it may not be possible to change a battery once it runs out of energy. As a consequence, the conservation of energy is of foremost concern for those networks. A good ad hoc routing protocol should therefore be energy efficient.

Routing protocols can broadly be classified into four major categories: proactive routing, flooding, reactive routing, and dynamic cluster-based routing (McDonald & Znati, 1999). Proactive routing protocols propagate routing information throughout the network at regular time intervals. This routing information is used to determine paths to all possible destinations. This approach generally demands considerable overhead-message traffic as well as routing-information maintenance. In a flooding approach, packets are sent to all destinations (broadcast) with the expectation that they will arrive at their destination at some point in time. While this means there is no need to worry about routing data, it is clear that for large networks, this generates very heavy traffic, resulting in unacceptably poor overall network performance. Reactive routing maintains path information on a demand basis by utilizing a query-response technique. In this case, the total number of destinations to be maintained for routing information is considerably less than flooding and, hence, the network traffic is also reduced. In dynamic cluster-based routing, the network is partitioned into several clusters, and from each cluster, certain nodes are elected to be cluster heads. These cluster heads are responsible for maintaining the knowledge of the topology of the network. As it has already been said, clustering

may be invoked in a hierarchical fashion.

Some of the specific approaches that have gained prominence in recent years are as follows: The dynamic destination-sequenced distance-vector (DSDV) routing protocol (Johnson & Maltz, 1999), wireless routing protocol (WRP; Murthy & Garcia-Luna-Aceves, 1996), cluster-switch gateway routing (CSGR; Chiang, Wu, & Gerla, 1997), and source-tree adaptive routing (STAR; Garcia-Luna-Aceves & Spohn, 1999) are all examples of proactive routing, while ad hoc on-demand distance-vector routing (AODV; Perkins et al., 1999), dynamic source routing (DSR; Broch, Johnson, & Maltz, 1999), temporally ordered routing algorithm (TORA; Park & Corson, 1997), relative-distance microdiversity routing (RDMAR; Aggelou & Tafazolli, 1999), and signal-stability routing (SSR; Ramanathan & Streenstrup, 1998) are examples of reactive routing. Location-aided routing (LAR; Haas & Liang, 1999) uses location information, possibly via GPS, to improve the performance of ad hoc networks, and global state routing (GSR) is discussed in Chen and Gerla (1998). The power-aware routing (PAR) protocol (Singh, Woo, & Raghavendra, 1998) selects routes that have a longer overall battery life. The zone-Routing protocol (ZRP; Haas & Pearlman, 2000) is a hybrid protocol that has the features of reactive and proactive protocols. Hierarchical state routing (Bannerjee & Khuller, 2001) and cluster-based routing (Amis, Prakash, Vuong, & Huynh, 2000) are examples of dynamic cluster-based routing.

FUTURE TRENDS AND CHALLENGES

MANET will continue to grow in terms of capabilities and applications in consumer as well as commercial markets. There are already quite useful applications of MANET in the military. Currently, it is not just an area of academic research, but also plays an important role in business applications for the future. This trend will continue in the future.

The usefulness of MANET also lies in how this technology will be integrated with the Internet and other wireless technologies like Bluetooth, WLAN, and cellular networks. Another important application of MANET will be in the area of sensor networks, where nodes are not as mobile as MANET but have the essential characteristic of MANET. We will continue to see more and more deployment of sensor networks in various places to collect data and enhance security. So, from that perspective, the future of MANET and its growth looks very promising along with its practical applications.

Although a great deal of work has been done, there are still many important challenges that need to be addressed. We summarize the important issues here.

- **Security and reliability:** Ad hoc networks use wireless links to transmit data. This makes MANET very vulnerable to attack. Although there is some work being done on the security issues of MANET, many important problems and challenges still need to be addressed. With the lack of any centralized architecture or authority, it is always difficult to provide security because key management becomes a difficult problem (Perkins 2002). It is also not easy to detect a malicious node in a multihop ad hoc network and to implement denial of service efficiently. Reliable data communications to a group of mobile nodes that continuously change their locations is extremely important, particularly in emergency situations. In addition, in a multicasting scenario, traffic may pass through unprotected routers that can easily get unauthorized access to sensitive information (as in the case with military applications). There are some solutions that are currently available based on encryption, digital signatures, and so forth

in order to achieve authentication and make the MANETs secure, but a great deal of effort is required to achieve a satisfactory level of security. The secure routing protocol (Papadimitratos & Haas, 2002) tries to make MANET more reliable by combating attacks that disrupt the route-discovery process. This protocol will guarantee that the topological information is correct and up to date.

- **Scalability:** Scalability becomes a difficult problem because of the random movement of the nodes along with the limited transmission radius and energy constraints of each node.
- **Quality of service (QoS):** Certain applications require QoS, without which communication will be meaningless. Incorporating QoS in MANET is a nontrivial problem because of the limited bandwidth and energy constraints. The success and future application of MANET will depend on how QoS will be guaranteed in the future.
- **Power management:** Portable handheld devices have limited battery power and often act as nodes in a MANET. They deliver and route packets. Whenever the battery power of a node is depleted, the MANET may cease to operate or may not function efficiently. An important problem is to maximize the lifetime of the network and efficiently route packets.
- **Interoperability:** Integrating MANETs with heterogeneous networks (fixed wireless or wired networks, Internet, etc.) seamlessly is a very important issue. Hosts should be able to migrate from one network to another seamlessly and make pervasive computing a reality.
- **Group membership:** In a MANET, sometimes a new node can join the network, and sometimes some existing nodes may leave the network. This poses a significant challenge for efficient routing management.
- **Mobility:** In MANETs, all the nodes are mobile. Multicasting becomes a difficult problem because the mobility of the nodes creates inefficient multicast trees and an inaccurate configuration of the network topology. In addition, modeling mobility patterns is also an interesting issue. Several researchers have been quite actively investigating this area of research.

CONCLUSION

The growing importance of ad hocs wireless network can hardly be exaggerated as portable wireless devices are now ubiquitous and continue to grow in popularity and in capabilities. In such networks, all of the nodes are mobile, so the infrastructure for message routing must be self-organizing and adaptive. In these networks, routing is an important issue because there is no base station that can be used for broadcasting.

Current and future research will not only address the issues described earlier, but will also try to find new applications of MANET. So far, the research community has been unable to find the killer app using MANET other than in military applications. So, the success of this technology will largely depend on how it will be integrated with the Internet, PANs, and WLANs. MANET will also play an important role in ubiquitous computing, when it will be able to seamlessly integrate with heterogeneous networks and devices, provide various services on demand, and offer secure and reliable communications.

REFERENCES

Aggelou, G., & Tafazolli, R. (1999). RDMAR: A bandwidth-efficient routing protocol for mobile ad hoc networks. *Proceedings of the Second ACM International Workshop on Wireless Mobile Multimedia (WoWMoM),* Seattle, WA.

Amis, A. D., Prakash, R., Vuong, T. H. P., & Huynh, D. T. (2000). Max-min D-cluster formation in wireless ad hoc networks. *Proceedings of IEEE INFOCOM*, Tel Aviv, Israel.

Bannerjee, S., & Khuller, S. (2001). A clustering scheme for hierarchical control in multi-hop wireless networks. *IEEE Infocom*, Anchorage, AK.

Bharghavan, V., Demers, A., Shenker, S., & Zhang, L. (1994). MACAW: A medium access protocol for wireless LANs. *Proceedings of ACM SIGCOMM '94*, Portland, Oregon.

Broch, J., Johnson, D. & Maltz, D. (1999). The dynamic source routing protocol for mobile ad hoc networks. *IETF, MANET Working Group*. Internet draft '03.

Chen, T.-W. & Gerla, M. (1998). Global state routing: A new routing scheme for ad-hoc wireless networks. *Proceedings IEEE ICC*, Atlanta, Georgia, 171-175.

Chiang, C. C., Wu, H. K., & Gerla, M. (1997). Routing in clustered multihop mobile wireless networks with fading channel. *Proceedings of IEEE Singapore International Conference on Networks*, Singapore.

Das, B., & Bharghavan, V. (1997). Routing in ad-hoc networks using minimum connected dominating sets. *Proceedings of the IEEE International Conference on Communications (ICC'97)*, 376-380.

Dhar, S., Rieck, M. Q., Pai, S., & Kim, E. J. (2004). Distributed routing schemes for ad hoc networks using d-SPR sets. *Journal of Microprocessors and Microsystems, Special Issues on Resource Management in Wireless and Ad Hoc Mobile Networks, 28*(8), 427-437.

Freebersyser, J., & Leiner, B. (2002). A DoD perspective on mobile ad hoc networks. In C. Perkins (Ed.), *Ad hoc networking.* Upper Saddle River, NJ: Addison Wesley.

Fullmer, C., & Garcia-Luna-Aceves, J. J. (1995). Floor acquisition multiple access (FAMA) for packet radio networks. *Computer Communication Review, 25*(4), 262-273.

Garcia-Luna-Aceves, J. J., & Spohn, M. (1999). Source tree adaptive routing in wireless networks. *Proceedings of IEEE ICNP*, Toronto, Canada.

Haas, Z., & Deng, J. (1998). Dual busy tone multiple access (DBTMA): A new medium access control for packet radio networks. *IEEE 1998 International Conference on Universal Personal Communications*, Florence, Italy.

Haas, Z.J. & Liang, B. (1999). Ad hoc location management using quorum systems. *ACM/IEEE Transactions on Networking, 7*(2), 228-240.

Haas, Z.J. & Pearlman, M. (2000). The zone routing protocol (zpc) for ad hoc networks. IETF, MANET Working Group, Internet draft '03. Retrieved from *http://www.ics.uci.edu/~atm/adhoc/paper-collection/haas-draft-ietf-manet-zone-zrp-00.txt*

Illyas, M. (2003). *The handbook of ad hoc wireless networks.* Boca Raton, FL: CRC Press.

Iwata, A., Chiang, C.-C., Pei, G., Gerla, M., & Chen, T. W. (1999). Scalable routing strategies for ad hoc wireless networks. *IEEE Journal on Selected Areas in Communications, 7*(8), 1369-1379.

Johnson, D. B, & Maltz, D. A. (1999). *The dynamic source routing protocol for mobile ad hoc networks* (IETF draft). Retrieved from *http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-03.txt*

Karn, P. (1992). MACA: A new channel access method for packet radio. *Proceedings of the Ninth ARRL/CRRL Amateur Radio Computer Networking Conference*, 134-140.

Liang, B., & Haas, Z. J. (2000). Virtual backbone generation and maintenance in ad hoc network

mobility management. *Proceedings of IEEE Infocom, 5,* 1293-1302.

McDonald, A. B., & Znati, T. (1999). A mobility-based framework for adaptive clustering in wireless ad-hoc networks. *IEEE Journal on Selected Areas in Communications, 17*(8), 1466-1487.

Murthy, S., & Garcia-Luna-Aceves, J. J. (1996). An efficient routing protocol for wireless networks. *ACM Mobile Networks and Applications, 1*(2), 183-197.

Papadimitritratos, P., & Haas, Z. (2002). Secure routing for mobile ad hoc networks. *Proceedings of CNDS,* San Antonio, Texas.

Park, V.D. & Corson, M.S. (1997). A highly adaptive distributed routing algorithm for mobile wireless networks. *Proceedsings IEEE INFOCOM,* 1405-1413.

Perkins, C. (2002). *Ad hoc networking.* Upper Saddle River, NJ: Prentice Hall.

Perkins, C. E., Royer, E. M., & Das, S. R. (1999). *Ad hoc on-demand distance vector routing* (IETF draft). Retrieved from *http://www.ietf.org/internet-drafts/draft-ietf-manet-aodv-04.txt*

Ramanathan, R., & Streenstrup, M. (1998). Hierarchically organized, multi-hop mobile wireless networks for quality-of-service support. *Mobile Networks and Applications, 3,* 101-119.

Rieck, M. Q., Pai, S., & Dhar, S. (2002). Distributed routing algorithms for wireless ad hoc networks using d-hop connected d-hop dominating sets. *Proceedings of the Sixth International Conference on High Performance Computing: Asia Pacific,* 443-450.

Singh, S., Woo, M., & Raghavendra, C. S. (1998). Power-aware routing in mobile ad hoc networks. *Proceedings of ACM/IEEE Mobicom,* 181-190.

Toh, C.-K. (2002). *Ad hoc wireless mobile networks.* Upper Saddle River, NJ: Prentice Hall Inc.

Weiser, M. (1993). Some computer sciences issues in ubiquitous computing. *Communications of the ACM, 36*(7), 75-84.

Wu, J., & Li, H. (2001). A dominating-set-based routing scheme in ad hoc wireless networks. *Telecommunication Systems, 18*(1-3), 13-36.

Wu, J., & Stojmenovic, I. (2004, February). Ad hoc networks. *IEEE Computer,* 29-31.

KEY TERMS

CSMA: Carrier-sense multiple access is a media-access control (MAC) protocol in which a node verifies the absence of other traffic before transmitting on a shared physical medium, such as an electrical bus or a band of electromagnetic spectrum. Carrier sense describes the fact that a transmitter listens for a carrier wave before trying to send. That is, it tries to detect the presence of an encoded signal from another station before attempting to transmit. Multiple access describes the fact that multiple nodes may concurrently send and receive on the medium.

GPS: It stands for Global Positioning System. It is an MEO (medium earth orbit) public satellite navigation system consisting of 24 satellites used for determining one's precise location and providing a highly accurate time reference almost anywhere on Earth.

MAC: Media-access control is the lower sublayer of the OSI (open systems interconnection reference model) data-link layer: the interface between a node's logical link control and the network's physical layer. The MAC sublayer is primarily concerned with breaking data up into data frames, transmitting the frames sequentially, processing the acknowledgment frames sent back by the receiver, handling address recognition, and controlling access to the medium.

MANET: A mobile ad hoc network is a system of wireless mobile nodes that dynamically self-organize in arbitrary and temporary topologies.

Peer-to-Peer Network: A peer-to-peer (or P2P) computer network is any network that does not have fixed clients and servers, but a number of *peer* nodes that function as both clients and servers to the other nodes on the network. This model of network arrangement is contrasted with the client-server model. Any node is able to initiate or complete any supported transaction. Peer nodes may differ in local configuration, processing speed, network bandwidth, and storage quantity.

Routing Protocol: Routing protocols facilitate the exchange of routing information between networks, allowing routers to build routing tables dynamically.

Ubiquitous Computing: This is a term describing the concept of integrating computation into the environment rather than having computers that are distinct objects. Promoters of this idea hope that embedding computation into the environment will enable people to move around and interact with computers more naturally than they currently do.

This work was previously published in Encyclopedia of Multimedia Technology and Networking, edited by M. Pagani, pp. 601-607, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.16
Convergence Technology for Enabling Technologies

G. Sivaradje
Pondicherry Engineering College, India

I. Saravanan
Pondicherry Engineering College, India

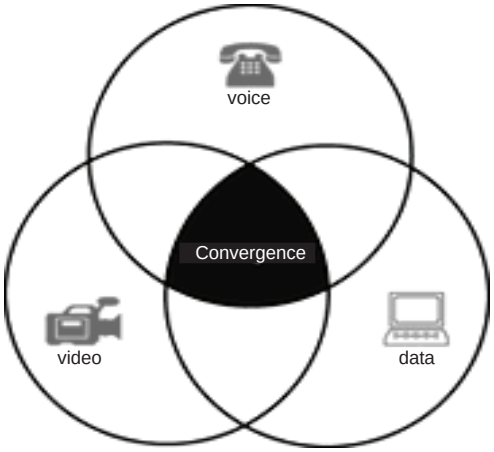
P. Dananjayan
Pondicherry Engineering College, India

INTRODUCTION

Today, we find a large number of wireless networks based on different radio access technologies (RATs). Every existing RAT has its own merits. Now the focus is turned towards the next-generation communication networks (Akyildiz, Mohanty, & Xie, 2005), which will seamlessly integrate various existing wireless communication networks, such as wireless local area networks (WLANs, e.g., IEEE 802.11 a/b/g and HIPERLAN/2), wireless wide area networks (WWANs, e.g., 1G, 2G, 3G, IEEE 802.20), wireless personal area networks (WPANs, e.g., Bluetooth, IEEE 802.15.1/3/4), and wireless metropolitan area networks (WMANs, e.g., IEEE 802.16) to form a converged heterogeneous architecture (Cavalcanti, Agrawal, Cordeiro, Xie, & Kumar, 2005). Seamless integration does not mean that the RATs are converged into a single network. Instead the services offered by the existing RATs are integrated as shown in Figure 1.

Convergence technology is a technology that combines different existing access technologies such as cellular, cordless, WLAN-type systems, short-range wireless connectivity, and wired systems on a common platform to complement each other in an optimum way and to provide a multiplicity of possibilities for current and future services and applications to users in a single terminal. After creating a converged heterogeneous architecture, the next step is to perform a common radio resource management (RRM) (Magnusson, Lundsjo, Sachs, & Wallentin, 2004). RRM helps to maximize the use of available spectrum resources, support mixed

Figure 1. Convergence of services



traffic types with different QoS requirements, increase trunking capacity and grade of service (GoS), improve spectrum usage by selecting the best RAT based on radio conditions (e.g., path loss), minimize inter-system handover latency, preserve QoS across multiple RATs, and reduce signaling delay. A typical converged heterogeneous architecture (Song, Jiang, Zhuang, & Shen, 2005) is shown in Figure 2.

CHALLENGES

The integration of different networks to provide services as a single interworking network requires many difficult challenges to be addressed. Because existing networks do not have fair RRM, the major challenge that needs to be addressed has to be mobility management. The heterogeneous network architecture will be based on IP protocol that will enhance the interoperability and flexibility. IETF Mobile IP protocol is used to support macro mobility management. But both IP protocol and mobile IP protocol (Pack & Choi, 2004; Montavont & Noel, 2002) was not basically designed to support the real-time applications. So, during the handoff between systems, users will experience the service discontinuity, such as long service time gap or network disconnection. Besides this service discontinuity, the different service characteristics of these interworked networks may degrade the quality of service (QoS).

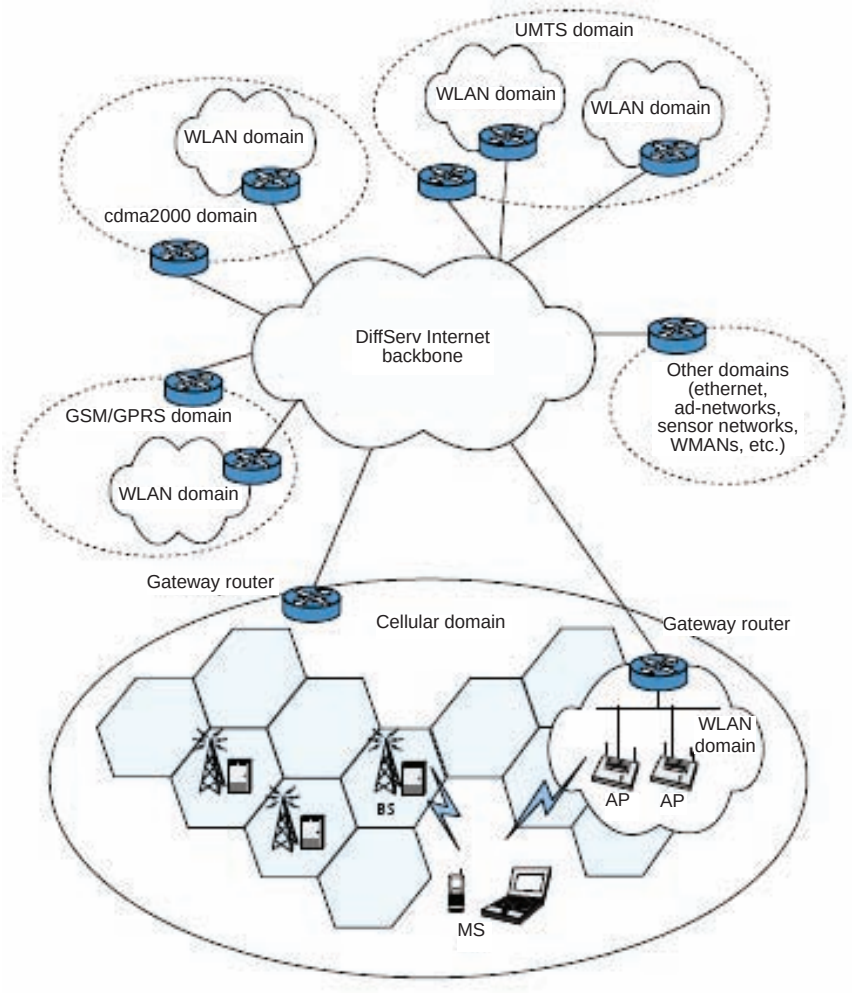
Some of the other challenges include topology and routing, vertical handoff management, load balancing, unified accounting and billing, and last but not least the protocol stack of mobile station (MS), which should contain various wireless air-interfaces integrated into one wireless open terminal so that same end equipment can flexibly work in the wireless access domain as well as in the mobile cellular networks.

PROTOCOL STACK

In a homogeneous network, all network entities run the same protocol stack, where each layer has a particular goal and provides services to the upper layers. The integration of different technologies with different capabilities and functionalities is an extremely complex task and involves issues at all the layers of the protocol stack. So in a heterogeneous environment, different mobile devices can execute different protocols for a given layer. For example, the protocol stack of a dual-mode MS is given in Figure 3.

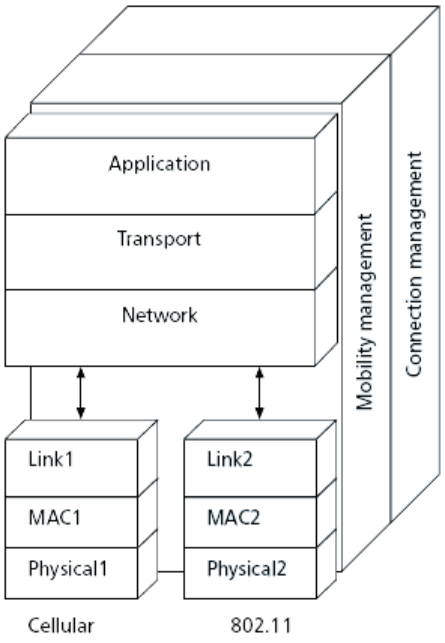
This protocol stack consists of multiple physical, data link, and medium access control (MAC) layers, and network, transport, and application layers. Therefore, it is critical to select the most appropriate combination of lower layers (link, MAC, and physical) that could provide the best service to the upper layers. Furthermore, some control planes such as mobility management and connection management can be added. These control planes can eventually use information from several layers to implement their functionalities. The network layer has a fundamental role in this

Figure 2. Converged heterogeneous network architecture



process, since it is the interface between available communications interfaces (or access technologies) that operate in a point-to-point fashion, and the end-to-end (transport and application) layers. In other words, the task of the network layer is to provide a uniform substrate over which transport (e.g., TCP and UDP), and application protocols can efficiently run, independent of the access technologies used in each of the point-to-point links in an end-to-end connection. Although there are issues in all layers, the network layer has received more attention than any other layer, and little integration-related work has been done at the lower layers. Indeed, integrated architectures are expected not to require modifications at the lower layers so that different wireless technologies can operate independently. However, this integration task is extremely complex, and it requires the support of integration architecture in terms of mobility and connection management. Seamless handoffs for "out of coverage" terminals and re-

Figure 3. Protocol stack of a dual-mode MS



source management can be provided by the two control planes.

ROUTING ISSUES

All RATS in the integrated architecture is considered as IPv6-based networks, and each element in the internetworking networks has a distinct ID number corresponding to the network routing address (Liu & Zhou, 2004). The infrastructure of a network is mapped into IPv6 addresses as shown in Figure 4. For example, the mapping of infrastructure of cellular network and IEEE 802.11 WLAN are shown in Figures 5 and 6. WLAN is given some reservation IDs, so that they can be utilized by mobile nodes under MANET mode.

VERTICAL HANDOFF MANAGEMENT

Vertical handoff is the handoff between different RATs. The major challenge in vertical handoff is that it is difficult to support a seamless service during inter-access network handoff (Wu, Banerjee, Basu, & Das, 2005; Ma, Yu, Leung, & Randhawa, 2004). The service interworking architecture and procedures, the way to provide the network and user securities, the control scheme for minimizing performance decrease caused by different service data rates, and the interworking network detection and selection methods are typical problems and to be addressed to provide stable and continuous services to users.

Unlike in the homogeneous wired networks, providing QoS for integrated architecture has some fundamental bottlenecks. This is because each radio access technology has different transmission-rate capacity over the radio interfaces, therefore the handoff between the two systems

Figure 4. Mapping infrastructure into IPv6 format

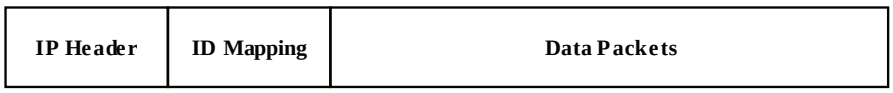
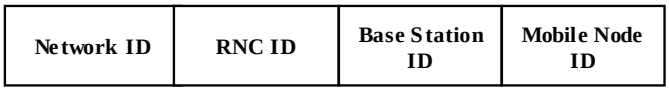


Figure 5. Mapping infrastructure of IEEE 802.11 WLAN into IPv6

Figure 6. Mapping infrastructure of cellular network into IPv6 address



makes the maintenance of QoS connection very hard. For example, WLAN can provide a transmission speed from 11Mb/s up to 54Mb/s theoretically, while UMTS has only 144kb/s at vehicular speed, 384 kb/s at pedestrian speed, and 2 Mb/s when used indoors. If we keep the QoS resource assigned by UMTS to a connection that is actually in a WAN hotspot, the advantage of the high speed of the WLAN is not fully taken. On the other hand, if we use a WLAN parameter for a station in the UMTS network, the connection may not be admitted at all (Zhang et al., 2003). Therefore, to maintain a sensible QoS framework, one has to consider the significant difference transmission capacity between two systems especially when user handover takes place.

APPLICATIONS

Convergence technology gives the possibility to combine audio and video, data, graphics, slides and documents, and Internet services in any way you like, so as to maximize the effectiveness of the communication. Integrating all traffic types enables more versatile and efficient ways of working, not just internally to the organization, but externally to customers, partners, and suppliers. It also creates a multi-system environment where a single service could be offered at different speeds at different locations/times via separate systems. The flexibility of convergence technology provides many applications and services to the user community. Some of the applications are:

- **Find-Me-Follow-Me:** This is a customizable service that makes it easy for callers to 'find' a user. Using a Web portal customers can choose how incoming calls should be handled. Options include ringing multiple phones simultaneously, or picking the order of phones to ring sequentially. Ubiquity's SIP A/S is used to dial out, in parallel or sequentially, to the user's contact numbers. Using IVR, the user can then accept the call or forward it to voicemail.
- **InfoChannels:** This is a multimedia content subscription application that pushes information and entertainment to users in real time. Users subscribe to content services through a Web portal, and new content is delivered to their designated device (mobile phone, PDA, PC browser) as soon as it is available.
- **Rich Media Conferencing:** Speak conference director is a highly scalable, carrier-class, IP conferencing application that enables conferencing service providers (CSPs) to offer hosted audio and Web conferencing services. This easy-to-use, browser-based solution offers a complete conferencing application feature set, as well as a Web portal for scheduling, initiating, managing, and terminating multi-party conferences.

Some of the services that the convergence provides to the user community are:

- **Unified Messaging:** Same inbox handling data, voice and fax.

- **Hosted IP Voice:** A complete, outsourced telephone service offering all PBX-type features.
- **IP Fax:** Delivery of e-mail to fax and fax to e-mail in a large number of countries.
- **IP Telephony:** A combination of quality transmission globally across the WAN and the LAN, with tailored consulting and end-to-end support.
- **Voice for IP VPN:** Integrated voice and data transmission, using a specific voice.
- **Video for IP VPN:** Point-to-point video transmission over the IP VPN network, using a specific class of service, called RT Vi.
- **Virtual Contact Center Services:** Optimization of agent resources while reducing costs, by allowing the routing of calls based on the agent's skills.
- **Voiceover Wi-Fi:** Full corporate mobility with a converged voice and data wireless solution.

CONCLUSION

This article provides features about convergence technology. The convergence of all existing networks will provide access to all available services using a single-user terminal. But there are many challenges to be addressed in converging the networks. In spite of converging the networks, management of the converged network is more challengeable. This article illustrates some of the challenges, and many are still open issues. Considering all the factors discussed, convergence technology is going to provide future flexibility to the wireless communication world. The complexity of this interesting technology must be addressed in the near future.

REFERENCES

Akyildiz, I. F., Mohanty, S., & Xie, J. (2005). A ubiquitous mobile communication architecture for next-generation heterogeneous wireless systems. *IEEE Radio Communications, 43*(6), S29-S36.

Cavalcanti, D., Agrawal, D., Cordeiro, C., Xie, B., & Kumar, A. (2005). Issues in integrating cellular networks, WLANS, and MANETs: A futuristic heterogeneous wireless network. *IEEE Wireless Communications, 12*(3), 30-41.

Liu, C., & Zhou, C. (2004). HCRAS: A novel hybrid internetworking architecture between WLAN and UMTS cellular networks. In *Proceedings of IEEE 2004* (pp. 374-379).

Ma, L., Yu, F., & Leung, V. C. M., & Randhawa, T. (2004). A new method to support UMTS/WLAN vertical handover using SCTP. *IEEE Wireless Communication, 11*(4), 44-51.

Magnusson, P., Lundsjo, J., Sachs, J., & Wallentin, P. (2004). Radio resource management distribution in a Beyond 3G Multi-Radio Access architecture. In *Proceedings of the IEEE Communications Society, Globecom* (pp. 3372-3477).

Montavont, N., & Noel, T. (2002). Handover management for mobile nodes in IPv6 networks. *IEEE Communications Magazine, 40*(8), 38-43.

Pack, S., & Choi, Y. (2004). A study on performance of hierarchical mobile IPv6 in IP-based cellular networks. *IEICE Transactions on Communication, E87-B*(3), 546-551.

Song, W., Jiang, H., Zhuang, W., & Shen, X. (2005). Resource management for QoS support in cellular/WLAN interworking. *IEEE Network, 19*(5), 12-18.

Wu, W., Banerjee, N., Basu, K., & Das, S. K. (2005). SIP-based vertical handoff between WWANS and WLANS. *IEEE Wireless Communications, 12*(3), 66-72.

Zhang, Q., Guo, C., Guo, Z., & Zhu, W. (2003). Efficient mobility management for vertical handoff between WWAN and WLAN. *IEEE Communication Magazine, 41*(11), 102-108.

KEY TERMS

Communication Network: Network of telecommunications links arranged so that messages may be passed from one part of the network to another over multiple links.

Grade of Service (GoS): A measurement of the quality of communications service in terms of the availability of circuits when calls are to be made. Grade of service is based on the busiest hour of the day and is measured as either the percentage of calls blocked in dial access situations or average delay in manual situations.

Heterogeneous Network: A network that consists of workstations, servers, network interface cards, operating systems, and applications from many vendors, all working together as a single unit.

Radio Access Technology (RAT): Technology or system used for the cellular system (e.g., GSM, UMTS, etc.).

Wireless Local Area Network (WLAN): Wireless network that uses radio frequency technology to transmit network messages through the air for relatively short distances, like across an office building or college campus.

Wireless Metropolitan Area Network (WMAN): A regional wireless computer or communication network spanning the area covered by an average to large city.

Wireless Personal Area Network (WPAN): Personal, short-distance area wireless network for interconnecting devices centered around an individual person's workspace.

Wireless Wide Area Network (WWAN): Wireless network that enables users to establish wireless connections over remote private or public networks using radio, satellite, and mobile phone technologies instead of traditional cable networking solutions like telephone systems or cable modems over large geographical areas.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 149-153, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.17
Document Management, Organizational Memory, and Mobile Environment

Sari Mäkinen
University of Tampere, Finland

INTRODUCTION

Wireless networks and new tools utilizing mobile information and communication technologies (ICTs) challenge the theories and practices of document management, in general, and records management, in particular. The impact of these new tools on document management as a part of organizational memory is as yet unexplored because the wireless and mobile working environment is a new concept. Recent studies of mobile environment have focused on mobile work itself or technologies used, and the aspect of document management, especially records management, has been ignored.

BACKGROUND

Records form one important part of the memory of an organization. From the organizational perspective, one method of managing intellectual resources is to augment the organization's memory. A standard connotation of organizational memory is a written record, although this is only one form of memory. Organizational memory has explicit and implicit forms and can be retained in several places like databases and filing systems, but also in organizational culture, processes, and structures (Ackerman, 1996; Walsh & Ungson, 1991). Megill (1997) specifies organizational memory to include all the active and historical information in an organization that is worth sharing, managing, and preserving for use. It is an important asset encompassing all types of documented and undocumented information that an organization requires to function effectively.

Digital documents and records can be found in every area of administration and business activities. Official records are produced in carrying out business or administrative processes, decision-making processes or procedures. These records

are vital and must be preserved for later use, as documentation and evidence and for cultural and historical reasons. Records are not preserved only for the use of the organization; they must be made accessible to individuals and customers (Young & Kampffmeyer, 2002). With a growing number of people using mobile tools, new kinds of problems are emerging. These problems arise because documents are created, processed, stored, managed, and shared through various mobile ICT tools and technologies. In a mobile working environment, it is essential that every piece of an organization's explicit memory is accessible, searchable, and preservable. This is vital, especially in the case of official and business records.

The literature on document management focuses mainly on the technologies used or the functionality of the document management systems created by practicing consultants. Academic research is rare (Bellotti & Bly, 1996; Eldridge et al., 2000; Luff, Heath & Greatbatch, 1992). Mobile working environment has been examined from the social-scientific and social interaction perspectives (Brown, Green & Harper, 2001; Katz & Aakhus, 2002). The mobile working environment in relation to the aspects of document management is an uninvestigated area and a new research topic.

ORGANIZATIONAL MEMORY

The concept of organizational memory is not new. Its roots go back to the organizational science and information-processing theories of the 1950s (Walsh & Ungson, 1991). Research on organizational memory increased especially in the 1990s in the field of information systems research. Understanding of the concept is limited, and the term is vague but commonly used. Mostly organizational memory is seen from the perspective of the organizational member. It refers to the stored information on the organization's history that can be brought to bear on present decisions (Walsh & Ungson, 1991).

The perspectives of information systems scientists on organizational memory are pragmatic, more often concentrating on the development of databases and information systems supporting organizational memory, since examining the contents of the concept is the focus of organizational scientists. Walsh and Ungson's (1991) classic study, in turn, is completely conceptual. Bannon and Kuutti (1996) claim that the concept of organizational memory does not belong exclusively to any particular research area or discipline and that a variety of definitions is available in such different fields as administrative science, organizational theory, change management, psychology, sociology, design studies, concurrent engineering, and software engineering. The viewpoint taken in archival science (see, e.g., Hedstrom, 2002; Yates, 1990, 1993) is on the historical mission of organizational memory. The purpose of archives is to retain and store the historical memory of an organization. Organizational memory research has been criticized for perceiving organizational memory as only a problem of information technology. The problem of how databases serve users is not the most essential (Koistinen & Aaltio-Marjosola, 2001).

On the basis of a through concept analysis, the definition of organizational memory is the organized knowledge of an organization, a process which is individual and distributed and past preserving, which has an effect on organizational learning, competitiveness and decision-making, and which can be supported by information technology. (Mäkinen & Huotari, 2004).

The preservation and use of organizational memory refer strictly to working life and information used in work-related settings. The empirical case studies on organizational memory pertain particularly to carrying out a task (Mäkinen & Huotari, 2004).

Schwartz, Divitini, and Brasethvik (2000) note that organizational memory has become a close

partner of knowledge management (KM), denoting the actual content that a knowledge management system purports to manage. They perceive knowledge as the key asset of the knowledge organization. They also argue that organizational memory amplifies this asset by capturing, organizing, disseminating, and reusing the knowledge. Generally, the purpose of KM is seen to make these resources available for use. This approach refers to knowledge as an object (Sveiby, 1996), and thus, brings KM close to the traditional role of information management.

Wilson (2002) argues that the information systems orientation dominates the approaches and implicit conceptions presented in the research papers, consulting practices and university curricula of KM. According to him, the theoretical foundation of this orientation is similar to that of information management research; that is, the term *knowledge* is in fact used to refer to information. Wilson argues that we cannot manage individual knowledge because it resides in human minds. Research on organizational memory information systems also supports this view by serving the needs of information retrieval and information seeking in the case of an explicit preserved form of organizational memory (Mäkinen & Huotari, 2004).

DOCUMENT MANAGEMENT IN MOBILE WORKING ENVIRONMENT

The issues of records management are not taken into account utilizing mobile tools for document management. The current need is to combine the perspectives of both document management and records management. For example, it has been suggested that about 12% of organizational knowledge is in its structured knowledge base and the majority (46%) lies scattered about organizations in the form of paper and electronic documents (Kikawada & Holtshouse, 2001). We can assume that the mobile working environment does not improve this situation.

Mobile devices can be defined in many ways. A mobile device can be described as an application of mobile technology—a technical device utilizing mobile technology and is designed to be mobile. Mobile devices, for example, include laptop computers, personal digital assistants (PDAs), mobile phones, and other handheld devices for data transfer and communication (Allen & Shoard, 2004; Weilenmann, 2003). Mobile technology is also about personal communication technologies (PCTs), which is a broader category and includes video cassette recorders, TVs, interactive voice response units (VRUs), beepers, and e-mail (Katz & Aakhus, 2002). The essential character of a mobile device is that it is mobile; it can be carried wherever you have to be and uses information and communication technology. The use of a mobile device is independent of time and space.

Even today, mobile professionals need to take paper documents with them when traveling. Paper is immediately viewable and is frequently used for ad hoc reading activities. This is still the case in spite of the amazing boom in mobile devices. The potential of combining, for example, mobile phone use with other kinds of information-related activities is being investigated in IT and telecommunications companies (O'Hara et al., 2002).

Mobile professionals have particular needs for technologies such as flexibility to accommodate their information needs in unpredictable circumstances. Mobile phone and paper documents respect this need and allow creative use while traveling (O'Hara et al., 2002).

For a mobile worker, the most important features of mobile document management are easy access, timely access, user interface, ubiquity, and compliance with security policies (Lamming et al., 2000). These features are also practical differences between document management using conventional ICT and mobile ICT. Current solutions in document management do not necessarily meet these requirements. The problems of

access are probably the most familiar to mobile workers: how to unpack and plug in a laptop in an unfamiliar environment, how to access remote documents, how to transfer a file, how to print a file, and how to secure a confidential file.

Organizational memory should be understood in a novel manner when its content, that is, documents, is managed in a wireless and mobile operating environment. The utilization of documents produced in mobile devices in knowledge processes and the problems caused by mobile environment to the lifecycle of these documents require attention regarding their creation, transfer, storage, dissemination, sharing, use, and disposal (Mäkinen, 2004).

The challenges of mobile document management and organizational memory augmentation become even more evident among communities of practice. This concept was introduced by Lave and Wenger (1991). Communities of practice are about relations among people, activity, and world in relation to other tangential and overlapping communities of practice. A newcomer learns from old-timers, and newcomers see communities of practice as an intrinsic condition for the existence of knowledge. It is a flexible group of professionals having common interests and interacting through independent tasks and embodying common knowledge (Davenport & Hall, 2002; Kimble, Hildreth & Wright, 2001). In mobile working environment communities of practice share knowledge through technological tools, but it has been argued that some types of knowledge are unsuitable for electronic storage and retrieval (Davenport & Hall, 2002).

FUTURE TRENDS

In recent years, there has been an explosion in mobile computing and telecommunications technologies. A lot of work is done outside the office in different and unpredictable locations (Allen & Shoard, 2004; Weilenmann, 2001). Mobile working environment poses challenges on organizational document management and augmentation of organizational memory. How do mobile produced documents become a part of organizational memory, and what is the relation of these documents to the intellectual capital of an organization?

The future research challenge is to increase understanding of the current state of document management and records management in mobile environments in relation to the development of organizational knowledge and intellectual capital. The focus of future research could be on the role and utilization of mobile documents produced in the joint knowledge processes and the problems caused by wireless and mobile environments, the lifecycle of these documents.

Another important research topic is the idea of access: what problems do mobile professionals have in accessing information sources of their organizations? Problems which a user encounters when trying to connect organizational information systems with mobile devices need to be studied. Using mobile devices and digital records, we also need to be convinced of the integrity of data, that it has not been modified or manipulated. If a document has been created and disseminated utilizing, for example, a mobile phone, what happens to the data when it is transferred to another information system, like document management system?

Social factors have an impact on document management practices. Wireless and mobile tools are technical innovations, but there may also be social innovations in use in the organizations when these tools are used. Organizational changes (flexible working hours), new services (use of Web pages in marketing), and new social arrangements (telework at home) are examples of social innovations. This relates to the concepts of intellectual capital and social capital.

The idea of studying communities of practice and mobile working environments provides new perspectives on mobile computing and joint value creation. It has been stated that really important

and useful information for improvement is too complex to put online. Workers might be afraid of job security and sabotage knowledge management systems (Davenport & Hall, 2002). Online communities of practice have the characteristics of material communities of practice, but they may be ephemeral, and the individuals involved may never have met.

CONCLUSION

The challenges of mobile devices and mobile working environment to document management and especially records management are varied and still largely unexplored. It is clear that the explosion of mobile computing will not improve or ease the augmentation of organizational memory, which is strictly connected to individuals.

The analysis of the concept of organizational memory suggests that its characteristics are contradictory, thereby reflecting the complex nature of the phenomenon. The explicit form of organizational memory is emphasized, but simultaneously, the individual and abstract nature of the concept are also underlined. Organizational memory, in recorded form, is concrete and palpable like paper records in an archive. However, organizational memory was also manifest implicitly and defined as invisible, mute, fuzzy, and easy to lose.

Understanding of the issues related to the management of an organizational memory is essential for enhancing the generative, productive, and representative knowledge processes in the joint value creation of different stakeholders. New knowledge is created in generative processes and with the new knowledge organization is able to provide new products and services. The new, generated knowledge is used in productive processes to provide the basis for products and services, and knowledge is transmitted to the customer as final products and services in representative processes (Huotari, 2000; Huotari & Chatman, 2001; Normann & Ramiréz, 1994).

The theoretical foundation of the organizational memory is more closely related to the multidisciplinary research area of KM and enhancement of knowledge construction based on organizational learning as a source of competitive capability than to information management. This indicates a shift from an individual organizational member's way of applying his/her own knowledge and use of information toward distributed knowledge, communication, and information and knowledge sharing, also through the use of information systems. This characteristic of the concept refers to the social nature of knowledge and information, implying that knowledge is socially constructed; that is, knowledge is a process, not an entity. The process perspective is rarely applied to studies on organizational memory, mostly in relation to an information system and its use (Ackerman & Halverson, 1998). The strategic perspective has gained more emphasis in economics (e.g., Hatami, Galliers & Huang, 2002).

REFERENCES

Ackerman, M. (1996). Organizational memory. Retrieved June 11, 2005, from *http://www.eecs.umich.edu/~ackerm/om.html*

Ackerman, M., & Halverson, C. (1998, November). Considering an organizational memory. *Proceedings of the Computer-Supported Cooperative Work (CSCW'98),* Seattle, Washington. Retrieved June 11, 2005, from *http://www.eecs.umich.edu/~ackerm/pub/98b24/cscw98.om.pdf*

Allen, D. K., & Shoard, M. (2004). Spreading the load: Mobile information and communication technologies and their effect on information overload. *Proceedings of the ISIC Conference*, Dublin, Ireland.

Bannon, L. J., & Kuutti, K. (1996, January 3-6). Shifting perspectives on organizational memory: From storage to active remembering. *Proceedings*

of the 29th Hawaii International Conference on System Sciences (HICSS-29) (pp. 156-167), Maui, Hawaii. Los Alamitos: IEEE Computer Press.

Bellotti, V., & Bly, S. (1996). Walking away from the desktop computer: Distributed collaboration and mobility in a product design team. *Computer Supported Cooperative Work '96,* Cambridge, MA (pp. 209-218).

Brown, B., Green, N., & Harper, R. (Eds.). (2001). *Wireless world: Social and interactional aspects of the mobile age.* London: Springer-Verlag.

Davenport, E., & Hall, H. (2002). Organizational knowledge and communities of practice. *Annual Review of Information Science and Techn ology, 36,* 171-227.

Eldridge, N., et al, (2000). Studies of mobile document work and their contributions to the satchel project. *Personal Technology, 4,* 102-112.

Hatami, A., Galliers, R. D., & Huang, J. (2002). Exploring the impacts of knowledge (re)use and organizational memory on the effectiveness of strategic decisions: A longitudinal case study. *Proceedings of the 36th HICSS.*

Hedstrom, M. (2002). Archives, memory and interfaces with the past. *Archival Science, 2,* 21-43.

Hofman, H. (1996, May 30-31). Lost in cyberspace – Where is the record? *Proceedings of the 2nd Stockholm Conference on Archival Science and the Concept of Record.*

Huotari, M.-L. (2000). Information behaviour in value constellation—An example from the context of higher education. *Swedish Library Research, 3/4,* 3-20.

Huotari, M.-L., & Chatman, E. (2001). Using everyday life information seeking to explain organizational behaviour. *Library and Information Science Research, 23*(4), 351-366.

Katz, J. E., & Aakhus, M. A. (2002). Conclusion: Making meaning of mobiles—a theory of Apparatgeist. In Katz & Aakhus (Eds.), *Perpetual contact: Mobile communication, private talk, public performance* (pp. 301-320). New York: Cambridge University Press.

Kikawada, K., & Holtshouse, D. (2001). The knowledge perspective in the Xerox Group. In I. Nonaka & D. J. Teece (Eds.), *Managing industrial knowledge: Creation, transfer and utilization* (pp. 283-314). London: Sage.

Kimble, C., Hildreth, P. & Wright, P. (2001). Communities of practice: Going virtual. In Y. Malhotra (Ed.), *Knowledge management and business model innovation* (pp. 216-230). Hershey, PA: Idea Group Publishing.

Koistinen, P., & Aaltio-Marjosola, I. (2001, July 5-7). Organizational memory in partnership. *Proceedings of the EGOS 2001 Conference*. Lyon, France.

Lamming, M., Eldridge, M., Flynn, M., Jones, C., & Pendlebury, D. (2000). Satchel: Providing access to any document, any time, anywhere. *ACM Transactions on Computer-Human Interaction, 7*(3), 322-352.

Lave, J., & Wenger, E. (1991). *Situated learning. Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Luff, P., Heath, C., & Greatbatch, D. (1992). Tasks-in-interaction: Paper and screen based documentation in collaborative activity. CSCW'92. Retrieved June 11, 2005, from *http://portal.acm.org*

Luff, P., & Heath, C. (1998). *Mobility in collaboration. Proceedings of the CSCW'98.*

Mäkinen, S. (2004, May 23-26). The use of mobile ICT in organizational document management in the context of organizational memory. *Proceeedings of the Information Resources Management Association International Conference IRMA2004,* New Orleans, Louisiana.

Mäkinen, S., & Huotari, M.-L. (2004, May 23-26). Organizational memory: Knowledge as a process or information as an entity. *Proceeedings of the Information Resources Management Association International Conference IRMA2004*, New Orleans, Louisiana.

Megill, K. (1997). *The corporate memory: Information management in the electronic age*. London: Bowker & Saur.

Megill, K. A., & Schantz, H. (1999). *Document management. New technologies for the information services manager*. London: Bowker & Saur.

Normann, R., & Ramiréz, R. (1994). *Designing an interactive strategy: From value chain to value constellation*. Chichester, UK: John Wiley & Sons.

O'Hara, K., Perry, M., Sellen, A., & Brown, B. (2001). *Exploring the relationship between mobile phone and document activity during business travel. Wireless World. Social and Interactional Aspects of the Mobile Age*. London: Springer-Verlag.

Schwartz, D.G., Divitini, M., & Brasethvik, T. (2000). On knowledge management in the Internet Age. In D. G. Schwartz, M. Divitini, & T. Brasethvik (Eds.), *Internet-based organizational memory and knowledge management* (pp. 1-23). Hershey, PA: Idea Group.

Sprague, R. H., Jr. (1995, March). Electronic document management: Challenges and opportunities for information systems managers. *MIS Quarterly.*

Sveiby, K.-E. (1996). What is knowledge management? *Quarterly, 19*(1), 29-49. Retrieved June 11, 2005, from *http://www.sveiby.com/articles/KnowledgeManagement .html*

Thomassen, T. (2001). A first introduction to archival science. *Archival Science, 1,* 373-385.

Walsh, J. P., & Ungson, G. R. (1991). Organizational memory. *Academy of Management Review, 16*(1), 57-91.

Weilenmann, A. (2001). Mobile methodologies: Experiences from studies of mobile technologies-in-use. *Proceedings of the 24th Information Systems Research Seminar in Scandinavia (IRIS 24).*

Weilenmann, A. (2003). Doing mobility: Towards a new perspective on mobility. *Proceedings of the 26th Information Systems Research Seminar in Scandinavia (IRIS 26).*

Wilson, T. D. (2002). The nonsense of "knowledge management". *Information Research, 8*(1), paper no. 144. Retrieved June 11, 2005, from *http://informationr.net/ir/8-1/paper144.html*

Yates, J. (1990). For the record: The embodiment of organizational memory, 1850-1920. *Business and Economic History, 2nd Series, 19,* 172-182.

Yates, J. (1993). *Control through communication: The rise of system in american management*. Baltimore: Johns Hopkins University Press.

Young, R., & Kampffmeyer, U. (2002). *Availability & preservation: Longterm availability & preservation of digital information* (AIIM Industry White Paper on Records, Document and Enterprise Content Management for the Public Sector). AIIM International Europe: Stephens & George Print Group.

KEY TERMS

Communities of Practice: A flexible group of professionals having common interests, interacting by independent tasks and embodying common knowledge (Davenport & Hall, 2002). Communities of practice are defined as a set of relations among people, activities, and the world (Lave & Wenger, 1991).

Document: Defined as a unit of recorded information structured for human consumption. Documents contain information in some structured way, and they are human creations. A document is created for a certain purpose (Megill & Schantz, 1999; Sprague, 1995).

Document Management: Covers the creation, modification, storage, and retrieval of documents required to meet users' needs and objectives (Megill & Schantz, 1999). Electronic Document Management (EDM) is the application of technology to save paper, speed up communications, and increase the productivity of business processes (Sprague, 1995).

Local Mobility: Refers to mobility within a certain space, as between rooms or floors.

Micro-Mobility: Refers to the way an artifact is mobilized and manipulated around a relatively circumscribed domain.

Mobile Device: Refers to an application of mobile technology, that is, to a technology which is designed to be mobile. Mobile devices, for example, include laptop computers, personal digital assistants (PDAs), mobile phones, and other handheld devices for data transfer and communication (Allen & Shoard, 2004; Weilenmann, 2003).

Mobility: Used here to signify the physical movement of nodes in a network or remote interaction between individuals who are far apart from each other using mobile technology. Mobility can be divided into micro mobility, local mobility, and remote mobility. (Luff & Heath, 1998; Weilenmann, 2001).

Organizational Memory: The organized knowledge of an organization, a process which is individual and distributed and past preserving, which has an effect on organizational learning, competitiveness, and decision making, and which can be supported by information technology (Mäkinen & Huotari, 2004).

Record: Regarded as process-bound information: a record is generated by work processes, structured and recorded by these work processes in order to be retrieved from the context of that work process (Thomassen, 2001). A record has four elements: recorded (physically), it contains information (content), it is an outcome of the process in which it was created (context), and it has a certain form or manifestation (structure) (Hofman, 1996). Contextual information is necessary for defining a document as a record. Unlike a document, a record needs to have contextual information. Records are also documentation of transactions, and they are preserved for evidential, historical, and cultural purposes.

Remote Mobility: Refers to remote users interacting with each other using technology.

This work was previously published in Encyclopedia of Communities of Practice in Information and Knowledge Management, edited by E. Coakes and S. Clarke, pp. 141-147, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.18
Business and Technology Issues in Wireless Networking

David Wright
University of Ottawa, Canada

INTRODUCTION

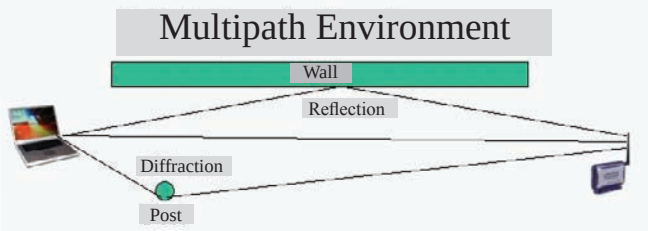
A major development in the enabling technologies for mobile computing and commerce is the evolution of wireless communications standards from the IEEE 802 series on local and metropolitan area networks. The rapid market growth and successful applications of 802.11, WiFi, is likely to be followed by similar commercial profitability of the emerging standards, 802.16e, WiMAX, and 802.20, WiMobile, both for network operators and users. This article describes the capabilities of these three standards and provides a comparative evaluation of features that impact their applicability to mobile computing and commerce. In particular, comparisons include the range, data rate in Mbps and ground speed in Km/h plus the availability of quality of service for voice and multimedia applications.

802.11 WiFi

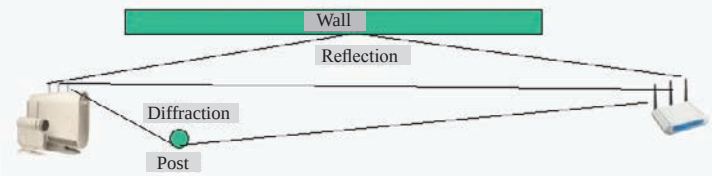
WiFi (IEEE, 1999a, 1999b, 1999c, 2003) was originally designed as a wireless equivalent of the wired local area network standard IEEE802.3, Ethernet. In fact there are many differences between the two technologies, but the packet formats are sufficiently similar that WiFi packets can easily be converted to and from Ethernet packets. Access points can therefore be connected using Ethernet and can communicate with end stations using WiFi.

WiFi can transport both real-time communications such as voice and video plus non-real time communications such as Web browsing, by providing quality of service, QoS, using 802.11e (IEEE, 2005). There are 2 QoS options. One provides four priority levels allowing real-time traffic to be transmitted ahead of non-real-time traffic, but with no guarantee as to the exact delay experienced by the real-time traffic. The other

Figure 1. (a) Receiver recovers a single signal from multiple incoming signals; (b) MIMO receiver recovers multiple signals using multiple antennas



(a)



(b)

allows the user to request a specific amount of delay, for example, 10 msecs., which may then be guaranteed by the access point. This is suited to delay sensitive applications such as telephony and audio/video streaming.

WiFi has a limited range of up to 100 metres, depending on the number of walls and other obstacles that could absorb or reflect the signal. It therefore requires only low powered transmitters, and hence meets the requirements of operating in unlicensed radio spectrum at 2.4 and 5 GHz in North America and other unlicensed bands as available in other countries.

WiFi is deployed in residences, enterprises and public areas such as airports and restaurants, which contain many obstacles such as furniture and walls, so that a direct line of sight between end-station and access point is not always possible, and certainly cannot be guaranteed when end stations are mobile. For this reason the technology is designed so that the receiver can accept multipath signals that have been reflected and/or diffracted between transmitter and receiver as shown in Figure 1(a). WiFi uses two technologies that operate well in this multipath environment: DSSS, Direct Sequence Spread Spectrum, which is used in 802.11b, and OFDM, Orthogonal Frequency Division Multiplexing, which is used in 802.11a and g (Gast, 2002). A key distinguishing factor between these alternatives, which is important to users, is spectral efficiency, that is, the data rate that can be achieved given the limited amount of wireless spectrum available in the unlicensed bands. DSSS as implemented in 802.11b uses 22 MHz wireless channels and achieves 11 Mbps, that is, a spectral efficiency of 11/22 = 0.5. OFDM achieves a higher spectral efficiency and is therefore making more effective use of the available wireless spectrum. 802.11g has 22 MHz channels and delivers 54 Mbps, for a spectral efficiency of 54/22 = 2.5 and 802.11a delivers 54 Mbps in 20 MHz channels, with a spectral efficiency of 54/20

Figure 2. WiFi handoff among access points

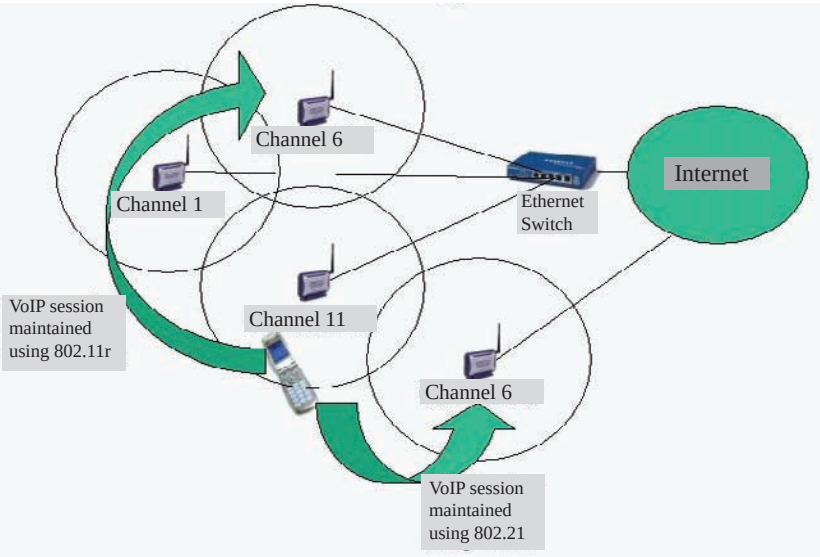
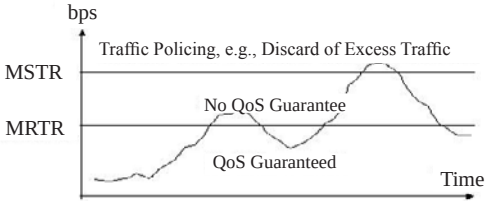


Figure 3. WiMAX traffic rate guarantees

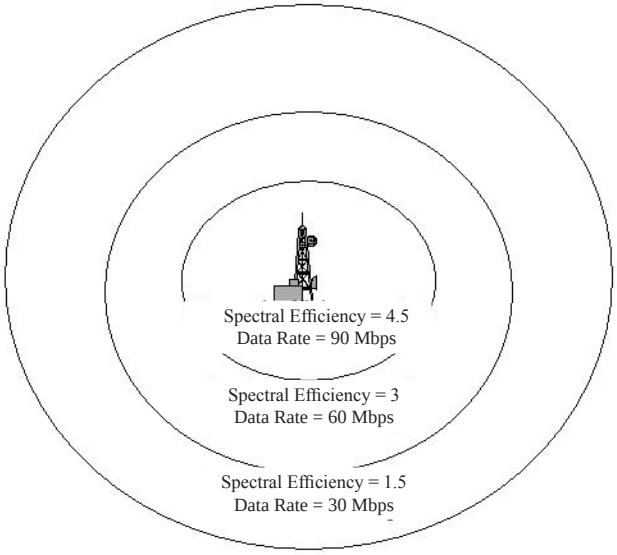


= 2.7. A recent development in WiFi is 802.11n (IEEE, 2006a), which uses OFDM in combination with MultiInput, MultiOutput, MIMO, antennas as shown in Figure 1(b). MIMO allows the spectral efficiency to be increased further by exploiting the multipath environment to send several streams of data between the multiple antennas at the transmitter and receiver. At the time of writing the details of 802.11n are not finalized, but a 4x4 MIMO system (with 4 transmit and 4 receive antennas) will probably generate about 500 Mbps in a 40 MHz channel, that is, a spectral efficiency of 500/40 = 12.5. 802.11n is suited to streaming high definition video and can also support a large number of users per access point.

The data rates in WiFi are shared among all users of a channel, however some users can obtain higher data rates than others. Network operators may choose to police the data rate of individual users and possibly charge more for higher rates, or they may let users compete so that their data rates vary dynamically according to their needs and the priority levels of their traffic. This provides considerable flexibility allowing many users to spend much of their time with low data rate applications such as VoIP, e-mail and Web browsing, with occasional high data rate bursts for audio/video downloads and data-intensive mesh computing applications.

Many deployments of WiFi use multiple access points to achieve greater coverage than the range of a single access point. When the coverage of multiple access points overlaps they should use different radio channels so as not to interfere with each other, as shown in Figure 2. For instance, in the North American 2.4 GHz band there is 79 MHz of spectrum available and the channels

Figure 4. Spectral efficiency and maximum data rates for WiMAX



of 802.11b and g are 22 MHz wide. It is therefore possible to fit 3 non-overlapping channels into the available 79 MHz, which are known as channels 1, 6 and 11. Other intermediate channels are possible, but overlap with channels 1, 6 and 11. In Figure 2, the top three access points are shown connected by Ethernet implying that they are under the control of a single network operator, such as an airport. As an end-station moves among these access points the connection is handed off from one access point to another using 802.11r (IEEE, 2006b), while maintaining an existing TCP/IP session. Movement can be up to automobile speeds using 802.11p (IEEE, 2006c). Standard technology, 802.21 (IEEE, 2006d), is also available to handoff a TCP/IP session when

Table 1. Comparative evaluation of technologies for mobile computing and commerce

	802.11, WiFi	802.16e, WiMAX	802.20, WiMobile
Range	100 metres	2-4 Km	2-4 Km
Coverage	Hot spots. Some city-wide deployments.	Designed for city-wide deployment	Designed for national deployment
Data Rate	11, 54, 500 Mbps flexibly shared among all users	Up to 90 Mbps flexibly shared among all users	> 1 Mbps per user
QoS	(a) Prioritization mechanism (b) data rate and QoS guarantees	Data rate and QoS guarantees	Data rate guarantees and QoS prioritization
Mobility Speed	100 Km/h	100 Km/h	250 Km/h
Cost	Very low unit cost access points. End-station interfaces built into phones, laptops, PDAs. Large number of access points required. Unlicensed spectrum.	Medium unit cost access points. End-station interfaces built into phones, laptops, PDAs. Licensed or unlicensed spectrum.	Medium unit cost access points. End-station interfaces built into phones, laptops, PDAs. Licensed spectrum.

a mobile end-station moves from an access point of one network operator to that of another, and this requires a business agreement between the two operators.

802.11 networks can therefore span extensive areas by interconnecting multiple access points, and city-wide WiFi networks are available in, for example, Philadelphia in the U.S., Adelaide in Australia, Fredericton in Canada and Pune in India. The broad coverage possible in this way greatly expands the usefulness of WiFi for mobile computing and electronic commerce. Enterprise users can set up secure virtual private networks from laptops to databases and maintain those connections while moving from desk to conference room to taxi to airport. A VoIP call over WiFi can start in a restaurant, continue in a taxi and after arriving at a residence.

The features of WiFi, IEEE 802.11, that are of particular importance for mobile computing and commerce are:

- Broad coverage achieved by handing off calls between access points, using 802.11r and 802.21, in cities where there are sufficient access points.
- Multimedia capability achieved by QoS, 802.11e.
- Flexibility in data rates achieved by allowing the total data rate of an access point to be shared in dynamically changing proportions among all users.
- Low cost achieved by using unlicensed spectrum, low power transmitters and mass produced equipment.

The downside to WiFi, IEEE 802.11, is limited coverage in cities that do not have extensive access point deployment.

802.16E WIMAX

802.16E (IEEE, 2006e) has a greater range than 802.11, typically 2-4 km and operates between base stations and subscriber stations. The initial IEEE standard 802.16 is for fixed applications, which compete with DSL and cable modems. Mobile applications including handoff capability among base stations, which we deal with here, are provided by 802.16E, and are based on similar but incompatible technology.

In 802.16E, WiMAX, mobility is limited to automobile speeds, up to about 100 Km/h so that it has limited use in high speed trains and aircraft. WiMAX uses the terminology "subscriber" stations, implying that customers are paying for a public service. Since the geographic range extends well into public areas, this is certainly one application. Another mobile application is a private campus network in which a central base station serves a business park or university campus. Initial deployment of WiMAX uses licensed spectrum, although low power applications in unlicensed spectrum are also specified in the standard.

WiMAX has sophisticated QoS capabilities, which allow customers to reserve capacity on the network including a reserved data rate plus quality of service. The data rate is specified by a minimum reserved traffic rate, MRTR, on which quality of service is guaranteed (Figure 3). The customer is allowed to send at a higher rate, up to a maximum sustainable traffic rate, MSTR, without necessarily receiving QoS, and above that rate, traffic will be policed by the network operator, that is, it may be discarded. The QoS parameters that can be specified by the customer are latency and jitter, plus a priority level, which is used by the base station to distinguish among service flows that have the same latency and jitter requirements. The combination of latency and jitter can be used to distinguish among service flows, and further detail on the performance of WiMAX is given by Ghosh et al. (2005).

Combinations of QoS parameters and data rates make WiMAX highly suited to mobile computing and commerce. Each subscriber can set up multiple service flows, for example, for Web browsing during a multimedia conference,

and use data rates that are quite different from those of other customers. The service provider can charge based on a combination of data rate and QoS.

WiMAX is based on OFDM, thus achieving a high spectral efficiency. There are a number of options within 802.16E for the channel widths and modulation techniques, resulting in a corresponding range of data rates and spectral efficiencies. It is important to recognize that the spectral efficiency depends on the distance between the base station and the subscriber station (Figure 4). As the signal degrades with distance it is not possible to encode so many bps within each Hz and 802.16E assigns encodings that take this into account. Closer to the base station the data rate is therefore higher. The exact distance depends on the operating environment since 802.16E uses multipath signals involving reflections and diffractions. The data rates shown in Figure 4 are the maximum achievable with the highest channel bandwidth allowed according to the standard—20 MHz—and can vary not only with distance but also according to how much forward error correction is used.

The features of 802.16E that are of particular importance for mobile computing and commerce are:

- Good range, enabling city-wide coverage with a reasonable number of base stations.
- Multimedia capability achieved by QoS, and guaranteed data rates.
- Flexibility in data rates achieved by allowing the total data rate of a base station to be shared in dynamically changing proportions among all users.

The downside to 802.16E is the cost of licensed spectrum.

802.20 WIMOBILE

At the time of writing, (1Q06), the specification of 802.20, (IEEE, 2006, f), is under development, so that less detail is available than for 802.11 and 802.16e. The key features of 802.20 are:

- It operates in licensed spectrum below 3.5 GHz.
- It is designed from the start for an all-IP environment and interfaces to IP DiffServ QoS service classes, (Grossman, 2002) which provide for prioritization of users' traffic.
- It interfaces to "Mobile IP" (Montenegro, 2001) as part of its mobility capability. Mobility includes not just automobile speed, but also high speed trains at up to 250 Km/h.
- It uses OFDM with MIMO antennas to achieve a very high spectral efficiency, so that large numbers of users can share access to a single base station.

COMPARATIVE EVALUATION

Mobile computing and commerce involves communicating from mobile devices for a variety of purposes including: data transfer for processing intensive applications and for Web browsing; voice and multimedia calls between human users; downloading audio, video and multimedia from a server, (a) streaming for real-time playout to human users and (b) file transfer for subsequent access on the mobile device. Each of these requires appropriate data rate and quality of service. Cost is also an important factor, since subscription may be required to a public network operator or an enterprise may need to build its own wireless network. Employees using mobile computing devices within a building require mobility only at pedestrian speeds. In public areas such as city streets, automobile speeds are required and between cities high speed trains may be used. The

type of mobile computing application determines which speed is appropriate. Table 1 provides a comparison among the three technologies described in this paper.

CONCLUSION

Mobile computing and commerce users have a wide range of emerging wireless communication technologies available: WiFi, WiMAX and WiMobile. Each of them offers high data rates and spectral efficiencies, and will therefore likely be available at low cost. They are the major enabling telecommunication technologies for mobile computing and are likely to be deployed in public areas and private campuses for in-building and outdoor use. WiFi is already extensively deployed and WiMAX is being deployed in Korea in 2006 and can be expected in many other countries in 2007. The WiMobile standard has not yet been specified (as of the time of writing 1Q06) and commercial equipment can be expected after WiMAX.

REFERENCES

Gast, M. (2002). *802.11 wireless networks: The definitive guide***.** O'Reilly.

Ghosh, A., Wolter, D. R., Andrews, J. G., & Chen, R. (2005). Broadband wireless access with WiMax/8O2.16: Current performance benchmarks and future potential. *IEEE Communications Magazine*, *43*(2), 129-136.

Grossman, D. (2002). *New terminology and clarifications for Diffserv.* RFC3260. Internet Engineering Task Force.

IEEE. (1999a). *802.11 wireless LAN: Medium access control (MAC) and physical layer (PHY) specifications*. New York: IEEE Publications.

IEEE. (1999b). *802.11a high-speed physical layer in the 5 GHz band*. New York: IEEE Publications.

IEEE. (1999c). *802.11b higher-speed physical layer (PHY) extension in the 2.4 GHz band*. New York: IEEE Publications.

IEEE. (2003). *802.11g further higher-speed physical layer extension in the 2.4 GHz band*. New York: IEEE Publications.

IEEE. (2005). *802.11e wireless LAN: Quality of service enhancements.* New York: IEEE Publications.

IEEE. (2006a). *802.11n wireless LAN: Enhancements for higher throughput* (In progress). Retrieved March 2006, from http://standards.ieee.org/board/nes/projects/802-11n.pdf.

IEEE. (2006 b). *802.11r wireless LAN: Fast BSS transition* (In progress). Retrieved March 2006, http://standards.ieee.org/board/nes/projects/802-11n.pdf.

IEEE. (2006c). *802.11p wireless LAN: Wireless access in vehicular environments.* (In progress). Retrieved March 2006, from http://standards.ieee.org/board/nes/projects/802-11p.pdf.

IEEE. (2006d). *802.21 media independent handover services.* (In progress). Retrieved March 2006, from http://grouper.ieee.org/groups/802/21/.

IEEE. (2006e). *802.16E-2005 air interface for fixed and mobile broadband wireless access systems: Amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands.* New York: IEEE Publications.

IEEE. (2006f). 802.20 mobile broadband wireless access systems. (In progress). Retrieved March 2006, from http://grouper.ieee.org/groups/802/20/.

Montenegro, G. (2001) *Reverse tunneling for mobile IP.* RFC3024. Internet Engineering Task Force.

KEY TERMS

Direct Sequence Spread Spectrum (DSS): A transmission technique in which data bits are multiplied by a higher frequency code sequence, so that the data are spread over a wide range of frequencies. If some of these frequencies fade, the data can be recovered from the data on the other frequencies together with a forward error correction code.

Mobile IP: An Internet standard that allows a mobile user to move from one point of attachment to the network to another while maintaining an existing TCP/IP session. Incoming packet to the user are forwarded to the new point of attachment.

Multipath: A radio environment in which signals between transmitter and receiver take several different spatial paths due to reflections and diffractions.

Orthogonal Frequency Division Multiplexing (OFDM): A transmission technique in which data bits are transmitted on different frequencies. The data transmitted on one frequency can be distinguished from those on other frequencies since each frequency is orthogonal to the others.

Quality of Service (QoS): Features related to a communication, such as delay, variability of delay, bit error rate and packet loss rate. Additional parameters may also be included, for example, peak data rate, average data rate, percentage of time that the service is available, mean time to repair faults and how the customer is compensated if QoS guarantees are not met by a service provider.

WiFi: A commercial implementation of the IEEE 802.11 standard in which the equipment has been certified by the WiFi Alliance, an industry consortium.

WiMAX: A commercial implementation of the IEEE 802.16 standard in which the equipment has been certified by the WiMAX Forum, an industry consortium.

WiMobile: Another name for the IEEE 802.20 standard which is in course of development at the time of writing (1Q06).

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 90-95, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.19
Mobile Phone Based Augmented Reality

Anders Henrysson
Norrköping Visualisation and Interaction Studio, Sweden

Mark Ollila
Norrköping Visualisation and Interaction Studio, Sweden

Mark Billinghurst
Human Interface Technology Laboratory, New Zealand

ABSTRACT

Mobile phones are evolving into the ideal platform for augmented reality (AR). In this chapter, we describe how augmented reality applications can be developed for mobile phones and the interaction metaphors that are ideally suited for this platform. Several sample applications are described which explore different interaction techniques. User study results show that moving the phone to interact with virtual content is an intuitive way to select and position virtual objects. A collaborative AR game is also presented with an evaluation study. Users preferred playing with the collaborative AR interface than with a non-AR interface and also found physical phone motion to be a very natural input method. This results discussed in this chapter should assist researchers in developing their own mobile phone based AR applications.

INTRODUCTION

In recent years, mobile phones have developed into an ideal platform for augmented reality (AR). The current generation of phones has full color displays, integrated cameras, fast processors, and even dedicated 3D graphics chips. It is important to conduct research on the types of AR applications that are ideally suited to mobile phones and user interface guidelines for developing these applications. This is because the widespread adoption of mobile phones means that they could be one of the dominant platforms for AR applications in the near future.

Traditionally AR content is viewed through a head mounted display (HMD). Wearing an HMD leaves the users hands free to interact with the virtual content, either directly or by using an input device such as a mouse or digital glove. However, for handheld and mobile phone based AR, the user looks through the screen and needs at least one hand to hold the device. The user interface for these applications is very different than those for HMD based AR applications. Thus, there is a need to conduct research on interaction techniques for handheld AR displays, and to produce formal user studies to evaluate these techniques.

In this chapter, we give an overview of the development path from mobile AR to mobile phone AR. We explain in detail how we developed an AR platform suited for mobile phones and discuss the uniqueness of mobile phone interaction for AR. We present sample applications and user studies performed to evaluate interaction techniques and metaphors.

RELATED WORK

The first mobile AR systems, such as Feiner's Touring Machine (Feiner, MacIntyre, & Webster, 1997), relied on bulky backpack worn computers and custom-built hardware. However, it was obvious that what was carried in a backpack would one day be held in the palm of the hand. Feiner showed the potential of mobile AR systems for outdoor context sensitive information overlay, while ARQuake (Thomas et al., 2002) showed how these same systems could be used for outdoor gaming.

At the same time these early mobile systems were being developed, Schmalstieg et al. (2002), Billinghurst, Weghorst, and Furness (1996), and Rekimoto (1996) were exploring early face-to-face collaborative AR interfaces. Billinghurst's Shared Space work showed how AR can be used to seamlessly enhance face-to-face collaboration (Billinghurst, Poupyrev, Kato, & May, 2000) and his AR Conferencing work (Billinghurst & Kato, 1999) showed how AR can be used to create the illusion that a remote collaborator is actually present in the local workspace. Schmalstieg's Studierstube (Schmalstieg et al., 2002) software architecture is ideally suited for building distributed AR applications, and his team has developed a number of interesting collaborative AR systems.

Using Studierstube, Reitmayr, and Schmalstieg (2001) brought the mobile and collaborative research directions together in a mobile collaborative augmented reality interface based on a backpack configuration. Prior to this, Höllerer, Feiner, Terauchi, and Rashid (1999) had added remote collaboration capabilities to the University of Columbia's touring machine, allowing a wearable AR user to collaborate with a remote user at a desktop computer. Piekarski and Thomas (2002) also added similar remote collaboration capabilities to their Tinmith system, once again between a wearable AR user and a colleague at a desktop computer. However Reitmayr's work was the first that allowed multiple users with wearable AR systems to collaborate in spontaneous ways, either face-to-face or in remote settings.

These projects showed that the same benefits that tethered AR interfaces provided for collaboration could also extend to the mobile platform, and new application areas could be explored, such as location based gaming.

Rekimoto's Transvision system explored how a tethered handheld display could provide shared object viewing in an AR setting (Rekimoto, 1996) (see Figure 1). Transvision consists of a small LCD display with a camera mounted on the back. Two users sit across the table and see shared AR content shown on the phone displays. The ARPAD interface (Mogilev, Kiyokawa, Billinghurst, & Pair, 2002) is similar, but it adds a handheld controller to the LCD panel. ARPAD decouples translation and rotation. A selected object is fixed in space relative to the LCD panel and can be moved by moving the panel. Rotation is performed using a trackball input device.

As significant computing and graphics power became available on the handheld platform, researchers have naturally begun to explore the use of personal digital assistants (PDAs) for AR applications as well. First, there was work such as the AR-PDA project (Geiger, Kleinnjohan, Reiman, & Stichling, 2001) and BatPortal (Ingram & Newman, 2001) in which the PDA was used as a thin client for showing AR content generated on a remote server. Then in 2003, Wagner and Schmalstieg (2003b) ported the ARToolKit (2005) tracking library to the PocketPC and developed the first self-contained PDA AR application. Unlike the backpack systems, handheld collaborative AR interfaces are unencumbering and ideal for lightweight social interactions.

Mobile phone-based AR has followed a similar development path. Early phones did not have enough processing power so researchers explored thin client approaches. For example, the AR-Phone project (Cutting, Assad, Carmichael, & Hudson, 2003) used Bluetooth to send phone camera images to a remote sever for processing and graphics overlay. However, Henrysson recently ported ARToolKit over to the Symbian phone platform (Henrysson & Ollila, 2003), while Moehring developed an alternative custom computer vision and tracking library (Moehring, Lessig, & Bimber, 2004). This work enables simple AR applications to be developed which run at 7-14 frames per second.

An additional thread that our work draws on is AR interaction techniques. As mobile AR applications have moved from a wearable backpack into the palm of the hand, the interface has changed. The first mobile AR systems used head mounted displays to show virtual graphics and developed a number of very innovative techniques for interacting with the virtual data. For example, in the Tinmith system (Piekarski et al., 2002), touch sensitive gloves were used to select menu options and move virtual objects in the real world. Kurata's handmouse system (Kurata, Okuma, Kourogi, & Sakaue, 2001) allowed people to use natural gesture input in a wearable AR interface, while Reitmayr et al. (2001) implemented a stylus based interaction method.

PDA-based AR applications do not typically use head mounted displays, but are based instead around the LCD display on the PDA or handheld device. At least one of the user's hands is needed to hold the PDA so some of the earlier mobile interaction techniques are not suitable. It is natural in this setting to use stylus input but there are other possibilities as well. In the AR-PAD project (Mogilev et al., 2002), buttons and a trackball on the display are used as input in a face-to-face collaborative AR game. Träskbäck and Haller (2004) use a tablet-PC and pen input for an AR-based refinery education tool. In Wagner's indoor navigation tool (Wagner & Schmalstieg, 2003c), user input is also a combination of stylus interaction and knowledge of display position from visual tracking of markers in the environment.

Handheld AR applications, such as the Invisible Train (Wagner, Pintaric, Ledermann, & Schmalstieg, 2005), also show an interesting combination of interacting with the AR content by interacting in the world and with the device itself. In this case, the user moves around in the real world to select the view of the virtual train set and then touches the screen with a stylus to change the position of tracks on the train set (see Figure 2). Similarly in Wagner's AR-Kanji collaborative game (Wagner & Barakonyi, 2003a), the user looks through the PDA screen to view real cards that have Kanji symbols printed on them. When the cards are seen through the screen, virtual models are seen corresponding to the translation of the Kanji characters. These can be manipulated by hand and the PDA shows the model from different viewpoints. There is very little stylus input required. These projects show that if the AR display is handheld, the orientation and position of the display can be used as an important interaction tool.

BRINGING AUGMENTED REALITY TO THE MOBILE PHONE

To bring AR to the mobile phone we had to develop a robust, lightweight tracking solution. Given the widespread adoption of built-in cameras in mobile phones, optical tracking was an obvious choice. There are various optical tracking techniques including fitting a projection of a 3D model onto detected features in the video image and matching a video frame with photos from known positions and orientations. However, we wanted to have a general tracking method suitable for interaction studies with minimal preparation. We choose to work with the ARToolKit software library, which provided a well-tested solution for optical tracking. ARToolKit can be used to calculate the 3D pose of a camera relative to a single square tracking marker.

In order to develop self-contained AR applications for Symbian based mobile phones we needed to port the ARToolKit tracking library to the Symbian operating system. The original implementation of ARToolKit uses double precision floating-points. However, both the mobile phones we are targeting and the PDA used by Wagner lack a floating-point unit, making floating-point arithmetic orders of magnitude slower than integer arithmetic. To overcome this, Wagner identified the most computational heavy functions and rewrote them to fixed-point using Intel's GPP library.

Fixed-point representations use the integer datatype to provide both range and precision. If great precision is required (e.g., for trigonometric functions), 28 of the 32 integer bits are used for precision. Since there was no equivalent fixed-point library featuring variable precision available for Symbian, we wrote our own. We did extensive performance tests to select the algorithms that ran the fastest on the mobile phone. The average speed-up compared to corresponding floating-point functions was about 20 times. We started out by porting the functions rewritten by Wagner and continued backwards to cover most of functions needed for camera pose estimation. The resulting port runs several times faster than the non fixed-point version of ARToolKit. This speed-up was essential for developing interactive applications.

To provide 3D graphics capabilities, we decided on OpenGL ES (OpenGL ES, 2002), which is a subset of OpenGL 1.3 suitable for low-power, embedded devices. To make it run on these limited devices some members of the Khronos group removed redundant APIs and functions. Memory and processor demanding functions such as 3D texturing and double precision floating-point values have been removed along with GLU. A 16:16 fixed-point data type has been added to increase performance while retain some of the floating-point precision. The most noticeable difference is the removal of the immediate mode in favor of vertex arrays. Since Symbian does not permit any global variables the vertex and normal arrays must be declared constant, which limits the dynamic properties of objects. While OpenGL ES takes care of the low level rendering there is still need for a higher-level game engine with ability to import models and organize the content into a scene graph. To import textured models from a 3D animation package we used the Deep Exploration tool from Right Hemisphere. This converts the exported model to C++ code with OpenGL floating-point vertex arrays, which are converted into OpenGL ES compatible fixed-point vertex arrays using a simple program we wrote. This conversion is not perfect since the exported OpenGL array indexing differs slightly from the OpenGL ES one.

Having this platform we were able to import complex 3D models and visualize them in an Augmented Reality application (see Figure 3). In this case, when the application recognizes the ARToolKit tracking marker a simple 3D model is overlaid on the live camera view. On a Nokia 6630 phone, this typically runs at 6-7 frames per second.

INTERACTION DESIGN FOR MOBILE PHONE AUGMENTED REALITY

In order to explore methods for virtual object manipulation in AR applications on a mobile phone, we need to consider the appropriate interaction metaphor. There are several key differences between using a mobile phone AR interface and a traditional head mounted display-based AR system. Obviously, the display is handheld rather than head worn meaning that the phone affords a much greater peripheral view of the real world. On the phone the display and input device are connected while they are separate for the HMD configuration. This means that with a mobile phone there is no need for a second device for interaction, configuration, or 2D menu browsing, which is the case for most HMD configurations.

These differences mean that interface metaphors developed for HMD-based systems may not be appropriate for handheld systems. For example, applications developed with a Tangible AR metaphor (Kato, Billinghurst, Poupyrev, Tetsutani, & Tachibana, 2001) often assume that the user has both hands free to manipulate physical input devices; this will not be the case with mobile phones. For phone-based AR applications, the user views the AR scene on the screen and needs at least one hand to hold the device.

These differences suggest that we look at the PDA applications for appropriate interface metaphors. However, there are also some key differences between a mobile phone and a PDA. The mobile phone is operated using a one-handed button interface in contrast to the two-hand stylus interaction of the PDA. It is therefore possible to use the mobile phone as a tangible input object itself. In order to interact we can move the device relative to the world instead of moving the stylus relative a fairly static screen.

Our approach is to assume the phone is like a handheld AR lens providing a small view into the AR scene. With this in mind, we assume that the user will more likely move the phone-display than change their viewpoint relative to the phone. The small form factor of the mobile phone lets us go beyond the looking-glass metaphor to an object-based approach. This means that input techniques can be developed largely based around motion of the phone itself, rather than keypad or button input on the phone.

For complex applications, we need 6 degree of freedom (DOF) manipulation. There have been many 6 DOF interface techniques developed for desktop applications, however there are a number of important differences between using a phone AR interface and a traditional desktop interface. Phone input options are limited since there is no mouse and keyboards are limited to a handful of high-end models. Limited screen resolution severely restricts the use of menus and multiple view-ports. We need input techniques that can be used one handed and only rely on a phone joypad and keypad input. Since the phone is handheld we can also use the motion of the phone itself to interact with the virtual object.

New opportunities in mobile phone interaction have emerged with the integration of cameras into the phones. Using simple image processing on the phone, it is possible to estimate the movement of the device, and implement 6 DOF manipulation. For example, we can fix the virtual object relative to the phone and then position and rotate objects by moving the phone relative to the real world. Bimanual interaction techniques can also be used; the dominant hand holding the phone and the non-dominant manipulating a real object on which AR graphics are overlaid.

SAMPLE APPLICATION: OBJECT MANIPULATION

To explore different object manipulation techniques, we have implemented tangible (isomorphic) interaction as well as keypad (isometric) interaction methods. In the tangible case, objects

are selected by positioning virtual cross hairs over them, clicking, and holding down the joypad controller. Once selected, the object is fixed relative to the phone and moves when the user moves the phone. When the joypad is released, the object orientation and position is set to the final phone orientation and position. These two transformations can also be handled separately where the orientation or position is reset upon release. The keypad interface allows the user to isolate one axis at the time.

To explore object rotation we implemented an ArcBall rotation technique (Chen, Mountford, & Sellen, 1988). The relative motion of the phone is used to rotate the currently selected object. The ArcBall allows the user to perform large 3DOF rotations using small movements. In a desktop implementation, the mouse pointer is used to manipulate an invisible ball that contains the object to be rotated. The resulting rotation depends on where on the ball the user clicked and in which direction the pointer was dragged. In our phone interface, the center of the object is projected into screen coordinates and a virtual crosshair acts as a mouse pointer rotating the object.

In the keypad/joypad method, the objects continuously rotate or translate a fixed amount for each fraction of a second while the buttons are pressed. In contrast, when the virtual object is fixed relative to the phone (tangible input), the user can move the object as fast as they can move the phone. Therefore, the user should be able to translate or rotate the objects faster using tangible input techniques than with keypad input.

To test the interaction methods and explore how they could be combined, we implemented a scene assembly application. The application consists of a minimal scene with two boxes and a ground plane (see Figure 4). The boxes can be moved freely above the ground plane. When selected, the object is locked to the phone and highlighted in white. The virtual model is fixed in space relative to the phone and so can be rotated and translated at the same time by moving the phone. When the keypad button is released the new transformation in the global (marker) space is calculated.

The keypad interface is used to modify all six degrees of freedom of the virtual objects. We chose to use the same buttons for both translation and rotation. To switch between the translation and rotation mode we implemented a semi-transparent menu activated by pressing the standard menu button on the joypad (see Figure 5). The menu layout consists of a 3 by 3 grid of icons that are mapped to the keypad buttons 1 to 9. Once the translation or rotation mode is entered the menu disappears.

In both modes, we are handling transformation in three dimensions corresponding to the x, y, and z-axes of the local object coordinate system. Since the joypad is 5-way and pressing it always means selection, it can only handle two of the dimensions. Therefore, to translate the object in the x-y plane we use the four directions of the joypad and complement it with the 2 and 5 keys for translation along the y-axis. For rotation, we use the joypad to rotate around the x and z-axis, while the 2 and 5 buttons rotate the object around the y-axis.

User Study: Object Manipulation

We performed a user study in order to test the usability of the manipulation techniques previously described. In the study, the users tried to position and orient blocks. The subject sat at a table, which had a piece of paper with a number of tracking makers printed on it. When the user looked through the phone display at the tracking marker, they saw a virtual ground plane with a virtual block on it and a wireframe image of a target block. The study was done in two parts, evaluating positioning and translating techniques separately. In the first, we tested the following three positioning conditions:

1. Object fixed to the phone (one handed).
2. Button and keypad input.
3. Object fixed to the phone (bimanual).

In each case, the goal was to select and move the block until it was inside the target wireframe block. In the second part of the experiment we tested the following rotation techniques:

1. ArcBall.
2. Keypad input for rotation about the object axis.
3. Object fixed to the phone (one handed).
4. Object fixed to the phone (bimanual).

For each condition, the virtual block was shown inside a wireframe copy and the goal was to rotate the block until it matched the orientation of the wireframe copy.

In the bimanual cases, the user was able to manipulate the tracking paper with one hand while moving or rotating the phone with the other, while in the other conditions the user wasn't allowed to move the tracking marker. When the block was positioned or rotated correctly inside the target wire-frame it changed color to yellow showing the subject that the trial was over. For each trial, we measured the amount of time it took the user to complete the trial and also continuously logged the position or rotation of the block relative to the target.

After three trials in one condition we asked the subject to subjectively rate his or her performance and how easy was it for him or her to use the manipulation technique. Finally, after all the positioning or orientation conditions were completed we asked the users to rank them all in order of ease of use.

Results

We recruited a total of nine subjects for the user studies, seven male and two female, aged between 22 and 32 years. None of the subjects had experience with 3D object manipulation on mobile phones but all of them had used mobile phones before and some of them had played games on their mobile phone.

Positioning

There was a significant difference in the time it took users to position objects depending on the positioning technique they used. Conditions **A** and **C** took less time than the keypad condition (condition **B**). Using a one factor ANOVA ($F(2,24) = 3.65$, $P < 0.05$) we found a significant difference in task completion times (see Table 1).

For each of the conditions, subjects were asked to answer the following questions:

Q1: How easy was it for you to position the object?
Q2: How accurately did you think you placed the block?
Q3: How quickly did you think you placed the block?
Q4: How enjoyable was the experience?

Using a scale of 1 to 7 where 1 = very easy, 7 = not very easy. Table 2 shows the average results.

The users thought that when the object was fixed to the phone (conditions **A** and **C**) it was easier to position the object correctly (**Q1**) but they could position the model more accurately (**Q2**) with the keypad input. A one factor ANOVA finds a near significant difference in the results for Q1 ($F(2,24) = 2.88$, $P = 0.076$) and Q2 ($F(2,24) = 3.32$, $P = 0.053$).

There was a significant difference in the other conditions. The users thought they could place the objects more quickly when they were attached to the phone (**Q3**) and the tangible interfaces were more enjoyable (**Q4**). A one factor ANOVA finds a significant difference in the results for Q3 ($F(2,24) = 5.13$, $P < 0.05$) and Q4 ($F(2,24) = 3.47$, $P < 0.05$).

The users were asked to rank the conditions in order of ease of use (1 = easiest, 3 = most difficult). Table 3 shows the average ranking. Condition **A** and **C** were the best ranked conditions. A one factor ANOVA gives a significant difference ($F(2,24) = 5.36$, $P < 0.05$).

Orientation

There was also a significant difference in the time it took users to orient objects depending on the technique they used. Table 4 shows the average time it took the users to rotate the virtual block to match the wireframe target. Conditions **A** (ArcBall) and **B** (keypad input) are on average twice as fast as the Tangible Input rotation conditions (**C** and **D**). A one-factor ANOVA finds a significant difference between these times ($F(3,32) = 4.60$, $P < 0.01$).

Subjects were also asked to answer the same survey questions as in the translation task, except **Q1** was changed "How easy was it for you to rotate the virtual object?" There were no significant differences between these survey responses. The subjects thought that the conditions were equally easy to use and enjoyable. The users were asked to rank the conditions in order of ease of use. There was also no significant difference between these results.

User Feedback

In addition to survey responses, many users gave additional comments about the experience. Several commented that when the virtual object was attached to the phone they felt like they were holding it. In contrast, when the keypad was used they felt that they were looking at a screen. They felt like they were more in control and they could use their spatial abilities when manipulating the virtual object with tangible input. In contrast, those that preferred the keypad liked how it could be used for precise movements and also how you didn't need to physically move yourself to rotate the object. Some users also commented on a lack of visual feedback about the rotation axis.

The block changed color when it was released inside the target but subjects thought it would have been good to change before it was released. They also felt visual cues showing the axis of rotation would be helpful, especially in the case of the ArcBall implementation. Those subjects that used two-handed input said that they felt they had more control because they could make gross movements with the camera and then fine tune the block position with small marker movements.

SAMPLE APPLICATION-FACE-TO-FACE COLLABORATIVE AR

To explore face-to-face collaborative AR we developed a simple two player game; AR Tennis. Tennis was chosen because it could be played in either a competitive or cooperative fashion, awareness of the other player is helpful, it requires only simple graphics and it is a game that most people are familiar with. For a multiplayer game, we needed a way to transfer data between phones. Since our game is a face-to-face collaborative application we chose Bluetooth and wrote a simple peer-to-peer communications layer that enables data to be shared between the phones.

Our tennis application uses a set of three ARToolKit markers arranged in a line. When the player points the camera phone at the markers they see a virtual tennis court model superimposed over the real world (see Figure 6). As long as one or more of these markers are in the field of view then the virtual tennis court will appear. This marker set is used to establish a global coordinate frame and both of the phones are tracked in this coordinate frame.

There is a single ball that initially starts on one of the phones. To serve the ball the player points their phone at the court and hits the "2" key on the keypad. Once the ball is in play, there is no need to use the keypad any more. A simple phys-

ics engine is used to bounce the ball off the court and respond to when the player hits the ball with their camera phones. The racket is defined as a circle centered on the z-axis in the xy-plane of the camera space. This means that holding the phone corresponds to holding a virtual racket. If there is an intersection between the racket plane and the ball, the direction of ball is reversed. The direction and position vectors of the ball are sent over to the other phone using Bluetooth. By sending the position the simulations will be synchronized each round. When receiving data the device switches state from outgoing to incoming and starts to check for collision with the racket. Both devices check for collision with the net and if the ball is bounced outside the court. If an incoming ball is missed the user gets to serve. Each time the ball is hit there is a small sound played and the phone of the person that hits the ball vibrates, providing haptic and audio multi-sensory cues.

In order to evaluate the usability of mobile phones for collaborative AR we conducted a small pilot user study. We were particularly interested in two questions:

1. Does having an AR interface enhance the face-to-face gaming experience?
2. Is multi-sensory feedback useful for the game playing experience?

To explore these questions we conducted two experiments, both using the AR tennis game we have developed.

Experiment One: The Value of AR

In this first study, we were interested in exploring how useful the AR view of the game was, especially in providing information about the other player's actions. Pairs of subjects played the game in each of the following three conditions:

A. Face-to-face AR: Where they have virtual graphics superimposed over a live video view.

B. Face-to-face non-AR: Where they could see the graphics only, not the live video input.

C. Non face-to-face gaming: Where the players could not see each other and also could see the graphics only. There was no live video background used.

In the face-to-face conditions (**A** and **B**) players sat across a table facing each other sharing a single set of tracking markers. In condition **C**, the players sat with a black cloth dividing them and each used their own tracking marker.

Players were allowed to practice with the application until they felt proficient with the game. Then they were told to play for 3 minutes in each of the conditions. The goal was to work together to achieve the highest number of consecutive ball bounces over the net. This was to encourage the players to cooperate together. After each condition the number of ball bounces was recorded and also a simple survey was given asking the subjects how well they thought they could collaborate together. Six pairs of subjects completed the pilot study, all of them male university staff and students aged between 21 and 40 years.

Experiment One Results

In general, there was a large variability in the number of ball bounces counted for each condition and there was no statistically significant difference across conditions. This is not surprising because pairs used many different strategies for playing the game. However, we did get some significantly different results from the subjective user surveys. At the end of each condition, subjects were asked the following four questions:

Q1: How easy was it to work with your partner?

Q2: How easily did your partner work with you?
Q3: How easy was it to be aware of what your partner was doing?
Q4: How enjoyable was the game?

Each questions was answered on a scale from 1 to 7 where 1 = Not Very Easy and 7 = Very Easy. Table 5 shows the average scores for each question across all conditions.

The users found each condition equally enjoyable (**Q4**). Interestingly enough, despite simple graphics and limited interactivity the enjoyment score was relatively high. However, there was a significant difference in response to the first three questions. The user felt that there was a difference between the conditions in terms of how easy it was to work with their partner (**Q1**) and how easily their partner worked with them (**Q2**). For question 1 (ANOVA $F(2,33) = 8.17$, $p < 0.05$) and for question 2 (ANOVA $F(2,33) = 3.97$, $p < 0.05$). The face-to-face AR condition was favored in both cases. Users felt that it was much easier to be aware of what their partner was doing (**Q3**) in the face-to-face AR condition with the live video background than in the other two conditions which had no video background (ANOVA $F(2,15) = 33.4$, $p < 0.0001$).

Subjects were also asked to rank the three conditions in order of how easy it was to work together. All but one of the users (11 out of 12) ranked the face-to-face AR condition first, confirming the results from the survey questions.

Experiment Two: Multi-Sensory Feedback

A second study was conducted to explore the value of having multi-sensory feedback in the collaborative AR application. In the game it was possible to play with audio and vibration feedback when the ball was hit. Players played the game in the following conditions:

A: Face-to-face AR with audio and haptic feedback.
B: Face-to-face AR with no audio feedback but with haptic.
C: Face-to-face AR with audio but no haptic feedback.
D: Face-to-face AR with no audio and no haptic feedback.

These four conditions were used to explore which of the audio and tactile options the players found most valuable. Each pair of players played in each condition for one minute, once again counting the highest number of consecutive ball bounces over the net and also completing a survey after each condition. The same six pairs who completed experiment one also completed experiment two. After finishing the conditions for experiment one they would continue to complete the conditions for experiment two, so that they were trained on the system.

Experiment Two Results

As with the first experiment, there was a wide variability in the average number of ball bounces counted and no statistical difference across conditions. However, we did get some significantly different results from the subjective user surveys. At the end of each condition subjects were asked the following three questions:

Q1: How easy was it to be aware of when you had hit the ball?
Q2: How easy what it to be aware of when your partner had hit the ball?
Q3: How enjoyable was the game?

Once again each questions was answered on a scale from 1 to 7 where 1 = Not Very Easy and 7 = Very Easy. Table 6 shows the average scores for each question across all conditions.

For awareness (**Q1** and **Q2**) the conditions using audio (**A** and **C**) were ranked the best. For

question 1 (ANOVA $F(3,44) = 11.1$, $p < 0.0001$) and for question 2 (ANOVA $F(3,44) = 6.59$, $p < 0.001$). They almost unanimously rated the condition that provided the most sensory output (audio, visual, haptic) as the most enjoyable (**Q3**) (ANOVA $F(3,44) = 6.53$, $p < 0.001$).

Subjects were also asked to rank the four conditions in order of how easy it was to work together. Almost all of the subjects ranked condition **A** best (10 out of 12 responses), followed by condition **C** (audio but no haptic feedback), then condition **B** (haptic but no audio feedback) and finally condition **D** (no audio or haptic feedback). Thus, they almost unanimously rated the condition which provided the most sensory output (audio, visual, haptic) as easiest to work in and also as the most enjoyable. There also appears to be a clear preference for audio only output over haptic output. This could be in part due to great awareness cue that audio provides for both the user and their partner when they hit the ball. With haptic only feedback, for the player that is not hitting the ball it is equivalent to having no feedback at all.

DESIGN RECOMMENDATIONS

Users found that the tangible interface metaphor provides a fast way to position AR objects in a mobile phone interface because they just have to move the real phone where the block is to go. The subjects also felt that it was more enjoyable.

However, there seems to be little advantage in using our implementation of a tangible interface metaphor for virtual object rotation. When the virtual object is fixed to the phone then the user often has to move themselves and the phone at the same time to rotate the object to the orientation they want, which takes time. Even when the person can use a second hand to rotate the tracking marker, this is still more time consuming than using the ArcBall or keypad input.

One of the main advantages of the keypad is that it just rotates the object around one axis at a time and so makes it easy for the user to understand what the rotation axis is and how to undo any mistakes. There is also a compromise between speed and accuracy that may affect performance. Tangible input techniques may be fast, but because they provide full six degree of freedom input, they may not be the best methods for precise input.

The collaborative AR game showed that face-to-face mobile games could benefit from combining computer graphics with views of the real world. The use of multi-sensory feedback, especially audio and visual is important for increasing game enjoyment. There are certain types of games that appear suitable for collaborative AR on mobile phones. If visual tracking is used then the ideal games have a focus on a single shared game space. This enables the players to easily see each other at the same time as the virtual content.

The screens on mobile phones are very small so collaborative AR games need only to use a limited amount of graphics and should mainly focus on enhancing the face-to-face interaction. For example in our tennis game a very simple ball, court, and net model was used, but this was enough to keep users happily engaged.

The use of an appropriate tangible object metaphor is also important for the usability of mobile phone AR applications. In our case we wanted the player to feel like the phone was a tennis racket hitting balls over a virtual net. This is why the phone vibrated when a ball was hit and a racquet sound was made. Once they understood this metaphor, it was easy for users to move the phone around the court space to hit the ball. Physical manipulation of a phone is very natural so it provides an intuitive interaction approach for collaborative AR games.

CONCLUSION AND FUTURE WORK

Mobile phones provide an interesting opportunity for augmented reality technology to move into the mainstream, used by millions of people. However before this happens more research has to be conducted on the best AR interaction metaphors and techniques for mobile devices.

In this chapter we present our experiences with mobile phone based AR. We have developed an optimized version of ARToolKit for the mobile phone, and then using that explored a tangible input metaphor where we use the real phone motion to interact with AR content. We developed a basic interaction application for 6DOF object manipulation, and the first collaborative AR game for mobile phones.

One of the main limitations of our platform is the tracking. To be able to track the phone position using ARToolKit, the complete marker pattern must be visible. We have begun to experiment with feature tracking to allow one corner of the marker square to be outside the viewfinder. Another problem is that the current ARToolKit tracking only works in a limited range. If the user is too close to the marker, one or more corners will fall outside of the viewfinder. Too far away and the resolution is too low for marker identification.

Though the focus will remain on optical tracking due to the widespread availability of camera phones, other tracking techniques might be commonly available and make the transition to wide-area mobile phone AR possible. Many 3G phones have built-in GPS, which enables outdoor positioning. Some phones have also electronic compasses and tilt sensors built-in. These sensors combined will make it possible to obtain the orientation and position of the device.

We will continue to explore mobile phone based augmented reality. In the future we would like to employ the 6DOF manipulation techniques in a collaborative set-up and conduct more in-depth user studies. Other applications will also be developed to explore other aspects of mobile phone AR such as content creation and interfacing with intelligent environments.

REFERENCES

ARToolKit (2005). ARToolKit Web site. Retrieved from www.hitl.washington.edu/artoolkit/

Billinghurst, M., & Kato, H. (1999). Real world teleconferencing. *Proceedings of CHI '99: CHI '99 Extended Abstracts on Human Factors in Computing Systems* (pp. 194-195). New York: ACM Press.

Billinghurst, M., Poupyrev, I., Kato, H., & May, R. (2000). Mixing realities in shared space: An augmented reality interface for collaborative computing. In *Proceedings of the Multimedia and Expo. IEEE International Conference* (Vol. 3, pp. 1641-1644). New York: IEEE Computer Society.

Billinghurst, M., Weghorst, S., & Furness, T. (1996). Shared space: Collaborative augmented reality. In *Proceedings of the Workshop on Collaborative Virtual Environments (CVE 96),* Nottingham, UK.

Chen, M., Mountford, S. J., & Sellen, A. (1988). A study in interactive 3D rotation using 2-D control devices. In *SIGGRAPH '88: Proceedings of the 15th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 121-129). New York: ACM Press.

Cutting, D., Assad, M., Carmichael, D. J., & Hudson, A. (2003, November 26-28). AR phone: Accessible augmented reality in the intelligent environment. In *Proceedings of OZCHI 2003.* Brisbane, Australia: University of Queensland.

Feiner, T. S., MacIntyre, B., & Webster, T. (1997, October 13-14). A touring machine: Protoyping 3D mobile augmented reality systems for exploring the urban environment. In *Proceedings of the*

1st IEEE International Symposium on Wearable Computers (ISWC 97) (pp. 74-81). Cambridge, MA: IEEE Computer Society.

Geiger, C., Kleinnjohan, B., Reiman, C., & Stichling, D. (2001). Mobile AR4ALL. In *Proceedings of the 2nd IEEE and ACM International Symposium on Augmented Reality (ISAR 2001).* New York: IEEE Computer Society.

Henrysson, A., & Ollila, M. (2003). Augmented reality on smartphones. In *Proceedings of the 2nd IEEE International Augmented Reality Toolkit Workshop.* Tokyo, Japan: Waseda University.

Höllerer, T., Feiner, S., Terauchi, T., & Rashid, G. (1999). Exploring MARS: Developing indoor and outdoor user interfaces to a mobile reality system. *Computers Graphics, 23*(6), 779.

Ingram, D., & Newman, J. (2001). Augmented reality in a WideArea sentient environment. In *Proceedings of the IEEE and ACM International Symposium on Augmented Reality (ISAR'01)* (pp. 77-85). Washington, DC: IEEE Computer Society.

Kato, H., Billinghurst, M., Poupyrev, I., Tetsutani, N., & Tachibana, K. (2001). Tangible augmented reality for human computer interaction. In *Proceedings of Nicograph 2001.*

Kurata, T., Okuma, T., Kourogi, T., & Sakaue, K. (2001). The hand-mouse: A human interface suitable for augmented reality environments enabled by VisualWearables. In *Proceedings of International Symposium on Mixed Reality (ISMR 2001),* Yokohama, Japan (pp. 188-189).

Moehring, M., Lessig, C., & Bimber, O. (2004). Video see-through AR on consumer cell phones. *Proceedings of the International Symposium on Augmented and Mixed Reality (ISMAR'04)* (pp. 252-253).

Mogilev, D., Kiyokawa, K., Billinghurst, M., & Pair, J. (2002). AR Pad: An interface for face-to-face AR collaboration. *Proceedings of CHI '02: CHI '02 Extended Abstracts on Human Factors in Computing Systems* (pp. 654-655). New York: ACM Press.

OpenGL ES (2002). *OpenGL ES Web site*. Retrieved from www.khronos.org/opengles

Piekarski, W., & Thomas, B. (2002). ARQuake: The outdoor augmented reality gaming system. *Communications of the ACM, 45*(1), 36-38.

Reitmayr, G., & Schmalstieg, D. (2001, October). Mobile collaborative augmented reality. In *Proceedings of the International Symposium on Augmented Reality 2001 (ISAR 2001)* (pp. 114-123). New York.

Rekimoto, J. (1996, September). Transvision: A hand-held augmented reality system for collaborative design. In *Proceedings of Virtual Systems and Multi-Media 1996 (VSMM '96),* Gifu, Japan (pp. 18-20).

Schmalstieg, D., Fuhrmann, A., Hesina, G., Szalavari, Z., Encarnacao, L., Gervautz, M., & Purgathofer, W. (2002). The Studierstube augmented reality project. *Presence: Teleoperators and Virtual Environments, 11,* 33-54.

Thomas, B., Close, B., Donoghue, J., Squires, J., Bondi, P. D., & Piekarski, W. (2002). First person indoor/outdoor augmented reality application: ARQuake. *Personal and Ubiquitous Computing, 6*(1), 75-86.

Träskbäck, M., & Haller, M. (2004). Mixed reality training application for an oil refinery: User requirements. In *Proceedings of the 2004 ACM SIGGRAPH International Conference on the Virtual Reality Continuum and its Applications in Industry (VRCAI '04)* (pp. 324-327). New York: ACM Press.

Wagner, D., & Barakonyi, I. (2003a). Augmented reality Kanji learning. In *Proceedings of the 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003)*

(pp. 335-343). Washington, DC: IEEE Computer Society.

Wagner, D., Pintaric, T., Ledermann, F., & Schmalstieg, D. (2005). Towards massively multi-user augmented reality on handheld devices. In *Proceedings of the 3rd International Conference on Pervasive Computing (Pervasive 2005)*, Munich, Germany.

Wagner, D., & Schmalstieg, D. (2003b). ARToolKit on the PocketPC platform. In *Proceedings of the 2nd IEEE International Augmented Reality Toolkit Workshop.* Tokyo, Japan: Waseda University.

Wagner, D., & Schmalstieg, D. (2003c, October 21-23). First steps towards handheld augmented reality. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC 2003)* (pp. 127-135). White Plains, NY: IEEE Press.

This work was previously published in Emerging Technologies of Augmented Reality: Interfaces and Design, edited by M. Haller, B. Thomas, and M. Billinghurs, pp. 90-109, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 3.20
Pen-Based Mobile Computing

Bernie Garret
University of British Columbia, Canada

INTRODUCTION

The original idea of a portable computer is credited to Alan Kay of the Xerox Palo Alto Research Center who suggested the idea in the 1970s (Kay, 1972a, 1972b; Kay & Goldberg, 1977). He envisioned a notebook-sized portable computer named the "Dynabook" that could be used for all of the user's information needs and using wireless network capabilities for connectivity.

BACKGROUND

Origins: Laptop Computers

The first actual portable "laptop" computers appeared in 1979: the Grid Compass Computer was designed in 1979 by William Moggridge for Grid Systems Corporation (Stanford University, 2003). The Grid Compass was one-fifth the weight of any model equivalent in performance and was used by NASA on the space shuttle program in the early 1980s. Portable computers continued to develop in the 1980s onwards, and most weighed around about 5 kg without any peripherals.

In 1984, Apple Computer introduced its Apple IIc model (Abbate, 1999), a true notebook-sized computer weighing about 5 kg without a monitor. The Apple IIc had an optional LCD panel monitor which made it genuinely portable and was therefore highly successful.

In 1986, IBM introduced its IBM Convertible PC with 256KB of memory; it was also a commercial success. By many, this is considered the first true laptop (mainly due to its clamshell design) that was shortly copied by other manufacturers such as Toshiba who were also successful with IBM laptop clones (Allen, 2001; Cringely, 1996). These devices retained the A4 size footprint, full QWERTY keyboards, and weighed between 3 and 4 kg (IBM, 2006). Following these innovations "tablet" PCs with a flat A4 footprint and a pen-based interface began to emerge in the 1990s.

There were several devices in the 1970s that explored the tablet, but in 1989 the Grid Systems GRiDPad was released, which was the world's first IBM PC-compatible tablet PC that featured handwriting recognition as well as a pen-based point-and-select system. In 1992, Microsoft released Microsoft Windows for Pen Computing, which had an application programming interface (API)

that developers could use to create pen-enabled applications. Focusing specifically on devices that use the pen as the primary input device, this interface has been most successfully adopted in the new breed of small, highly portable personal digital assistants (PDAs).

Personal Digital Assistants

In 1984 David Potter and his partners at PSION launched the "PSION Organiser" which retailed for just under £100 (Troni & Lowber, 2001). It was a battery-powered, 14 x 9cm, block-shaped unit with an alphabetic keyboard and small LCD screen, with 2K of RAM, 4KB of applications in ROM, and a free 8KB data card (which had to be reformatted using ultraviolet light for reuse). Compared to the much larger notebook computers of the time, it was a revolutionary device, but because of its more limited screen size and memory, it fulfilled a different niche in the market and began to be used for personal information management and stock inventory purposes (with a plug-in barcode reader).

In the late 1980s and throughout the 1990s, PSION continued to develop commercially successful small computing devices incorporating a larger LCD screen, and a new fully multi-tasking graphical user interface (before even Microsoft had got Windows up and running). These small devices were truly handheld. The PSION 3c (launched in 1991) dimensions were 165 x 85 x 22 mm, with a 480 x 160 pixel LCD screen, and the device weighed less than 400 g. A small keyboard and innovative touch-sensitive pad provided control of the cursor, and graphical icons could be selected to start applications/functions and select items from menus. The small keyboard proved difficult to use however, and the following 5c model in 1997 used an innovative foldout miniature QWERTY keyboard. These genuinely "handheld" devices with their interface innovations and ability to synchronize data with a host personal computer made the PSION models particularly successful and firmly established the personal digital assistant as a portable computing tool for professionals.

Pen-Based Interfaces for the PDA

The limitations of keyboard-based data entry for handheld devices had been recognized, and following PSION's lead, Apple Computers introduced the Newton Message Pad in 1993. This device was the first to incorporate a touch-sensitive screen with a pen-based graphical interface and handwriting-recognition software. Although moderately successful the device's handwriting recognition proved slow and unreliable, and in 1998 Apple discontinued its PDA development. However, the PDA market was now becoming firmly based upon devices using pen-based handwriting recognition for text entry, and in mid-2001, PSION, with dwindling sales and difficulties with business partnerships, ceased trading. US Robotics launched the "Palm Pilot" in 1996 using its simple "Graffiti" handwriting recognition system, and Compaq released the "iPAQ" in 1997 incorporating the new Microsoft "Windows CE/Pocket PC" operating system with the first PDA color screen.

Microsoft's relatively late entry into this market reflected the considerable research and development it undertook into developing a user-friendly pocket PC handwriting recognition interface. This remains a highly competitive field, and from November 2002 PalmSource (the new company owning the Palm Operating System) replaced the Graffiti system with Computer Intelligence Corporation's JOT as the standard and only handwriting software on all new Palm-powered devices. Computer Intelligence Corporation (CIC) was founded in conjunction with the Stanford Research Institute (SRI) based on research conducted by SRI on proprietary pattern recognition technologies (CIC, 1999). The original Graffiti system relied on the user learning a series of special characters, which while simple was irksome to

many users. The CIC JOT and Microsoft Pocket PC systems have been developed to avoid the use of special symbols or characters and allow the user to input more naturally by using standard upper and lowercase printed letters. Both systems also recognize most of the original Palm Graffiti-based special characters. In 2006 Palm introduced the Windows Mobile (Pocket PC) operating system on its own high-end devices.

The Thumb Board Text Interface

The arrival of the short messaging service (SMS), otherwise known as text messaging for cellular phones, in the late 1990s led several PDA manufacturers to adopt an alternative Thumb Board interface for their PDAs. SMS allows an individual to send short text and numeric messages (up to 160 characters) to and from digital cell phones and public SMS messaging gateways on the Internet. With the widespread adoption of SMS by the younger generation, thumb-based text entry (using only one thumb to input data on cell phone keypads) became popular (Karuturi, 2003). Abbreviations such as "C U L8er" for "See you later" and "emoticons" or "smileys" to reduce the terseness of the medium and give shorthand emotional indicators developed. The rapid commercial success of this input interface inspired the implementation of Thumb Board "keyboards" on some PDAs (such as the Palm Treo 600) for text interface. Clip-on Thumb Board input accessories have also been developed for a range of PDAs.

Tablet Format PCs

The tablet PC provides a small (usually 10 x 12" screen size) rectangular format device equipped with a sensitive screen designed to interact with a device-specific pen. The pen is used directly to write or tap on the screen. It can be used in place of a keyboard or mouse for data entry; to select, drag, and open files; to draw on the screen; and to handwrite notes and communications. Tablet PCs also incorporate handwriting recognition and conversion to text software. Unlike a touch-sensitive screen, the Tablet PC screen only receives information from the device-specific pen. It will not take information from pressure applied to the screen, so users can rest their hands on the screen and write in a more natural way. Most Tablet PCs also come with optional attachable keyboards and docking stations so they can be used in the same way as a desktop computer.

A pen-based interface for the PC was developed in the early 1990s and was originally envisaged as a challenge to the mouse. Microsoft launched "Pen Extensions for Windows 3.1" in 1991 calling it "Windows for Pen Computing." The system was designed to use plug-in slate and pen systems. However, pen-based systems would take another 10 years to become established. Shortly after its launch a number of companies introduced hardware to support it. Among them were Samsung, Fujitsu, Compaq, Toshiba, and IBM. The original IBM ThinkPad was designed as a pen-based computer. However, these pen-based systems were not well received, as many users found the Windows interface difficult to use with the stylus, and by 1995 sales of pen-based systems failed to support their further mainstream development. Bill Gates remained a strong supporter of the interface, and Microsoft decided to reintroduce pen computers as the "Tablet PC" in 2002. This time the Tablet PC specification was more successful as the use of touch-screen technologies for the pen (not well developed in the 1990s), handwriting recognition, and better integrated smaller devices made the portable tablet more acceptable for consumers.

The tablet PC has proved popular for specialist uses such as in the classroom, for creative artistic use, or more recently as the platform of choice for electronic flight planning/mapping software in aviation. A growing number of manufacturers are now producing Tablet PC hardware. However, the format still retains a far smaller proportion of the mobile PC market compared to laptops and PDAs.

MULTIMEDIA AND WIRELESS INTEGRATION

Current developments in pen-based computer interfaces are exploring the use of multimedia, voice recognition, and wireless connectivity. The expansion of memory capabilities and processor speeds for mobile computing devices has enabled audio recording, digital music storage/playback, and now digital image and video recording/playback to be integrated into these devices. This and the integration of wireless network and cellular phone technologies have expanded their utility considerably.

One of the mobile computer user. Audio is attractive for mobile applications because it can be used when the user's hands and tablet interfaces remains the output display, it can be used in conditions of low screen visibility, and it may consume less power than text-based input in the PDA. The latest PDA interface innovations include voice command and dictation recognition (voice to text), voice dialing, image-based dialing (for cell phone use, where the user states a name or selects an image to initiate a call), audio memo recording, and multimedia messaging (MMS). Several devices (e.g., the new Carrier Technologies I-Mate and Palm Treo) also incorporate a digital camera.

Wireless connectivity has enabled Internet connectivity, enabling users to access e-mail, text/graphical messaging services (SMS and MMS), and the Web remotely. These developments are gradually expanding the PDA's functionality into a true multi-purpose tool.

FUTURE TRENDS

One of the key limitations of PDA and tablet interfaces remains the output display screen size, brightness, and resolution. Issues of resolution and brightness continue to hinder many potential applications for this technology. As input technologies improve, and voice and handwriting recognition come of age, then attention to the display capabilities of these devices will need to be addressed before their full potential can be realized.

Coding PDA applications to recognize handwriting, speech, and incorporate multimedia requires additional code beyond traditionally coded interfaces. PDA application design and development environments need to support this functionality more effectively in order to promote the development of more complex mobile applications.

Data and device security are key areas for highly portable networked PDAs, and the first viruses for PDAs have started to emerge (Melnick, Dinman, & Muratov, 2004; BitDefender 2004). As multimedia interfaces develop, the specific security issues that they entail (such as individual voice recognition and prevention of data corruption of new file formats) will also need to be addressed.

CONCLUSION

Since the early models, manufacturers have continued to introduce smaller and improved portable computers, culminating in the latest generation of powerful handheld PDAs offering fast (400 MHz and faster) processors, with considerable memory (64MB of ROM and 1GB of RAM or more). This area of technological development remains highly competitive, and by necessity, the user interface for these devices has developed to fulfill the portable design brief, including the use of pen- and voice-based data input, collapsible LCD displays, wireless network connectivity, and now cell phone integration. Modern PDAs are much more sophisticated, lightweight devices and are arguably much closer to Kay's original vision of mobile computing than the current laptop or tablet computers, and possibly have the potential to replace this format with future interface developments. Indeed, if the

interface issues are successfully addressed, then it is probable that these devices will outsell PCs in the future and become the major computing platform for personal use.

REFERENCES

Abbate, J. (1999). Getting small: A short history of the personal computer. *Proceedings of the IEEE, 87*(9), 1695-1698.

Allen, R.A. (2001). *A history of the personal computer: The people and the technology III* (pp. 11-20). London; Ontario, Canada: Allen Publishing.

BBC. (2004). *First pocket PC virus discovered*. Retrieved July 17, 2006, from http://news.bbc.co.uk/1/hi/technology/3906823.stm

BitDefender. (2004). *Proof-of-concept virus hits the last virus-resistant Microsoft OS*. Retrieved July 17, 2004, from http://www.bitdefender.com/bd/site/presscenter.php?menu_id=24&n_id=102

CIC. (1999). *Economic assessment office report: Computer recognition of natural handwriting.* Retrieved August 8, 2004, from http://statusreports-atp.nist.gov/reports/90-01-0210.htm

Cringely, R. X. (1996). *Accidental empires: How the boys of Silicon Valley make their millions, battle foreign competition and still can't get a date* (pp.164-167). New York: Penguin Books.

IBM. (2006). *ThinkPad: A brand that made history*. Retrieved August 8, 2006, from http://www.pc.ibm.com/us/thinkpad/anniversary/history.html

Karuturi, S. (2002). *SMS history*. Retrieved August 8, 2006, from http://www.funsms.net/sms_history.htm

Kay, A. (1972a, August). A personal computer for children of all ages. *Proceedings of the ACM National Conference* (pp. 370-376).

Kay, A. (1972b, November). A dynamic medium for creative thought. *Proceedings of the National Council of Teachers of English Conference* (pp. 121-124).

Kay, A., & Goldberg, A. (1977). Personal dynamic media. *IEEE Computer,* (March), 31-41.

Melnick, D., Dinman, M., & Muratov, A. (2004). *PDA security: Incorporating handhelds into the enterprise* (pp. 129-131). New York: McGraw-Hill.

Stanford University. (2003) *Human computer interaction: Designing technology*. Retrieved August 10, 2006, from http://hci.stanford.edu/cs547/abstracts/03-04/031003-moggridge.html

Troni, P., & Lowber, P. (2001). *Very portable devices (tablet and clamshell PDAs, smart phones and mini-notebooks: An overview.* Retrieved August 10, 2004, from http://cnscenter.future.co.kr/resource/rsc-center/gartner/portabledevices.pdf

KEY TERMS

Audio Memo: An audio recorded message of speech digitally recorded as an audio file on a PDA.

Laptop: A portable personal computer small enough to use on your lap.

Media Player: A device or software application designed to play a variety of digital communications media such as compressed audio files (e.g., MPEG MP3 files), digital video files, and other digital media formats.

Multimedia: Communications media that combines multiple formats such as text, graphics, sound, and video.

Multimedia Messaging Service (MMS): An emerging cellular phone service that allows the sending of multiple media in a single message,

with the ability to send a message to multiple recipients. As such it can be seen as an evolution of SMS, with MMS supporting the transmission of additional media types, including: pictures, audio, video, and combinations of the above.

Palmtop: A portable personal computer which can be operated comfortably while held in one hand.

Pen Computing: A computer that uses an electronic pen (or stylus) rather than a keyboard for data input. Pen-based computers often support handwriting or voice recognition so that users can write on the screen or vocalize commands/dictate instead of typing with a keyboard. Many pen computers are handheld devices. Also known as pen-based computing.

Personal Digital Assistant (PDA): A small handheld computing device with data input and display facilities with a range of software applications. Small keyboards and pen-based input systems are commonly used for user input.

Personal Information Manager (PIM): A software application (such as Microsoft Outlook) that provides multiple ways to log and organize personal and business information such as contacts, events, tasks, appointments, and notes on a digital device.

Short Message Service (SMS): A text message service that enables users to send short messages (160 characters) to other users. A popular service amongst young people, with 400 billion SMS messages sent worldwide in 2002 (GSM World 2002).

Smart Phone: A term used for the combination of mobile phone and PDA.

Synchronization: The harmonization of data on two (or more) different digital devices so that both contain the same data. Data is commonly synchronized on the basis of the date it was last altered.

Tablet PC: A newer type of format for personal computers. The Tablet PC provides all the power of a laptop PC, but without a keyboard for text entry. Tablet PCs use pen-based input, and handwriting and voice recognition technologies as the main form of data entry, and commonly have an A4-size footprint.

Texting: Sending short text messages by SMS.

Wireless Connectivity: The communication of digital devices between one another using data transmission by radio waves.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 754-757, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.21
The Smart Card in Mobile Communications:
Enabler of Next-Generation (NG) Services

Claus Dietze
The European Telecommunications Standards Institute (ETSI), France

ABSTRACT

This chapter gives an introduction into the smart card technology and its history by outlining the role of the smart card in mobile communications systems. The role of the smart card as a key enabler for services requiring or utilizing unambiguous user-identification is outlined. These services include multimedia and high-security services such as mobile commerce or mobile banking. Smart cards containing the described mechanisms provide the user with privacy and the capabilities to use information, personalized according to his needs, in a wide-spread system with a virtually unlimited number of services. Furthermore, the capabilities of the smart card to enhance services, to secure the issuers' revenues and to increase the usage of the services by providing a trustful platform for the user are described. Future evolutions and further developments of the smart card are illustrated, including how they pave towards new types of applications and services.

INTRODUCTION

The smart card in mobile communications is used both as a service platform and as a marketing instrument for the network operator. The (Universal) Subscriber Identity Module-(U)SIM—is the network operator's "business card" that is handed out to the end-user. The design of the artwork printed on the smart card, the packaging, and the functionality directly influence the positioning of the operator's brand in the market. The smart card as used in mobile communications enjoys a high reputation and is very important for the network operators. It does not only provide security and trust thus securing the revenues of the network operator, but is also a platform for value added services. Its importance for the network operator is

impressively expressed by one of the world-leading network operators: they included the shape of the SIM into their corporate identity and use it within their logo and advertisement. Why this is absolutely justifiable will be outlined in the following chapter.

This chapter is divided into the following seven sections:

- The first section gives a brief introduction into the structure of the chapter and subject;
- The following section derives a dedicated definition for the term "smart card in mobile communications" to create a common understanding for the remainder of the chapter;
- The next section briefly lists and describes the main different specifications for smart cards used in today's mobile communications systems;
- The next section describes the technological and commercial evolution of the early SIM towards the next generation smart card (UICC, USIM, ISIM) used for 3G and further generations. Issues such as the technological constraints as well as the enhancements of the smart card are described and their impact on the market is highlighted;
- We then illustrate the role of standardizing organizations and explain the importance of standards for the success of a mobile communications system and the smart card in particular;
- The following section details the key capabilities of current and future smart cards and describes their importance for the creation of successful mobile services;
- And finally, we give an outlook on future evolutions of the smart card in mobile communications.

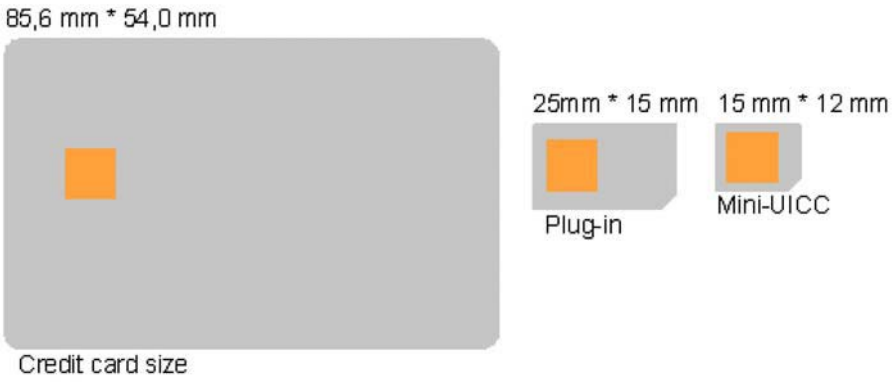
DEFINING THE "SMART CARD IN MOBILE COMMUNICATIONS"

When searching the internet or other technical literature for definitions and explanations of the term "smart card", the following can be found: "The smart card is a credit-card size plastic card containing a micro-processor". Please also refer to the Smart Card Handbook for further information on smart card technology in general. For the context of this chapter and for the usage of the smart card in mobile communications, this definition is only to some extent true and need to be modified. A more appropriate definition of what the smart card in mobile communications actually is, is developed below by examining the features and applications implemented on and executed by it. The first indication on the purpose of a particular product may in many cases be derived from its name. This also holds for a smart card in mobile communications. As, of course, everybody using a Global System for Mobile communications (GSM™) phone knows, it has been called the Subscriber Identity Module or simply the "SIM". In fact, the capability to uniquely and securely identify one single user within the network has been one of the key features for the SIM since the beginning. How this feature was extended during the evolution of the SIM will be outlined later in this chapter.

Coming back to the above cited definition of the "smart card", the following precision are made below that focus on the use of the smart card in the area of mobile communications. The first precision concerns the first part of the definition, that is, "The smart card is a credit-card size...".

Simply looking at a SIM reveals that the actual size is much smaller than the size of a regular credit card. This reduction in size was felt necessary already at a very early stage in order to allow the smart card to be inserted in smaller and smaller devices, i.e. mobile terminals. In the respective specifications and standards this small size SIM is called Plug-in or ID-000. A further reduction

Figure 1. Size reduction of SIM card



in size was introduced into the standards in the beginning of 2004 (Mini-UICC) and show that the size of the smart card should not be part of the definition.

Another physical characteristic of the SIM or the next generation smart card for telecommunication is even more important. It was one of the crucial factors for the success of the SIM. The SIM is a token that can be removed from one terminal and easily put into another one. This allows the user to transport all personal as well as end user subscription related data from one terminal to another, for example when buying a new terminal. Even in the days of tri-band terminals allowing end-users to perform calls in almost every part in the world, new access technologies arise that again benefit from the "removableness" of the SIM. Wireless Local Area Networks (WLANs) could be mentioned as just one example of where the smart card may need to be removed from the mobile terminal and put into a WLAN device. Another solution will allow the smart card in the mobile terminal to be used for the authorization of the WLAN session that runs on a different piece of hardware.

The second modification of the definition is related to the part "... size plastic card...". Due to the reduced size of the SIM (see above) only a small piece of plastic is used to hold the module containing the micro-processor. From this point of view the material that is used to hold the module should not have any significance in the definition and could be left out.

The smart card is also described as "...containing a micro-processor" for the execution of functions implemented on it. Even though the micro-processor of the smart card is its heart, the soul of the smart card is or are the applications implemented on it. The applications characterize the smart card and make it useable in dedicated markets. In addition to the micro-processor, more and more memory capacity is required in the smart card. This memory is needed in order to contain multiple applications and value added services as well as complex configuration and provisioning parameters for services such as Multimedia Messaging Service, General Packet Radio Service (GPRS) connectivity or others. Rather than defining a smart card through its possession of a micro-processor, the smart card

in mobile communications should be defined through its capabilities to execute applications and to manage specific types of data such as data which is personalized according to the individual users' needs.

As a result of the above observations the following is offered as a more accurate definition of the "smart card used in mobile communications": "The smart card in mobile communications is an individually personalized and removable authentication token. It is used to execute dedicated applications and manages specific data within the mobile communications system."

SMART CARD IN MOBILE COMMUNICATIONS SYSTEMS

Having defined the "smart card in mobile communications", we may now consider its role and the respective specifications for different mobile communications systems.

Besides the already frequently mentioned SIM used in GSM/EDGE Radio Access Network (GERAN) or USIM/UICC used in the Universal Terrestrial Radio Access Network (UTRAN), smart cards are also specified for other mobile communications systems—with the difference that in these systems the smart card is an optional component whereas it is mandatory in GERAN and UTRAN systems. The SIM specification, GSM TS 11.11, is the mother of almost every specification that was developed for other mobile communications systems. It was used as a basis for a smart card used in Terrestrial Trunked Radio (TETRA) systems that focus on emergency services as well as for the smart card used in the Digital Enhanced Cordless Telecommunications (DECT™) system and the Code Division Multiple Access (CDMA) system, just to mention three.

System Architecture

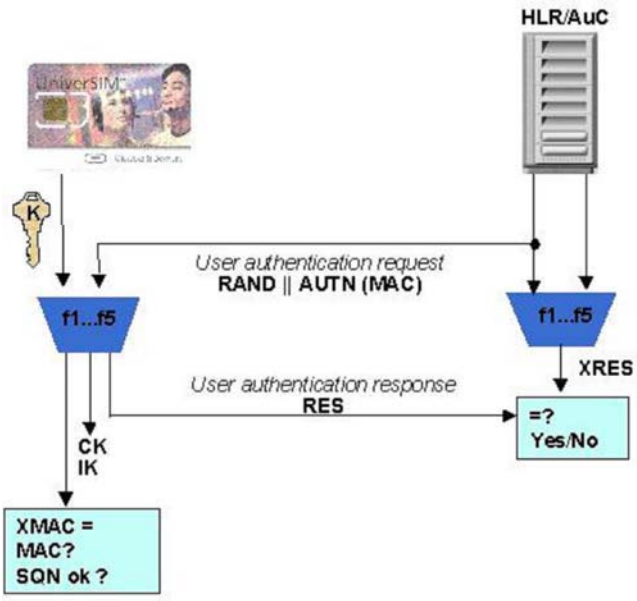
The smart card in mobile communications represents one crucial component in the network infrastructure. It mainly plays two different roles: to provide secure access and provisioning to the network and to provide additional value added services for the end user and/or the network operator. Both roles are outlined in the following two subsections.

Network System Component

The smart card as a network system component is primarily used to authenticate the subscriber to the network as in GSM or to mutually authenticate both the subscriber and the network as in UTRAN. The following Figure 2 shows the simplified authentication procedure in a UTRAN system and illustrates the role of the smart card. All authentication relevant computations on the user side are executed inside the smart card. The secret key K used for the computations never leaves the smart card and is safely stored in the secure memory area of the chip. The smart cards' counterpart for the authentication in the network is the Authentication Centre (AuC). The AuC also possesses the secret key K and is therefore able to calculate the expected response of the smart card (RES). By performing the respective calculations (using functions f1 to f5) on both the user and the network side and by comparing the results of the calculations with the values submitted by the respective counterpart, a decision on the network access can be made by both parties.

Connectivity parameters such as Multimedia Message Service (MMS) parameters and service related information such as preferred network identities are stored on the smart card. This information is used by the handset to access the appropriate networks and services. Putting the connectivity parameters on the smart card allows the user to access the network operators' services independent of the terminal used.

Figure 2. Authentication procedure

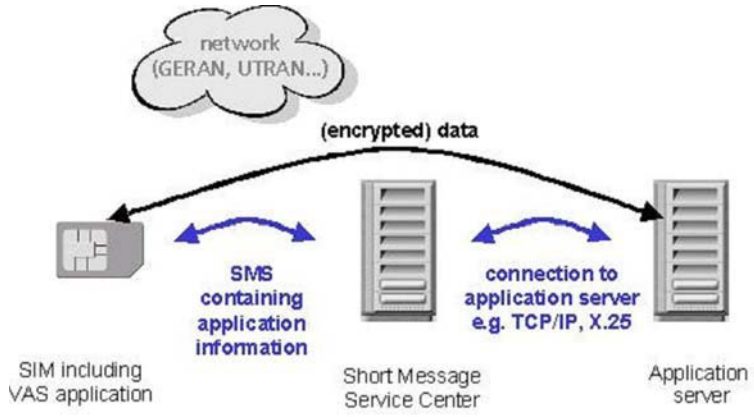


Value Added Services System Component

The Card Application Toolkit (CAT) provides a common set of commands that was derived from the SIM Application Toolkit. This framework enabled the development of additional applications that can be put on the smart card. These applications are in general linked to a network entity in the background server system. This server system is not necessarily located at the network operators premises and could also be operated and maintained by third party service providers such as news content providers or banks. An end-to-end (server-to-card) communication channel for applications stored and executed on the smart card can be established. Dedicated security mechanisms that are defined in the respective specifications (see TS 03.48 and TS 102 127 for further information) provide a secure channel between the smart card and the network entity. Application relevant data can be encrypted and will only be available for the application server and the application in the smart card.

Figure 3 illustrates today's system architecture for SMS based services in a 2G network. The SIM issues a short message that contains dedicated and application specific information. This information can optionally be encrypted. The data part contained in the short message is then routed via the Short Message Service Center (SMSC) to the application server. The application server (optionally decrypts and) interprets the contents of the message and triggers the relevant behaviour such as downloading further information to the SIM.

Figure 3. Simplified value added services architecture



GSM/EDGE Radio Access Network (GERAN) and Universal Terrestrial Radio Access Network (UTRAN)

The main (and most successful) mobile communications system that currently involves a smart card is certainly GERAN (GSM). The Subscriber Identity Module—the SIM—is specified as a mandatory component in the whole system. Therefore, all mobile terminals have to include a smart card reader that is able to contain the smart card. The technical core specifications that are being maintained and further developed by the respective standards organizations are:

- TS 11.11 the "Specification of the SIM-Mobile Equipment Interface" that started in the early days and was developed further until Release 99; and
- TS 51.011 the "Specification of the SIM-Mobile Equipment Interface" which is the corresponding document that only exists in Release 4.

The members of the 3rd Generation Partnership Project (3GPP™ concluded to freeze the SIM specifications at Release 4 and to include further enhancements to the SIM into the evolutional counterpart of the SIM in UTRAN—the Universal Subscriber Identity Module (USIM). The main specifications for the USIM are:

- TS 31.101 defining the "UICC-Terminal Interface; Physical and Logical Characteristics" of the smart card from Release 99 onwards; and
- TS 31.102 specifying the "Characteristics of the USIM Application" from Release 99 onwards.

A comprehensive overview of how the two smart card applications, the SIM and the USIM, interwork and how they could be combined on one single smart card that could be used for GERAN as well as UTRAN is also part of these specifications. A related technical report, TR 31.900 "SIM/USIM Internal and External Interworking Aspects", describes these aspects in detail. It

outlines further the key role that the smart card plays for a network operator that migrates from 2nd Generation to 3rd Generation networks.

Code Division Multiple Access (CDMA)

The second major mobile communications player in the industry is CDMA. 3GPP2, the equivalent to 3GPP for the specification of CDMA, is responsible for the definition and maintenance of the specifications for CDMA2000. 3GPP2 is a partnership project consisting of the following partners: Association of Radio Industries and Businesses (ARIB-Japan), China Communications Standards Association (CCSA-China), Telecommunications Industry Association (TIA-North America), Telecommunications Technology Association (TTA-Korea) and The Telecommunication Technology Committee (TTC-Japan).

In CDMA networks the smart card is optional, that is, all parameters such as subscription data, network settings, and security functions are stored and personalized into the handset. Due to the absence of the smart card CDMA has been facing the following issues:

- No roaming to GSM/GERAN networks for the subscriber, which means a new card and a new terminal is required when travelling abroad;
- Difficult handset exchange due to difficult transfer of personal and subscription related data from the existing to the new terminal;
- Difficult manufacturing process for terminals due to personalization of each of the terminals with user individual data; and
- No SIM Application Toolkit based services and applications available that can be easily transferred from one terminal to another.

These limitations lead to strong requests for a smart card. This request was also supported by the CDMA Development Group (CDG). Therefore 3GPP2 specified--also based on the SIM specification in TS 11.11—the requirements for the Removable User Identity Module (R-UIM) in technical specification C.S0023-0. The R-UIM is an extension of the Subscriber Identity Module (SIM) capabilities, to enable operation in a radiotelephone environment. Examples of this environment include, but are not limited to, analogue CDMA. The specification is based on the SIM specification and includes additional commands and responses necessary within the context of CDMA. The introduction of the R-UIM allows subscriber to “plastic roam” (by switching the smart card) between CDMA and GSM networks.

Digital Enhanced Cordless Telecommunications (DECT)

The European Telecommunications Standards Institute (ETSI) has developed a total of more than 30 publications (technical specifications, technical reports or technical base for regulation) for the Digital Enhanced Cordless Telephony (DECT). The first system became operative in 1992. The DECT Authentication Module (DAM), based on the SIM specification TS 11.11, was specified in the early 1990’s. This enabled a smart card to be used as an authentication token for the end user to be introduced and several specifications were approved:

- ETSI ETS 300 331 ed.1 (1995-11): Digital Enhanced Cordless Telecommunications (DECT); DECT Authentication Module (DAM); and
- ETSI ETS 300 825 ed.1 (1997-10): Digital Enhanced Cordless Telecommunications (DECT); 3 Volt DECT Authentication Module (DAM).

Mobile network operators that also operate as fixed line operators have experienced the

advantages of a smart card in a mobile communications system and are seeking to adapt services that are being used in the mobile world also to the home and fixed line environment. This is an interesting phenomenon. Even though the fixed line telephony has been used for far longer than mobile telephony, the standards for the features and the look and feel of terminals are being set by the mobile industry. This trend will also impact services being used in the mobile world that are going to be introduced in the fixed line environment. Features such as short messages and multimedia messages are being introduced into the "regular" phones, which are mainly DECT phones. The inclusion of a smart card adding further advantages to them seems to be a logical evolution of today's DECT phones. A DECT Local Area Network (DECT LAN) also appears to be an area where secure user authentication and therefore the DAM could be essential.

TErrestrial Trunked RAdio (TETRA)

TETRA is an open digital standard developed by ETSI that describes a common mobile radio communications infrastructure. This infrastructure is targeted primarily at the requirements and needs of public safety groups such as police and fire departments. The requirements comprise the need to rely on fast and accurate file communication even if no network coverage is given. These groups have been high-end users of private/professional mobile radio (PMR) or public access mobile radio (PAMR) technology. Based on digital, trunked radio technology, TETRA is targeted to be the next-generation architecture and standard for current, analogue PMR and PAMR markets. As TETRA is targeted to public safety groups, privacy and confidentiality of the data and voice communication is essential.

Again based on the SIM specification TS 11.11, a smart card, the TETRA SIM, was specified for the usage in the TETRA system. The TETRA SIM used for user authentication and storage of configuration data and phonebooks was specified by the respective ETSI technical body in the mid 1990s in:

- ETR 295 "Radio Equipment and Systems (RES); Trans-European Trunked Radio (TETRA); User requirements for Subscriber Identity Module (SIM)". This Technical Report describes the high level requirements that have to be fulfilled by the TETRA SIM. ETR 295 was published in 1996 and indicates that the SIM is an optional device within TETRA Mobile Stations (MS): thus this ETR does not preclude the implementation of MS without a SIM.
- ETS 300 812 "Terrestrial Trunked Radio (TETRA); Security aspects; Subscriber Identity Module to Mobile Equipment (SIM - ME) interface", Edition 1, was the first version of the TETRA SIM specification. It was published in 1998.
- ETS 300 812 Edition 1 was revised to EN 300 812 version 2.1.1, published in 2001.
- ES 200 812 part 1 (Physical and logical characteristics) and part 2 (Characteristics of the TSIM application), published in 2002, are the Edition2/Release 1 versions of TS 100 812-1 and TS 100 812-2 (see chapter 3.2) and are technically identical. They are intended to ensure a smooth transition to TETRA Release 2.

In line with the evolution of the smart card to a multi-application platform, the TETRA SIM specification in TETRA Release 2 split off the physical characteristics and concentrated on the definition of a TETRA SIM application. The aim was to bring convergence with the Universal SIM (USIM), to meet the needs for TETRA specific services whilst gaining the benefits of interworking and roaming with public mobile networks such as GSM, GPRS and UMTS™ TETRA Release 2 started in September 2000 in order to enhance the services and facilities of TETRA. Release 2

of the TETRA SIMs (TSIM) aligns the TETRA SIM application with 3GPP. Release 2 was previously known as Edition 3 of the TETRA SIM specifications.

EVOLUTION OF THE SMART CARD IN MOBILE COMMUNICATIONS

Evolution Of Smart Card Hardware

In 1988 the idea of introducing a smart card in mobile communications led to the first conception of the SIM. In these early days microprocessor chips that could be embedded in a smart card had only a very limited amount of memory that did not allow storage of large data sets. Therefore the first functionality implemented on the SIM focussed on the authentication algorithm. From the beginning the smart card was used to provide the end user with a secure token that enabled her/him to access the GSM service. For network operators the SIM allowed from the very beginning to manage billing and other information about their subscribers. The security features of the smart card in mobile communications were constantly enhanced and developed.

Data storage capability was discussed as early as May 1988. It was introduced for the provision of services including Short Message Service, Advice of Charge, Abbreviated Dialling Numbers and Public Land Mobile Network (PLMN) selection, but memory was a major constraint. Initially, the total memory capacity of a SIM was about 10 kB for both the operating system (read only memory—ROM) and data storage (programmable memory—EEPROM). Only in the mid-1990's did larger chips with 8 kB of programmable memory become available. Today, chips provide 128 kB to accommodate (programmable) data and applications, and at least the same amount for the operating system and other ROM-based applications. New technologies such as flash or floating EEPROM technology will soon enable the mass deployment of even larger chips, with 1 MB or more of programmable memory—today's expectations with today's available technology reach to an estimated maximum of about 16MB memory within the today's (U)SIM.

Considering the fact that comparing the 128KB with a 1MB smart card is already an increase in available memory of about 800%, the question arises how to actually use up all memory. The answer is given by the network operators: The SIM is the property and under full control of the network operator. Therefore all important information and network connectivity parameters as well as service related information should be stored and managed by the SIM.

Also, the separation of programmable and read only memory could be dissolved in favor of the programmable memory. New technologies allow the storage of the operating system in a special one-time programmable memory that can be loaded onto the smart card during the production process. This has the advantage that packages of features, tailored to the network operators' needs, could be loaded onto the smart card rather than burning the complete set of features into the ROM part of the chip. Figure 4 illustrates the traditional split of the memory into ROM and EEPROM. It shows the separation of the memory into an operating system area (ROM) and the applications and data area (EEPROM). The provisioning of the memory is fix and cannot be changed after the production of the chip. New technologies such as flash or "floating EEPROM", as shown in Figure 5, provide a more flexible memory management. The operating system, applications and data share one common memory pool. This pool is not split and can be managed according to the network operators needs.

Features that resided in ROM by default and that are not required by the network operator could be removed, freeing more memory for additional applications that are requested by the network operator or their customers. Based on this technology development cycles and thus time to

Figure 4. Traditional memory separation scenario

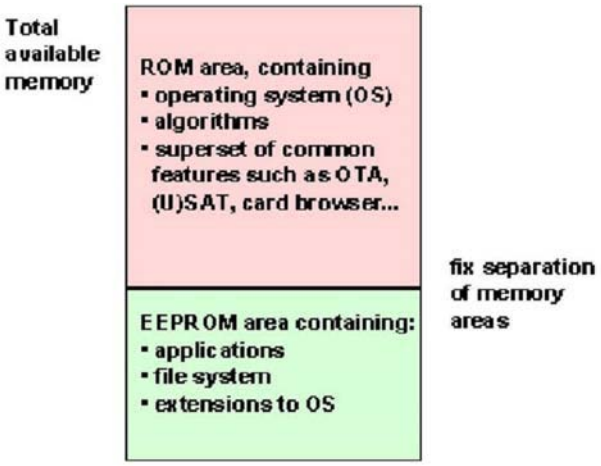
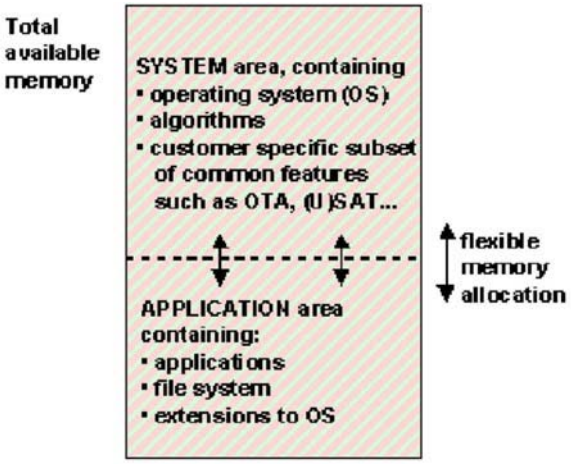


Figure 5. Flexible memory separation scenario



market will be reduced. Providing new operating systems and features in the ROM mask of a chip is a time consuming task. From the finalization of the development by the smart card manufacturer until the reception of first commercial samples of the chip takes between three and six months. Making the development of these components independent of the silicon manufacturer saves valuable time for the introduction of new features and enhancements for the network operator.

Along with the evolution of the memory, SIM processors have followed a similar growth pattern. From the first eight bit processors to today's chips which have the computing power of 16 bit and 32 bit processors. Dedicated crypto co-processors allow the powerful execution of asymmetric crypto-algorithms. Asymmetric

crypto-algorithms enable the smart card to play a major role in the application of Public Key Infrastructures (PKI) to regulate the use of certificates for authentication in e-transactions.

Evolution of Software and Operating System Architecture

In parallel to the evolution of the chip hardware the operating system as well as the capabilities of the operating system evolved. SIM development reached a major milestone in 1996 when ETSI approved the first technical specification for the SIM Application Toolkit (Technical Specification TS 11.14). This specification defined a set of commands and procedures to enable the card to contain applications specific to the issuer (the network operator), allowing the operator to introduce a wide range of new services including information and location based services, banking and Internet access. Today, the smart card in mobile devices operates and manipulates menus and services and authenticates users for service access. The high level of security which the SIM offers means it can secure financial transactions over a mobile phone, enabling mobile commerce. Work continues to introduce additional features. The introduction of the SIM Application Toolkit was the first step towards the opening of the SIM for additional and customer, that is network operator-specific applications. These applications had to be developed by specialists that had access to the operating system of the card and that were able to low level code the application (hard code) into the EEPROM of the SIM card. Therefore development was time consuming and expensive. Only smart card manufacturers were able to implement such applications on their cards. As different smart card vendors had their own (different) smart card operating system the development had to be done as often as the network operators have smart card suppliers.

A need for an Application Programming Interface (API) arose that, as for other computer systems, allowed applications to be developed by almost anybody and rapidly. It should be possible to run these applications interoperable on any smart card, independently of the smart card vendor. To fulfil this request, ETSI approved a high level requirements specification to introduce such an API in TS 02.19. The first API fulfilling the requirements of TS 02.19 was approved in 1998 (TS 03.19) and was based on the Java CardTM specifications developed by the Java CardTM forum. The new Java CardsTM were supposed to allow development of applets in a quick and cost efficient way. However, reality showed that even though JavaTM seemed to be the right way to go, all aims could actually not be achieved in the beginning. Interoperability on a 100% technical level needed to be slowly established and proven in the market. Using JavaTM technology does not automatically guarantee interoperability.

Smart card vendors offer development kits for their Java CardTM products (such as Sm@rtCafe Professional of JavaTM Mobile Application Designer (JMAD) from Giesecke & Devrient, Cyberflex from axalto or GemXPlore from Gemplus). New services and applications can comparatively easy be developed, leading to an increased demand of the smart card vendors' Java CardsTM. Actually, only very few application developers could profitably enter the market and offer new smart card applications to network operators. The reason is rather simple: network operators have been used to getting applications almost for free from the smart card vendor. Due to the history of application development on SIM cards and the competitive situation in the market, most of the smart card vendors continued to offer the applications for free or at least for a far less cost than an independent application developer. Nevertheless JavaTM proved to be very important within the industry especially with regards to time to market and flexibility for the network operator. It allows the one time development of services utilizing SIM Application Toolkit commands. After perambulating the learning curve

and due to the help of external organizations such as SIMalliance and ETSI, the major smart card vendors managed to provide truly interoperable Java CardsTM. That means that applets that have been developed by one party should run equally on cards of any other party that also followed the appropriate specifications for the development of the applet and the Java CardTM.

With the advent of 3G, the SIM has evolved to become the "USIM" (the Universal Subscriber Identity Module). Whereas the SIM is the definition of a complete smart card including the physical and logical characteristics (i.e., the plastic and the chip), the USIM is defined as being an application. The USIM resides on a smart card that is to be implemented according to the technical specifications for the smart card platform, the UICC. Figures 6 and 7 illustrate the different concepts. Figure 6 shows the traditional SIM architecture whereas Figure 7 illustrates the new modular concept. A smart card for 3G mobile communications consists of the UICC containing at least one USIM application.

This new approach of separating the physical and logical characteristics from the functions and applications enables the smart card to become multi-application capable. It is like having a PC with a basic operating system (being the UICC) and the Internet explorer managing the access to the network (being the USIM). The USIM provides features which equip it to play a key role in crucial aspects of 3G such as managing security access, virus intrusion, customer profiles, mutual authentication, downloading, and a new phonebook allowing the management of additional information such as fax numbers and e-mail addresses. The USIM also has the ability to store applications for network services, offering, for example, pre-paid service activation and control, information services, directory services, mobile banking, and ticketing. See later in this chapter for further information.

The UICC allows users access to global roaming by means of their smart card, irrespective of the radio access technology used. It is able to contain multiple applications, allowing smooth roaming and interworking between different services and networks, whether GSM, the new Wideband Code Division Multiple Access (W-CDMA) or other networks; the handset will be able to access a portfolio of services and applications available to users via their user profiles.

The UICC's revolutionary ability to handle true multi-applications, providing the platform for independent applications which can even run in parallel, present an interesting test of both the ingenuity of marketing experts and the ability of different market sectors and manufacturers to co-operate in the deployment of services. For instance, telecommunication operators are able to issue UICCs containing both a USIM and an electronic purse.

The UICC also contains new features such as enhanced security, further Application Programming Interfaces (APIs), new form factors, enhancement of the interface speed, access to shared multimedia sessions through the Internet Protocol Multimedia Subsystem (IMS), and, of course, backwards compatibility for network operators, allowing them a smooth transition from 2G to 3G.

Evolution Milestones

Figure 8 provides an overview on major steps achieved during the evolution and development of the smart card in mobile communications, in particular the SIM and USIM.

Supplementary information on the evolution of the SIM can be found in the article "The Subscriber Identity Module, Past, Present and Future" of Dr. Klaus Vedder in [1].

Further evolution of the smart card in mobile communications is certain as the market for GERAN and UTRAN and therefore the market for the smart card used in such systems continues to grow. To give an indication on the market size:

Figure 6. Single application smart card

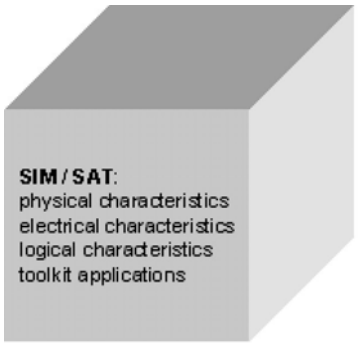


Figure 7. Multiapplication smart card

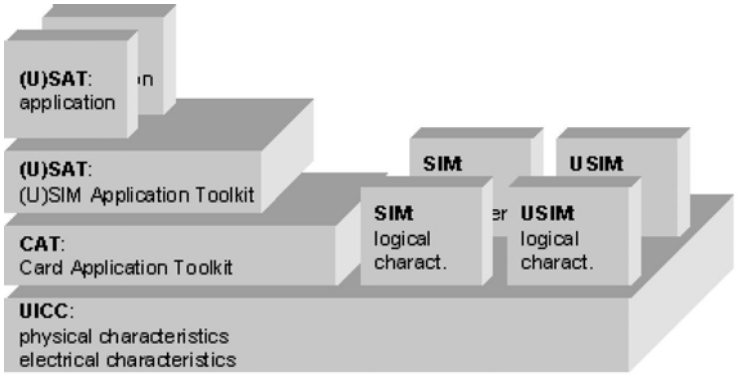
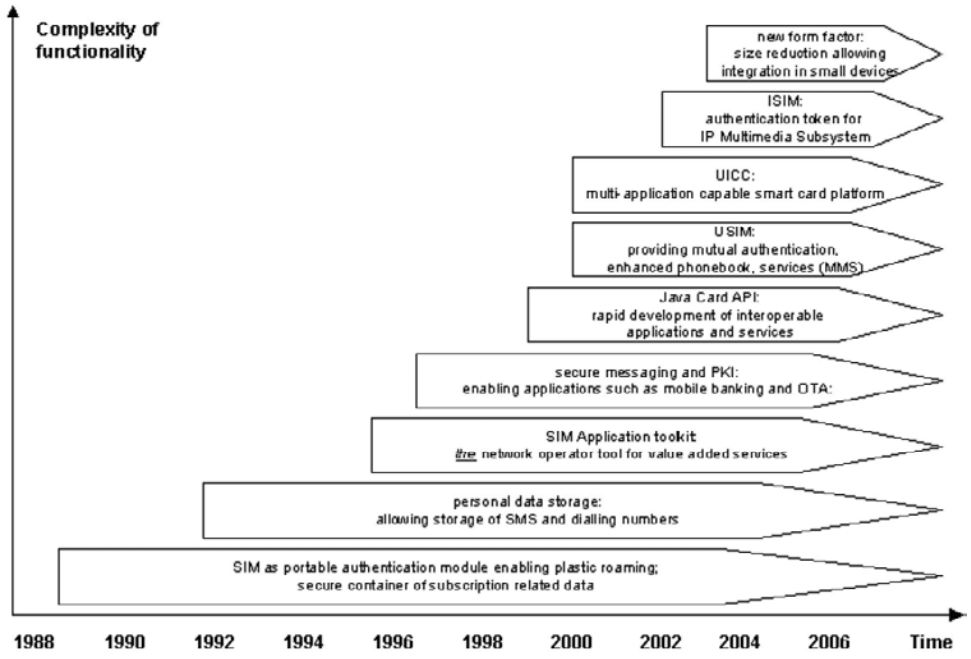


Figure 8. Evolution of the SIM



from its first delivery to the present, far more than 2 billion SIM cards have been delivered.

STANDARDIZATION

When talking about markets, features, services and business one of the main drivers for the business is in most of the cases underrepresented: standardization. Standardization is a subject that is not very often discussed in connection with services, marketing, and business but is actually one of the most important factors in the game. Standards aim to establish systems that make interoperable solutions among different manufacturers possible and therefore also reduce cost. This drives the acceptance of the system in the market. Without standardization systems like mobile communications could not have become reality on such a large scale. One can without doubt say that standardization has been the driver for the success of GSM and the SIM.

Today's purchasing decisions are more and more based on whether the solution or product is implemented according to a particular set of standards. The availability of specifications and standards was a key factor for the success of the SIM as well as the ability to provide the network operator with a standardized subscriber authentication method. The revenue of the network operator depends to a large extent on the security of the authentication and thus the billing system. It is crucial to rely on a defined method for the subscriber authentication and in addition a set of harmonized and standardized security features. The production of the relevant standards has been undertaken by ETSI.

ETSI is the recognized European Standardization Organization for telecommunications and related fields of broadcasting and information technology. From its inception in 1988, the Institute has been at the leading edge in setting security standards. It achieved an outstanding success with the standardization of the Global System for Mobile communications (GSM), which included authentication, anonymity and customer privacy. This represented the first full, worldwide, commercial deployment of encryption and smart cards, and ETSI's standardization of the SIM for GSM has helped make it the most widely deployed smart card ever.

With the closure of ETSI committee SMG9 (Special Mobile Group 9) in the year 2000, which was responsible for specifying the SIM, and the establishment in December 1998 of the 3rd Generation Partnership Project (3GPP), of which ETSI is a founding partner, ETSI's work on the SIM application, i.e. the non-platform and non-generic part, was transferred to 3GPP's Technical Specification Group TSG-T3. T3's task is to further evolve the SIM application to meet the demands of the new 3rd generation (3G) mobile network.

Further smart card-related work continues within ETSI's Smart Card Platform Project (EP SCP), which was founded in 2000 as the successor of SMG9. EP SCP is the focal point in ETSI for the standardization of the common Integrated Circuit (IC) card platform for 2G (e.g. GSM) and 3G mobile communications systems, the UICC. (As described earlier, the UICC comprises the platform specifications implemented on the smart card, together with all resident applications based on them. It also contains the USIM as an application for access to the 3GPP system, and/or the R-UIM application for access to the 3GPP2 system.) The work of EP SCP provides a common platform on which others, including organizations from the financial sector, can base their system-specific applications.

In addition, EP SCP has worked to make the specifications for GSM independent of the bearer network and, as part of the process, new deliverables have been approved. These specifications provide standardized security mechanisms for the interface between a network entity (e.g. a toolkit application) and an entity in the UICC. They also make available a standardized method for the secure, remote management of files and applica-

tions on the UICC. A requirements specification for a generic API, the UICC-API, was approved in 2002, and the work on a corresponding functional and architectural specification was completed in May 2003. This allows the rapid development of interoperable card-based applications (applets).

The upshot of these developments is that, while the SIM retains its original function of authentication, it has evolved to become both a service platform offering multiple value added services and a multi-application platform providing interoperability and interworking between different access technologies.

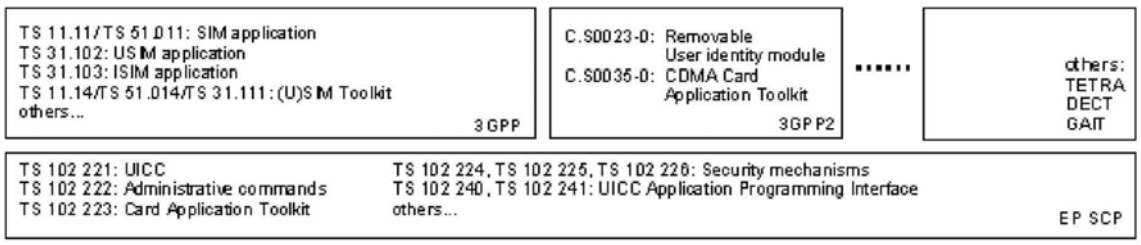
Further advancements of the UICC platform and the applications based on it will create new possibilities. For example, the new form factor specified by EP SCP allows the development of smaller devices, for example, for data transmission only, and offer additional communication and financial applications. New chip technology and the continuation of standardization work will advance the capabilities of smart cards to the point where they are really personal mini-computers. This development in turn will enable new applications, which will drive the growth of other new technologies, particularly 3G mobile.

The following Figure 9 illustrates the responsibility and interdependencies of the different standards producing bodies. ETSI project Smart Card Platform (EP SCP) is the initiator of common and application independent platform specifications. It is responsible for the creation and maintenance of these platform specifications on which other committees such as 3GPP and 3GPP2 base their system and application specifications such as the SIM, USIM and CDMA Card Application Toolkit.

Institutes such as ETSI, and partnership projects (such as 3GPP and 3GPP2) between different institutions, are important for the generation of specifications and standards. It is very easy to imagine that a set of specifications for a system such as 3G is a huge effort and that the different parties that are going to base their products and services on these specifications heavily contribute to such standards. Hundreds of member companies from all over the world develop their products and services based on the standards they produce under the umbrella of the institutes and partnership projects.

Clearly, decision-making and Intellectual Property Rights (IPRs) are or could become quite an issue among the member companies. For these reasons ETSI and 3GPP respectively, created working procedures and established an IPR policy under which the specifications are produced, with the aim of minimizing these issues. This framework is accepted by the different parties when they become a member and sign the membership agreement of the institute.

Figure 9. Organisation of standards and responsibility of standards bodies



The development of specifications/deliverables is consensus driven. The way decisions are made is well defined in the rules and working procedures. Given the diversity of interests, it is remarkable that, in the long history of standardization, the use of the ultimate decision-making tool where no consensus can be reached, — the vote — is very rare. In the case of standardization for smart cards in telecommunication, up to now only one single vote has been; this was in December 2003 in ETSI Project Smart Card Platform, and concerned the introduction of the new form factor for the smart card.

Last but not least, the fruitful collaboration between the different standardization organizations and their collaboration with the industry partners ensures that the standards meet market requirements.

2G, 3G AND NG SERVICES BASED ON THE SMART CARD

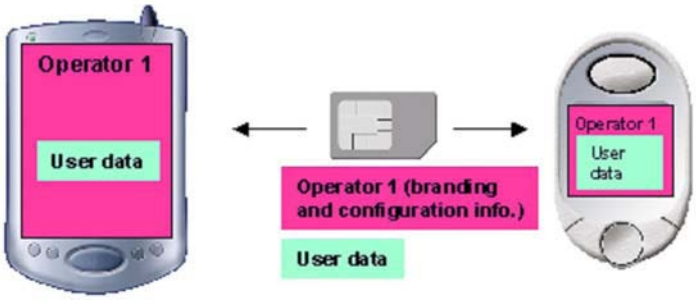
Smart card technology has consistently and reliably provided solutions to current and future requirements and challenges. The smart card evolution cycles become ever shorter, whilst the capabilities of the smart card increase as dramatically as for any other product in the information technology domain. Even though those capabilities increase, the key features of the smart card in mobile communications remain the same, yet are subject to enhancements and evolution. This section explores those key features and discusses the services and applications that utilize these key features. It concludes with a review of some of the work areas that are currently being considered by the smart card specification groups. They are listed and briefly described to illustrate just some example areas of further enhancements and future trends.

Security Mechanisms

One of the main features of the smart card is to provide security: security when authenticating the subscriber to the network (and also the network to the subscriber) by performing cryptographic operations as well as security mechanisms to protect security relevant data from unauthorized access. Security means trust and the level of trust depends on the relevance of the performed transaction (and the connected value). The end user need to trust the stability and security of the system when performing dedicated operations that in particular have some financial impact. These services include: digital signature to sign commercial transactions; mutual authentication of subscribers and service providers to get access to special (with cost) services; storage of secret user data such as keys or PINs to grant access to the secret or personal data on the smart card; secure authentication to access Virtual Private Networks for company subscriptions; providing features to secure copyrights (Digital Rights Management) when for example transferring valuable content from one terminal to another terminal; preventing theft of mobile terminals by connecting the smart card and some secret data to the terminal; and many more.

As the smart card is the token which uniquely identifies a subscriber in the network it is also an ideal container for all subscriber individual data that is not related to the authentication procedure. Such data could either be personal data stored on the smart card by the end user or data that is used by specific applications (e.g. application toolkit services) running on the smart card and managed by the card or application issuer. Personalized applications that behave according to the individual needs or characteristics of every single end user can be envisaged. Personalized applications are commonly used within the Internet. Once subscribed to a bookstore and as soon as some articles have been purchased, an individual customer profile is created. This customer profile

Figure 10. Personalisation of terminal by smart card



will then initiate the application to behave in a personalized way when visiting the bookstore again. Starting with a personal greeting reaching to a tailored product offering for the customer to more easily find the appropriate book. Services of this kind are also introduced to the mobile world. The creation of Web pages for the terminal that can be displayed on the screen mirror the Internet behaviour on the mobile world.

As smart card applications are used for billing services the security of the smart card has also been subject to attacks. Appropriate countermeasures were developed by the industry to fend off such attacks, for instance, power analysis on the chip and measurement of the required power consumption during dedicated operations could be used to determine security relevant data. Random waiting periods to the processor and the use of processors with constant and steady power consumption were some of the introduced countermeasures.

Personal Identity and Data Management

The smart card is a removable device allowing an easy transfer of personal data (such as phone numbers) from one mobile terminal to another and of course to allow plastic roaming (in different networks, see also sections above).

Life cycles of terminals and smart cards, and the necessity to change the terminal from time to time for reasons of functionality, forces the end user to also transfer all personal data from the existing terminal to the new terminal. This has to be done whilst keeping all relevant authentication parameters and mechanisms, as access to the network still needs to be provided. In order to ensure a smooth migration from one terminal to another without any tools, the smart card and all personal data stored on it can be easily ported from one terminal to another by simply taking it out of the reader and putting it into the reader of the new terminal. With the increasing number of services and a resulting increasing amount of personal data to be stored, the migration became more and more inconvenient and complex for the user if the data was stored in the terminal. Dedicated connectivity parameters and settings for the network and some services can easily be ported as well as additional keys and PINs for high security applications such as mobile stock trading and virtual private network access. The same applies to personal data such as phone numbers, email addresses and templates stored on the smart card. Equally, applications that are stored on the

smart card remain available to the end user when the exiting terminal is replaced by a new one. For the network operator, branding information and connectivity parameters that are stored on the smart card are available on any terminal.

Value Added Service Platform

The smart card is a platform for value added services. As mentioned above, a standardized framework to develop services exists for the smart card. This framework is supported by virtually all terminals in an harmonized way. The framework comprises the Card Application Toolkit (SIM application toolkit in 2G and USIM application toolkit in 3G) allowing the creation of applications that utilize the capabilities of the terminal with regards to the user interface (keyboard, display...) and the radio interface and bearers (set up call, set up GPRS connections...). Value added services starting with simple information services such as weather information or horoscope can be rolled out. But also more complex applications involving 3rd parties (such as a financial institution or a WLAN operator) are generated and can be easily put on the smart card. The SIM Application Toolkit and/or USIM Application Toolkit provide mechanisms to also retrieve location information and allow the development of location-based services. Dedicated routines and procedures on the smart card are defined for the execution of call control features that allow the restricted or dedicated use of subscriptions related to the calls that can be set up. In order to offer flexible solutions, a dynamic approach exists for the interpretation or execution of applications on the smart card. A smart card browser was defined that allows development of applications that are stored on a Web server and that are dynamically downloaded to the smart card and interpreted accordingly. Besides the evolution of the smart card hardware, the toolkit is one of the areas that has developed most rapidly in order to reflect the new technologies and make them available from and for smart card applications. All major smart card vendors offer development tools to generate applications based on the card application toolkit standards by even providing a graphic interface to the developer. Adding the Java CardTM API to the smart card allowed an even faster and easier way to develop applications for the smart card (see also section 4.2).

Today's smart cards offer true multi-application capabilities. New generation smart cards are multi-application capable. This means that completely independent applications can run on the same physical smart card. Whereas in the "old" days the smart cards contained one application and the SIM was a specification of both the logical and the physical characteristics, the new concept is to separate the physical aspects from the application specific characteristics. This allows the combination of completely independent applications on the same smart card platform. For example, a credit card application could be combined with a USIM application on the same smart card. These applications are completely independent from each other. That means that the credit card application (e.g. EMV) does not rely on the mechanisms provided by the USIM application. Whereas today a banking application uses the bearers offered by the mobile communication system, an infrared interface could be used for the payment transaction if the credit card application was separate. The SIM application (remember, the SIM is now an application and not a complete smart card anymore) or any other application may be located on the same card. The maximum number of applications running on the same smart card is only limited by the available memory space. The multi-application capability of the smart card enables network operators to prepare for a smooth transition from 2G to 3G by providing both the 2G and the 3G application on the same smart card. This concept will permit the network operators to also smoothly migrate from 3G to any further generation system. Considering the technical capabilities and the increase in memory available on the smart card, concepts are

developed to rent out a specific part of the memory on the card to a third party. This allows them to offer their services to the customer base of the network operator. Any third party could provide applications such as ticketing or banking in case the network operator itself decided to concentrate on other core business rather than offering all services and applications by himself.

Service Deployment and Management

The smart card is owned by the network operator. Even though the smart card is physically possessed by the end user, it is still the property of the network operator. As the smart card is the only token being under full control of the network operator, the network operator is independent from the type and model of mobile terminal used by the customer. The smart card enables the network operator or service provider to offer applications according to standardized methods on all different types of mobile terminals. These services will then be available to practically all customers. Operators are putting one generic service enabling application on the smart card that can be managed and enhanced remotely. This generic service application is a flexible tool that adjusts itself depending on the terminal in use and depending on the subscribers needs. It is a platform for specific applications such as infoservices, banking or any other type of application. This ensures that the network operator can base at least some or parts of their most important services on the smart card. So, as more and more smart terminals and personal digital assistants (PDAs) with mobile communications capabilities enter the market, the smart card with its features will still be the property of the network operator and remains independent of the terminal.

Services are easily deployed. Service deployment need to be fast and cost efficient. Bringing new applications and services into the market is essential and has direct impact on revenue. The Java CardTM concept allows different parties to develop applications that can be placed on the smart card for mobile communications. Rapid development times represent one important factor for the rapid supply of new services. More complex, however, is the provisioning and management of these services in the field. Services need to be brought to the customer (see also section 4.2). Three principles for the deployment of services are:

- Delivery of smart cards already containing the new application;
- Update of the smart card in the shop / at the point of sale; this update is done either offline (by an autonomous PC system) or online (connected to a background server system); and
- Update of the smart card by using over the air mechanisms.

Logistics for the shipment of smart cards is well established in the industry. In general, network operators order only a limited number of different smart card types. This makes the deployment of applications based on the smart card rather simple. It can be done by using the established way of delivering the smart cards to the end user via the network operator's shop or by fulfilment services offered by the smart card manufacturer. Application could be loaded onto the smart card already during the production phase and delivered to the end user already containing all relevant services. In case new applications need to be brought onto cards that are already deployed in the field, they can be updated and applications can be uploaded on the smart card on the air interface (over the air — OTA) by using standardized mechanisms (e.g. TS 03.48/TS 23.048) and available "bearers" such as SMS or GPRS. The configuration of the smart card can also be done at the point of sale using dedicated tools to update the card in a card reader. At the point of sale the smart card is removed from the terminal and inserted into

a card reader connected to a personal computer. The employee in the network operators' shop is then able to configure the card according to the subscribers' requirements and the available new applications.

OTA capabilities of the smart card include the possibility to remotely manage the file system and the applications of the smart card. By means of existing bearers (such as SMS and GPRS) an OTA server is able to create, delete and modify files on the smart card. The smart card contents can be read out and sent to the background server, whilst new file contents can be written on the smart card from the background server. The OTA mechanisms can be utilized to update provisioning information for services such as MMS and to perform adjustment of parameters for bearers such as GPRS. Additionally, in the area of user equipment management (UEM) some relevant terminal related information could be stored and maintained on the smart card. The OTA capabilities even permit entire applications to be downloaded onto, and managed on, the smart card. Management of applications include provision of the application to the customer, update of the application or the application status, activation or deactivation due to business or technical reasons, re-activation, and finally deletion from the card. The OTA server system is capable of managing each individual card by storing information on the applications and services loaded, the memory space available for further applications any many more. This provides the marketing departments of the network operator a powerful tool to determine which kind of service is accepted, how frequently it is being used and what the effort was to deploy a new service to which category of subscriber.

Interworking Aspects

Platform for the interworking of different access technologies such as WLAN and 3GPP networks. The smart card offers a secure authentication platform to access networks. As wireless communication is a huge success in the market new types of wireless communication systems have been developed and further technologies may exist in the future. WiFi hotspots in public areas require secure authentication just as it is required in 2G and 3G networks. Billing for the usage of a service is only possible if there is a way to secure the network against unauthorized access. As the smart card is already used for authentication of the subscriber by the 2G or 3G operator, an enhancement of either the smart card, the authentication system or both is a natural step. This enhancement could then be used to authenticate the user and to grant him access to the WLAN access network. Combining the authentication functionality of different systems on one single smart card and using the same terminal to access different access technologies (e.g. whichever is cheaper to fulfil the end users needs) makes the smart card an interworking token. Furthermore (as mentioned previously) the smart card enables the issuer to prepare for a smooth transition from 2G to 3G and beyond to nG systems by providing the related authentication application (WLAN, 2G, 3G...) on the smart card. Interworking aspects between the WLAN systems and the existing 2G/3G networks are currently investigated by standardization groups. For the smart card two options are considered. Firstly, to use the existing SIM and/or USIM application to provide secure access to WLANs, for example, by implementing the IETF standards Extensible Authentication Protocol EAP-SIM or EAP-AKA. Secondly, to develop a new independent application that reside in parallel to the SIM and the USIM application on the UICC, the smart card platform. Which ever way will be decided, the smart will be an integral part within the system.

IP Multimedia Systems as specified by 3GPP allow IP based services to be used within the 3G context in the mobile world. The authentication to this kind of IP based services such as multimedia video streaming is being done by means of an application based on the smart card. The IMS Sub-

scriber Identity Module (ISIM), as defined in TS 31.103 "Characteristics of the ISIM application" allows secure access to IM services. This opened the way of the smart card into the authentication for IP based services and underlines the importance of the smart card for secure authentication of both subscribers and networks.

Last but not least, the smart card is fully standardized and harmonized. All required specifications for the implementation of services based on the smart card and for the smart card itself are available and mature. This allows an interoperable and harmonized implementation of services. Harmonization can be interpreted as at least twofold. The first interpretation is the harmonization of applications within the network operator group. Core applications can be defined for each subsidiary of the network operator that can then be easily adjusted according to individual local needs. The second interpretation is the harmonization of services among terminals. Different terminals (and PDAs etc.) behave in the same specified way and execute the applications stored on the smart card in the same way. As there is a much broader diversification of terminal operating systems compared to only one smart card operating system, network operators would have to develop their services for each different type of terminal operating system.

Current Work Areas and Next Generation Smart Card

The above described set of features is not exhaustive, it covers the most distinctive ones, allowing the derivation of an unlimited number of services and applications. In addition to the features described so far, some concrete enhancements are currently being discussed in the respective standardization groups, and are briefly outlined below:

- Multimedia Broadcast/Multicast Service (MBMS) security: In order to protect the Multimedia Broadcast/Multicast Service some security mechanisms need to be implemented to prevent unauthorized users to get access to the MBMS service. As the smart card is a proven token for containing security functionality, these mechanisms and features are being enhanced in order to also provide the requested level of security for the new service.
- Voice Group Call Services: The one-to-one communication channel within the 2G and 3G network need to be extended to a many-to-many communication. The related network authentication of the multiple members of the group that want to communicate to each other at the same time as well as the authentication of the individual members within the group is crucial. The classic authentication procedure based on the smart card needs to be enhanced to allow users to join or perform a group call. This service is especially important for emergency services and is already available in systems other than 2G or 3G, such as TETRA or TETRApol.
- UICC Security Services Module (USSM): The UICC may contain several applications, each dealing with keys and realizing crypto-services. Some keys might be shareable even though there are no standardized mechanisms to share keys and indicate allowed functions to authorized applications. To allow applications on the UICC to use shared security objects, it is essential to introduce standardize mechanisms to administer and to use these shared objects on the UICC. The USSM will consist of security objects (keys, PINs, etc.) including information on allowed functions and authorized applications, an API for administrative functions to administer objects of the USSM, and an API for cryptographic functions (non-administrative) to be used by applications.
- Advanced Communication: The demands on the classic smart card communication

channel are driving it beyond its design and intents. Bearer independent protocols allow UICC applications to access communication channels whose native speed is greater than needed for classic terminal/UICC traffic. Higher level protocols such as network and transport protocols are starting to appear on some bearers. The channel is multiplexed between multiple applications using a number of different techniques. This work item considers the evolution of the smart card communication channel with respect to transfer rate, size and protocol.

- Large Files: Applications such as multimedia or identification applications require data storage capabilities that are reaching the current file size maximum of 65,535 bytes. Increasing the maximum file size beyond this limit impacts a wide range of size fields, parameters and commands within the standards. This work item will upgrade the standard in a synchronized and harmonized manner by providing backward compatibility.
- Reduced voltage class: The aim of this work item is to respect the requirement to prolong the lifetime of the terminal's battery. In order to address this need the electrical characteristics of a new (1.2V) UICC-ME interface (the Class D operating conditions) has been defined and is awaiting final agreement.
- Next Generation UICC: This work item identifies and evaluates commercially-viable hardware and software technologies needed to define a next generation smart card platform. The scope of the work item includes, but is not limited to, the possible role of memory management units, ASIC co-processors, proof-carrying code, new memory architectures, natural clocks, multi-tasking operating systems, embedded electrical sources, free-running oscillators, integrated biometrics sensors, universal byte codes, alternative form factors, new chip carriers, and high-speed communication channels. One of the essential characteristics of these new technologies that will be catalogued is their impact—positive and negative—on the security of the UICC platform.

OUTLOOK

In Information Technology memory space and processing power has never been enough. This holds for the personal computer as well as for almost every component in the system. Displays have been too small, the resolution not good enough, battery lifetime was too short and the size of the battery too large. The smart card was therefore also seen as offering too little memory and not enough processing power. Today's multimedia applications and services demand more and more memory space for the storage of application or service related data such as pictures, movies and configuration data. But one thing needs to be remembered when demanding more memory, more processing power or more whatever: the smart card today already has the processing power of the early personal computers (PCs) concentrated on just a few square millimeters of silicon. And it is not the size but the effectiveness and the clever design of applications and services that make the service successful (and, of course, its market relevance).

The capabilities of the smart card will increase both physically (hardware) and logically (features implemented as software). From a hardware point of view, which has been the main constraining factor for the smart card, new technologies will enable a noticeable increase of memory capacity. Memory sizes of megabytes seem to be possible in near term and continued chip development appears to promise even multiple megabyte capacity in the mid term. The development of memory cards for digital and video cameras, together with the size reduction that is envisaged for these com-

ponents, gives a good indication of the potential for smart cards.

As in the PC world, chip technologies and processor capacities continue to evolve. We can foresee smart cards with the capability to execute complex security calculations such as asymmetric en- and decryption. Equally, the development costs for new high-end smart cards can be compared with those in the PC area. The prices of the PCs are more or less always stable, whilst the capabilities of the system, such as memory, processing power and software packages included, steadily increase.

Much interest is shown in smart cards with a contact-less interface for areas such as public transportation. The combination of contact and contact-less technology in mobile communications devices would also enable future services to be deployed by means of the mobile terminal. New standards such as near field communication (NFC) where an active component in the terminal could act as a card reader for contact-less cards could dramatically enhance the uses of the smart card.

Where the hardware capabilities of the smart card permit, and suitable opportunities exist, merging markets are foreseen. Transferring applications that reside on physically different smart cards onto the new multi application smart card makes sense for a number of applications, especially those that require the possibility to communicate with external entities or a personal card reader. Such applications are the ideal candidates to be incorporated into the smart card for mobile communications.

Mobile communication is a growing market: new applications and new areas of use are established regularly. Niche markets or markets that traditionally could not be served due to physical limitations of current technology will be served by new products. Telematics, just to give one example, is one of the areas that could not be served with a full range capability due to the limitation of the temperature range of existing hardware and chip products. The continuous development of these components is expected to allow increases in the temperature range to a degree that is acceptable in telematic systems.

Further reductions in size or even a completely new design and architecture of the smart card is part of the investigations carried out by the ETSI Project Smart Card Platform. The UICC next generation is a work item that has been set up by the committee to search for a smart card solution and architecture that will meet the future requirements. As well as possible size reductions, the definition of new communication protocols, new file systems and other enhancements are being considered as required by future markets. A new file system that reflects the capabilities of the smart card to store megabytes of data will enable the smart card to ease the management of data storage on the card. This will make it much easier for the issuer of the card to maintain and to manage the data related to the services on the card.

The evolution from a one-application card to a multi-application card is paralleled by evolution of the operating system. Multi-tasking operating systems and other state of the art personal computer features are required and may be added one by one to the smart card. This allows the addition of further applications that run in parallel or in background mode.

A further trend concerns the existence of multiple different communications systems such as 2G, 3G, WLAN, DECT and others. These will lead to a merger of the applications for the end user, and thus for end user devices. The token containing the authentication data and performing the trusted operations in the network for the end user will contain different authentication and network access applications. These applications may share some of the information stored on the smart card. This puts the smart card in a role of providing a smooth roaming from one access technology to another. The smart card acts as the medium that keeps the subscriber connected and that provides the means of interworking between

the different networks. Smart card ownership in such a scenario is crucial as it means that the churn of subscribers could be reduced: the subscriber could be more easily added as a customer for a new access technology by simply putting the required information on the smart card to access the new network. The network operator issuing the SIM and also operating WiFi hotspots could very easily enable the subscriber to access the WLAN at the airport by simply re-using the SIM card for the authentication. This makes it more difficult for new providers to offer such a service.

The pressure in the industry to generate additional revenues is extremely high. Companies have invested massively in the purchase of licenses and the setting up of new generation networks. Further enhancements to the network infrastructure and the introduction of further generations of communication systems will be equally expensive, so even more revenue has to be generated in order to be prepared for those future systems. This revenue will be generated by those companies who succeed in providing the appropriate services that are accepted by the consumers. Including the power of the smart card in the concept of such services may well be one major step towards achieving that goal.

REFERENCES

SCP-010141 work item description on "Advanced communication"

SCP-010142 work item description on "Large files"

SCP-010265 work item description on "Introduction of a new voltage class"

SCP-020185 work item description on "UICC next generation"

SCP-030281 work item description on "USSM (UICC Security Services Module)"

TS 11.11 " Specification of the SIM-ME Interface"

TS 31.101 "UICC-Terminal Interface; Physical and Logical Characteristics"

TS 31.102 "Characteristics of the USIM Application"

TR 31.900 "SIM/USIM internal and external interworking"

Klaus, V. (2001). In Hillebrand, F. (Ed.). GSM and UMTS: The Creation of Global Mobile Communication. Wiley Europe.

Rankl, W. & Effing, W. (2003). Smart Card Handbook. London: John Wiley & Sons.

ENDNOTES

Trademark Information

This work was previously published in Mobile and Wireless Systems Beyond 3G: Managing New Business Opportunities, edited by M. Pagani, pp. 221-253, copyright 2005 by IRM Press (an imprint of IGI Global).

Chapter 3.22
Unobtrusive Movement Interaction for Mobile Devices

Panu Korpipää
Finwe Ltd., Finland

Jukka Linjama
Nokia, Finland

Juha Kela
Finwe Ltd., Finland

Tapani Rantakokko
Finwe Ltd., Finland

ABSTRACT

Gesture control of mobile devices is an emerging user interaction modality. Large-scale deployment has been delayed by two main technical challenges: detecting gestures reliably and power consumption. There have also been user-experience-related challenges, such as indicating the start of a gesture, social acceptance, and feedback on the gesture detection status. This chapter evaluates a solution for the main challenges: an event-based movement interaction modality, tapping, that emphasizes minimal user effort in interacting with a mobile device. The technical feasibility of the interaction method is examined with a smartphone equipped with a sensor interaction cover, utilizing an enabling software framework. The reliability of detecting tapping is evaluated by analyzing a dataset collected with the smartphone prototype. Overall, the results suggest that detecting tapping is reliable enough for practical applications in mobile computing when the interaction is performed in a stationary situation.

INTRODUCTION

The source of innovations in a mobile device user interface lies in combinations of input and

output technologies that match the user's needs. In the mobile context, movement sensing, and haptic feedback as its counterpart, offers a new dimension to multimodal interactions. There are use cases where traditional interaction modalities are insufficient, for example, when the device is placed in a pocket or a holster, or if the user is wearing gloves. In these situations the user cannot press or see buttons to interact with the device. Instead, small motion gestures can be used as a limited, but convenient, control modality. The movement of the device can be captured with a 3-axis accelerometer, and the resulting acceleration signal can be used to detect the movement patterns for controlling the device.

One of the main questions in the application of a movement-based interface is how to distinguish gesture movements the user performs from those movements that are produced by various other activities while carrying and using the device. Reliability can be argued to be the most important challenge in developing a mobile device gesture interface. This chapter presents a reliability evaluation of an unobtrusive event-based gesture interface by analyzing a multiuser dataset collected with a smartphone prototype. Another main challenge has been the relatively high power consumption from the continuous measurement of acceleration, which is not acceptable in mobile devices. Novel accelerometers are capable of producing interrupts based on exceeded thresholds; therefore, the detection, initiated by a hardware interrupt, can be implemented as event based and low power. The technical feasibility of event-based tapping detection is examined with a smartphone equipped with a sensor interaction cover, Figure 1, and an enabling software framework. Furthermore, the chapter addresses the issue of flexibly customizing the gesture interface and feedback modalities relevant to aiding the user.

There are various ways of implementing a gesture interface. This chapter focuses on analyzing the tapping interaction, which shows potential as a significant application of accelerometers in future mobile devices. More specifically, the chapter addresses the movement pattern where the

Figure 1. Smartphone prototype equipped with the sensor interaction cover

user taps the device twice consecutively, which is called a double tap. With an implementation based on abstractions initiated by sensor-driven interrupts, the aim is a low-power, reliable, and customizable user interaction modality.

Gestures can be detected either from a continuous stream or discrete segments of sensor data. In detection from discrete segments, gesture start and end are explicitly marked with a button instead of a continuous flow of device movements. From the usability perspective, interaction without explicit marking is preferred, in general, since it requires less attention from the user. However, continuous data streaming and execution of the gesture detection algorithm requires continuous data processing, which normally consumes battery power.

The development in digital acceleration sensor technology enables the integration of programmable interrupt-based solutions that can operate with low current consumption. Such sensors generate interrupts when acceleration on a spatial axis is over or below a set threshold level. Hence, movement detection algorithms, initiated by an exceeded threshold, can be implemented as event based instead of continuously processing a stream of data. The processing load at the mobile device side is similarly reduced since the operating system is woken up less frequently. This development opens up new possibilities for practical application of the technology in mass products such as mobile phones.

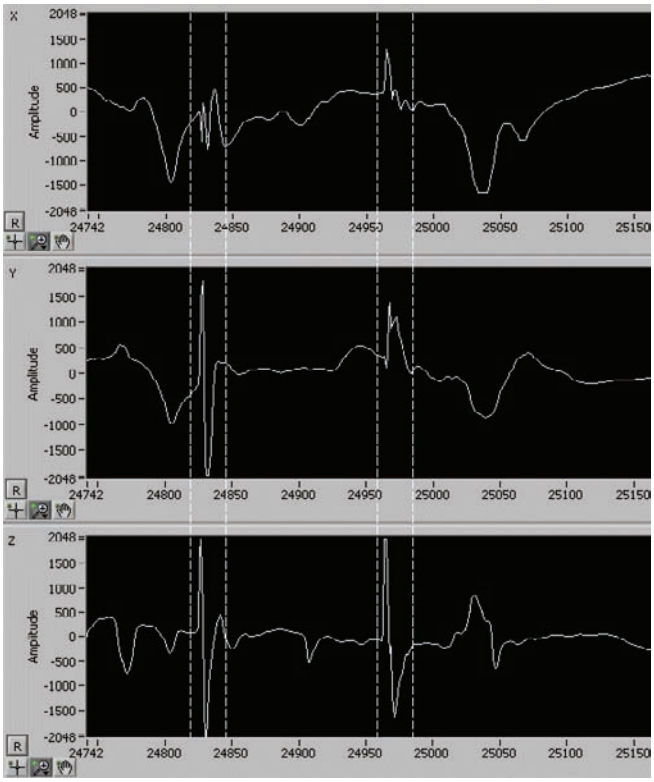
The distinguishable form of the tapping pattern, processed after the event threshold, is the basis for the potential reliability of detecting them, even when the detection process is continuously active, Figure 2. By contrast, free-form gesture recognition has a much wider problem setting, requiring a more complex model of the gesture and thus, heavier processing load, making continuous processing much more challenging, especially in mobile devices.

This chapter publishes the first statistical performance evaluation based on a dataset that characterizes the reliability of user-independent tapping interaction in mobile phones. Moreover, the sensitivity of the method to misrecognitions is evaluated with scenarios consisting of various activities. As an introductory topic, applying a smartphone equipped with sensor interaction cover, customization, and feedback of the addressed interaction modality are discussed.

BACKGROUND

In acceleration sensor-based gesture recognition, gestures are detected either from a continuous stream or from discrete segments of sensor data. While this chapter addresses the detection of movement patterns from a continuous stream, there are a lot of studies in the literature on gesture recognition from discrete segments (Feldman, Tapia, Sadi, Maes, & Schmandt, 2005; Mäntyjärvi, Kela, Korpipää, & Kallio, 2004). Specifically, acceleration sensors have been applied in user-trainable and pretrained machine-learning-based gesture recognition systems (Kallio, Kela, Korpipää, & Mäntyjärvi, 2006; Kela, Korpipää, Mäntyjärvi, Kallio, Savino, Jozzo, & Di Marca, 2006). Free-form gesture recognition still has a limitation; it requires an explicit marking of the gesture, for example, with a button, and longer duration gestures to increase the recognition accuracy. Hence the interaction requires more user effort, and gesturing can be socially obtrusive. However, despite the possible obtrusiveness when applied in public places, free-form gestures also have a wide range of potential uses in other settings, such as games, home electronics control, and so forth, where social acceptance does not limit the use of the modality. The social aspect, distinctively important in the mobile usage context, has been addressed by Linjama et al. (Linjama, Häkkilä, & Ronkainen, 2005), Rekimoto (2001), and Ronkainen et al. (Ronkainen, Häkkilä, Kaleva, Colley, & Linjama, 2007). Based on the literature, it can be extrapolated that, when

Figure 2. Three channels of acceleration data on a double tap performed while walking. The Z axis has two distinguishable spikes in this double tap.



performed with a mobile device such as a phone, smaller gestures are considered more socially acceptable than large ones.

This chapter especially advocates the unobtrusiveness of the interaction; gestures as small and as unnoticeable as possible are preferred, assuming they are more acceptable by the users (Linjama et al., 2005). Examples of possibly useful small-scale gestures include shaking the device, for example, Levin and Yarin (1999), and swinging it from side to side (Sawada, Uta, & Hashimoto, 1999). However, both of these interaction methods can be considered quite noticeable, regardless of scale. Shaking also raises the question of how many repetitions of the shake movement are required until a shake is recognized. A simple accelerometer-based tilting control has been discussed in the literature in many studies over the years, for example, Rekimoto (1996), but also recently, for example, combining tilt and vibrotactile feedback (Oakley, Ängeslevä, Hughes, & O'Modhrain, 2004), scrolling, and switching between landscape and portrait display orientations (Hinckley, Pierce, Horvitz, & Sinclair, 2005). Tilting is another potentially unobtrusive, and very simple to implement, movement-based interaction modality to be applied in carefully selected use cases in mobile computing.

A minimalist extreme in hand gestures is tapping the mobile device, first introduced in Linjama and Kaaresoja (2004). Tapping only requires a small scale of device movement, and can be performed by finger or palm. The technological benefit is that tapping can be relatively straightforwardly captured with a 3-D accelerometer, since the resulting movement pattern has a distinguishable sharp spike form. The detection problem can be narrowed down by applying a small, predefined fixed set of movement patterns: tap events.

The unique usability benefit of the tap interaction is that it is discreet and can be used if the mobile device is located in a pocket or a backpack, since explicit marking is not needed. Furthermore, the user is not required to hold the device or see the keyboard to interact. A good example of a use case where tapping is useful can be found in the Nokia 5500 phone (Nokia, 2006): when a text message arrives, the user has 30 seconds to tap the phone twice and the message will be read aloud to the user. It is useful when the phone is in a pocket or on a belt, or the user is wearing gloves; the message can be read without first taking the phone into the hand and opening the keypad lock. Furthermore, tapping can be used as an additional modality. For instance, phone music player commands, such as play next or previous song, can be controlled by tapping on either side of the phone, which is convenient when the device is worn on a belt or in a pocket. Again, the user does not have to take the phone, open keypad lock, and press a button to perform the control action.

SENSOR INTERACTION COVER

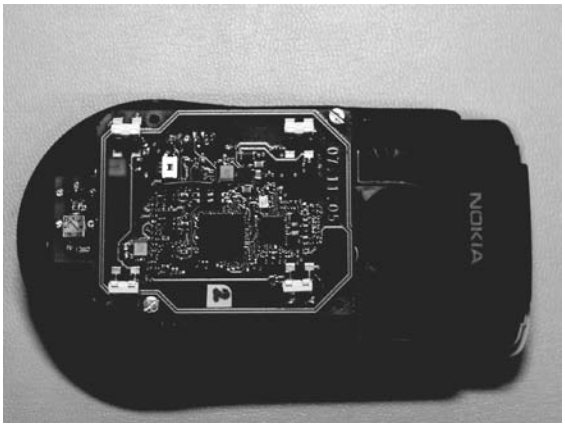
Interrupt-initiated abstracting of movement patterns can be performed using a separate microcontroller, or, ideally, it can be directly integrated in the sensor chip. A sensor interaction test platform was developed to experiment with the interaction concept. The platform consists of a Symbian S60 phone (Nokia 6630) equipped with a sensor and feedback cover attached to the back of the smartphone, Figure 3.

Inside the cover, the hardware includes a 3-D acceleration sensor (STMicroelectronics LIS3LV02DL), a microcontroller (Atmel), an NFC reader, blue LEDs, a buzzer, and a vibra motor, Figure 2. The board is two sided. Tap detection parameters and feedback configuration can be set to the microcontroller from the phone software. The tap detection algorithm and the feedback processing are performed in the cover microcontroller, and the cover transmits recognized tap events to the phone through USB. Thus, the communication between the cover and the phone, as well as power consumption, is minimized.

INTERACTION CUSTOMIZATION

Once sensor events are abstracted by the microcontroller and sent to the phone through the USB, they should activate the desired actions

Figure 3. Sensor interaction cover hardware

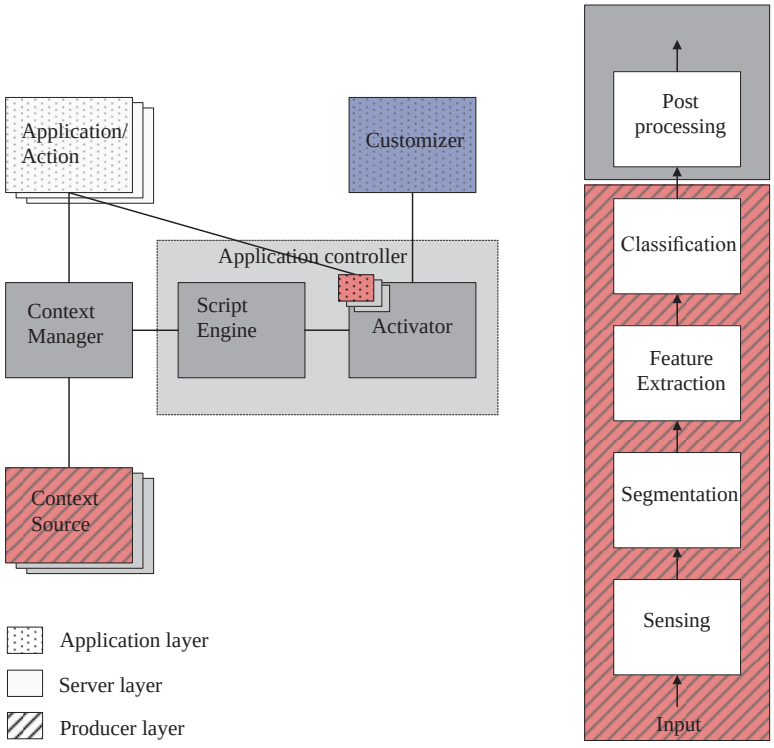


in the mobile phone. Flexibly connecting the abstracted sensor events to various application actions requires supporting middleware on the phone side. Instead of connecting an application directly to a device driver, the data is abstracted into a uniform representation applied through Context Framework.

Context Framework (CF) is a blackboard-based software framework for enabling and customizing situation-aware and sensor-based mobile applications (Korpipää 2005; Korpipää, Mäntyjärvi, Kela, Keränen, & Malm 2003). All interaction-related information, including implicit and explicit sensor-based inputs, is treated as context objects within the framework, expressed with a uniform vocabulary. An implemented instantiation of the framework is illustrated in Figure 4 (left-hand side). In this case, the sensor signal abstracting process functionality is on the microcontroller side, illustrated in Figure 4 (right-hand side flow diagram). Sensing, feature extraction, and classification are performed at the cover's microcontroller. Classified movement (context) events are sent over the USB to the phone side, where CF enables controlling any available application action based on the events.

The user can create desired context-action behavior with a mobile phone by creating

Figure 4. Context Framework (CF) architecture example instantiation (left), and pattern recognition flow (right)



XML-based rule scripts with the graphical UI of the Customizer. CF handles the background monitoring of context events and the triggering of actions according to the rules. The Application Controller facilitates the application control inference on behalf of the user or application, in other words, provides an inversion of control. The framework completely separates the management of sensor-based context events from application code and the hardware. Hence, by applying CF, no changes need to be made to existing mobile phone applications when they are augmented with sensor-based features.

In the case of tapping input, the events are abstracted into context objects by the sensor cover of the phone and delivered to CF. The application developer or the user interface designer can use the Customizer tool to define which application actions are executed by which abstracted sensor events. The definable actions include available feedback modalities, such as tactile, auditory, and visual indications. By creating rules with the Customizer tool, the user can define actions on an operating system level, or for a specific application, by setting a condition part of a rule to include a specific foreground application. For instance, the following accelerometer-based features were defined and executed simply as XML-based rule scripts:

- Playing the next or previous song in music player using double tap
- Activating display illumination using tap
- Unlocking the keypad using double tap

Figure 5 presents a series of screenshots from the Customizer tool, illustrating the definition of a rule that enables the user to unlock the keypad by double tapping the phone.

In Figure 5a, the user selects an action for the rule by navigating through the action type Phone. Keypad and selecting the action value KeypadUnlock. In Figure 5b, the user selects a trigger for the rule by navigating through context type Gesture and selecting context value DoubleTap. The first screenshot in Figure 5c shows the complete rule after the user has selected the elements. After the user selects the option Done, the rule script is generated, and the rule is activated and functional in the context framework. The second screenshot in Figure 5c shows the main rule view with the list of active rules. When the rule conditions are met, the Context Framework automatically performs the action.

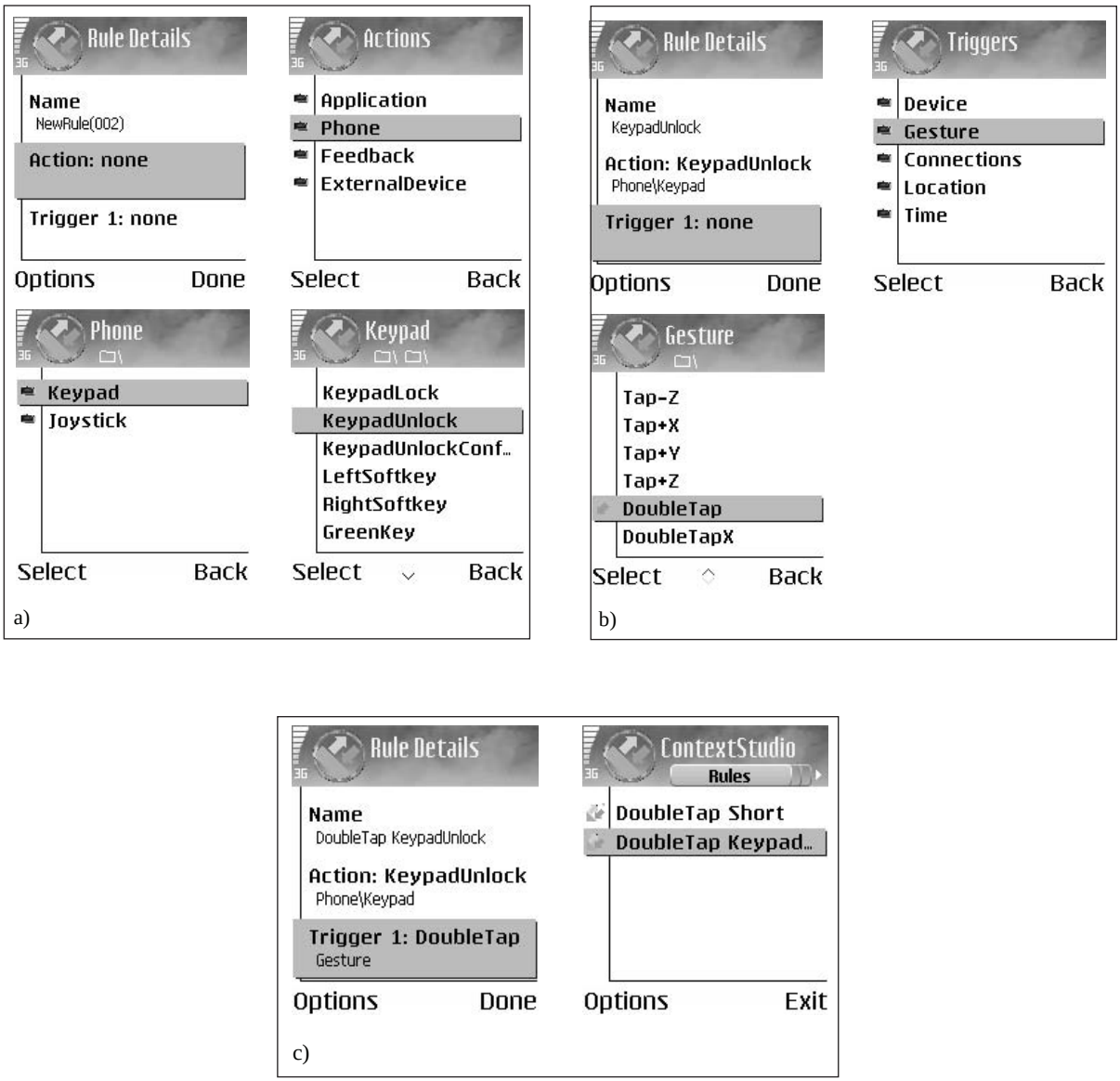
USE CASES AND USABILITY

Evaluating the general usability is an essential aspect in developing tapping interaction, in addition to evaluating the reliability and technical feasibility. As an extensive topic, however, it cannot fit within the scope of this chapter. The purpose of this section is to briefly discuss a few usability-related points as an overview of the experiments studying the usability of the novel interaction modality.

New interaction modalities, like tapping, have certain application areas where they add value, both in terms of utility (usability) and joy (fun of use). The only way of evaluating these aspects is to try the interaction elements in practice, with real hardware and applications. The smartphone sensor interaction cover serves this purpose. It enables the rapid prototyping and iterative development of interaction concepts and demonstrations. User experiences and feedback can be collected during various stages of development, early concept tests, peer evaluations, and end user tests in the lab and in the field.

A number of formal user tests, to be published separately, have been conducted with the smartphone prototype. The tests measure the potential usefulness of tap interaction with a standard Symbian Series 60 phone user interface. For instance, controlling a phone music player with tapping commands, in addition to existing keypad controls, has been studied. The results

Figure 5. Series of screenshots illustrating how to program the phone to open the keypad lock with a double tap.



from the studies indicate that it is very important to maintain consistency in case there are multiple different uses of tap interaction in several applications of the same device. Users may get confused if tapping is used for too many different purposes, such as muting the phone audio in one application and selecting the next or previous music track in another application. Furthermore, using different

tap directions, for example, tapping on the device top or the side for activating different controls in different applications, requires delivering specific instructions to the users.

User satisfaction, joy of use, has also been addressed in the usability tests. The enjoyability of the user experience is largely determined by the very details of the interaction; what kind of feedback elements support the user interaction. What is the metaphor behind the observed device behavior that the user learns when using new interaction modalities? All sense modalities in multimodal interaction must be addressed together.

CONTINUOUS DETECTION RELIABILITY: EXPERIMENTS

This study focuses on analyzing the reliability of detecting double taps in various usage situations. The experiments to be presented next aim to answer how accurately double taps can be detected in a general mobile usage setting and how many misrecognitions occur. The results should reflect an essential part of how feasible this interaction method could become, from the reliability viewpoint, when used in mobile phone applications. Detection accuracy is quantitatively analyzed based on acceleration data collected from users performing the interaction, and the results are discussed.

There were 11 users performing the interaction and the scenarios; 7 of the users were male and 4 were female, aged from 25 to 36 years. The subjects were selected randomly from acquaintances of the authors. The subjects were not interviewed and no subjective opinions were collected, only acceleration signals. Therefore, the limited variability of the subjects in the user group was assumed not to bias the results significantly.

There are two categories of use cases for continuous detection of movement events. In the first category, the detection process is initiated by a specific application or a situation, and is active for a certain time. In the other category, the detection process is always active. In the first category, the use cases can be designed such that misrecognitions, false positives, have a minimal effect. In the latter category, false positives usually have a more negative effect since they may result in incorrect operation. In both cases, the sensitivity of detection should yield enough correct recognitions, true positives, to be acceptable for the users.

This section describes the experiments aiming at evaluating how well the tapping interface performs from a statistical point of view, based on collected data. Detecting tapping events is a type of pattern recognition problem (Duda, Hart, & Stork, 2001), although not a very complex one. The aim of the data analysis during the development process was to reach optimal recognition of a double tap pattern, that is, to find detection algorithm parameters that produce a minimal number of false positives while maintaining a high percentage of true positives. The primary goal was to minimize misrecognitions. The algorithm should give the best results as an average when performed by multiple users, not just one specific user. In other words, the aim is to reach optimal user-independent detection accuracy. The experiment involved collecting a dataset on the target patterns performed by several users in controlled stationary conditions. Furthermore, data from several real-world scenarios containing various daily activities was collected to find out how often misrecognitions occurred.

Data Collecting

In order to evaluate the tapping detection reliability statistically, a sufficiently large dataset is required. Dataset size and the variation it contains are in direct relation to the evidence to support generalization. Data was collected in three stages with the sensor cover-equipped smartphone prototypes. The first stage involved exploring a wide set of activities by a user carrying 1-2 proto-

types to find out whether there were any specific activities that produced a lot of false positives. The dataset was collected by one user, and the total duration of the activities in the dataset was 5 hours 8 minutes.

The second stage involved having several users perform the target patterns in stationary controlled conditions involving no other activities. This dataset consisted of 11 users performing double-tap patterns. Data was collected in three categories, arranged by the user's skill level and the given advice. The user groups were beginner, people who had never heard about tapping, and advanced, people who knew or were informed about how the tapping user interaction works. There were six users in the beginner group. In the beginner group, the users were only given one piece of advice: to perform the tapping with their hand(s), not by tapping the phone on the table. The second group, five advanced users, were first told to use one hand for tapping and next to use both hands, that is, hold the phone with one hand and tap with the other. Figure 6 shows an example of both ways of tapping interaction.

There were five users, the same ones, in both of the advanced groups. In the three categories, each user performed a double tap 18 times, resulting in total target of 288 repetitions in the dataset. Repetitions were performed in phases of three repetitions and a break, during which the device was put on the table to avoid a routine speed-up and fixation on a certain way of interaction.

The third stage involved having several users perform scenarios involving real-world activities while carrying the prototype in their pocket. The purpose of this dataset was to find the occurrence of false positives during the scenarios, on average over multiple users. There were four to five users in each of the scenarios. The total length of the activity dataset was approximately 54 minutes. The tapping pattern has a sharp spike-form shape, and proper detection requires a relatively high sampling rate. Hence, the total amount of raw data collected for this experiment was approximately 68 megabytes.

Figure 6. Tapping performed with one hand (a) and with both hands (b)

EXPERIMENT RESULTS

The collected acceleration data was used to analyze the tapping interaction from multiple aspects. The experiments focused on a specific form of tapping, a double tap. Double tap means performing two consecutive taps within a certain short time span, much like a double click with a mouse. Each aspect of this interaction studied with the collected dataset is described in detail in this section. The experiments produced numerical measurements of the system's tapping detection accuracy. The measurements are briefly introduced here before presenting the results and analyzing them.

The first experiment was an initial pilot test, which was designed to count the number of double-tap patterns detected where they should not exist. In other words, the experiment measures the occurrence of false positives, which can be reported as a number per time unit. For example, the aim could be that there is no more than one false double-tap detection per hour.

False positives can also be represented in relation to how many patterns could be falsely detected from a dataset. The relative number of false positives in a dataset can be given by dividing the occurrence of all detected false positive patterns with all segments of data where there should not be a detected pattern. Here a segment is defined as the maximum time span required to detect one pattern. For example, for double tap pattern the maximum allowed duration is 1.1 seconds. This is due to the algorithm wait time for the second tap to appear after the first one. For instance, in a dataset of 110 seconds, there are 100 segments that could potentially contain a double tap. One false double tap in that dataset would result in one percentage of false positives.

True positive means a correctly detected pattern, for example, a double tap is detected correctly when it is performed by the user. The relative occurrence of true positives can be given by dividing all detected true positive patterns by all actually performed true patterns in a dataset.

Pilot Test

The goal of the pilot test was to explore whether some of the randomly selected ordinary daily activities would produce a high occurrence of false positives. This experiment did not contain any actual double taps performed by the user. The user was assigned to carry one or two prototypes in a pocket during various daily activities, for example, random outdoor activities (cleaning the yard, commuting, driving a car, walking, jogging, biking, cross-country skiing, and roller-skating). The users were free to select which clothes to wear and which pockets to carry the devices in. The tasks were given as, for example, "take the phone with you and go jogging." Table 1 summarizes the results of this test.

There were several activities that did not produce any false positives, such as jogging, various outdoor activities, biking, going for lunch, and roller-skating. The activity that produced the most false positives was cross-country skiing.

Overall, the test indicated that potential problem areas are accidental tapping by hand, ski stick, backpack, and so forth, and when the phone is laying freely on a moving and trembling surface such as a car dashboard. After the pilot test, the detection algorithm and parameters were adjusted to reduce the misrecognitions.

Stationary Conditions

Next, an experiment was performed in controlled stationary conditions. The purpose of the experiment was firstly to gain validation of how well the target patterns are recognized in a stationary situation when there are no external disturbances. Secondly, it is important to know whether there are differences between two groups of users when one has no idea what a tapping interface is and the other has prior knowledge of how to interact with

Table 1. Occurrences of double-tap false positives during random daily activities

Activity	**Phone numbers, placement**	**Duration (min)**	**False positives**
Commuting (dressing, driving, walking, stairs up, stairs down, office)	2, left and right lower jacket pocket	28	2
Travel by car, tarmac road	1, dashboard	70	1
Travel by car, rough gravel road	1, dashboard	20	2
Jogging	2, left and right jacket chest pocket	5	0
Cross-country skiing (walking, changing, skiing, walking, undressing)	2, jacket pocket, backpack	75	7
Outdoor activities (removing snow, walking, putting bike in storage)	2, left and right jacket chest pocket	5	0
Biking (gravel road and tarmac road)	2, jacket pockets	10	0
Going for lunch (stairs down, lunch, walking, stairs up)	1, jeans pocket	25	0
Roller skating	1, loose short trousers front pocket	35	0
Roller skating with sticks	1, pants front pocket	35	1
Total		**5 hours 8 min**	**13**

tapping. The results indicate different variations in the first-time use of tapping in terms of gesture signal waveform and the detection accuracy. Thirdly, the results show whether there are major differences between individual users, whether the interaction is equally assimilated by all users or if there are some individuals that cannot use the method as well. Finally, the interaction by tapping can be performed by using one hand or both hands, and the results indicate which is preferred from the reliability point of view with the evaluated algorithm.

The results can be calculated in two ways: the interaction can be allowed from any of the three axes, or from one selected axis only. In most single application use cases, the direction of tapping is known in advance and can be restricted. For example, music player next and previous commands can only be initiated with a tap on either side of the phone, by utilizing only the x-axis while disregarding the others. Hence, depending on the use case, it is feasible to filter the data from one or two other axes and apply the signal from one axis only. The results are first presented for 3-axis detection, Table 2.

The results show that double taps can be detected fairly well in stationary conditions, except in the beginner group. The difference between the beginner and advanced user groups is quite large, which suggests that first-time users may

Table 2. Recognition rate in stationary situation for various user groups in 3-axis detection

User group	Users	True positive %
Beginner	6	55.2
Advanced one hand	5	90.6
Advanced both hands	5	90.2

have trouble when starting to apply the method if they are not properly informed. There were also distinct differences between the individual beginner users, Table 3.

The data from the beginner users that produced low accuracies revealed that they performed the taps too lightly. Half of the beginner users chose to perform the tapping with one hand, and half with both hands. One beginner user tapped the top of the device and one the bottom, others from the side. The two beginner users that tapped with one hand had the zero results. The recollection from the actual test situation and data visualization confirm that the two one-hand users having a zero result only touched the device very lightly instead of properly tapping it. In other words, the first-time users' low performance is partly an algorithm sensitivity issue, but most importantly it is due to the lack of information the user has on how to do the tapping in the first place. The results can be improved by modifying the parameters to be more sensitive, but then the false positives

Table 3. Recognition rate in a stationary situation for each individual user in 3-axis detection

User group	User1	User2	User3	User4	User5	User6	Total
Beginner	94.4	0	77.8	88.9	0	73.3	**55.2**

User group	User1	User2	User3	User4	User5	Total
Advanced one hand	88.2	73.3	100	88.2	100	**90.6**
Advanced both hands	100	50.0	100	100	100	**90.2**
Total	**94.1**	**61.7**	**100**	**94.3**	**100**	**90.4**

tend to increase. The most straightforward way to improve the result is simply to advise the beginner users to tap with the correct intensity. Feedback is one way of giving immediate information to the user.

It must be noted that this experiment produced no information on the learning curve; it simply provides data on how differently first-time users may perform the gesture. There was no feedback or interaction in the test to guide the user on how to improve. In this sense it was a "blind" blank test to examine different users' approaches to performing a double tap, as interpreted from the signal waveform and the resulting detection accuracy. In a normal usage situation, the user would learn that too light taps do not cause the desired operation, and would likely either modify their behavior or abandon the method. In this test the users did not know that they tapped too lightly and thus, could not know how to change their tapping style.

The results in Tables 2 and 3 present the results for a setting where double taps from any direction are allowed. Table 4 presents results where only one predetermined axis signal is applied to detect a double tap. A significant increase in detection accuracy is evident. Furthermore, it is likely to reduce the occurrence of false positives, although it was not tested in this study. In light of the results, it is preferable to restrict the detection axis whenever is possible.

The results indicate that tapping is detected slightly more accurately when performed with one hand in 3-axis detection. In 1-axis detection, the accuracy is slightly better when performed with both hands. However, statistically, a conclusion cannot yet be drawn with this dataset on which way of tapping is more reliable.

Table 4. Advanced user recognition rate in a stationary situation in 1-axis detection

User group	Users	True positive %
Advanced one hand	5	95.3
Advanced both hands	5	98.8
Total	**5**	**97.0**

False Positives - Multiple Users

The purpose of the experiment with mobile scenarios was to find the occurrence of false positives during the selected common daily activities: walking, walking up stairs, jogging, and roller-skating. Furthermore, the scenarios were performed by multiple users in order to address the issue independent of the user. The scenarios in the experiment were designed to address a usage situation where the phone is in the user's pocket and the user could tap the phone from any direction. The users wore their own clothes and were free to select where to put the phone during the test. No other hard objects were allowed in the same pocket.

The results show that the number of double tap false positives was zero during the total of 54 minutes of activity data. By adjusting the algorithm parameters to more sensitive (which also increased true positives in the stationary test), false positives started to occur. The most false positives occurred on stairs. However, the parameter set that produced zero misrecognitions was generally perceived as sensitive enough, even though there were beginner users who would have benefited from increased sensitivity.

Summary of Results

Overall, the results based on the collected data, Table 5, indicate that detection is reliable enough for practical applications in mobile computing when the user performs the interaction in a

Table 5. Overview of the test results

Test	Users	True positive %
Beginner	6	55.2
Advanced one hand 3 axis	5	90.6
Advanced both hands 3 axis	5	90.2
Advanced one hand 1 axis	5	95.3
Advanced both hands 1 axis	5	98.8

stationary situation. Moreover, the number of false positives is low enough for types of mobile applications with at least a restricted scope. The results have significance for commercial applications built on use cases that have a clear usability advantage from the tapping interaction.

The results also show that there is room for improvement. This especially concerns the usability aspect of first-time use. An important question is how to give instruction on using the interface. This experiment took a worst-case scenario where the user was given almost no information, much like when the user does not even read the manual before starting to use the device. In a real learning situation, however, the user may sometimes even look for instructions in the manual, or someone will demonstrate how to use the feature. Thus, the results could be different. Furthermore, unlike in this test, the user would get feedback if the device did not respond to the interaction. Analyzing the learning curve, which is another relevant topic, requires a different experiment setup.

Having zero misrecognitions from four activities performed by four to five users with a total of 54 minutes of data does not yet statistically allow a strong generalization statement, although it is a good result, and shows that practical application is certainly feasible. To gain even wider evaluation, the next phase is to perform longer tests by equipping the users with prototypes for use in their normal daily lives.

FUTURE TRENDS

Although this study did not specifically discuss the user experience side of movement-based interaction, there is one aspect we would like to briefly address when viewing future trends: feedback. This aspect is still often found insufficient in novel user interfaces. While the presented experiments evaluated the reliability of the double-tap detection, future work includes analyzing the learning curve, the best type of feedback, and its effect on the user experience.

The user experience and learning curve for new interaction modalities can potentially be improved with suitable feedback. For example, if the beginner user makes too light taps in a tutorial mode, the device can indicate this with feedback. In general, feedback gives an indication of the state of the system and guides the users in how to use it. As suggested by O'Modhrain (2004), a key to the design of successful touch and haptic-based mobile applications is in ensuring a good mapping between the tasks, the required sensory cues, and the capabilities of the system on which the application is to be implemented. With the Customizer tool, introduced earlier, developers and user interface designers can easily experiment with different multimodal input and output combinations to find the most suitable and enjoyable solution for their application needs.

Different combinations of the feedback patterns (vibration, LED, sound) available in the interaction cover were implemented in this study. The option of using direct cover feedback in addition to phone vibra in the interaction had the benefit of avoiding possible latencies in feedback generation on the phone side. The vibra feedback was thus precisely adjustable to the desired parameters. Even though experiments on feedback supporting usability were not presented in this chapter, it

can be predicted that utilizing minimalist gesture control, together with related haptic feedback elements, has great potential in a mobile device usage and technology context. Haptic content fidelity can be rather low if it is designed to be multimodal; visual and haptic content are applied synchronously to support each other. The interaction and content design are used to promote the adoption of the technology among users.

Continuous detection of small sharp movement events also facilitates forms of gestures other than double tap. As an analogy to mouse control, there is a click and a double click. Obviously, single taps can be utilized for many purposes. However, single taps are more sensitive to various disturbances, such as accidental knocking, dropping, quick swings, turning, and so forth, that can produce a similar sharp pattern to the data and thus, a false positive. Another interesting gesture that feels natural is to swing the device. There are many other possible movement patterns to utilize in the future.

Several research questions remain, such as how to inform the user about the correct intensity of the tapping, and what kind of learning curve the tapping has. Many of the misrecognitions in the beginner group, as well as in the group that used only one hand, were due to too light a touch when tapping the device. In the beginner group, the gestures were even confused with touching in a user's approach. From the detection algorithm point of view, there is a trade-off: the parameters cannot be set too sensitively to avoid increasing the occurrence of false positives. Even though a lighter tap is viewed as more satisfying by some users, this usability increase cannot cost the reliability too much.

Yet another relevant research problem is to examine the recognition accuracy of target patterns during various activities in mobile usage. This study addressed the stationary situation and false positives during scenarios. A relevant question is what happens if the user performs the interaction while jogging, for example, without stopping to do it. Future work includes examining whether and how the continuous gesture interaction algorithms should adapt to the movement situation of the device.

CONCLUSION

Gesture control is increasingly being applied in mobile interaction. Widespread movement interaction application in mobile devices has been delayed by research challenges such as reliably detecting gestures, power consumption, and user experience-related issues such as obtrusiveness and increased effort. This chapter has focused on analyzing and evaluating the reliability of an event-based gesture interaction modality that emphasizes minimal user effort in interacting with a mobile device. The technical feasibility of the interaction modality was examined with an implementation in a smartphone environment. The reliability of continuous detection of sharp movement events produced by the user by lightly tapping the phone was evaluated by analyzing a dataset collected with the prototype.

The results show that for five informed users performing 36 repetitions of double taps in controlled stationary conditions, the target pattern was recognized with 90.4% accuracy for 3-axis detection and 97.0% for 1-axis detection. In four mobile scenarios containing 54 minutes of daily activities, each performed by four to five users carrying the prototype, there were no false positive detections of the pattern. Overall, the results based on a statistical analysis of the collected acceleration data suggested that double-tap detection is reliable enough for practical applications in mobile computing when the user performs the interaction in a stationary situation. Furthermore, it was found that the occurrence of false positives is low enough for application, presuming carefully selected usage situations where possible misrecognitions are not critical.

The contribution of this work has significance for commercial utilization.

Several research questions remain to be addressed as future work. These include how to inform the user about the correct intensity of tapping; there were users with too light a touch in the experiments. From the detection algorithm point of view, a balance needs to be found as the parameters cannot be set too sensitive to avoid increasing the occurrence of false positives. Another important research problem is to examine the recognition accuracy of target patterns during various activities in mobile usage. This study addressed the stationary situation and false positives during scenarios.

As to the movement interaction detection performance in general, the trend of development firmly aims toward increased reliability. As a result, the restricted application-specific use cases are likely to be followed by more general platform-level operations, where movement can be used as an additional interaction modality complementary to the existing ones. With emerging commercial utilization, it is easy to see the beginnings of wider adoption of the new interaction modality in mobile computing, while not forgetting that there is still further work to be done.

ACKNOWLEDGMENT

We would like to acknowledge the work of Arto Ylisaukko-oja in the hardware development, Hannu Vasama for designing the cover casing, and other contributors at Finwe, Nokia, and VTT for their kind collaboration.

REFERENCES

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification* (2nd ed.). John Wiley & Sons.

Feldman, A., Tapia, E. M., Sadi, S., Maes, P., & Schmandt, C. (2005). ReachMedia: On-the-move interaction with everyday objects. In *Proceedings of IEEE International Symposium on Wearable Computers (ISWC'05)* pp. 52-59.

Hinckley, K., Pierce, J., Horvitz, E. & Sinclair, M. (2005). Foreground and background interaction with sensor-enhanced mobile devices. *ACM Transactions on Computer-Human Interaction (TOCHI), 12*(1), 31-52.

Kallio, S., Kela, J., Korpipää, P., & Mäntyjärvi, J. (2006). User independent gesture interaction for small handheld devices. *Special Issue on Intelligent Mobile and Embedded Systems of IJPRAI, 20*(4), 505-524.

Kela, J., Korpipää, P., Mäntyjärvi, J., Kallio, S., Savino, G., Jozzo, L., & Di Marca, S. (2006). Accelerometer based gesture control for a design environment. *Personal and Ubiquitous Computing* (pp. 1-15). Online First Springer.

Korpipää, P. (2005). *Blackboard-based software framework and tool for mobile device context awareness.* Ph.D dissertation. VTT Publications 579. Retrieved from http://www.vtt.fi/inf/pdf/publications/2005/P579.pdf

Korpipää, P., Mäntyjärvi, J., Kela, J., Keränen, H., & Malm E-J. (2003). Managing context information in mobile devices. *IEEE Pervasive Computing Magazine*, *2*(3), 42–51.

Levin, G. & Yarin, P. (1999). Bringing sketching tools to keychain computers with an acceleration-based interface. In *Proceedings of the CHI 98* (pp. 268-269). ACM: New York.

Linjama, J., Häkkilä, J., & Ronkainen, S. (2005, April 3-4). Gesture interfaces for mobile devices—Minimalist approach for haptic interaction. Position paper in *CHI 2005 Workshop "Hands on Haptics."* Portland, Oregon. Retrieved from http://www.dcs.gla.ac.uk/haptic/sub.html

Linjama, J., & Kaaresoja, T. (2004). Novel, minimalist haptic gesture interaction for mobile

devices. In *Proceedings of the NordiCHI2004* (pp. 457-458). ACM Press.

Mäntyjärvi, J., Kela, J., Korpipää, P., & Kallio, S. (2004). Enabling fast and effortless customization in accelerometer based gesture interaction. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM)* (pp. 25–31). ACM.

Nokia Corporation. 5500 phone. (2006). Retrieved from http://europe.nokia.com/link?cid=EDITORIAL_8657

Oakley, I., Ängeslevä, J., Hughes, S., & O'Modhrain, S. (2004). Tilt and feel: Scrolling with Vibrotactile Display. In *Proceedings of Eurohaptics* (pp. 316-323).

O'Modhrain, S. (2004). Touch and go - Designing haptic feedback for a hand-held mobile device. *BT Technology Journal*, *22*(4), 139-145.

Rekimoto, J. (1996). Tilting operations for small screen interfaces. In *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology* (pp. 167-168).

Rekimoto, J. (2001). GestureWrist and GesturePad: Unobtrusive wearable interaction devices, In *Proceedings of the Fifth International Symposium on Wearable Computers (ISWC)* (pp. 21-27).

Ronkainen, S., Häkkilä, J., Kaleva, S., Colley, A., & Linjama, J. (2007). Tap input as an embedded interaction method for mobile devices. In Proceedings of the First Tangible and Embedded Interaction (pp. 263-270). ACM: New York.

Sawada, H., Uta, S., & Hashimoto, S. (1999). Gesture recognition for human-friendly interface in designer - consumer cooperate design system. In *Proceedings IEEE International Workshop on Robot and Human Interaction* (pp. 400-405). Pisa, Italy.

KEY TERMS

Accelerometer: 3-D accelerometer is a sensor capable of measuring object acceleration along three spatial axes.

Double Tap: Double tap is a form of movement interaction where the user performs two consecutive taps on a mobile device with a finger or palm, each producing a sharp spike waveform in an accelerometer signal measured with a high sampling rate.

Gesture Interaction: Gesture interaction here refers to explicit movements made with a mobile device while holding it in a hand in order to perform any tasks with the device.

False Positive %: False positive percentage is the relative number of falsely detected patterns, given by dividing the occurrence of all detected false positive patterns by all segments of data where a detected pattern in a dataset should not exist.

Pattern Recognition: Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. Objects can be, for example, signal waveforms or any type of measurement that needs to be classified. These objects are here referred to using the generic term "patterns."

Smartphone: A smartphone is an advanced multifunctional mobile phone with a platform open to third-party software.

True Positive %: True positive percentage is the relative number of correctly detected patterns, given by dividing all detected true positive patterns by all actually performed true patterns in a dataset.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 507-523, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.23
Positioning Technologies for Mobile Computing

Michael J. O'Grady
University College Dublin, Ireland

Gregory M. P. O'Hare
University College Dublin, Ireland

INTRODUCTION

Mobility is, as the name suggests, the defining characteristic of mobile computing and the primary differentiator between it and other computer usage paradigms. Traditionally, computers were used in what may be termed a static context. However, when computers are used in a mobile context, a number of difficulties that challenge traditional assumptions emerge. Not least amongst these are those difficulties that arise in delivering a service that is relevant and consistent with the situation in which the end-user find themselves. Should a person be waiting at a bus stop, he or she does not wish to go online and browse a bus timetable. Rather, he or she wishes to know when the next bus will stop at his or her particular stop. Thus location and time would be fundamental to the provision of such a service. Capturing time provides no major difficulties. However, identifying the physical location of a service subscriber may prove problematic.

In this review, we summarize some of the key technologies that enable the position of a mobile computer user to be determined.

BACKGROUND

Research in mobile computing and associated disciplines (Vasilakos, 2006) began in earnest the 1990s as the feasibility of the paradigm became increasingly clear. As the various research issues began to crystallize, researchers became aware of the desirability of using additional known facts of the end user's prevailing circumstances as a basis for customizing or personalizing the service for the individual end user. The term *context-aware computing* was coined to conceptualize these ideas. Pioneering research in this area was conducted at Xerox Parc in California by Schilit Adams,

and Want (1996). The Oxford Concise dictionary defines context as "the interrelated conditions in which something exists or occurs." Intuitively, everybody understands what context is. Almost paradoxically, this has made the derivation of an agreed definition almost impossible, leading some researchers to reconsider its philosophical roots (Dourish, 2004) and inherently dynamic nature (Greenberg, 2001). One issue commonly agreed is that a person's location or physical position forms an indispensable aspect of his or her context—so much so that Schmidt, Beigl, and Gellerson (1999) almost remind researchers that there are other aspects of context that should be considered. The reasons for researchers' enthusiasm are understandable. In the mid-1990s, the global positioning system (GPS) was deployed, making it possible to determine position to within 100 meters for those people equipped with a GPS receiver. Thus the technological issues were being addressed in a meaningful way. However, it was developments in wireless telecommunications that provided the spur for the upsurge in business interest in what would be termed location-aware computing (Patterson, Muntz, & Pancake, 2003).

In 1996, the Federal Communications Commission (FCC) in the United States announced the E-911 directive. In brief: this obliged public telecommunication network operators to provide the position of those people making emergency calls, thus enabling police, medical, and other personnel to react quicker. It soon became clear that this facility could have other uses for commercial purposes as, in principle at least, the location of any subscriber could be identified. Thus an era of location-aware services was anticipated. This era has yet to materialize, but as outstanding technological issues are continually being addressed, it is only a matter of time before a suite of location-aware services are available for subscribers.

To deliver location-aware services, it is necessary that an appropriate technology be selected that will provide a subscriber's position within a certain range. In the next section, some of the principal technologies for determining position are described.

TECHNOLOGIES

Various technologies and techniques are described in the academic literature for determining user position. Naturally, each has its respective advantages and disadvantages. For the purposes of this discussion, it is useful to classify them as satellite techniques, cellular network techniques, and hybrid. Each classification is now considered briefly.

Satellites Technologies

Trilateration is the basic principle for determining position using satellites. In short, the time taken for a signal to travel from a satellite at a known position to a receiver is calculated. This process is repeated for three satellites and a solution can be generated. In practice, a fourth measurement is necessary to account for the lack of synchronization between the atomic clocks on the satellite and the receiver's internal clock. The accuracy of the resultant calculation may vary due to a number of factors, including atmospheric conditions and the satellite constellation configuration. However, a reading within 20 meters of the receiver's exact geographic position may be realistically expected.

At present, there are two satellite systems in operation that broadcast signals:

1. *Global positioning system* was deployed in 1996, covers the entire earth, and is freely available. It remains under the control of the United States military. It is currently the de facto standard with specialized receivers on the market for all kinds of purposes including aviation, maritime, and leisure. To use GPS, a mobile computer user would

acquire a receiver, usually in the form of a Compact Flash (CF) card. More recently, receivers are sold as separate devices that can interface with any device that supports the Bluetooth protocol stack. Interestingly, a significant number of mobile phones on the market support Bluetooth, thus offering one scenario for providing location-aware services to mobile phone users.

2. *GLONASS* was developed and deployed by the former USSR in competition to GPS. For a number of years, it was not adequately maintained. However, this situation has changed recently, and GLONASS is currently being overhauled and restored to its former state. There are very few commercial products available that use GLONASS at present.

A third satellite navigation system is scheduled for launch in 2008. *GALILEO* is an initiative by the European Union (EU) that seeks to deliver a similar service to GPS and GLONASS, but with adequate guarantees regarding signal reliability. It is designed for purely civilian and commercial use, and unlike GPS and GLONASS, it is not controlled by defense or military groups. However, the signal broadcast will be compatible with GPS and GLONASS, and it is hoped that receivers that can utilize all three systems will developed.

Cellular Network Techniques

E-911 obligated network operators and, implicitly, telecommunications equipment manufacturers to facilitate the determination of a subscriber's position within an emergency call context. A number of *cellular network techniques* were proposed as a result of ongoing research, and Zhao (2002) provides a useful overview of these. The Third Generation Partnership Project (3GPP) proceeded to standardize on four different techniques for third-generation (3G) UMTS (Universal Mobile Telephone Networks) networks (3GPP, 2005):

1. In *cell-ID*, the geographic coordinates of the base station serving the subscriber are identified. The position of the subscriber must be within the radius of this cell. Though this method is easy to implement, its principle limitation concerns the variability in cell size. Thus the precision with which the subscriber's position is calculated may range from tens to hundreds of meters.
2. *Observed time difference of arrival (OTDOA)* requires the handset to measure the time taken for a signal to arrive from three separate base stations. Hyperbolic curves must be constructed, and their intersection indicates the position of the subscriber. Though computationally expensive, a particular difficulty involves guaranteeing that the subscriber can see three base stations simultaneously. OTDOA is highly susceptible to fading and interference.
3. *Assisted GPS (A-GPS)* involves the handset measuring GPS signals from satellites. Initially, the handset is informed as to where to look for the signals, thus minimizing delay in signal acquisition. The signal measurements are then returned to the appropriate component on the network where the position is calculated. Though increasing power consumption on the device, users can expect position readings comparable with GPS.
4. *Uplink time difference of arrival (UTDOA)* is similar in principle to OTDOA, but in this case, the signals are generated at the handset and measured at a number of base stations. As the geographic positions of the base stations are known, the position of the subscriber can be calculated using hyperbolic trilateration.

With the exception of A-GPS, the accuracy of a position obtained using these techniques is variable and unpredictable. In the case of the cell-ID method, urban areas will have a concentration of base stations so the method may work well. In

contrast, the diameter of cells in rural areas may be several kilometers, thus rendering the method ineffective. In the case of OTDOA and UTDOA, accurately measuring the time it takes the signal to travel between the subscriber's handset and surrounding base stations, and vice versa, is essential. Yet the signal may be subject to interference and fading, depending on the vagrancies of the immediate physical environment.

Hybrid Techniques

A scenario can be envisaged where a number of techniques may be combined, with each remedying their respective deficiencies in certain situations. For example, in an urban area, base stations are relatively plentiful, and in certain cases, a number may be deployed in individual streets. Thus techniques like cell-ID, OTDOA, and UTDOA will function reasonably well. In contrast, GPS—and implicitly, A-GPS—may not perform satisfactory, as the high nature of the surrounding buildings, so-called urban canyons, can result in satellites being obscured. In rural areas, the sparsity of base stations may render techniques based on the topology of the cellular network redundant. However, a clear view of the sky is likely, thus GPS and A-GPS should both function satisfactorily.

It should be noted that A-GPS itself could be arguably considered a hybrid technology. However, its close association with and standardization in the telecommunications world result in it being generally considered as a cellular network technique.

The Indoor Scenario

Determining the position of people in an indoor scenario raises particular issues and difficulties. Traditionally, satellite technologies have not operated indoors, as the signal is weak and is subject to additional reflection and fading problems when tracked indoors. A new generation of receivers promises to address this deficiency, with each succeeding generation being incrementally more sensitive. However, the key issues of accuracy and precision remain. This continues to be the case when cellular network techniques are considered, thus making the provision of guarantees concerning the quality of the calculated position exceedingly difficult.

If it is necessary to track a person in an indoor environment with confidence; it is almost essential to consider deploying a dedicated infrastructure, expensive and time-consuming as this may be. However, the required accuracy is a significant determinant. For example, it may be only necessary to track a person to room level. Alternatively, in a museum or art gallery setting, it may be necessary to determine the visitor's position to within one meter so as to determine which artifact is nearest to him or her.

Hightower and Borriello (2001) and Pahlavan, Xinrong, and Makela (2002) provide useful overviews of the issues involved in indoor tracking and positioning. A common approach is to tag the person and place a network of sensors throughout a building. This approach was adopted by Want, Hopper, Falco, and Gibbons (1992) in the pioneering active badge project, and the feasibility of the approach was verified. Systems that use a similar approach today include Cricket (Priyantha, Chakraborty, & Padmanabhan, 2000) and Ubisense (Cadman, 2003). Indeed, given the increased interest in Radio Frequency IDentification (RFID), one can easily envisage a solution involving a fixed network of RFID readers and RFID-tagged personnel.

FUTURE TRENDS

One of the key developments currently taking place concerns the deployment of *satellite-based augmentation systems (SBAS).* Such systems are a satellite-based implementation of the well-known differential GPS (DGPS) method of improving

GPS positions to within a few meters. A number of SBAS systems are being deployed, including the European Ground Navigation Overlay Service (EGNOS) and the Wide Area Augmentation System (WAAS) in the United States. More SBAS satellites are expected to be launched for other areas of the world in the coming years. Two methods for accessing SBAS are of interest. The easiest way is to incorporate an appropriate chip in a GPS receiver. In this way, the position is augmented seamlessly and transparently to the user. A second method involves the Internet, via which SBAS signals can also be broadcast. SISNet (Chen, Toran-Marti, & Ventura-Traveset, 2003) is one example of such a system. Indeed, when A-GPS is reconsidered, it can be seen that integrating this approach with a system such as SISNet is relatively straightforward.

Indoors, the situation is more complex. One approach receiving increasing attention by the research community concerns pseudolites (Wang, 2002). Pseudolites (pseudo-satellites) are placed throughout a building and mimic the GPS signal. Naturally, the pseudolite network should be calibrated for the building in question. However, the important issues of interoperability and standardization—issues that have so far been neglected—must also be addressed.

CONCLUSION

A significant choice of technologies is available for aspiring providers of location-aware services. The required accuracy and precision of the resultant subscriber position is a key determinant of the choice of technology. Attitudes of network operators toward independent small businesses seeking to deploy new services are also of critical importance. It is essential that such operators provide an open and transparent mechanism for accessing subscriber position information. Should the operator adopt an attitude of restricting access or charging excess fees for such information, the potential of location-aware services will be compromised. Overtime, it can be anticipated that a number of mobile phones with integrated GPS and SBAS technologies will be launched on the market. Ultimately, however, it beholds those people designing for mobile users to judiciously consider the merits of the respective positioning technologies in the context of both the application domain and target audience. Only in this way can they be reassured that the needs and expectations of their customers will be addressed.

REFERENCES

Cadman, J. (2003). Deploying commercial location-aware systems. *Proceedings of the Workshop on Location-Aware Computing (held as part of UbiComp)* (pp. 4-6).

Chen, R., Toran-Marti, F., & Ventura-Traveset, J. (2003). Access to the EGNOS signal in space over mobile-IP. *GPS Solutions, 7*(1), 16-22.

Dourish, P. (2004). What we talk about when we talk about context. *Personal & Ubiquitous Computing, 8,* 19-30.

Greenberg, S. (2001). Context as a dynamic construct. *Human-Computer Interaction, 16,* 257-268.

Hightower, J., & Borriello, G. (2001). Location systems for ubiquitous computing. *IEEE Computer, 34*(8), 57-66.

Pahlavan, K., Xinrong, L., & Makela, J.P. (2002). Indoor geolocation science and technology. *IEEE Communications Magazine, 40*(2), 112-118.

Patterson, C. A., Muntz, R. R., & Pancake, C. M. (2003). Challenges in location-aware computing. *IEEE Pervasive Computing, 2*(2), 80-89.

Priyantha, N. B., Chakraborty, A., & Padmanabhan, H. (2000). The cricket location support system. *Proceedings of the 6th ACM International*

Conference on Mobile Computing and Networking (MOBICOM) (pp. 32-43).

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *Proceedings of the Workshop on Mobile Computing Systems and Applications* (pp. 85-90). Santa Cruz, CA.

Schmidt, A., Beigl, M., & Gellersen, H.-W. (1999). There is more to context than location. *Computers and Graphics, 23*(6), 893-901.

3GPP. (2005). *3GPP TS 25.305, Technical Specification Group Radio Access Network; Stage 2 Functional Specification of User equipment (UE) positioning in UTRAN (Release 7).*

Vasilakos, A., & Pedrycz, W. (2006). *Ambient intelligence, wireless networking, ubiquitous computing.* Norwood, MA: Artec House, Inc.

Wang, J. (2002). Pseudolite applications in positioning and navigation: Progress and problems. *Journal of Global Positioning Systems, 1*(1), 48-56.

Want, R., Hopper, A., Falco, V., & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems, 10*(1), 91-102.

Zhao, Y. (2002). Standardization of mobile phone positioning for 3G systems. *IEEE Communications Magazine, 40*(7), 108-116.

KEY TERMS

GPS: Global positioning system.

OTDOA: Observed time difference of arrival.

Pseudolite: Pseudo satellite.

SBAS: Satellite-based augmentation system.

SISNet (Signal In Space through the Internet): An initiative by the European Space Agency (ESA) to broadcast corrections to the standard GPS signal through the Internet and in real time.

3GPP: Third Generation Partnership Project.

Trilateration: A method of determining the position of an object using the known position of at least three reference points.

UMTS: Universal mobile telephone system.

UTDOA: Uplink time difference of arrival.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 769-772, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.24
Emerging Mobile Technology and Supply Chain Integration:
Using RFID to Streamline the Integrated Supply Chain

Richard Schilhavy
University of North Carolina at Greensboro, USA

A. F. Salam
University of North Carolina at Greensboro, USA

ABSTRACT

This chapter explores how a mobile tracking technology is able to further streamline the integrated supply chain. Previous technologies which have attempted to integrate suppliers, manufactures, distributors and retailers have lacked the flexibility and efficiency necessary to justify the prohibiting costs. Radio frequency identification (RFID) technology however enables various organizations along the supply chain to share information regarding specific products and easily remotely manage internal inventory levels. These applications are only a sample of what RFID is able to accomplish for the integrated supply chain, and this chapter seeks to explore those applications.

INTRODUCTION

This chapter sets forth to provide a holistic view of how a recently adopted wireless identification technology, specifically radio frequency identification (RFID) tags, could potentially revolutionize the integrated supply chain. Companies are able to become more flexible and efficient by using a combination of mobile technologies and RFID to provide for remote inventory control and real-time, information-rich tracking of shipments in the distribution channel (Lapide, 2004). Although this technology has several hindrances currently blocking it from mass usage (Thompson, 2003), recent advancements in the technology have increased the viability of RFID for widespread organizational use, increasing the capacity and

strength while decreasing the size and cost. RFID now rests in a unique position wherein large organizations are strongly considering its viability in a variety of applications to streamline the supply chain.

Organizations have already begun considering its application in the realm of supply chain management, attempting to further streamline the process. However, while many authors have discussed the benefits of RFID tags for parts of the supply chain, this insight has only focused on a localized level, such as inventory management in retail outlets (Atkinson, 2004; Lapide, 2004; Kinsella, 2003; Schindler, 2003). Much of this discussion is centered on reducing costs for those isolated parts of the supply chain. For example, several large manufacturers are pushing the technology by actively conducting trials in manufacturing, distribution, and even retail. These companies include Proctor and Gamble, Gillette, Unilever, and retail giant, Wal-Mart (Kinsella, 2003). These RFID trials have been limited to single stages of the supply chain, focusing on the reduction of costs as the ultimate goal. Although cost reduction is commendable, true improvements in value for industry and consumers come through a unified effort to improve the entire supply chain network, reducing costs and improving accuracy and efficiency for all companies integrated into the network.

This chapter will first provide an overview of the technical aspects of RFID. Following this, an analysis of two perspectives of the integrated supply chain will be framed in light of the current and possible future applications of RFID in each area and the relationships between those areas. Finally, RFID will be framed in a holistic view of both the integrated supply chain as well as the demand chain, addressing some inter-organizational issues.

BACKGROUND

The RFID Tag and Reader

The core of RFID technology consists of two components, the identification tag and the tag reader. The identification tag itself is composed of a small antenna and a microchip, which stores a small amount of information pertinent to the object tagged (Rappold, 2003). Although the information stored may take a wide variety of forms, for many objects a simple code would be sufficient to identify the item. Asif and Mandviwalla (2005) identified five types of RFID tags in their RFID Applications Framework, including active, semi-passive, passive, chipless, and sensor. Tag readers may also be stationary or mobile, depending on the application. Of those tags which contain chips, RFID tags may either be active, passive, or a combination of the two. Active tags are powered by some external power source, such as a small battery. Passive tags, on the other hand, have no individual power source and receive power from the electromagnetic waves the tag reader uses to access the information from the tag. Some tags may use a combination of the two strategies, where an active tag containing a battery is recharged by the transmission used to read it. Chipless tags have the lowest power consumption, range, as well as cost of all the types of RFID tags since they do not contain either a battery or a silicon chip. Information storage is also significantly less, often only enough to store a simple product code.

The tag reader uses electromagnetic waves in the radio frequency band to transmit the data stored on the identification tags to the reader and, in some cases, power the identification tags. The reader may be a mobile or stationary unit depending on the application, and an organization could easily employ both. Mobile readers naturally benefit from being able to change location; however, power limitations have a severe impact on the range and may even become an issue if passive

RFID tags are used extensively. The effective range of a tag reader is a function of the frequency the tag reader is operating on and the power output available. At lower frequencies, range is severely diminished; however, power output is minimal. At higher frequencies, identification tags are able to be accessed from further away, but require significantly more power. Here, active RFID tags may need to be utilized to increase the range along with increasing the power output of the tag reader. Finally, sensor tags combine a small sensor targeted at a particular purpose, such as measuring temperature, viscosity, movement, and so forth, with an antenna, chip, and battery to store and transmit information from the sensor to a reader or network.

Not unlike any other wireless technology, RFID comes with a few limitations or issues affecting communication reliability (Angeles, 2005). The use of radio frequencies becomes a significant issue since these bands are often open to a multitude of other devices, such as wireless phones, computes, radios, and other office equipment. These common workplace devices may cause interference with reading a RFID tag and should not be overlooked when problems arise. Another problem common among mobile technologies is the increased collision of packets when more senders and receivers (in this case, more tags) are present. Since the reader is not limited to line of sight, a reader may pick up a multitude of tags in any direction of the reader, and if a large amount is present surrounding the reader, the transmissions between identification tag and reader may become interrupted due to such collisions. Anti-collision technologies are currently being developed to confront this problem with collisions (Angeles, 2005).

The RFID Network

Thanks to the Auto-ID Center at MIT, recent developments in RFID technologies have expanded RFID technology to create a holistic product identification system that consists of four components (Rappold, 2003; Smith & Konsynski, 2003; Asif & Mandviwalla, 2005). The identification tag and tag reader, again, transmit and read the information stored on the identification tag attached to the particular physical object. Stored on the identification tag is an electronic product code (EPC), which identifies the particular object or the state of that object. The EPC is a 96-bit identification code similar to conventional bar codes which uniquely identifies a product. The object name server (ONS) is a local or remote server that acts as a directory service, mapping the EPC to additional information about the physical object. This additional information in Physical Markup Language (PML) provides a standard format for describing products and storing other information about them (Smith & Konsynski, 2003). For example, the PML documents may contain information about the product manufacturer, source, and destination, or simply more detailed information about the product itself (Rappold, 2003). By mapping the EPC code to the PML documents containing information about the product, the tag may be significantly smaller since all the information is not required to be stored on the tag itself.

RFID Applications

Companies are able to become more flexible and efficient by using a combination of mobile technologies and RFID to provide for remote inventory control and real-time, information-rich tracking of shipments in the distribution channel (Lapide, 2004). Although this technology has several hindrances currently blocking it from mass usage (Thompson, 2003), the potential long-term benefits are astounding for both the integrated supply chain and other mobile technology applications as well. RFID technology is able to provide item- and product-specific information which remains with the physical object. Since no line-of-sight is required to read a tag and multiple tags may be

read simultaneously, inventories may be tabulated quickly with little manual labor and items may be tracked regardless of their location in the range of the readers. Rich information can be stored on the tags themselves, or simply mapped to the tags via an EPC, allowing this new technology to be easily mapped into current systems. PML provides additional information through a standardized markup language, providing additional interoperability between existing systems and the systems of other organizations. The technology is small, flexible, and relatively inexpensive. In the following sections, we will look more closely at the applications of RFID—specifically in the supply chain—and analyze the relative costs and benefits for each.

RFID AND THE SUPPLY CHAIN

One of the problems of current implementations of RFID tags and readers in the supply chain is that they have largely been efforts of a single company operating independently in their area of the supply chain. Technology improvements in the supply chain which are isolated to a single stage, such as manufacturing or retail, are limited to minor improvements in costs. To further illustrate this point, the following paragraphs will explain potential implementations of RFID in each stage of the supply chain considered entirely in isolation. Therefore, suppliers will be considered apart from manufacturers, retailers apart from distributors and consumers, and so on. Relative costs and benefits will be weighed with each implementation, as well as the possible risks and rewards in undertaking the endeavor.

Table 1.

	Costs	***Benefits***
Supplier	Moderate.	Improved inventory management and control.
	High for complex systems, such as monitoring devices.	Improve demand forecasting.
Manufacturer	Moderate, increasing with the complexity of the product or equipment.	Improved inventory management and control.
		Improve demand forecasting.
Distributor	Moderate. Package-level tagging.	More accurate tracking information.
	High costs when integrating with tracking systems.	En-route location tracking of packages.
Retailer	High. Item-level tagging required for pervasive implementation.	Improved inventory management and control.
		Reduce stock-outs.
Consumer	Variable. Low for luxury items to extremely high for low priced goods.	Improved shopping experience.
		Reducing in price.

Suppliers

A common theme among many of the stages of the supply chain is inventory management, even when considered in isolation. Technology improvements in inventory management allow for significant improvements in labor and capital costs, and accuracy over traditional inventory systems which require manual operation. Suppliers are no different, requiring the maintenance of large amounts of raw and processed materials in various forms. However, there are unique considerations for each stage of the supply chain in regards to inventory management which needs to be addressed. In particular to suppliers, some materials and parts require significant specialization and complexity, such as composition requirements or well-defined specifications, which can be maintained through the information stored in RFID tags or using EPCs to map the product to the information stored in a database.

Materials which require constant monitoring of temperature, viscosity, or other physical qualities could also benefit. This also applies to those materials which are heavily time dependent in regards to time-to-disposal, time-to-shipment, and so forth. RFID tags could wirelessly transmit updated information of the state of the material in real time, without human interaction or the costs of installing and/or maintaining an infrastructure based upon a physical connection. While the wireless monitoring devices would require maintenance in case of failure, the overall complexity of the system and of the maintenance would significantly decrease.

Naturally, there are significant costs involved with such implementations of RFID technology in the supplier's world. Compared with manufacturing or distribution, the information necessary to store is uniquely different. The cost of identification tags and tag readers are a common theme among all of the stages in the supply chain—an unavoidable cost. The supplier does have a slight advantage in this regard, in that relatively few tags are necessary in comparison to the other stages in the supply chain. However, if the materials required highly precise specifications or other physical properties, the complexity of the system required to maintain the information could increase the cost exponentially. Monitoring the state of the materials poses even more problems, requiring specialized identification tags attached to sensors. Additionally, simple EPC codes may no longer be sufficient to monitor the possible states of the materials and therefore require a more specialized system for each individual monitoring tag. In this regard, semantic mark-up languages similar to PML may become incredibly useful in such applications.

Manufacturers

Similar to suppliers, manufactures have much to gain in regards to simplifying inventory management and increasing the robustness and richness of the inventory system on the whole. However, what is unique to manufactures is the need to maintain large amounts of data on highly specialized or customized parts or products. RFID tags will provide item-level information unique to the particular part or product, which remains with the individual part of product. Manufacturers will find a significant reduction in the effort necessary to manage the inventory of parts or products on hand and find an increase in the accuracy of that inventory.

Manufacturers often require many complex pieces of equipment for specific applications which, in a large bustling factory, may become lost or simply difficult to find. TransAlta found that tagging pieces of equipment across the 600-foot plants made finding and maintaining the equipment easier and more flexible (Malykhina, 2005). Using Wi-Fi and Bluetooth wireless technologies in conjunction with RFID to blanket the entire facility, TransAlta was able to locate equipment regardless of its location in the company's large facilities. Active RFID tags were used to elimi-

nate the need for manual operation and provide real-time information about the location of the equipment and specific metrics from temperature gauges, vibration probes, and a variety of other peripherals (Malykhina, 2005).

If suppliers could have a problem due to complexities in inventory management, manufacturers have an overabundance of them. Implementing RFID throughout the manufacturing process requires that the individual raw materials and parts from other manufacturers be tagged, and the ultimate product to be shipped out the door also be tagged, either individually or as a package. The system becomes exorbitantly more complex when the manufacturing process requires multiple steps where information of the part or product at each stage is required.

Here, at the manufacturing stage, managers will be first posed with a difficult question when considering implementing RFID technology at their site. The question is whether package level or item level will be more economical for the specific application. For larger products or specialized equipment, tagging individual items is an economical choice. However, if the factory produces millions of widgets per month with little or no variation in those widgets, tagging them on an individual level would be a foolish choice. In most cases, manufacturers have little need for individual tagging of parts and products that come off the assembly lines, leaving package-level tagging a more prudent choice.

Distribution

In isolation, distributors are able to see significant benefits. One of the most significant proposed implementations of the technology in distribution channels is the ability to locate in real time each individual shipment, regardless of its location, and to provide information about its shipment location, destination, content, and so forth. This can be accomplished through combining several other mobile technologies. One implementation suggested the use of GPS technology to locate and identify individual vehicles, then remotely transmit the information obtained by scanning the individual shipment tags, thus providing remote access to the current contents of the vehicle, regardless of its location. Such a system could provide some value to the business and consumers alike. Richer tracking information allows interested parties to know exactly where a highly important package was last scanned, even if the package or shipment is "en route."

Wireless mobile technologies are used throughout distribution channels for varieties of business benefits, such as geographic positioning systems (GPS), for tracking and monitoring vehicles in distribution channels (Faber, 2001; Schindler, 2003). However, one issue distributors have with current mobile tracking technologies is that en-route information of vehicle contents is almost impossible to monitor, which is where RFID technology has significant promise.

However, several obstacles hinder the practical feasibility of such a system. First, current tracking systems are easily able to provide similar information, but nevertheless lack the remote, real-time tracking such a system would offer. The practical issue comes with the fact that the cost involved with implementing such a system still surpasses any benefit, even considering how the cost of RFID per tag and per reader has become more reasonable.

Retail and Consumers

Similar to suppliers, if the product retail is selling is dependent upon time—that is, perishable with a limited shelf life—those must be sold, moved, or discarded by a particular date. By outfitting each individual item sold within the store, single, outdated items will not find themselves sitting on then shelves for extended periods of time. A simple system could automatically read the tags and inform the employees which items have or will soon pass the date in question.

There have been several technologies over the years which have been considered as replacement for the aging UPC standard. However, few have provided sufficient benefits over and above UPC labels. Electronic tags were considered as a replacement, but the benefits over the current standard were so minimal that the small cost associated with each tag was too great to justify the switch. On the other hand, RFID tags provide richer information about the product over and above the current UPC standard, which simply identifies the product. Since the RFID tag can be read via a wireless connection, many items can be scanned and identified in seconds, whereas the present UPC standard requires line-of-sight reading through an optical scanner. This application could further be extended into the often-dreamed automatic checkout machines.

Current theft-deterrent technologies are an independent system from the limited UPC standard. The technology is often unreliable, resulting in countless false positives, and requires that the electronic tag passes through a small area before detection. RFID, on the other hand, would be able to provide for both the UPC label functionality discussed previously, as well as a rudimentary theft deterrent system similar to the ones currently used in retail outlets. Through RFID, retailers would be able to know exactly what products are entering and leaving the store, and which have been purchased and which have not. However, cost barriers are significant in this application since item-level tagging is necessary for it to be effective. Until the cost of individual RFID tags drops substantially in comparison to the price or quantity of the product, in the realm of fractions of a cent, it is not likely we will see widespread, item-level application of RFID. However, some retailers have begun tagging larger, higher priced items for this purpose. Both Wal-Mart and Woolworth's in the United Kingdom have begun tagging items considered high risk, such as CDs, mobile phones, computer accessories, and other electronic goods (Smith & Konsynski, 2003).

However, retail has a particularly difficult time when isolated from the rest of the supply chain in the implementation of RFID technology. Although the obstacles from the distribution side of retail are not insurmountable—requiring the contents of palettes and packages to be tagged at the distribution centers, similar to package-level distribution or manufacturing inventory, before being sent to individual retail stores—on the consumer side however, implementation becomes a much more difficult task. To be effective, individual items must be tagged. While other stages could have survived with a small amount of identification tags and tag readers, retailers cannot avoid the substantial costs associated with the thousands upon thousands of tags to cover individual items in some RFID applications. Inventory management and automatic checkouts require each individual item in the entire store to be appropriately tagged, with no assurance that the tags could be reused. Effectively, the costs of tagging individual items would be directly added to the wholesale cost of each item, and at the current five cent-mark—a substantial portion of many items' wholesale costs—further reducing the already meager margins in retail. Some organizations have implemented theft-prevention systems utilizing RFID technologies, but it is a small fraction of merchandise, particularly high-risk or high-cost items. However, retail systems are particularly simple in comparison to those found in the other stages. Since all items already possess UPC labels, mapping EPC to individual items via RFID tags may not be a daunting task and would provide some technological redundancy in case of system failure.

STREAMLINING THE SUPPLY CHAIN

Implementing technologies in the supply chain ultimately creates value when each organization at each stage of the supply chain vertically inte-

grates, standardizing on the single technology. RFID is no different. Using the same technology from the manufacturing of a product to the sale of the product throughout the entire supply chain substantially reduces costs and provides business value for everyone (Poirier, 1999). However, this does not always occur, for example, when an integral part of the supply chain chooses not to implement or share information, or organizations force their supply or demand chain to implement a particular technology, largely at the cost of those implementing (Kinsella, 2003). In the following section, how RFID can provide value to a more integrated supply chain through the implementation of RFID technologies will be discussed, ultimately culminating on a view of the entire supply chain. While many of the applications, costs, and benefits have been covered in previous sections, how their applications tie together in the supply chain and provide value will be the focus.

Suppliers to Manufacturers

The real power and value of RFID technology in supply chain management sadly comes later in the supply chain, although many benefits can be realized between suppliers and manufacturers. Materials are cultivated, packaged, distributed, received, and processed by the respective manufacturers. Throughout this entire process, RFID is able to track the inventories of materials, the source of the materials (the supplier), and the destination (the manufacturer). Manufacturers then benefit by knowing what inventory they have on hand of pre-production materials and from what vendors those materials originated. This improves the production process at manufacturing facilities greatly if such information is already accounted for by the source of the materials. However, because of the transforming nature of the manufacturer, the same RFID tags are not valuable to the remainder of the supply chain. Once the materials are transformed into products and finally head downstream to distributors and retailers, they must be retagged. In fact, after every transformation of the product, the nature of the product changes and tags must be reapplied.

Manufacturers to Distributors

Here, at the manufacturing phase, we begin to see the real potential of implementing RFID technologies and the scope of their effect on the integrated supply chain. Package-level tagging at the manufacturing level before distribution occurs both helps maintain inventories at the manufacturing sites and aids distribution and other inventory management and control systems to be equally as streamlined. Incoming materials from suppliers are able to be logged, and inventories updated and maintained throughout the manufacturing process. However, the beauty occurs when item-level tagging is implemented. As the product is produced, the item may be tagged with manufacturing information and other specifications particular to the product. Other organizations down the supply chain will be able to access this information even when the package it was contained in was dismantled and the contents strewn across retail stores and the consumer population. Additionally, the benefits from improved forecasting comes from information downstream, at the retail level, where manufactures are able to determine which products are sold, at what locations, in specific quantities. For retailers to provide this information, with the RFID network previously discussed in place, would be far from an insurmountable feat, and the value coming downstream from the manufacturers would be to their benefit. While manufacturers ultimately may bear much of the implementation cost, they will receive equally in benefits, with even more significant benefits for organizations downstream from the manufacturer.

Distributors to Retailers

Leaving the distribution centers are countless cases and palettes of merchandise heading to different retail stores with varying quantities of thousands of different products. Managing what products are leaving or being received and where they are going can become a daunting task, even with some of today's technologies. For distributors, RFID technologies provide some of the more impressive benefits, even in isolation with relatively smaller increases in cost. However, having package-level inventories tagged by suppliers and manufacturers before entering the distribution channels improves the efficiency of the logistics systems for both those parties. As the tagged packages move through the distribution channels, retailers ultimately will benefit as well as the packages move through their receiving centers, actively managing the incoming inventory at individual stores. However, between manufacturing and retail, distributors must retag if the packages themselves are repacked. Luckily, if this is not the case, the RFID network system provides separate semantic information for each of the EPCs associated with the packages. Again, the beauty is at the item level. Since each individual product is now tagged, even repackaging the products does not require new tags to be placed on them. In fact, distributors need not retag any items whatsoever, only change the information associated with the corresponding EPC, which is unique for all the items entering and leaving the distribution channel, regardless of the location.

Tesco, a large UK retailer, recently implemented an RFID system of significant size, totaling 20,000 identification tags for their stores and distribution centers, with 4,000 tag readers and 16,000 antennas to receive the identification tag signals (Sullivan, 2005). RFID tags have been installed in order to track the merchandise cases and palettes which grace the docks at the distribution centers and receiving doors at retail stores. Unlike the retail giant Wal-Mart, however, Tesco made the investment in RFID themselves independent of their suppliers, in hopes that they perceive similar cost benefits as Tesco.

In some situations, logistics systems may need to make a sudden re-route of product or material in case of a sudden stock-out. When a product is in demand, having no inventory of a product on hand means lost sales for retailers or lost production time for manufacturers. In conjunction with GPS and cellular technologies (Schindler, 2003), distributors now may locate items en-route between destinations and calculate precise inventories of those vehicles. If a reroute is economically reasonable, the vehicle is able to be informed of where the reroute is located and what products the reroute are for. All in all the distribution system becomes more flexible and capable of providing for the retail and customer base.

Retailers to Consumers

One of the significant benefits from implementing RFID technology at the retail level is the reduction of labor costs from managing inventory, which now can be accounted for and monitored with little or no manual operation. Reduction in labor costs provides for two potential outcomes that benefit consumers. First, the additional labor capacity can be used to improve customer service of the retail establishment for customers. Or, the additional savings in labor costs not rerouted to another activity could be brought directly to the consumer in the form of lower prices, ultimately an increased value for the customers.

Merchandise Security

RFID technologies also move the retail and consumer relationship to the Holy Grail, the market of one. Prada of New York is one of the first retailers to use RFID technologies to revolutionize the shopping experience of its customers. Each item sold in the store is tagged with an RFID tag. Naturally, this provides additional security in the

case of stolen merchandise; however, the interesting aspect of the implementation occurs when the customer re-enters the store. Whether carrying or even wearing the previously purchased garment, the tag readers at the entrance to the store scan for an RFID tag; if found, information pertinent to that garment appear on large flat-panel displays around the store. For example, items matching that garment may appear, in reasonable sizes to the item purchases, directing the customer. In addition, richer marking information is obtained through this system, as item purchases are now tied together with how frequently the customer visits the store, what items are purchased in combination or sequences, and so forth. However, consumer privacy concerns have already arisen in regards to this implementation.

From the information gathered here, at the retail and consumer level, suppliers and manufacturers are now better able to forecast demand and control inventories, sending the business value back upstream (Lapide, 2004). The value invested earlier in the supply chain to tag either packages or items leaving manufacturing facilities returns to them through this improved ability to forecast demand.

CONCLUSION AND FUTURE DIRECTION

RFID technology is a fairly simple wireless technology, composed of a small antenna and microchip, and able to streamline the mobile supply chain. Technologies surrounding the RFID technology, such as EPC and PML, improve the interoperability, transparency, and flexibility of implementing RFID systems with current inventory management and distribution systems. The mobile nature of the technology incorporates additional advantages only found with more complex, higher cost systems. However, important cost considerations must be given, as choosing between the costs and benefits for package level and item level becomes an important decision. While it provides substantial value for organizations downstream, it requires significant investment upstream. As additional implementations appear throughout the supply chain, the cost of the technology will fall and the relative benefits will increase. If standardized on RFID technology, regardless of package- or item-level implementation, the entire supply chain benefits from a standard mechanism to identify objects moving up and down the supply chain, through distribution channels, and off the shelves at retail stores. RFID is poised to revolutionize the supply chain by streamlining operations, providing flexible, transparent communication between organizations.

REFERENCES

Asif, Z., & Mandviwalla, M. (2005). Integrating the supply chain with RFID: A technical and business analysis. Communications of the AIS, 15, 393-426.

Angeles, R. (2005). RFID technologies: Supply-chain applications and implementation issues. Information Systems Management, 22(1), 51-65.

Anonymous. (2003). Supply chain technologies—At Woolworth's. Work Study, 52, 44-46.

Atkinson, W. (2004). Tagged: The risks and rewards of RFID technology. Risk Management, 51, 12-19.

Kinsella, B. (2003). The Wal-Mart factor. Industrial Engineer, 35, 32-36.

Lapide, L. (2004). RFID: What's in it for the forecaster. Journal of Business Forecasting Methods and Systems, 32(2), 16-19.

Leary, D. E. O. (2000). Supply chain processes and relationships for electronic commerce. In M. Shaw, R. Blanning, T. Stradder, & A. Whinston

(Eds.), Handbook on electronic commerce (pp. 431-444). Berlin: Springer-Verlag.

Malykhina, E. (2005). Active RFID meets Wi-Fi to ease asset tracking. Information Week, 1022, p. 38.

Poirier, C. C. (1999). Advanced supply chain management. San Francisco: Berrett-Koehler.

Rappold, J. (2003). The risks of RFID. Industrial Engineer, 35, 37-38.

Schindler, E. (2003). Business: The 8th layer: Location, location, location. netWorker, 7(2), 11-14.

Smith, H., & Konsynski, B. (2003). Developments in Practice X: Radio frequency identification (RFID)—An Internet for physical objects. Communications of the AIS, 12, 301-311.

Sullivan, L. (2005). UK retailer goes in RIFD shopping spree. Information Week, 1022, p. 36.

Yang, B. R. (2000). Supply chain management: Developing visible design rules across organizations. In M. Shaw, R. Blanning, T. Stradder, & A. Whinston (Eds.), Handbook on electronic commerce (pp. 445-456). Berlin: Springer-Verlag.

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 859-869, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.25
Content Personalization for Mobile Interfaces

Spiridoula Koukia
University of Patras, Greece

Maria Rigou
University of Patras, Greece and
Research Academic Computer Technology Institute, Greece

Spiros Sirmakessis
Technological Institution of Messolongi and
Research Academic Computer Technology Institute, Greece

INTRODUCTION

The contribution of context information to content management is of great importance. The increase of storage capacity in mobile devices gives users the possibility to maintain large amounts of content to their phones. As a result, this amount of content is increasing at a high rate. Users are able to store a huge variety of content such as contacts, text messages, ring tones, logos, calendar events, and textual notes. Furthermore, the development of novel applications has created new types of content, which include images, videos, MMS (multi-media messaging), e-mail, music, play lists, audio clips, bookmarks, news and weather, chat, niche information services, travel and entertainment information, driving instructions, banking, and shopping (Schilit & Theimer, 1994; Schilit, Adams, & Want, 1994; Brown, 1996; Brown, Bovey, & Chen, 1997).

The fact that users should be able to store the content on their mobile phone and find the content they need without much effort results in the requirement of managing the content by organizing and annotating it. The purpose of information management is to aid users by offering a safe and easy way of retrieving the relevant content automatically, to minimize their effort and maximize their benefit (Sorvari et al., 2004).

The increasing amount of stored content in mobile devices and the limitations of physical mobile phone user interfaces introduce a usability challenge in content management. The physical mobile

phone user interface will not change considerably. The physical display sizes will not increase since in the mobile devices the display already covers a large part of the surface area. Text input speed will not change much, as keyboard-based text input methods have been the most efficient way to reduce slowness. While information is necessary for many applications, the human brain is limited in terms of how much information it can process at one time. The problem of information management is more complex in mobile environments (Campbell & Tarasewich, 2004).

One way to reduce information overload and enhance content management is through the use of *context metadata*. Context metadata is information that describes the context in which a content item was created or received and can be used to aid users in searching, retrieving, and organizing the relevant content automatically. Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves (Dey, 2001). Some types of context are the *physical context*, such as time, location, and date; the *social context*, such as social group, friends, work, and home; and the *mental context*, which includes users' activities and feelings (Ryan, Pascoe, & Morse, 1997; Dey, Abowd, & Wood, 1998; Lucas, 2001).

By organizing and annotating the content, we develop a new way of managing it, while content management features are created to face efficiently the usability challenge. Context metadata helps the user find the content he needs by enabling single and multi-criteria searches (e.g., find photos taken in Paris last year), example-based searches (e.g., find all the video clips recorded in the same location as the selected video clip), and automatic content organization for efficient browsing (e.g., location-based content view, where the content is arranged hierarchically based on the content capture location and information about the hierarchical relationships of different locations).

DATE, TIME, LOCATION, AND PROXIMITY

While context can be characterized by a large number of different types of attributes, the contribution of context attributes to content management is of great importance. We focus on a small number of attributes, which are considered the most important in supporting content management and also have the most practical implementations in real products, such as date, time, location, and proximity (nearby Bluetooth devices). Bluetooth is a short-range wireless technology used to create personal area networks among user mobile devices and with other nearby devices.

The first two attributes, date and time, are the most common in use in a wide range of applications. They are used to organize both digital and analog content, and offer an easy way of searching and retrieving the relevant content automatically. For example, many cameras automatically add the date and time to photographs. Furthermore, the location where content is created is another useful attribute for searching the content (e.g., home, workplace, summer cottage). Mobile devices give users the possibility to create content in many different locations. Users can associate the location with the equivalent content in order to add an attribute to it that will enable them to find it easier. Finally, proximity also plays an important role in content management, as nearby Bluetooth devices can provide information both in social and physical context. While each Bluetooth device can be uniquely identified, information can be provided on nearby people by identifying their mobile phones. An example for physical context is the case of a Bluetooth-based hands-free car kit that can be used to identify that the user is in a car.

USABILITY ISSUES AND PROBLEMS

The expansion of the dimension of context information in order to include location, as well as proximity context, can be of benefit to users while they are able to store, access, and share with others their own location-based information such as videos and photos, and feel the sense of community growing among them (Kasinen, 2003; Cheverist, Smith, Mitchell, Friday, & Davies, 2001). But when it comes to proximity to be included in context information, the problem of *privacy* emerges. It appears that users are willing to accept a loss of privacy when they take into account the benefits of receiving useful information, but they would like to control the release of private information (Ljungstrand, 2001; Ackerman, Darrel, & Weitzner, 2001).

While context metadata is attached to content, when users share content, they have to decide if they share all the metadata with the content or they filter out all or some part of them. The cost for memory and transmission of metadata, as it is textual information, is not an important factor to influence this decision. When the user receives location and proximity information attached to content, he or she may also find out where and with whom the creator of the content was when the content was created. As a result, both the location of the content creator and the location of nearby people are shared along with the content information. If this information is private, the sharing of it could be considered as a privacy violation. This violation may be 'multiplied' if the first recipient forwards the content and the metadata to other users.

However, users seem to be willing to share context metadata attached to content, as it would be convenient if context metadata were automatically available with the content (so that users do not have to add this information manually). Furthermore, it would be very helpful for the recipient if the received content was annotated with context metadata so that the recipient does not have to annotate it manually and be able to manage the content more easily. For example, in the case of image and video content, the filtering of context metadata such as location and people could be useless, since these same items appearing in the image or video can be identified visually from the image content itself.

But what is meaningful information to the end user? It seems that users want meaningful information, but they are not willing to put too much effort in creating it, unless this information is expected to be very useful. In the case of location, it would be difficult for users to type the name of the place and other attributes manually, since it would require their time and effort. Thus it would be important if meaningful context metadata, which include the required information, are automatically generated.

Proximity information also needs to be meaningful. In this way, meaningfulness is important when attaching information on nearby devices in the form of metadata. If the globally unique Bluetooth device address and the real name of the owner of the device could be connected, this functionality would give meaningful information to the user.

It is hard to determine which information is useful, while what is useful information in one situation might be totally useless in another. For example, when looking at photo albums, what is thought to be useful information varies a lot. When one is looking at family pictures taken recently, it is needless to write down the names of the people, since they were well known and discernable. But it is different looking at family pictures taken many years ago: the same people may not be that easily recognizable.

It appears that useful information depends on a user's location, what the information is used for, and in which time span. In order to create meaningful information, users need to put much effort into getting the data, organizing it, and annotating it with context metadata. Ways to

minimize their effort and maximize their benefit should be developed.

CONCLUSION

The increasing amount of stored content in mobile devices and the limitations of physical mobile phone user interfaces introduce a usability challenge in content management. The efficient management of large amounts of data requires developing new ways of managing content. Stored data are used by applications which should express information in a sensible way, and offer users a simple and intuitive way of organizing, searching, and grouping this information. Inadequate design of user interface results in poor usability and makes an otherwise good application useless. Therefore, it is necessary to design and built context-aware applications.

Issues of usefulness and meaningfulness in utilizing context metadata need to be further investigated. Usefulness depends on the type of metadata. As far as location and proximity are concerned, it appears that the more time has passed since the recording of the data, the more accurate the information needs to be. Furthermore, in the case of location information, the closer to one's home or familiar places the data refers to, the more detailed the information needs to be. A main usability challenge is the creation of meaningful context metadata automatically, without users having to add this information manually. There exist many ways for automatic recording of information about a user's context, but the generated information is not always meaningful.

Another field that requires further research is privacy. It seems that users are willing to accept a loss of privacy, provided that the information they receive is useful and they have control over the release of private information. Content management provides users with a safe, easy-to-use, and automated way of organizing and managing their mobile content, as well as retrieving useful information efficiently.

REFERENCES

Ackerman, M., Darrel, T., & Weitzner, D. J. (2001). Privacy in context. *Human Computer Interaction, 16,* 167-176.

Brown, P. J. (1996). The stick-e document: A framework for creating context-aware applications. *IFIP Proceedings of Electronic Publishing '96,* Laxenburg, Austria, (pp. 259-272).

Brown, P. J., Bovey, J. D., & Chen, X. (1997). Context-aware applications: From the laboratory to the marketplace. *IEEE Personal Communications, 4*(5), 58-64.

Campbell, C., & Tarasewich, P. (2004). What can you say with only three pixels? *Proceedings of the 6th International Symposium on Mobile Human-Computer Interaction*, Glasgow, Scotland, (pp. 1-12).

Cheverist, K., Smith, G., Mitchell, K., Friday, A., & Davies, N. (2001). The role of shared context in supporting cooperation between city visitors. *Computers &Graphics, 25,* 555-562.

Dey, A. K., Abowd, G. D., & Wood, A. (1998). CyberDesk: A framework for providing self–integrating context-aware services. *Knowledge Based Systems, 11*(1), 3-13.

Dey, A. K. (2001). Understanding and using context. *Personal & Ubiquitous Computing, 5*(1), 4-7.

Kaasinen, E. (2003). User needs for location-aware mobile services. *Personal Ubiquitous Computing, 7,* 70-79.

Kim, H., Kim, J., Lee, Y., Chae, M., & Choi, Y. (2002). An empirical study of the use contexts and usability problems in mobile Internet. *Proceed-*

ings of the 35th Annual International Conference on System Sciences (pp. 1767-1776).

Ljungstrand, P. (2001). Context-awareness and mobile phones. *Personal and Ubiquitous Computing, 5,* 58-61.

Lucas, P. (2001). Mobile devices and mobile data—issues of identity and reference. *Human-Computer Interaction, 16*(2), 323-336.

Ryan, N., Pascoe, J., & Morse, D. (1997). Enhanced reality fieldwork: The context-aware archaeological assistant. In V. Gaffney, M. v. Leusen, & S. Exxon (Eds.), *Computer applications in archaeology.*

Schilit, B., & Theimer, M. (1994). Disseminating active map information to mobile hosts. *IEEE Network, 8*(5), 22-32.

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *IEEE Proceedings of the 1st International Workshop on Mobile Computing Systems and Applications,* Santa Cruz, CA, (pp. 85-90).

Sorvari, A., Jalkanen, J., Jokela, R., Black, A., Kolil, K., Moberg, M., & Keinonen, T. (2004). Usability issues in utilizing context metadata in content management of mobile devices. *Proceedings of the 3rd Nordic Conference on Human-Computer Interaction,* Tampere, Finland, (pp. 357-363).

KEY TERMS

Bluetooth: A short-range wireless technology used to create personal area networks among user devices and with other nearby devices.

Content Management: Ways of organizing and annotating content in order to retrieve and search it more efficiently.

Context: Any information that can be used to characterize the situation of an entity.

Context Metadata: Information that describes the context in which a content item was created or received.

Entity: A person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves.

Location: The place where content is created by the user.

Usability: The effectiveness, efficiency, and satisfaction with which users can achieve tasks in the environment of mobile devices.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 116-118,

Chapter 3.26
Distributed Mobile Services and Interfaces for People Suffering from Cognitive Deficits

Sylvain Giroux
Université de Sherbrooke, Canada

Hélène Pigot
Université de Sherbrooke, Canada

Jean-François Moreau
Université de Sherbrooke, Canada

Jean-Pierre Savary
Division R&D CRD, France

ABSTRACT

The mobile device presented here is designed to offer several services to enhance autonomy, security, and communication for the cognitively impaired people and their caregivers. Two mobile devices are linked through a server; one is dedicated to the patient, the other one to the caregivers. The services fill three functions for patients: a simplified reminder, an assistance request service, and an ecological information gathering service. Three services are available for the caregiver: monitoring patients'ADLs, informing system and colleagues of an intervention, and planning patients' ADLs.

INTRODUCTION

The number of people suffering from cognitive impairments (Alzheimer disease, head injury, schizophrenia, etc.) is growing continually. For one, the aging of the society plays an important role in this trend. The always increasing needs for resources generates a huge stress on human and economical actors. Thus in Canada and United

States, as in most western countries, social and medical policies try to keep people suffering from cognitive deficits at home (Hareven, 2001). Progress in cognitive rehabilitation also augments the number of semi-autonomous people that would be able to stay at home if light assistance is provided. Of course, this results in higher expectations and even more demand on resources. But most of the time, families have to take responsibility for care without access to appropriate resources. Too often, this situation then turns to an exhausting weight. Therefore, natural and professional caregivers urge for help.

Recent advances in mobile technology can provide affordable solutions to lessen the burden and anxiety put on caregivers and to collect reliable information. For many years, numerous devices have been designed to counterbalance physical and sensory deficits. Nowadays progress in technology set high hopes for cognitive orthotics that address cognitive decline (Lange, 2002; Pollack, 2004). Thanks to their small size and their versatility, mobile devices can offer a personalized assistance anytime anywhere. Mobile devices then become portable cognitive prosthesis, for instance acting as pervasive system to remind people activities of daily living (ADL) to perform when needed. They can also foster sense of security by keeping people and caregivers directly in touch. In its simplest form a direct call button may be used to request immediate assistance. Geo referenced data may also be used to detect crisis of schizophrenia or when a patient suffering from Alzheimer disease is lost (Médical Intelligence, 2005). Mobile devices can also provide non-intrusive remote supervision by caregivers. Besides mobile devices can gather ecological data compulsory to adapt or fine-tune diagnosis and treatments. Nevertheless, the design of devices and mobile services must put a very careful attention to user interfaces used to deliver information. Population suffering from cognitive deficits has often severe limitations and constraints.

This chapter discusses features and implementation of mobile services for cognitive assistance and remote supervision. Mobile devices provide practical solutions to the issues presented above. First, we review benefits and limitations of current reminder systems that can assist patients and/or caregivers. Then we present the needs of the population regarding the assistance necessary to stay safely at home. The multiple mobile services, designed especially for that population, are also described. We show how supervision can safely foster autonomy. Next, we describe cognitive assistance to patients and remote supervision by professionals or relatives, and then we go further and gather ecological data to foster better treatments. Finally, outdoors, safety and security of the patients is relying on geo-localization features. Implementation specific details are also sketched.

MOBILE PROSTHETIC SYSTEMS

The management of ADL is a central issue for people suffering from cognitive deficits. In the process of rehabilitation, occupational therapists provide a patient with a paper agenda as a tool he has to master to manage his life and autonomy. A lot of research projects and commercial applications also targeted electronic adaptation of specific functionalities of these agendas. In this section, we review the advantages and limitations of current reminder and/or agenda systems.

Some systems are device specific and are explicitly designed to be used inside the patient home and then could not help outside (Helal et al., 2003; Visions, 2005). Another category of systems is designed specifically for mobile devices in such a way that their use is not restricted to a specific location. These applications are usually running on PDAs or smart phones to either assist for needs specifically related to mobility (Patterson et al., 2004) or as general reminder/agenda systems or electronic organizers (Gorman, Dayle, Hood,

& Rumrell, 2003; Haberman, Jones, & Mueller, 2005; Neuropage; Szymkowiak, Morrison, Prveen Shah, Evans, & Wilson, 2005). They are usually provided with acoustic alert or remote communication.

Activity Compass project intends to provide compensatory aid for outdoor courses for patient suffering from Alzheimer disease (Patterson et al., 2004). It consists basically of a PDA equipped with a GPS. Its role is to guide patient towards their destination by a mean of an arrow indicating the direction to take.

Neuropage is a reminder system that uses radio technology to send reminders of things to do (Neuropage). When the message arrives, the pager beeps (or vibrates), one button is pressed and the message can be read from the screen. Messages can be regular events or single-time message. Messages are added or removed by contacting Neuropage office by phone, e-mail, letter, or fax.

The ISAAC system acts as a cognitive prosthesis used by people experiencing dysfunctions in autonomy due to different kinds of brain injury (Gorman et al., 2003). It provides a checklist reminder to execute safely the ADL. Its hierarchical structure permits to navigate easily throughout the different pieces of advice displayed. Personalization and adaptation of pieces of advice enable ISAAC to evolve according to the patient cognitive evolution.

Providing a reminder helps the cognitively impaired patient to increase his autonomy. But the patient himself and his caregiver express needs for safety. Both want to keep in touch and be confident that activities will be performed in time. So a mean of communication must be added to the current reminders. In the next section, we review the cognitive deficits encountered and the needs that ensue.

NEEDS OF COGNITIVELY IMPAIRED PEOPLE

Cognitive impairments encountered in schizophrenia, head trauma and during the early stages of the Alzheimer disease provoke similar losses in autonomy that justify exploring a common approach to support, assistance and remediation by technology (Pigot, Savary, Metzger, Rochon, & Beaulieu, 2005). Cognitively impaired people present lacks of initiation. For instance, they remain for long periods without undertaking actions. They all call for frequent reminders to remember what to do and sometimes how to do. Alzheimer disease strikes elders and one's situation evolves from relative autonomy to an unrelenting dependence while trauma injury and schizophrenia appear most often during adulthood and one's situation evolves from dependence to relative autonomy.

Unless severely affected, cognitively impaired people are usually quite autonomous in performing basic ADL such as eating, dressing, and washing themselves. But they at times forget to do them and then need continuous recalls in order to initiate ADLs. Taking medication is a critical issue, as it requires a good short-term memory in order to fulfill the prescription correctly: which pills, how many and when. Without a reliable short-term memory, it can be forgotten or taken twice. At home oblivion or bad discernment in the use of domestic appliances often cause bath and sink flooding, fire, burn, cut...Beside these hazardous situations, risk of malnutrition, bad hygiene, and isolation are also real.

Cognitive impairments experienced at home have also manifest effects outside. Moreover, cognitively impaired people exhibit more difficulties in an unusual context. Oblivion leads to problems in finding one's way to a destination, in remembering the shopping list or even the reason of the trip.

Loss of autonomy requires assistance. Caregivers lavish attention to the patients days and

nights long. Frequent recalls and safety are two factors especially important in the overall burden coming from constant supervision and assistance. Therefore, an electronic companion could alleviate the "threatening" recalls and improve the relationship. Safety prevents a caregiver from letting the patient alone.

MOBILE AND PERVASIVE COMPUTING

Although existing services are invaluable for patients and caregivers, overall they suffer from some drawbacks. These services are not integrated in a common portal. They must run on a specific patient device. They are either closely linked to a given location or, if they are location independent, they can not benefit from the casual presence of other devices in the environment, for instance sensors on the kitchen cupboards.

In next sections, we present the first steps of an on-going research project at DOMUS Laboratory. Mobile computing is coupled with pervasive computing (Weiser, 1991). Mobile computing will enable patients and caregivers to use services wherever they are. Pervasive computing will allow them to access services whatever the device is and to benefit from existing devices, sensors, and effectors present in their current environment. This approach then promotes a tighter integration of devices and services to the environment when possible while preserving mobility and independence of the patients.

Currently, a common portal and many services were implemented and are ready to be evaluated in situ. Through this portal, a patient gets access to services with a simplified adapted user interface. Caregivers also have access to their counterpart of services but the interface is more complex. For the moment, devices used are PDAs, but we intend to port them on e-mate, a platform that renders deployment of services independent of devices they are deployed on (Giroux, Carboni, Paddeu, Piras, & Sanna, 2003). So phones, laptops, TVs could be used to access services.

ASSISTANCE AND SUPERVISION SERVICES

A mobile device is designed to offer several services to enhance autonomy, security, and communication to cognitively impaired people and their caregivers. PDAs are linked through a server. Some are dedicated to patients. Others run applications presenting the caregivers point of view on information. Patient services fill four functions: a simplified reminder, a prompter to request immediate assistance, a service to gather ecological information, and a location-based health information service (Boulos & Maged, 2003) to help patient in case of crisis. On the other side, three services are available for the caregiver: monitoring patients' ADLs, informing system and colleagues of an intervention, and planning patient' ADLs. Services also enable patients and caregivers to communicate. The client component running on a patient's PDA is kept very simple. The patient benefits on an auto login procedure in order to reduce the cognitive load.

An Enhanced Agenda as a Mean for Cognitive Assistance

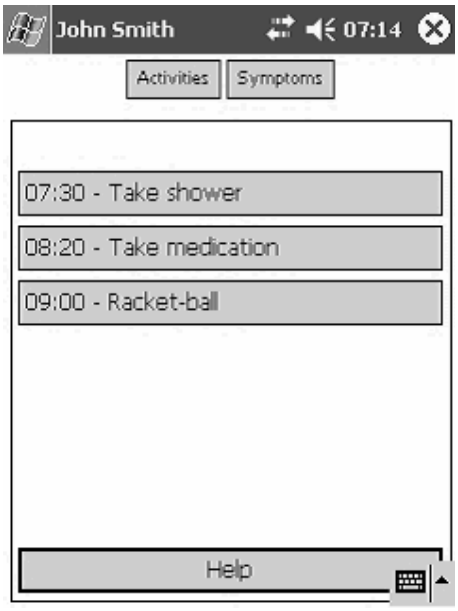
A basic paper agenda is designed to help remembering appointments, to fill a timetable and to prevent time conflicts. Our aim here is quite different. The agenda acts mainly as a cognitive prosthesis. Typically, the home of a patient suffering from head injury is covered with numerous notices, for instance written on a post-it. Each one indicates something to do. On the contrary, the electronic diary can show solely the activities often forgotten, and moreover just when it is necessary. By default, our prototype displays on the PDA single window the next three activities to perform in a three hours time bracket (Figure

1). Depending on the patient's cognitive abilities the number of activities displayed and the time slot can be easily reduced or increased. When an activity is performed the patient clicks on it to indicate its completion to the caregivers. The activities monitored on the agenda are either events occurring just one time, as appointments, or recurring events as ADLs often forgotten.

Keeping in Touch

If help is necessary, the patient clicks on the button "Help" at the bottom of the screen to warn caregivers. This button is always visible and active. This remote supervision ensures permanent contact between them. It provides a sense of safety for both patients and caregivers. Caregivers feel insecure to let patients alone without the opportunity to be reached. For patients, the phone is often too complicated to use. Carrying everywhere a PDA offers permanent assistance at a single click.

Figure 1. Patient agenda



Monitoring ADL

Previous sections described the patient side of the system. In this section and the next two, we depict the caregiver view: how ADLs are monitored, how cooperation can be foster between caregivers, and how to manage a patient list of ADLs.

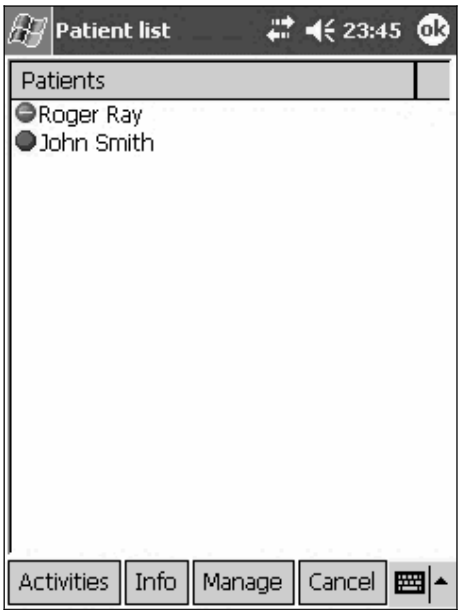
The role of the patient PDA is to foster his or her autonomy. From time to time, he or she does not perform given activities in time which may compromise his or her ability to stay alone. Equilibrium must be maintained between a patient intimacy and his or her safety. For instance, forgetting one day to take a shower does not justify an immediate intervention. However, after several days without showers, caregivers urge him or her to do it. On the other side, forgetting once to take medication could be harmful for the patient's health. Without a tool that helps to monitor ADLs, a caregiver surveys constantly the patients asking for the ADLs completion.

The current prototype supports remote supervision of ADLs to perform (Figure 2). Caregivers then know when an ADL is completed without bothering residents. On his or her mobile device, a caregiver can monitor every patient under his or her responsibility. On his or her patients list, colored icons indicate the status of current ADLs for each patient: green if all is OK, yellow if the patient is near to forgot to perform an ADL, and red if time planned for the ADL is over (Figure 2a). For each patient, a caregiver can know: which activities are under monitoring (Figure 2b) and the time of completion of a given ADL. The same color code prevails.

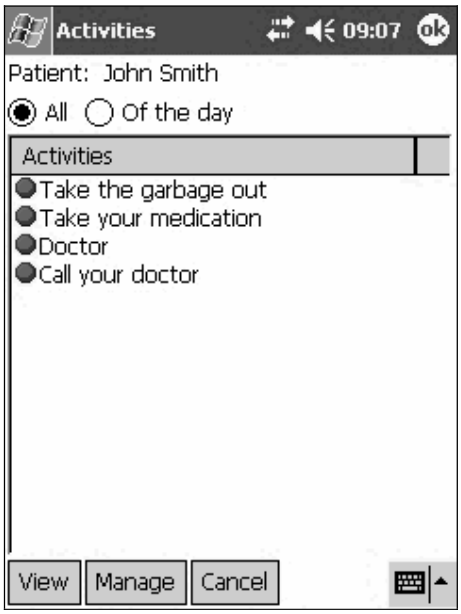
Coordination Between Caregivers

Since many caregivers can be in charge of a patient, coordination is important. Several caregivers can take care of many patients. When a patient does not perform a task in time, all responsible caregivers are informed. It is then essential to

Figure 2. Caregiver's ADL monitoring a) Supervised patients b) One patient ADLs



(a)



(b)

coordinate their actions, for instance in order to avoid all of them going to see the patient. As soon as a caregiver notifies his intervention to the system, this information is sent to all the mobile devices owned by the other caregivers. Caregivers can also exchange responsibility of patients, for instance when a working shift occurs.

Implementation section explains in more details how the information is updated on each PDA.

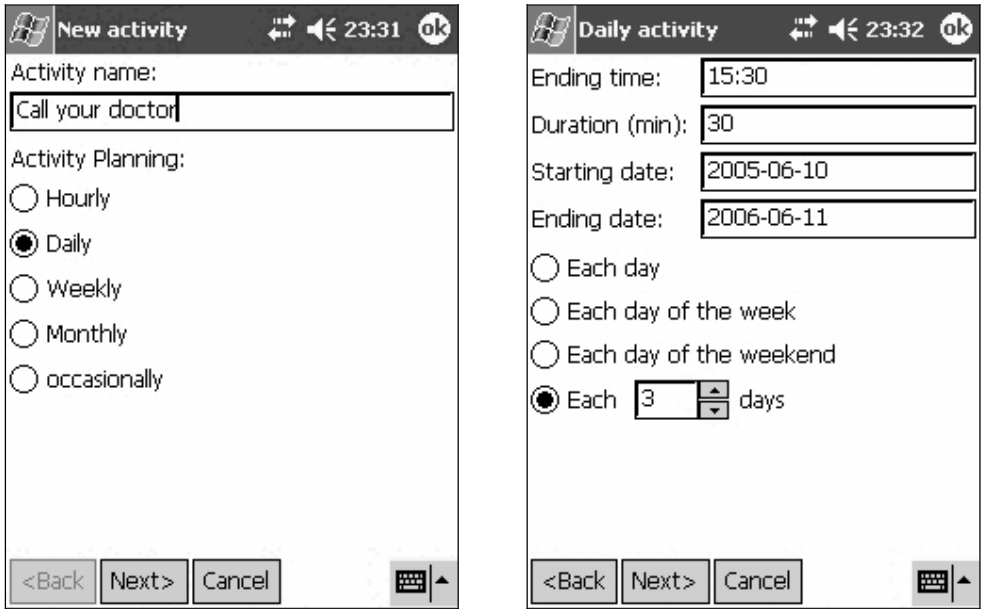
Planning ADL

In addition to ADL monitoring, caregivers must have also simple means to update and change rapidly the ADLs and events that have to be monitored. Like a usual organizer, the caregiver mobile device enables to register an activity occasionally or on a periodic basis ranging from hours to months (Figure 3a). The caregiver has to specify duration and completion time and for how long the activity has to be monitored (Figure 3b). This last information is particularly suitable during the rehabilitation period or to update a medication prescription that is regularly changed by the physician.

Ecological Data: Gathering Information for Better Treatments

Collecting ecological information may be of tremendous value for delivering better treatments. For instance to control the schizophrenic symptoms, a physician prescribes neuroleptics which induce severe side effects (Shriqui & Nasrallah, 1995). The physician has then to adjust medication

Figure 3. Planning activity a) Activity period basis b) Time activity information



(a) (b)

during years to limit side effects and to avoid the recurrence of hallucinations. He or she needs valid information about the patient feelings.

On his or her mobile device, the patient can notify symptoms (or other information) as soon as he or she feels them (Figure 4). An intensity scale permits him or her to estimate the level of severity. The PDA is also equipped with a GPS to register automatically date, time, and position. These geo-referenced data are sent to the server on a batch mode. Data can be analyzed later to improve diagnosis and treatments.

IMPLEMENTATION

The basis of the implementation is a client-server architecture (Figure 5). Patients and supervisors use iPAQs to connect to the server through wireless networks. The system is available both indoors and outdoors. Indoors, the wireless connections to the server use IEEE 802.11b,g. Outdoors, the wireless connection is made through a Sierra Wireless AirCard®. Each PDA connects to the server through TCP sockets. The server allocates a different thread to each connection. If the connection is lost, the mobile device tries to restore it. The PDAs and the server exchange serialized objects through sockets. Outdoors, a GPS gives the patient location. To ensure rapid communication the traffic is kept low. When a patient confirms an ADL has been done or when a caregiver let know about his intervention, a message is sent to the server. The server first updates the database, and then sends a message to each PDA to inform a change happens. At that moment, some PDAs

Figure 4. Patient symptoms

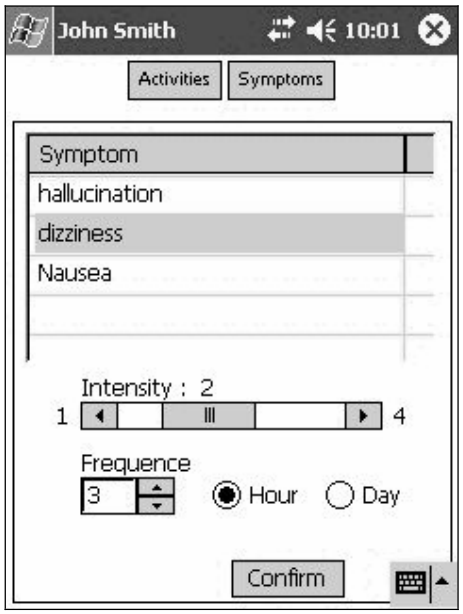
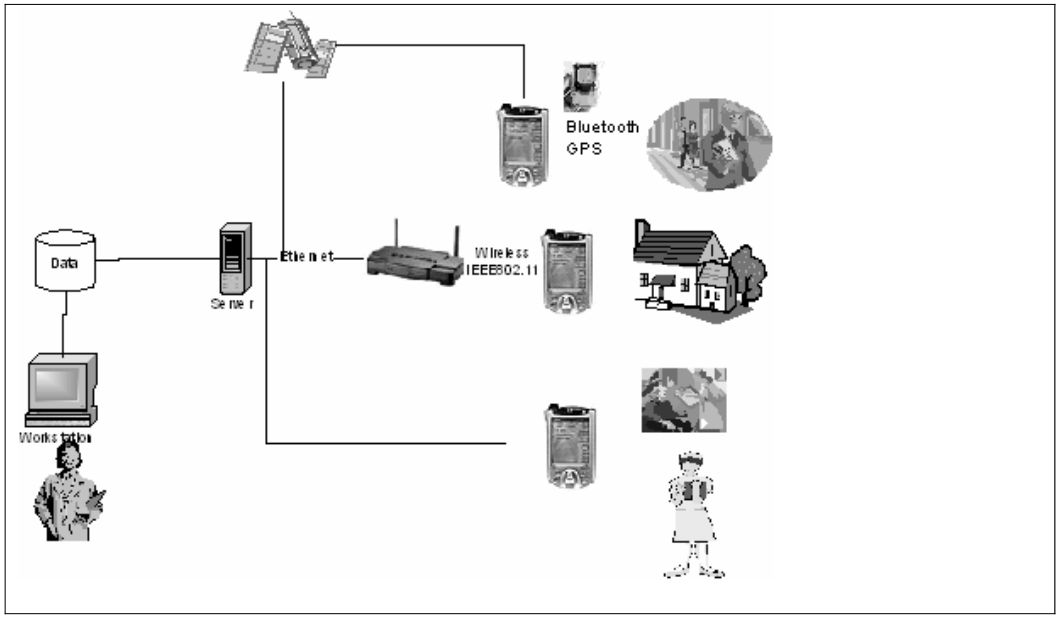


Figure 5. Connection infrastructure



may want to display the information under change. They request all the information needed to update their screen displayed.

Different types of communication are used between the PDAs and the server. Inside, when the connection is established the transmission is in direct mode. Outside or when the connection is out, the communication is in batch mode. Messages are saved in a queue waiting until a connection becomes available. Finally, a configuration file is used to personalize the application: auto-login, font, font size, message strings, icons, and number of activities displayed.

The implementation is in standard Java (J2SE) for the server part. PersonalJava is used for the client part on PDAs. On the PDAs, user interfaces were implemented with SWT. On the PDAs, the virtual machine was J9 for PocketPC. The system was tested on HP iPAQ 4155 and HP iPAQ 1945 running Microsoft PocketPC OS.

DISCUSSION

Specific needs are addressed by our system for cognitively impaired people in order to foster their autonomy. Functions of an electronic agenda are redesigned to limit cognitive load, especially the burden put on memory. Presently only the patient name is displayed. We are evaluating the relevance of adding other information such as the patient address and caregiver phone number. It may be helpful when a patient is lost but it could raise the cognitive load by requiring one more screen.

The help function is germane to a panic button provided by various companies. A panic button is generally linked to a remote controller that intervenes when the button is pressed. The use of such panic systems remains generally difficult especially for elders suffering Alzheimer disease. First, we hope that providing on the PDA a help button they use frequently is more likely to be used in case of emergency. Secondly, making explicit the help button on the PDA by an icon or by writing the function is a guarantee they could understand its use.

Providing geo-referenced data is another critical issue. Knowing a patient position may be worthful and even vital in many situations and for a range of purposes:

- To send directly somebody at the patient location if he is in crisis or in danger
- To forbid specific zones
- To detect crisis states by analyzing paths and detecting pathologic patterns of movement of patients suffering from schizophrenia
- To predict the goal of a patient combining path history and analysis

When a patient is outdoors, a GPS connected to his PDA get his or position. This position is sent at predefined time interval to the server. Path analysis can then be performed. Currently client and server components are under development to specify and manage forbidden zones.

Mobile services are actually designed to offer the activities monitoring by professional caregivers. Families need too a similar system to be informed of difficulties. They feel insecure to let alone the patient. They could exploit geo-referenced data in the timetable information to know if it is conform to the one expected. For instance, a patient suffering from schizophrenia shows disorganization symptoms before experiencing a new crisis. Sending no information about the activity performed is a premonitory sign. Applying it to the Alzheimer disease may help to detect falls or other health problems.

CONCLUSION

In this chapter, we presented a client-server-based mobile system for patients suffering from cognitive deficits and their caregivers. Users will use mobile devices. They help especially to remember the activities a patient have to perform.

We adapted traditional agenda to decrease the cognitive prerequisites and to provide a permanent communication between a patient and his caregiver. All along the development process, attention was paid on the real needs expressed by patients and caregivers. A sociological questionnaire and study had established previously the needs among patients suffering Alzheimer disease, head injury and schizophrenia (Pigot et al., 2005). Periodic meetings with caregivers have ensured conformity of the design of services to the needs expressed. Finally, a clinical validation is under process among people suffering from schizophrenia.

Future works will integrate mobile services to a helpful environment designed for people suffering from cognitive impairments in the spirit of pervasive computing. For instance instead of asking a patient to notify when an activity is completed, the environment will be able to recognize the activities performed by means of sensors (Pigot, 2004). The home will adapt itself to alleviate autonomy, remembering when necessary what needs to be done and how to do it. Such smart homes are an answer to the growing demand of cognitively impaired people to stay safely at home.

REFERENCES

Boulos, K., & Maged, N. (2003). Location-based health information services: A new paradigm in personalised information delivery. International Journal of Health Geographics 2003, 2(2). Retrieved from http://www.ij-healthgeographics.com/content/2/1/2

Giroux, S., Carboni, D., Paddeu, G., Piras, A., & Sanna, S. (2003). Delivery of services on any device: From Java code to user interface. The 10th International Conference on Human Computer Interaction 2003, June 22-27, 2003, Crète, Grèce. In C. Stephanidis & J. Jacko (Eds.), Human computer interaction (Vol. 1-2). Laurence Erlbaum Associates.

Gorman, P., Dayle, R., Hood, C. A., & Rumrell, L. (2003). Effectiveness of the ISAAC cognitive prosthetic system for improving rehabilitation outcomes with neurofunctional impairment. NeuroRehabilitation, 18, 57-67.

Haberman, V., Jones, M., & Mueller, J. (2005). Mobile technology, compensatory aids, and usability evaluations. Pervasive Computing, 4(2), 82-83.

Hareven, T. K. (2001). Historical perspectives on aging and family relations. In R. H. Bintock, & L. K. George (Eds.), Handbook of aging and the social sciences (pp. 141-159). New York: Academic press.

Helal, S. et al. (2003). Smart phone based cognitive assistant. UbiHealth 2003: The 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications, Seattle, Washington, October 12, 2003.

Lange, M. L. (2002). Technology and occupation: Contemporary viewpoints. The future of electronic aids to daily living. American Journal of Occupational Therapy, 56(1), 107-109.

Médical Intelligence. (2005). A simple solution to prevent disappearance. Retrieved from http://www.medicalintelligence.ca/en/columba.html

Neuropage. The Oliver Zangwill Centre Princess of Wales Hospital, Cambridgeshire, UK. Retrieved from http://www.neuropage.nhs.uk/

Patterson, D. J., Liao, L., Gajos, K., Collier, M., Livic, N., Olson, K.,Wang, S., Fox D., & Kautz, H. (2004, October). Opportunity knocks: A system to provide cognitive assistance with transportation services. In N. Davies, E. Mynatt, & I. Siio (Eds.), Proceedings of UBICOMP 2004: The 6th International Conference on Ubiquitous Computing (LNCS 3205, pp. 433-450). Springer-Verlag.

Pigot, H., Lefebvre, B., Meunier, J. G., Kerhervé, B., Mayers, A., & Giroux, S. (2003, April 2-4). The role of intelligent habitats in upholding elders in residence. The 5th International Conference on Simulations in Biomedicine, Slovenia (pp. 497-506).

Pigot, H., Savary, J. P., Metzger, J. L., Rochon, A., & Beaulieu, M. (2005). Advanced technology guidelines to fulfill the needs of the cognitively impaired population. The 3rd International Conference on Smart Homes and Health Telematic. Canada. (In press.)

Pollack, M. E. (2004). Special committee on aging. United States Senate hearing on assistive technology for aging populations. Retrieved from http://www.eecs.umich.edu/~pollackm/Pollack-web_files/senate-testimony.pdf

Shriqui, C. L., & Nasrallah, H. A. (1995). Contemporary issues in the treatment of Schizophrenia. Gilmore Academic Psychiatry.

Szymkowiak, A., Morrison, K., Prveen Shah, P. G., Evans, J. J., & Wilson, B. A. (2005). A memory aid with remote communication using distributed technology. Personal and Ubiquitous Computing, 9, 1-5.

Visions. (2005). The visions system. Retrieved from http://www.thevisionssystem.com/

Weiser, M. (1991). The Computer for the 21st Century. Scientific American.

KEY TERMS

Batch Mode: The batch mode indicates the transmission mode where the data are sent to the receptor a moment after the transmittor command.

Cognitive Assistance: It is the assistance provided by the environment to compensate the cognitive deficits.

Cognitive Deficits: They are the deficits encountered by a person following a brain lesion. The memory losses and the lacks of planning, initiation, and attention are examples of cognitive deficits.

Direct Mode: The direct mode indicates the transmission mode where the data are sent to the receptor immediately after the transmitter command.

Mobile Services: Mobile services refer to the software applications available on portable devices such as PDA or smart phone.

Pervasive Computing: Pervasive computing is the next generation computing environments with information and communication technology everywhere, for everyone, at all times.

Reminder: A reminder is a prosthetics aid which displays the common agenda functions and information on memories or advices to be followed.

Remote Monitoring: See smart home.

Sensors: The sensors are electrical devices that gather information on the state of the environment and on the localization and the activities performed by the inhabitant. For instance some sensors detect the presence of the inhabited or which door is open.

Smart Home: The smart home is the home equipped with sensors and effectors that could react according to the inhabitant actions. The smart home could warn persons outside by the remote monitoring or give advices to the inhabitant.

This work was previously published in Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 544-554, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.27
Context-Aware Mobile Capture and Sharing of Video Clips

Janne Lahti
VTT Technical Research Centre of Finland, Finland

Utz Westermann[1]
VTT Technical Research Centre of Finland, Finland

Marko Palola
VTT Technical Research Centre of Finland, Finland

Johannes Peltola
VTT Technical Research Centre of Finland, Finland

Elena Vildjiounaite
VTT Technical Research Centre of Finland, Finland

ABSTRACT

Video management research has been neglecting the increased attractiveness of using camera-equipped mobile phones for the production of short home video clips. But specific capabilities of modern phones — especially the availability of rich context data — open up new approaches to traditional video management problems, such as the notorious lack of annotated metadata for home video content. In this chapter, we present MobiCon, a mobile, context-aware home video production tool. MobiCon allows users to capture video clips with their camera phones, to semi-automatically create MPEG-7-conformant annotations by exploiting available context data at capture time, to upload both clips and annotations to the users' video collections, and to share these clips with friends using OMA DRM. Thereby, MobiCon enables mobile users to effortlessly create richly annotated home video clips with their camera phones, paving the way to a more effective organization of their home video collections.

INTRODUCTION

With recent advances in integrated camera quality, display quality, memory capacity, and video compression techniques, people are increasingly becoming aware that there mobile phones can be used as handy tools for the spontaneous capture of interesting events in form of small video clips. The characteristics of mobile phones open up new ways of combining traditionally separated home video production and management tasks at the point of video capture: The ability of mobile phones to run applications allows video production tools that combine video capture and video annotation. The classic approach of using video annotation tools to provide metadata for the organization and retrieval of video long after capture lacks user acceptance leading to the characteristic lack of metadata in the home video domain (Kender & Yeo, 2000). Context data about video capture available on mobile phones can be exploited to ease annotation efforts, which users try to avoid even at the point of capture (Wilhelm, Takhteyev, Sarvas, van House, & Davis, 2004). Time, network cell, GPS position, address book, and calendar can all be used to infer events, locations, and persons possibly recorded.

Furthermore, mobile phone-based video production tools can combine video capture with video upload and video sharing. With the ability to access the Internet via 2G and 3G networks from almost anywhere, phone users can directly load their clips to their home video collections stored on their PCs or by service providers disencumbering the limited memory resources of their phones. They also can share clips instantly with their friends via multimedia-messaging services. Digital rights management platforms like OMA DRM give users rigid control over the content they share preventing unwanted viewing or copying of shared clips.

However, video management research so far has mainly regarded mobile devices as additional video consumption channels. There has been considerable work concerning mobile retrieval interfaces (e.g., Kamvar, Chiu, Wilcox, Casi, & Lertsithichai, 2004), the generation of video digests for mobile users (e.g., Tseng, Lin, & Smith, 2004), and adaptive video delivery over mobile networks (e.g., Böszörményi et al., 2002), but a comprehensive view that considers the use of mobile phones as video production tools is still missing.

In this chapter, we present MobiCon: a context-aware mobile video production tool. Forming a cornerstone of the Candela platform, which addresses mobile home video management from production to delivery (Pietarila et al., 2005), MobiCon allows Candela users to record video clips with their camera phones and to semi-automatically annotate them at the point of capture in a personalized fashion. After recording, MobiCon extracts context data from the phone and passes it to an annotation Web service that derives reasonable annotation suggestions. These do not only include time- or position-based suggestions such as the season, city, or nearby points of interest possibly documented by the video; they also include personal calendar- and address book-based suggestions such as likely documented events and known locations like a friend's house. Besides these suggestions, the user can select concepts from a personal ontology with little manual effort or enter keywords for additional annotation.

MobiCon is further capable of uploading clips and their annotations to the users' private video collections in Candela's central video database directly after capture and permits users to immediately share these clips with friends, granting controlled access via OMA DRM.

Thus, MobiCon enables mobile phone users to create and share richly annotated home video clips with little effort, paving the way towards the more effective organization of their home video collections. The extensible architecture of the annotation Web service allows us to embrace and incrementally integrate almost any method for the generation of annotation suggestions based

on context without having to change the MobiCon application.

In the following, we first illustrate the use of MobiCon in an application scenario. We then relate MobiCon to state-of-the-art mobile home video production tools. After a brief coverage of the Candela platform, we provide a technical description of the MobiCon tool. We provide a discussion and outline future developments, before we come to a conclusion.

MOBICON APPLICATION SCENARIO

In this section, we want to provide an intuitive understanding of MobiCon by illustrating its usage for home video clip production and sharing in a typical application scenario.

In the scenario, MobiCon is used to produce two video clips of a birthday barbecue and sauna party. Figure 1 depicts a sequence of screenshots of the basic steps involved when using MobiCon to capture, annotate, and share a video clip showing some guests having a beer outdoors; Figure 2 shows a similar sequence for an indoor clip showing guests leaving the sauna that is created by a different user, who also wants to restrict the playback of the shared clip via DRM protection. After the capture of both video clips (Figure 1(a) and Figure 2(a)), the users can immediately annotate them. MobiCon gathers context data from each phone and passes it to an annotation Web service operated by the Candela platform. Based on this data, the Web service infers possible annotations that are suggested to the users (Figure 1(b) and Figure 2(b)). Suggestions do not only include rather simple ones inferred from the capture time like “April” and “evening” (Figure 2(b)); when a mobile phone is connected to a GPS receiver that MobiCon can access, they also include location annotations like “Oulu” (town) and “Peltokatu” (the street name) the Web service derived from the GPS position of the capture using a reverse-geocoder (Figure 1(b)). The availability of a current

Figure 1. Basic video capture, annotation, and sharing with MobiCon

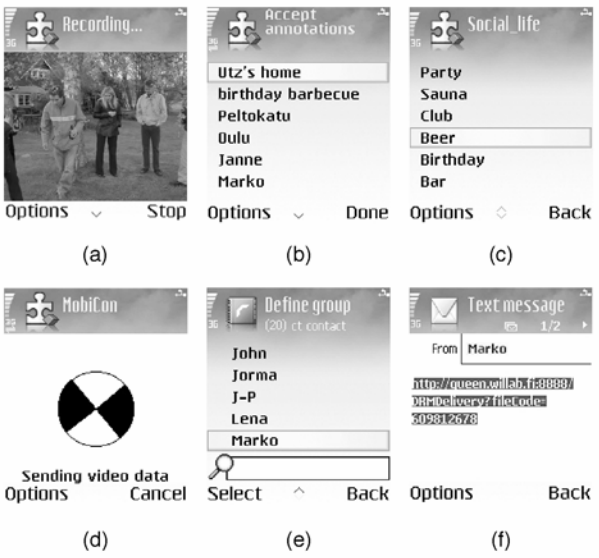
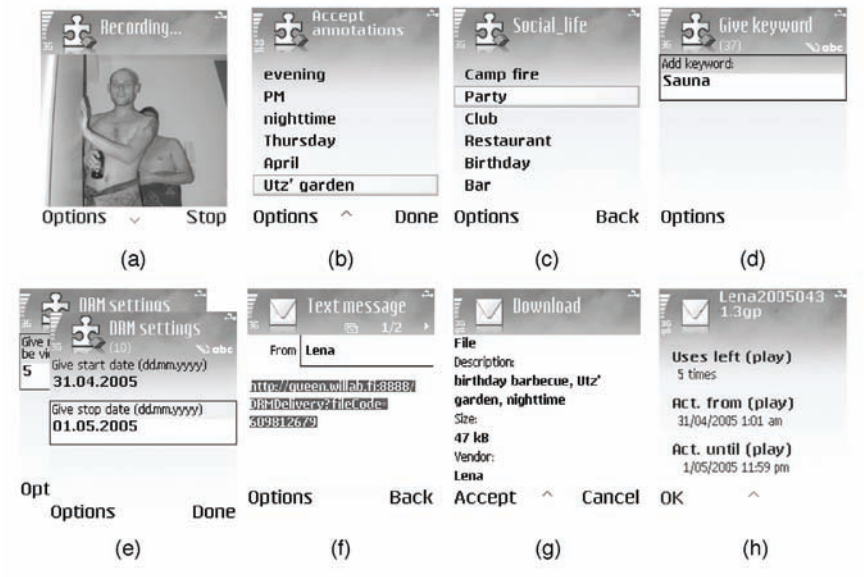


Figure 2. DRM-protected sharing of clips



GPS position also suggests that a clip covers an outdoor event (not shown in Figure 1(b)).

There are further highly personalized suggestions derived from phone address books and calendars, which can be synchronized with the Web service. Matching derived location information from the entries in a user's address book, the Web service can suggest known locations like "Utz's home" as annotations (Figure 1(b)); matching the capture time with the entries in a user's calendar, the Web service can suggest documented events like "birthday barbecue" (Figure 1(b)) along with event locations like "Utz's garden" (Figure 2(b)) and participants like "Janne" and "Marko" (Figure 1(b)) provided with the calendar entries. Users can correct the suggestions of the annotation Web service. In Figure 1(b), for instance, the user can remove the name "Marko" because he does not appear in the video.

In addition to the automatically generated annotation suggestions, MobiCon allows users to provide personalized manual clip annotations. Users can select concepts from personal, hierarchically organized home video ontologies that cover the aspects of their daily lives that they frequently document with video clips. The creator of the first video clip likes to have beers with friends, so his personal ontology contains the concept "beer" as a sub concept of "social life" (Figure 1(c)) that he can simply select for the annotation of his clip. The ontology of the creator of the second clip can contain different concepts due to different interests, such as the concept "camp fire" depicted in Figure 2(c). For the annotation of situations not covered by a user's personal ontology, MobiCon permits the entry of arbitrary keywords with the phone's keyboard as a last resort (Figure 2(d)).

After annotation, MobiCon uploads video clips and annotations to the users' personal video collections on the Candela platform (Figure 1(d)). Furthermore, MobiCon allows users to share freshly shot clips with contacts from their phone address books (Figure 1(e)). MobiCon then sends a text message with a link pointing to the shared clip in the user's collection to each selected contact, as depicted by Figure 1(f). When the recipient selects the link, the phone will download and play the clip. The second video clip shows the somewhat delicate situation of two party guests coming out of the sauna. While the creator of this clip still wants to share it with a friend, she wants to impose usage restrictions. Utilizing MobiCon's DRM support, she restricts playback of the shared clip to five times within the next 24 hours on the phone of her friend (Figure 2(e)). MobiCon makes the Candela platform prepare a copy of the clip that encodes these limitations using OMA DRM. The link to the video contained in the text message that is then sent to the friend points to the DRM-protected copy (Figure 2(f)). After selecting the link, the recipient sees a description of the clip and is asked for permission to download Figure 2(g)). If download is accepted, the OMA-DRM-compliant phone recognizes and enforces the restrictions imposed upon the clip and displays the corresponding DRM information before starting playback (Figure 2(h)).

RELATED WORK

The previous section illustrated MobiCon's different functionalities from a user's perspective in a typical application scenario. We now compare MobiCon to existing approaches in the field of mobile video production tools, thereby showing how it exceeds the state-of-the-art. In particular, we relate MobiCon to mobile video capture tools, mobile video editing applications, mobile video annotation tools, and tools for mobile content sharing.

Mobile Video Capture Tools

Probably every modern mobile phone with an integrated camera features a simple video capture tool. MobiCon goes beyond these tools by not only allowing the capture of a video clip but also allowing for immediate annotation for later retrieval, its immediate upload to the user's home video clip collection, as well as its immediate sharing controlled via OMA DRM.

Mobile Video Editing Tools

Mobile video editing tools like Movie Director (n.d.) or mProducer (Teng, Chu, & Wu, 2004) facilitate simple and spontaneous authoring of video clips at the point of capture on the mobile phone. Unlike MobiCon, the focus of these tools lies on content creation and not on content annotation, uploading, and sharing.

Mobile Video Annotation Tools

While there are many PC-based tools for video annotation as a post-capturing processing step (e.g., Abowd, Gauger, & Lachenmann, 2003; Naphade, Lin, Smith, Tseng, & Basu, 2002), mobile tools like MobiCon permitting the annotation of video clips at the very point of capture, when users are still involved in the action, are rare. M4Note (Goularte, Camancho-Guerrero, Inácio Jr., Cattelan, & Pimentel, 2004) is a tool that allows the parallel annotation of videos on a tablet PC while they are being recorded with a camera. Unlike MobiCon, M4Note does not integrate video capture and annotation on a single device. Annotation is fully manual and not personalized; context data is not taken advantage of for suggesting annotations. M4Note does not deal with video upload and sharing.

Furthermore, mobile phone vendors usually provide rudimentary media management applications for their phones that — compared to MobiCon and its support for annotation sug-

gestions automatically derived out of context data and personalized manual annotation using concepts from user-tailored ontologies and keywords — offer only limited video annotation capabilities. As an example, Nokia Album (n.d.) allows the annotation of freshly shot clips with descriptive titles. As a form of context-awareness, Nokia Album records the time stamps of video captures but does not infer any higher-level annotations out of them.

The lack of sophisticated mobile video annotation tools constitutes a contrast to the domain of digital photography. Here, research has recently been investigating the use of context data such as time and location to automatically cluster photographs likely documenting the same event (Cooper, Foote, Girgensohn, & Wilcox, 2003; Pigeau & Gelgon, 2004) and to automatically infer and suggest higher-level annotations, such as weather data, light conditions, etc. (Naaman, Harada, Wang, Garcia-Molina, & Paepcke, 2004). Compared to MobiCon, these approaches do not present the inferred annotation suggestions to users at the point of capture for immediate acceptance or correction; inference takes place long afterwards when the photographs are imported to the users' collections.

For the annotation of photographs at the point of capture, Davis, King, Good, and Sarvas (2004) have proposed an integrated photo capture and annotation application for mobile phones that consults a central annotation database to automatically suggest common annotations of pictures taken within the same network cell. Apart from its focus on video, MobiCon mainly differs from this approach by offering a different and broader variety of derivation methods for context-based annotation suggestions and by addressing content upload and sharing.

Mobile Content Sharing Tools

Mobile content sharing applications like PhotoBlog (n.d.), Kodak Mobile (n.d.), and MobShare (Sarvas, Viikari, Pesonen, & Nevanlinna, 2004) allow users to immediately share content produced with their mobile phones, in particular photographs. Compared to MobiCon, there are two major differences. Firstly, these applications realize content sharing by uploading content into central Web albums, in which users actively browse for shared content with a Web browser. In contrast, MobiCon users view shared content by following links in notification messages they receive. Also, MobiCon gives users more control over shared content by applying DRM techniques.

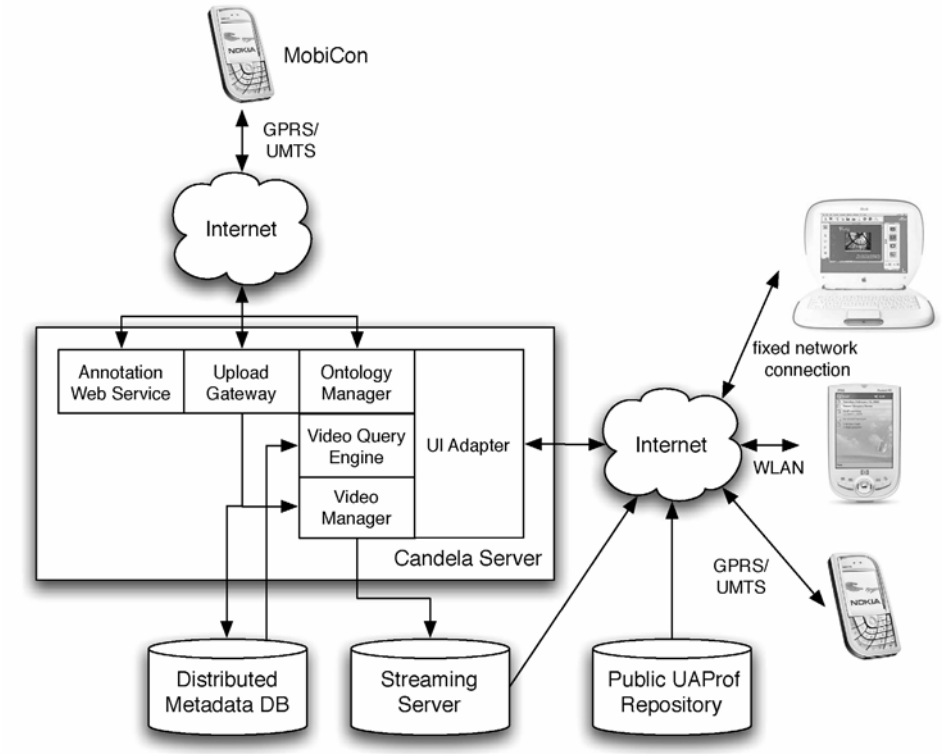
Secondly, current content sharing systems offer rather restricted means for content annotation, mainly allowing content to be manually assigned to (usually flat) folder structures and attaching time stamps for folder- and timeline-based browsing. Nokia Lifeblog (n.d.) goes a bit beyond that by automatically annotating content with the country where it has been created, which is obtained from the mobile network that the phone is currently logged in to. But compared to MobiCon, these still constitute very limited forms of context-based annotations.

THE CANDELA PLATFORM

Facing the increasingly popular use of mobile devices for home video production, we have developed the Candela mobile video management platform. Incorporating MobiCon, it provides support for all major process steps in the mobile home video management chain, ranging from mobile video creation, annotation and sharing to video storage, retrieval, and delivery using various mobile and stationary terminals connected to the Internet via various types of networks like GPRS/EDGE, 3G/UMTS, WLAN, and fixed networks. In the following, we briefly describe the platform's key elements and their relationship to MobiCon.

Figure 3 illustrates the interplay of the different components of the Candela platform. As explained

Figure 3. Candela platform architecture



before, the MobiCon mobile phone-based video production application permits the integrated capture, personalized, context-aware annotation, upload, and DRM-controlled sharing of video clips. To this end, MobiCon interacts closely with the central Candela server, namely with its ontology manager, annotation Web service, and upload gateway components.

The RDF-based ontology manager stores the personal home video ontologies of Candela's users. When MobiCon starts for the first time, it loads the ontology of the current user from the manager so that its concepts can be used for the personalized annotation of videos.

The annotation Web service is called by MobiCon during clip annotation, passing context data such as capture time, GPS position, and user information. The Web service derives annotation suggestions based on this data, which MobiCon then presents to the user.

The upload gateway is used to transfer clips and their annotation after capture from MobiCon to the users' video collections. The gateway receives the clips in 3GP format and clip metadata including user annotations and context data in MPEG-7 format. The clips are passed on to the video manager for storage and transcoding into suitable formats for the video players of different devices and for different network speeds. The

video manager also prepares OMA DRM-enhanced clip variants when MobiCon users define usage restrictions for the video clips that they are about to share. The clip metadata is stored in a database implemented on top of the Solid Boost Engine distributed relational database management system for scalability to large numbers of users and videos.

Via its UI adapter, video query engine, and video manager components, the Candela server also provides rich video retrieval facilities. While MobiCon is a standalone mobile phone application, the video retrieval interfaces of the Candela platform are Web browser-based. Thus, we can apply Web user interface adaptation techniques to give users access to their video collections from a variety of user terminals and networks.

The UI adapter is implemented on top of an Apache Cocoon Web-development framework. Using XSLT stylesheets, it generates an adaptive video browsing and retrieval interface from an abstract XML-MPEG7 content, considering the capabilities of the user devices obtained from public UAProf repositories. For example, when using a PC Web browser, the adapter creates a complex HTML interface combining keyword queries, ontology-based video browsing, as well as the display and selection of query results into a multi-frame page. When using a mobile phone browser, the adapter splits the same interface into several HTML pages.

For performing video browsing and content-based retrieval, the UI adapter interacts with the video query engine, which supports the use of time, location, video creators, and keywords as query parameters. The video query engine translates these parameters into corresponding SQL statements run on the metadata database and returns a personalized ranked result list in MPEG-7 format, which the UI adapter then integrates into the user interface. The engine interacts with the ontology manager for personalized keyword expansion. For example, the search term "animal" will be expanded to all subconcepts of "animal," (e.g., "cat" and "dog") in querying user's personal ontology.

When a video clip is selected for viewing, the video manager takes care of its delivery. It selects the format and compression variant most appropriate to the client device and network, again exploiting the device capability profiles in the public UAProf repositories–especially the information about screen size, and the video manager supports HTTP-based download of a clip as well as streaming delivery via the Helix DNA streaming server.

MOBICON

MobiCon is a Java 2 Micro Edition/MIDP 2.0 application that runs on Symbian OS v8.0 camera phones with support of the Mobile Media, Wireless Messaging, and Bluetooth APIs. We now provide details on the video production and management tasks—video capture, annotation, upload, and sharing—combined by MobiCon.

Video Capture

When MobiCon is started for the first time, the user is authenticated by the Candela platform. Upon successful authentication, MobiCon receives the user's personal ontology from the ontology manager and stores it along with the user's credentials in the phone memory for future use making use of MIDP record management, as it is assumed that the user stays the same. MobiCon still permits re-authentication for a different user.

After successful login, users can start capturing clips. For this purpose, MobiCon accesses the video capture tool of the mobile phone via the Mobile Media API. The captured content is delivered in 3GP format, using AMR for audio encoding and H.263/QCIF at 15 frames per second and 174x144 pixels resolution for video encoding. MobiCon stores the captured video clip in the phone's memory. Users can view the

captured or another stored clip, capture another clip, or start annotating a stored clip as explained in the following.

Video Annotation

For the annotation of video clips, MobiCon provides automatic, context-based annotation suggestions as well as the option to manually annotate clips with concepts of personal home video ontologies or keywords. We now provide more details on the generation of context-based annotation suggestions and the use of personal ontologies for annotation.

Context-Based Annotation Suggestions

For the generation of appropriate annotation suggestions, MobiCon gathers context data that is available about the capture of a video clip on the mobile phone. In particular, MobiCon collects the username, capture time, and duration of the clip. Additionally, MobiCon is able to connect via the Bluetooth API to GPS receivers that support the NMEA protocol. If such a receiver is connected to the phone, MobiCon polls for the current GPS position and stores it along with a timestamp as a measure for its age. Given these context data, MobiCon invokes the annotation Web service running on the Candela server as a Java servlet via an HTTP request, opening a connection to the Internet via UMTS or GPRS if not yet established.

The reasons for outsourcing the derivation of annotation suggestions to a Web service are mainly ease of prototyping and deployment. We can incrementally add new methods for annotation suggestions to the Web service while keeping the MobiCon client unchanged, thus saving on update (re)distribution costs. Also, a Web service allows the reuse of the context-based annotation suggestion functionality on devices other than mobile phones.

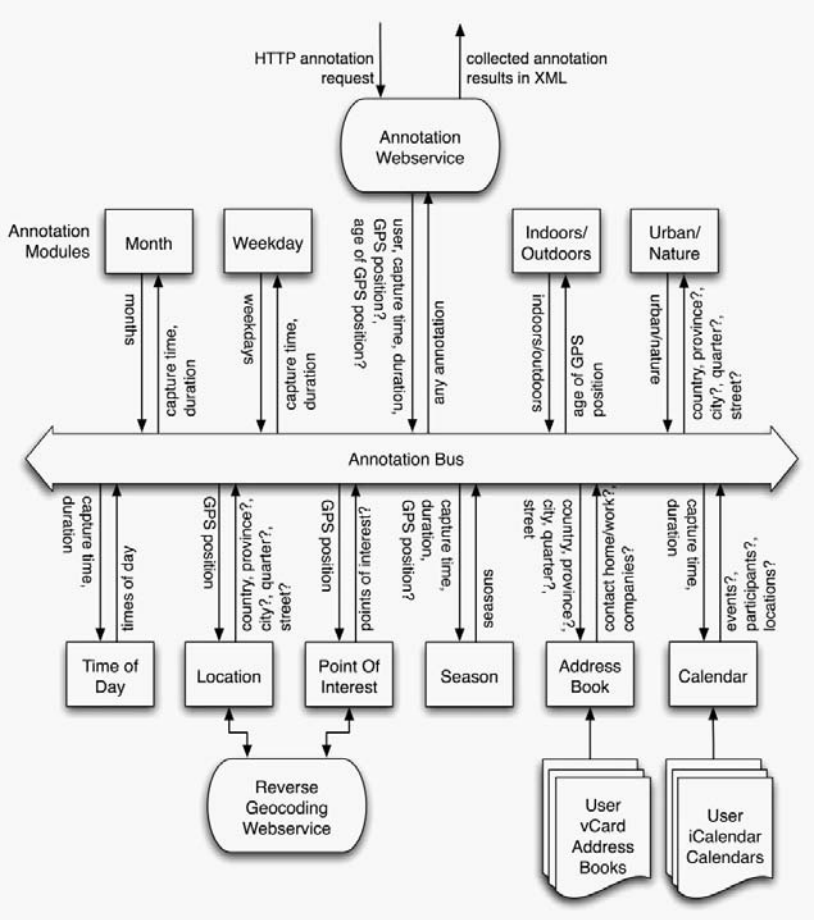
A drawback of this design is the costs incurred by remotely invoking a Web service from a mobile phone. But given the costs accrued anyway by uploading and sharing comparably high-volume video clips, these are negligible. A further problem is how to provide the Web service with access to personal user data for the generation of annotation suggestions, such as phone calendars or address books; passing the whole address book and calendar of a user as parameters to the Web service with each invocation is certainly not feasible. Leaving privacy issues aside, we circumvent this problem by allowing users to upload their calendars and address books to a central directory on the Candela server in iCalendar and vCard formats via a MobiCon menu option. From this directory, this data can be accessed from the Web service with user names as keys.

Figure 4 presents an overview of the design of the annotation Web service. When the Web service receives an annotation request, it publishes the context data carried by the request on the annotation bus. The annotation bus forms a publish/subscribe infrastructure for annotation modules that are in charge of actually deriving annotation suggestions. The annotation modules run concurrently in their own threads, minimizing response times and maximizing the utilization of the Web service's resources when processing multiple annotation requests.

The annotation modules listen to the bus for the data they need for their inferences, generate annotation suggestions once they have received all required data for a given annotation request, and publish their suggestions back to the bus, possibly triggering other annotation modules. The annotation Web service collects all suggestions published to the bus for a request, and, once no more suggestions will be generated, returns the results to MobiCon.

This results in a modular and extensible design: the annotation modules used for the generation of annotation suggestions can be selected to suit the needs of an individual application and new

Figure 4. Annotation Web service design



modules can be dynamically added to the system as they become available without having to reprogram or recompile the Web service.

Figure 4 also provides information about the annotation modules currently implemented, along with the types of data on which they base their inferences and the types of suggestions they publish. In the following, we highlight some of the more interesting ones:

The location and point of interest annotation modules suggest address and points of interests probably captured by the clip being annotated based on GPS position utilizing the commercial ViaMichelin reverse-geocoding Web service. The calendar annotation module searches the user calendar for events that overlap with the capture time, suggesting event names, locations, and participants as annotations. The address book annotation module searches the user address book for the home or work addresses of contacts or company addresses matching the address data derived by any other annotation module, suggesting them

as location annotations. The indoors/outdoors annotation module suggests whether a clip has been shot outdoors or indoors, utilizing the fact that GPS signals cannot be received indoors and thus the age of the GPS position will exceed a threshold in this case. Depending on the level of detail of address data derived by other modules, the urban/nature annotation module suggests whether a clip shows an urban environment or nature. If information about a city or street is missing, it suggests nature, otherwise an urban environment is assumed.

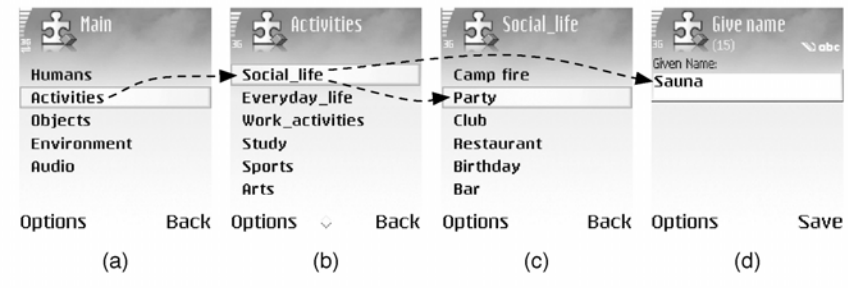
Ontology-Based Annotations

MobiCon permits an inexpensive manual annotation of content using hierarchically structured ontologies with concepts from the daily lives of users. Instead of having to awkwardly type such terms with the phone keyboard over and over again, users can simply select them by navigating through MobiCon's ontology annotation menu as illustrated in Figure 5 (a-c). Without imposing a single common ontology onto every user, MobiCon permits each user to have a personal ontology for home video annotation, merely predefining two upper levels of generic concepts that establish basic dimensions of video annotation (Screenshots (a) and (b) of Figure 5). Below these levels, users are free to define their own concepts, such as those depicted in Screenshot (c). MobiCon's user interface permits the entry of new concepts at any level at any time during the annotation process in Screenshot (d).

The rationale behind this approach is as follows: firstly, it allows users to optimize their ontologies for their individual annotation needs, so that they can reach the concepts important to them in few navigation steps and without having to scroll through many irrelevant concepts on a small phone display on the way. Our experiences from initial user trials indicate that precisely because users want to keep annotation efforts low, they are willing to invest some efforts into such optimization.

The concepts that are important for clip annotation differ very much between people: a person often enjoying and documenting sauna events might introduce "sauna" as a subconcept of "social life" to his or her ontology, whereas an outdoor person might need a subconcept "camp fire", and so on. Differences also occur in the hierarchical organization of concepts: users frequently visiting bars might consider the concept "bar" as a subconcept of "social life" (like in Screenshot (c)),

Figure 5. MobiCon ontology user interface



while a bar's owner might see it as a subconcept of "work activity."

Secondly, by imposing a common set of top-level concepts (used for representation of profiles of users' interests) onto the personal ontologies of the users, we establish a common foundation for the querying and browsing of video collections, making it easier to find interesting clips also in the collections of other users.

MobiCon receives the personal ontology of a user from the ontology manager in RDF format after successful authentication and caches it for successive use in the phone's memory.

Video Upload and Storage

After annotation, MobiCon gives the user an opportunity to upload the video clip and its annotations to his or her video collection on the Candela server via the upload gateway. As already explained, the video clip is handed over to the video manager which transcodes it to different formats at different bit rates in order to provide a scaleable service quality for different devices and network connections: Real Video, H.264, and H.263 encodings are used for delivering video content to mobile devices, as well as MPEG4 for desktop computers. In the future, scalable video codecs will remove the need of transcoding.

The clip metadata is represented in MPEG-7 format that mainly constitutes a profile of the video and video segment description schemes defined by the standard. Figure 6 gives a sample of this format. It incorporates context data about the clip's capture including the creator's name, GPS position, region and country, date and time of day, and length of the video clip, as well as the clip annotations embedded in free text annotation elements. This includes the suggestions generated by the annotation Web service, the concepts selected from the user's personal home video ontology, and the keywords manually provided by the user.

Video Sharing

Users can share uploaded clips with the contacts in their address book, defining usage restrictions according to the OMA DRM standard if desired. The standard offers three approaches to content protection: forward-lock, combined delivery, and separate delivery. Forward-lock thwarts the forwarding of content to a different device, while combined delivery allows one to impose further restrictions, such as a limited number of playbacks or a permissible time interval for playback. In both approaches, the protected content is embedded by the content provider in a DRM packet along with the specification of the usage restrictions. Under separate delivery, the restrictions and the content are delivered separately and integrated on the playback device.

MobiCon supports the protection of video clips via forward-lock and combined delivery. For reasons of implementation, usage complexity, and the requirements imposed onto client devices, we have chosen not to support separate delivery at this stage.

When the user has specified the desired usage restrictions for a clip being shared, MobiCon uses a secure connection to contact the video manager, which employs the Nokia Content Publishing Toolkit to put a copy of the video clip into a DRM packet with the specified restrictions. The video manager also creates a key pair for each recipient of the clip. One key of every pair remains with the DRM packet, while the other is returned to MobiCon.

Using the Wireless Messaging API, MobiCon then sends a text-message to each recipient containing URL-link with a key pointing to the DRM protected clip. When the recipient of the message selects the link, the phone establishes an HTTP connection to the video manager. Using the recipient's key, the video manager checks whether access to the DRM protected clip can be granted by pairing the key with the right clip. If a matching clip is found, a download descrip-

Figure 6. The MobiCon metadata format

```
<?xml version="1.0" encoding="iso-8859-1"?>
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
       xmlns:xsi="http://.../XMLSchema-instance"
<Description xsi:type="ContentEntityType">
  <DescriptionMetadata>
    <Creator>
      <Role href="creatorCS">
        <Name>Utz</Name>
      </Role>
    </Creator>
    <CreationLocation>
      <GeographicPosition>
        <Point latitude="65.0057"
               longitude="25.4654"/>
      </GeographicPosition>
      <Region>FI</Region>
      <Country>Finland</Country>
    </CreationLocation>
    <CreationTime>
      2005-04-07T20:36:00+03:00
    </CreationTime>
    <Instrument>
      <Tool>
        <Name>MobiCon v1.0</Name>
      </Tool>
    </Instrument>
  </DescriptionMetadata>
  <MultimediaContent xsi:type="VideoType">
    <Video>
      <MediaLocator>
        <MediaUri>
          mobileUpload/Utz200543118.3gp
        </MediaUri>
      </MediaLocator>
      <MediaTime>
        <MediaTimePoint>
          T00:00:00:0F30000
        </MediaTimePoint>
        <MediaDuration>
          PT0H0M24S0N30000F
        </MediaDuration>
      </MediaTime>
      <TemporalDecomposition>
      <VideoSegment>
        <TextAnnotation confidence="1"
                        relevance="1"
                        type="scene">
          <FreeTextAnnotation>
            Peltokatu
          </FreeTextAnnotation>
          <FreeTextAnnotation>
            Janne
          </FreeTextAnnotation>
            ...
          <FreeTextAnnotation>
            beer
          </FreeTextAnnotation>
        </TextAnnotation>
        <MediaTime>
          <MediaTimePoint>
            T00:00:00:0F30000
          </MediaTimePoint>
          <MediaIncrDuration>
            ...
          </MediaIncrDuration>
        </MediaTime>
      </VideoSegment>
      </TemporalDecomposition>
    </Video>
  </MultimediaContent>
</Description>
</Mpeg7>
```

tor with basic information about the clip like creator, length, and description is returned to the recipient's mobile phone and the used key pair is removed, in order to prevent re-usage. After deciding to really download the packet, the user can finally watch the protected video clip, but only on the paired device and within the limits of the usage restrictions.

DISCUSSION

Having given a technical description of the MobiCon application for the combined production, context-aware annotation, and sharing of home video clips with mobile phones at the point of capture, we now provide a critical discussion and outline future developments.

The ways in which the annotation Web service can utilize temporal and spatial context data for the generation of annotation suggestions are not limited to those described in the previous section: weather or light conditions probably documented by a video can be obtained from meteorological databases given capture time and location (Naaman et al., 2004), annotations from other videos shot at the same time and place can be suggested using clustering methods (Davis et al., 2004; Pigeau & Gelgon, 2004), and much more. We want to support these uses for time and location context

data with MobiCon as well. For that purpose, we benefit from the extensible design of the annotation Web service, as it enables us to incrementally develop and integrate modules for these kinds of annotation suggestions without having to modify the MobiCon application itself.

Reasonable annotation suggestions cannot only be derived from context data, from content analysis, or a combination of both. We plan to integrate an audio classifier that is capable of identifying segments of speech, music, and different kinds of environmental noises within videos with high degree of reliability. The results of such an audio classification can be used to enhance our simplistic indoors/outdoors and urban/nature annotation modules, which so far are solely based on the age of the last available GPS position and the level of detail of the address returned by the reverse-geocoder for that position.

Integrating content analysis with the current centralized annotation Web service design is problematic. As an annotation module using content analysis methods needs access to the full video clip being annotated, the clip has to be uploaded to the Web service before any suggestions can be created. The incurring delay will hamper the capture and annotation process. Therefore, we want to distribute the annotation Web service, permitting annotation modules to run on the server and on the mobile phone. This will not only allow us to perform content analysis on the mobile phone avoiding upload delays; we will also be able to perform annotations based on sensitive personal data like address books and calendars directly on the phone, avoiding the privacy issues raised by moving such data to a central server as done currently.

Beyond improving the generation of annotation suggestions, MobiCon's user interface for annotating video clips on the basis of personal ontologies will also require some improvement. So far, users only have very limited means of modifying their ontologies in the middle of the video capture and annotation process, merely being able to add new subconcepts. Larger modifications must be performed outside of MobiCon using Candela's Web front-end. Moreover, MobiCon's DRM-based video sharing functionality is limited, allowing the sharing of clips only right after capture. We are currently investigating the integration of a user interface into MobiCon that allows users to share any clip existing in their collections.

Finally, we want to improve the video capturing and editing functionalities of MobiCon by integrating it with a mobile video editing application.

CONCLUSION

This chapter has introduced MobiCon, a video production tool for mobile camera phones that exploits specific characteristics of mobile phones— in particular the ability to run applications, the availability of context data, and access to the Internet from almost anywhere — to integrate traditionally separated home video production and management tasks at the point of video capture. MobiCon assists mobile phone users in capturing home video clips, uses context data after capture to suggest reasonable annotations via an extensible annotation Web service, supports personalized manual annotations with user-specific home video ontologies and keywords, uploads video clips to the users' video collections in Candela's central video database, and facilitates the controlled sharing of clips using OMA.

Initial experiences we have been able to gain so far from our personal use of MobiCon are encouraging. With MobiCon, the provision of useful annotations for home video clips is largely automatic and not overly intrusive to the general video capturing process, effectively resulting in the better organization of home video clips without much additional overhead. We are in the process of subjecting this personal experience towards a user study.

This work was done in the European ITEA project "Candela", funded by VTT Technical Research Centre of Finland and TEKES (National Technology Agency of Finland). Support of Finnish partners Solid Information Technology and Hantro Products is greatly acknowledged.

REFERENCES

Abowd, G. D., Gauger, M., & Lachenmann, A. (2003). The family video archive: An annotation and browsing environment for home movies. Proceedings of the 11th ACM International Conference on Multimedia, Berkeley, CA.

Böszörményi, L., Döller, M., Hellwanger, H., Kosch, H., Libsie, M., & Schojer, P. (2002). Comprehensive treatment of adaptation in distributed multimedia systems in the ADMITS project. Proceedings of the 10th ACM International Conference on Multimedia, Juan-les-Pins, France.

Cooper, M., Foote, J., Girgensohn, A., & Wilcox, L. (2003). Temporal event clustering for digital photo collections. Proceedings of the 11th ACM International Conference on Multimedia, Berkeley, CA.

Davis, M., King, S., Good, N., & Sarvas, R. (2004). From context to content: Leveraging context to infer multimedia metadata. Proceedings of the 12th ACM International Conference on Multimedia, New York.

Goularte, R., Camancho-Guerrero, J. A., Inácio Jr., V. R., Cattelan, R. G., & Pimentel, M. D. G. C. (2004). M4Note: A multimodal tool for multimedia annotations. Proceedings of the WebMedia & LA-Web 2004 Joint Conference, Ribeirão Preto, Brazil.

Kamvar M., Chiu P., Wilcox L., Casi, S., & Lertsithichai, S. (2004). MiniMedia Surfer: Browsing video segments on small displays. Proceedings of the 2004 Conference on Human Factors and Computing Systems (CHI 2004), Vienna, Austria.

Kender, J. R., & Yeo, B. L. (2000). On the structure and analysis of home videos. Proceedings of the 4th Asian Conference on Computer Vision (ACCV 2000), Taipei, Taiwan.

Kodak Mobile (n.d.). Retrieved May 3, 2005, from http://www.kodakmobile.com

Movie Director (n.d.). Retrieved May 3, 2005 from http://www.nokia.com/nokia/-0,6771, 54835,00.html

Naaman, M., Harada, S., Wang, Q. Y., Garcia-Molina, H., & Paepcke, A. (2004). Context data in geo-referenced digital photo collections. Proceedings of the 12th ACM International Conference on Multimedia, New York.

Naphade, M., Lin, C. Y., Smith, J. R., Tseng, B., & Basu, S. (2002). Learning to annotate video databases. Proceedings of the SPIE Electronic Imaging 2002 Symposia (SPIE Volume 4676), San Jose, California.

Nokia Album (n.d.). Retrieved May 3, 2005, from http://www.nokia.com/nokia/-0,6771, 54835,00.html

Nokia Lifeblog (n.d.). Retrieved May 3, 2005, from http://www.nokia.com/lifeblog

PhotoBlog (n.d.). Retrieved May 3, 2005, from http://www.futurice.fi

Pietarila, P., Westermann U., Järvinen, S., Korva J., Lahti, J., & Löthman, H. (2005). Candela — storage, analysis, and retrieval of video content in distributed systems — personal mobile multimedia management. Proceedings of the IEEE International Conference on Multimedia & Expo (ICME 2005), Amsterdam, The Netherlands.

Pigeau, A., & Gelgon, M. (2004). Organizing a personal image collection with statistical model-based icl clustering on spatio-temporal camera

phone meta-data. Journal of Visual Communication & Image Retrieval, 15(3), 425-445.

Sarvas, R., Viikari, M., Pesonen, J., & Nevanlinna, H. (2004). MobShare: Controlled and immediate sharing of mobile images. Proceedings of the 12th ACM International Conference on Multimedia, New York.

Teng, C. M., Chu, H. H., & Wu, C. I. (2004). mProducer: Authoring multimedia personal experiences on mobile phones. Proceedings of the IEEE International Conference on Multimedia & Expo (ICME 2004), Taipei, Taiwan.

Tseng, B. L., Lin, C. Y., & Smith, J. R. (2004). Using MPEG-7 and MPEG-21 for personalizing video. IEEE MultiMedia, 11(1), 42-52.

Wilhelm, A., Takhteyev, Y., Sarvas, R., van House, N., & Davis, M. (2004). Photo annotation on a camera phone. Proceedings of the 2004 Conference on Human Factors and Computing Systems (CHI 2004), Vienna, Austria.

KEY TERMS

3GP Format: Mobile phone video file format produced by mobile phone video recording applications.

Annotation: Extra information or note associated with a particular object.

Candela: A two-year EUREKA/ITEA project researching content analysis, delivery, and architectures.

DRM: Digital rights management is a method for licensing and protecting digital media.

GPS (Global Positioning System): A global satellite-based navigation system.

Metadata: Metadata is the value-added information of data, for example, describing a content of picture, video, or document.

MIDP 2.0 (Mobile Information Device Profile Version 2.0): A Java runtime environment for mobile devices.

MPEG-7 (Multimedia Content Description Interface): MPEG-7 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) to describe multimedia content.

OMA DRM (Open Mobile Alliance's Digital Rights Management): A standard developed by the OMA organization for the management of digital rights in mobile phones.

Ontology: A description of the concepts and relationships of objects in a formal way using a controlled vocabulary.

ENDNOTE

[1] This work was carried out under the tenure of an ERCIM fellowship.

This work was previously published in Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 340-356, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.28
From CCTV to Biometrics through Mobile Surveillance

Jason Gallo
Northwestern University, USA

INTRODUCTION

Surveillance is the act or process of observing, tracking, or recording personal details for the purpose of exercising control over the individual or population being watched. Control in this context can mean many things, from directly influencing the behavior of the observed to the use of gathered information for the purpose of management or governance.

Mobile surveillance can be defined as two distinct, yet related, practices. The first is the ability to observe the physical movement of an individual through space. This is most often accomplished through documenting their interaction with a surveillance network. The object of surveillance is tracked from one node of the network to another, providing a record of behavior. The second practice is often referred to as dataveillance, or the ability to monitor an individual's behavior through studying a trail of personally identifiable data, including credit card purchases, mobile phone calls, and health records.

Mobile surveillance employs an array of technologies including video and photography cameras, visual recognition software, radio frequency identification (RFID), global positioning receivers (GPS), information and communication technologies (ICTs), and biometrics. Examples of mobile surveillance networks include the dense deployment of closed-circuit television (CCTV), video, and photographic technologies in a distinct geographic space to monitor activity, the tracking of automobiles and mobile phones via GPS, and radio frequency sensing that records motion as identity chips pass through a distributed network of receivers. As these networks proliferate, individuals are the exposed to overlapping layers of surveillance. Although many of these surveillance networks are deployed for limited purposes, the increasing ability to save and store personally identifiable information in searchable databases, and the ability to mine information from multiple sources raises privacy concerns for the individual. This is especially true in advanced capitalist societies that rely on sophisticated data gathering

to track, model, and predict consumer behavior, as well as for citizen management.

BACKGROUND: SURVEILLANCE, BUREAUCRACY, AND THE STATE

Surveillance has been an integral part of human social interaction since the need for oversight and management of collective endeavors was first realized. As the scope and complexity of these endeavors grew, the need for more reliable information increased accordingly. Surveillance has long been an important method for dealing with risk (Lyon, 1994, 2002, 2003a, 2003b), as the advanced knowledge of aberrant behavior can help minimize the threat or upheaval caused by the unusual events or actions. Therefore, surveillance is often a positive feature of governance, allowing those in power to manage against risk in order to protect public welfare. Nevertheless, surveillance regimes are also employed by the state out of a "desire to more completely manage populations (Lyon, 2003b, p. 20)," identifying and sorting out individuals who's behavior is deemed threatening to the majority. It is of little surprise that the fields of law enforcement and national security and intelligence gathering are the sites of some of the most sophisticated surveillance practices as well as the targets of social concern over privacy and the power of the state.

In Discipline and Punish, Foucault (1977) examines the rise of the surveillance society by utilizing Jeremy Bentham's Panopticon prison as a model for the exercise of power in modern society. The architecture of the Panopticon exerts power over the incarcerated body by making it constantly visible to an invisible central observer. The prospect of persistent observation is used to ensure compliance with the disciplinary rules of the institution, therefore making the simple awareness of surveillance a means of exerting power over the watched individual.

Foucault (1977) notes the historic extension of surveillance architecture from the prison to other social institutions such as schools, hospitals, mental institutions, and the workplace, which increasingly relied on the specter of persistent observation in order to exert control over their subjects. In addition to the direct surveillance enabled by panoptic architecture, the rise of bureaucratic organizations, especially in the West, lead to an institutionalization of mechanisms for the capture, retention, and processing of personally identifiable data.

The direct and indirect surveillance employed by public libraries in Victorian Britain (Black, 2001) serves as a historical example of this phenomenon. Libraries have been at the forefront of efforts to manage, catalogue, and retrieve information since the sorting, and storing of information is central to their mission. To this end, libraries have employed increasingly sophisticated surveillance mechanisms to track, record, and monitor the habits of their users and their interaction with the library's collections. While the hierarchical systems of knowledge and the tracking of library users' habits employed in Victorian libraries did not necessarily originate as a means of coercive control but often as an effort to provide enhanced service, their existence often placed the librarian in a position of social power over those observed (Black, 2001, p. 74).

Surveillance is a central feature of the rational bureaucratic organization in modern society, and the explosion of surveillance is intertwined with the historical development and growth of bureaucratic organizations (Beniger, 1986; Dandeker, 1990; Foucault, 1977; Giddens, 1987; Lyon, 1994; Weber, 1968). Dandeker describes the symbiotic relationship between capitalist organizations and the modern state, declaring that their activities are focused on both the internal exigencies of managing a system of administrative control over subject populations and the problems attendant upon monitoring and managing external relations with other organizations. This theme has been central

in providing a framework in terms of which the growth of bureaucratic surveillance in modern societies can be explained. (p. 195)

In Control Revolution, Beniger (1986) writes that "bureaucratic organization serves as the generalized means to control all large social systems, tending to develop whenever collective activities need to be coordinated toward some explicit and impersonal goal, that is, to be controlled" (p. 390). As the complexity of operations required to control the functioning of a bureaucratic organization increases, so to does the need for advanced technologies to manage information throughput (Beniger, 1986, p. 424). Historically, bureaucratic organizations have utilized technological advances to exert control over the volume of information vital to the functioning of their operation, often to automate data gathering, record keeping, and record retrieval.

Dandeker (1990, p. 40) provides an excellent four point schema for evaluating the surveillance capacity of organizations. This model evaluates the size of the files held in a surveillance system, the centralization of those files, the speed of information flows, and the points of contact between the system and its subject population. The escalating use of automated surveillance technologies, sorting software, and searchable computer databases has lead to increases in all four of these areas and has greatly enhanced the surveillance capacity of organizations, making the practice of mobile surveillance possible. The ability of organizations to utilize information and computer technologies in order to search and cross-reference personally identifiable information from a variety of independently established databases has greatly expanded the scope of their surveillance, and has enabled the tracking of individual through digital data profiles compiled from records stored in computer databases.

MOBILE SURVEILLANCE

As social relationships have become more fluid and individual mobility increases, surveillance technologies have developed to keep up with the mobile subject. They are increasingly capable of tracking subjects on the move, and across various media, and through a variety of environments, casting a continual and inescapable gaze upon their subject (Lyon, 2003b). This is accomplished in a variety of ways. Perhaps the surveillance regime that most clearly illustrates the capabilities of mobile surveillance, and embodies the extension of Foucault's panopticism into society at large is CCTV. A CCTV system consists of a network of cameras that provide optical surveillance of a specific geographic area and transmits the visual data to a central location for analysis.

CCTV is most often employed by law enforcement in high-crime areas as a method for identifying criminal behavior, as well as a deterrent factor. Additionally, Norris and Armstrong (pp. 43-51) note the use of CCTV surveillance in residential areas, schools, banks, shops, workplaces, hospitals, schools, and train stations, as well as to regulate automobile traffic and police football stadia. The ubiquity of CCTV in Britain has lead to authors to conjecture that for a British urban dweller it is nearly impossible to move through public and, to some extent, private space without being photographed and recorded (Norris & Armstrong, 1999, p. 2). Increasingly these systems are being automated to work with face recognition software to look for "known" individuals and track their movement from camera to camera throughout the network.

While CCTV surveillance is directly concerned with the local observation of movement, the rise of dataveillance is critical for the observation of what Lyon refers to as "disappearing bodies" (Lyon, 2002). As transactions occur over longer distances, often with the aid of information and communication technologies, the physical body "disappears" and is replaced with personally

identifiable data that represents the individual (Gandy, 1993). Mechanisms such as security numbers, banking codes, and telephone numbers are recorded to provide a record of the interaction, which is often stored in computer databases. This information can be mined and analyzed by software using sophisticated algorithms to detect information patterns and assign a relative value to an individual, or what Gandy refers to as the "panoptic sort."

Automation through the employment of information and communications technologies and advances in surveillance hardware and software have expand the scope and speed of surveillance systems, enabling these systems to increasingly observe and record real-time activity and physical mobility at often exceptional distances. The growing reliance on information and communications technologies to conduct and coordinate surveillance has lead to the increasing importance of codes in the surveillance process (Lyon, 2003b). Codes are not only critical for the efficient operation of computerized systems, but they are also embedded with politics (Lessig, 1999). Programming establishes the rules that guide the functioning of computer codes, determining what information is stored and sorted, which individuals are tracked, and whose data-profile is flagged for review. To this end, the choices that are made during the programming and implementation of surveillance systems generate the set of laws that govern the operation of those systems.

In the wake of the September 11th attacks on New York and Washington, DC, the U.S. government has been on the forefront of bureaucratic uses of mobile surveillance technology and systems. Examples include the recently discontinued Terrorism Information Awareness (TIA) program (formerly Total Information Awareness) that was under development by the Defense Advanced Research Projects Agency (DARPA). The goal of the program was to preempt terrorist attacks by examining a variety of independently collected data sources in order to build comprehensive data profiles of potential terrorists. To accomplish this goal, DARPA was developing software that would have enabled intelligence officials to mine a virtual database that would consist of government, financial, education, medical and housing records from around the globe (Swartz, 2003, p. 6). Although the program was later abandoned, extensive data collection and data mining operations will almost certainly continue to be developed by national governments wishing to hedge against the risks inherent in an increasingly globalized world characterized by global flows of information, finance, and population (Castells, 1996).

FUTURE TRENDS

In Surveillance After September 11, David Lyon (2003a) provides us with three key issues that have emerged during the U.S.-lead "war on terror," namely suspicion, secrecy, and the mobilization of citizens as spies. He asserts that suspicion has been harnessed by local and national governments to broaden the scope of who may legitimately become a target of state-sponsored surveillance, while also being used to justify the secrecy of new or enhanced surveillance regimes under the rubric of "national security." Finally, the culture of suspicion that has arisen in the U.S. and to varying degrees in other societies around the world has increased acceptance for enhanced surveillance activities.

Mobile surveillance must be viewed through the prism of Lyon's three-part schema. If the culture of suspicion persists as a major motivating factor in bureaucratic implementation of enhanced surveillance capabilities, we can expect that new mobile surveillance technologies will be at the forefront of research, development and implementation, as they provide authorities with the ability to track and sort individuals and populations in real time. In the wake of the September 11 attacks, and the subsequent string of attacks around the globe including the bombings in Madrid and

Bali, a number of national governments have made upgrading their surveillance capabilities a priority. Additionally, technological advances will decrease the need for the human supervision of surveillance systems, creating a fully automated surveillance apparatus.

Three technologies in particular will enhance the surveillance capacity of the bureaucratic organization in the future: GPS, RFID, and biometrics. The inclusion of GPS receivers in mobile phones, often at lawmakers' request to provide assistance in locating missing individuals, allows for monitoring the precise location of the phone in real time, whether a call is being made or not. RFID tagging, a bonus for merchandisers keen to increase logistical efficiency, is also an ideal technology for bureaucratic management. It will likely become a permanent feature of future identification cards, as the miniaturized tags are capable of storing personally identifiable information and transmitting it wirelessly to strategically placed receivers in airports, and other access-restricted locales. Finally, biometrics, the practice of identifying an individual based on physiological characteristics, seems poised to be the next big field of personally identity. Coupled with optical surveillance, biometrics can be used to further automate CCTV systems, providing a reliable method for identifying individuals.

Serious questions must be asked about the architecture of current and future systems and the codes that govern them. Who is being tracked, how they are being tracked, and why they are being tracked are important design questions that will influence human outcomes. While it is hard to argue against the use of efficient systems that enable authorities to prevent a small handful of individuals from doing great harm to large number of innocent people, a balance must be struck between the rights of the individual and the safety of the majority. Many nations have a legal and legislative framework in place to wrestle with balancing these two responsibilities. However, the push for secrecy inspired by a climate of suspicion may be the single greatest variable to consider when examining the future of surveillance. A move toward greater state secrecy serves to obscure the existence and operation of the surveillance apparatus, limiting transparency, and diminishing the possibility of legislative and public oversight.

Looking forward, it is important to remember that technological development must be coupled with legislative action and social awareness. State sponsored surveillance that is designed to protect the public from harm should be at least minimally transparent and ideally subjected to oversight in order to protect against abuses. Technological advances in data gathering, sorting, storage, and retrieval, coupled with complimentary advances in computing and mobile ICTs will enhance the surveillance capacity of large organizations. They will be increasingly able to tap into vast stores of personally identifiable information from multiple sources through refined data-mining practices.

CONCLUSION

The regulatory framework in which these organizations operate will help determine which surveillance practices are available to the state and which are not. Legislative bodies will need to set guidelines that simultaneously encourage technological growth and positive uses of surveillance, while demanding state accountability and balancing the rights of the citizen and individual to privacy. Despite increasing globalization and surveillance regimes aimed at minimizing the risks that arise from global population flows, it is import to remember that concept of privacy and privacy regulations vary greatly from state to state. A patchwork of laws governs the surveillance of globally mobile bodies, as individuals pass borders and therefore into and out of the gaze of the state surveillance apparatus. Post-September 11th agreements between nations have lead to greater international security cooperation and informa-

tion sharing, subjecting the actions of citizens of one nation to the gaze of another. What right to privacy does the global citizen have from the government of a foreign nation, for whom he or she cannot vote?

The ability of the citizenry and legislatures in democratic nations to oppose, alter, and eliminate surveillance regimes should not be underestimated. Despite an intense "culture of suspicion" following the September 11, U.S. citizens and lawmakers were able to halt two of the more controversial government programs designed to enhance the nation's surveillance capacity. A key surveillance passage of the USA PATRIOT Act, which granted federal authorities almost unchecked power to collect personally identifiable data, was struck down as unconstitutional by the courts, and the Terrorist Information Awareness (TIA) program, which was attempting to build highly advanced data-mining software to cull vast amount data in order to build predictive models of terrorist behavior, had its funding denied by the Congressional committee charged with its oversight after intense public scrutiny. These two local successes do not signal a victory for transparency and oversight, but rather point to a possible trend of ad hoc coalitions formed to resist specific instances of particularly intrusive state surveillance.

REFERENCES

Beniger, J. R. (1986). The control revolution: Technological and economic origins of the information society. Cambridge, MA: Harvard University Press.

Black, A. (2001, January). The Victorian information society: Surveillance, bureaucracy, and public librarianship in 19th-century Britain. Information Society, 17(1), 63.

Castells, M. (1996). The rise of the network society. Malden, MA: Blackwell.

Dandeker, C. (1990). Surveillance, power and modernity: Bureaucracy and discipline from 1700 to the present day. Cambridge, UK: Polity Press.

Foucault, M. (1977). Discipline and punish: The birth of the prison (1st American ed.). New York: Pantheon Books.

Gandy, O. H. (1993). The panoptic sort: A political economy of personal information. Boulder, CO: Westview.

Giddens, A. (1987). The nation-state and violence (Contemporary critique of historical materialism, Vol 2). Berkeley: University of California Press.

Lessig, L. (1999). Code : And other laws of cyberspace. New York: Basic Books.

Lyon, D. (1994). The electronic eye : The rise of surveillance society. Minneapolis: University of Minneapolis Press.

Lyon, D. (2003a). Surveillance after September 11. Malden, MA: Polity Press.

Lyon, D. (2003b). Surveillance as social sorting: Privacy, risk, and digital discrimination. London: Routledge.

Lyon, D. (2002). Surveillance society: Monitoring everyday life. Buckingham, UK: Open University Press.

Norris, C., & Armstrong, G. (1999). The maximum surveillance society: The rise of CCTV as social control. Oxford, UK: Berg.

Swartz, N. (2003). Controversial surveillance system renamed. Information Management Journal, 37(4), 6.

Weber, M. (1968). Economy and society; an outline of interpretive sociology. New York: Bedminster Press.

KEY TERMS

Biometrics: Biometrics is the science and practice of verifying individual identity based on the analysis of unique physiological or behavioral characteristics. Examples include the analysis of fingerprints, retinas scanning, voice pattern analysis, facial patterns, and analysis of an individual's walking gait.

CCTV: Closed Circuit Television is a technological system of video surveillance that employs a closed network of cameras to provide a visual observation of a targeted area. CCTV has been used extensively in high crime areas not only as a means of fighting crime but also as a deterrent. The use of CCTV is increasingly being combined with face recognition software to create automated video surveillance networks that can operate with limited human interaction.

Data Mining: Data mining, also known as knowledge discovery in databases, is the practice of extracting targeted information from large databases through the use software technology utilizing algorithms to detect patters.

Face Recognition Software: A software package that is designed to identify individuals in crowds based on distinguishing facial characteristics. This software must be used in conjunction with visual surveillance systems such as CCTV. The facial characteristics of target individuals are loaded into computer systems that analyze the visual data captured by the visual surveillance apparatus and alter system users when a target individual is spotted.

GPS: Global positioning system is a satellite navigation system that is able to provide extremely accurate time and position data through the tracking of user-held receivers. The system was developed and maintained by the United States Department of Defense and is available free of charge to nonmilitary users. The system operates through transmissions between user-held and earthbound receivers to a network of satellites, whereby a receiver's exact location is determined through the process of trilateration and time is determined by the coordinated atomic clocks of the satellites.

Identity Documents: Identity documents take many forms from drivers licenses to passports to national identity cards. These cards are government issued documents that contain personally identifiable information often including a photograph, date of birth, place of residence, gender, physical characteristics such as height, weight, and eye and hair color, and include a unique identifier number specific to the card holder. Increasingly identity documents include personal information encoded in magnetic strips that can be read through the use of a scanner.

RFID: Radio frequency identification technology is a type of wireless automatic identification system that collects data and transmits it directly to a computer database using radio waves. A typical RFID system consists of a radio frequency tag that transmits identifiable data when in proximity of a reader that is then recorded to a database.

This work was previously published in Encyclopedia of Digital Government, edited by A. Anttiroiko and M. Malkia, pp. 841-845,

Chapter 3.29
Discovering Multimedia Services and Contents in Mobile Environments

Zhou Wang
Fraunhofer Integrated Publication and Information Systems Institute (IPSI), Germany

Hend Koubaa
Norwegian University of Science and Technology (NTNU), Norway

ABSTRACT

Accessing multimedia services from portable devices in nomadic environments is of increasing interest for mobile users. Service discovery mechanisms help mobile users freely and efficiently locating multimedia services they want. The chapter first provides an introduction to the topic service discovery and content location in mobile environments, including background and problems to be solved. Then, the chapter presents typical architectures and technologies of service discovery in infrastructure-based mobile environments, covering both emerging industry standards and advances in the research world. Their advantages and limitations, as well as open issues are discussed, too. Finally, the approaches for content location in mobile ad hoc networks are described in detail. The strengths and limitations of these approaches with regard to mobile multimedia services are analyzed.

INTRODUCTION

Recently, the advances in mobile networks and increased use of portable devices deeply influenced the development of multimedia services. Mobile multimedia services enable users to access multimedia services and contents from portable devices, such as laptops, PDAs, and even mobile phones, at anytime from anywhere. Various new applications, that would use multimedia services on portable devices from both the fixed network backbone and peer mobile devices in its proximity, are being developed, ranging from entertainment

and information services to business applications for M-Commerce, fleet management, and disaster management.

However, to make mobile multimedia services become an everyday reality, some kinds of service infrastructures have to be provided or enhanced, in order to let multimedia services and contents on the network be discovered and utilized, and simultaneously allow mobile users to search and request services according to their own needs, independently of the physical places they are visiting and the underlying host platforms they are using. Particularly, with the explosive growth of multimedia services available in the Internet, automatic service discovery is gaining more and more significance for mobile users. In this chapter we focus on the issue of discovering and locating multimedia services and contents in mobile environments. After outlining necessary background knowledge, we will take an insight into mobile multimedia service discovery. Major service discovery architectures and approaches in infrastructure-based networks and in mobile ad hoc networks will be investigated. We present also a detailed analysis of their strengths and limitations with regard to mobile multimedia services.

DISCOVERING MOBILE MULTIMEDIA SERVICES AND CONTENTS IN INFRASTRUCTURE-BASED ENVIRONMENTS

Overview

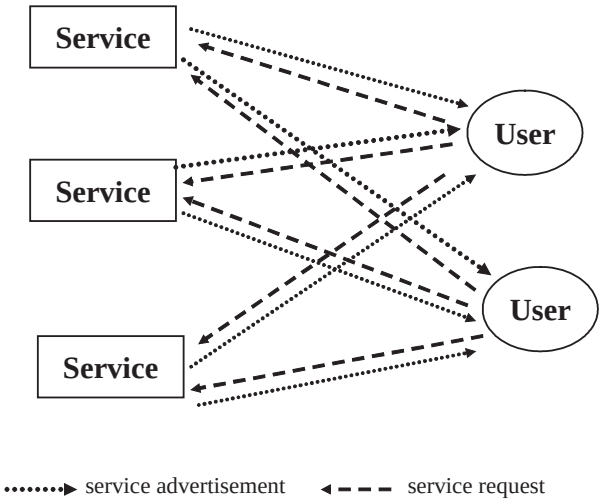
In order to use various multimedia services on the network, the first necessary step is to find the exact address of service providers that implement the service. In most cases, end users might only know what kind of service (service type) and some service characteristics (e.g., data format, cost) they want, but without having the server address. Currently, browsing is one often-used method to locate relevant information. As the number and diversities of services on the network grow, mobile users may be overwhelmed by the sheer volume of available information, particularly in an unacquainted environment. On the other side, user mobility presents new challenges for service access. Mobility means that users probably change their geographic locations frequently. Consequently, services available to users will appear or disappear dynamically while users move here and there. Moreover, mobile users are often interested in the services, (e.g., malls, restaurants) in the close proximity of his or her current place. Therefore, unlike classical distributed environments where location is often kept transparent, applications often need to dynamically obtain information that is relevant to their current location. The service search procedure should be customized according to user's context, (e.g., in terms of when (i.e., time) and where (i.e., location) a user is visiting).

Since most current multimedia services are designed for stationary environments, they do not address these issues. Recently, a number of service discovery solutions are developed. These solutions range from hardware-based technologies such as Bluetooth SDP, to single protocols, (e.g., SLP and SDS) to frameworks such as UPnP and Jini. From architectural point of view, we observed three models are used to discover services in different network environments (Wang, 2003): the broadcast model, the centralized service directory model, and the distributed service directories model. Next, we will investigate these paradigms in detail.

Broadcast Model

The simplest architecture for service discovery is using broadcast to locate services and contents. The conceptual scheme of the broadcast model

Figure 1. Broadcast model



is depicted in Figure 1. In this model, clients and servers talk directly with each other through broadcast or multicast.

According to who initiates the announcement and who listens, two strategies are differentiated. The first strategy is the *pull strategy* where a client announces his requests, while all servers keep listening to requests. The servers that match the search criteria will send responses (using either unicast or multicast) to the client. The other strategy is the *push strategy.* The servers advertise themselves periodically. Clients who are interested in certain types of services listen to the service advertisements, and extract the appropriate information from service advertisements. Of course, hybrid strategies are applied by some approaches.

The **simple service discovery protocol** (**SSDP**) is one typical approach based on the broadcast model (Goland, Cai, Leach, Gu, & Albright, 1999). The SSDP builds upon HTTP and UDP-multicasting protocols, and employs a hybrid structure combining client announcement and service announcement. When a device is newly added to the network, it multicasts an "ssdp:alive" message to advertise its presence. Similarly, when a client wants to discover services, it multicasts a discovery message and awaits responses.

The broadcast model works well in small simple networks, such as home and small office. The primary advantage of such systems is that they need "zero" or little configuration and administration. Besides, they accommodate well to frequent service join/leave actions in a dynamic environment. However, they usually generate heavy network traffic due to broadcast, and thus have only minimal scalability.

In order to improve scalability and performance, an additional entity, service directory, is introduced. Two different models use the service directory: the centralized service discovery model and the distributed service directories model. Both models will be presented in the following sections.

Centralized Service Directory Model

The conceptual scheme of the centralized directory model is shown in Figure 2. The service directory becomes the key component in the search discovery architecture, because it stores information about all available services.

The service discovery procedure consists usually of the following steps:

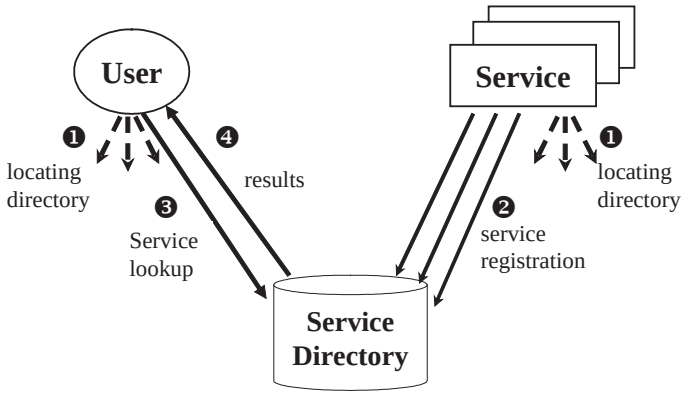
1. **Locating directory:** Either clients or servers should determine the address of the service directory before they utilize or advertise services. The directory could be located by manual configuration, by querying a well-known server, or through broadcast/multicast requests/replies.
2. **Service registration:** Before a service can be found by clients, it must be registered in the appropriate directory. A service provider explicitly initiates a registration request to the directory, and the directory stores the service data in its database. The service description data include service type, service attributes, server address, etc.
3. **Service lookup:** As a client searches for a particular service, he describes his requirements, e.g. service type and desired characteristics, in a query request, and sends it to the directory.
4. **Searching:** The directory searches services in its database according to the criteria provided by the client. When services are found, the server addresses and other information of qualified services are sent back to the client.

The centralized directory model has been used by several service discovery approaches. In this section we will examine some of them.

Service Location Protocol (SLP)

The service location protocol (SLP) is an example of centralized directory-based solution, and is now an IETF standard (Guttman, Perkins, Veizades, & Day, 1999). The current version is SLP Version 2 (SLPv2). The SLP uses DHCP options, or UDP-based multicasting to locate the service directory (known as directory agent (DA)), without manual

Figure 2. Centralized directory model



configuration on individual clients and services (known as user agents (UAs), service agents (SAs) respectively). A multicast convergence algorithm is adopted in SLP to enhance multicast reliability.

Service registration and lookup are performed through UDP-based unicast communication between UAs/SAs and DAs. In addition, SLP can operate without DAs. In this mode, SLP works in the same way as the broadcast model. A service in SLP is described with service type in the form of a character string, the version, the URL as server address, and a set of attribute definitions in the form of key-value pairs.

To improve performance and scalability, more DAs can be deployed in network. However, SLPv2 does not provide any synchronization mechanisms to keep DAs consistent, but leaves this responsibility to SAs which should register with each DA they detect. Recently, (Zhao & Guttman, 2000) proposed a mesh enhancement for DAs to share known services between one another. Each SA needs to register only with a single DA, and its registration is automatically propagated among DAs.

Generally, SLP is a flexible IP-based service discovery protocol which can operate in networks ranging from a single LAN to an enterprise network. However, it is intended to function within networks under cooperative administrative control, and thus does not scale for the Internet.

JINI

Sun's JINI provides a similar architecture as SLP for delivering services in a network (Sun Microsystems Inc., 2003), but it is tightly bound to the Java environment and needs Java Virtual Machine (JVM) support. The protocols in JINI are implemented as Java APIs. For this reason, the JINI client is not as lightweight as the SLP client. However, JINI is more than a discovery protocol. It provides further facilities for service invocation, for transaction, and for distributed events.

INS

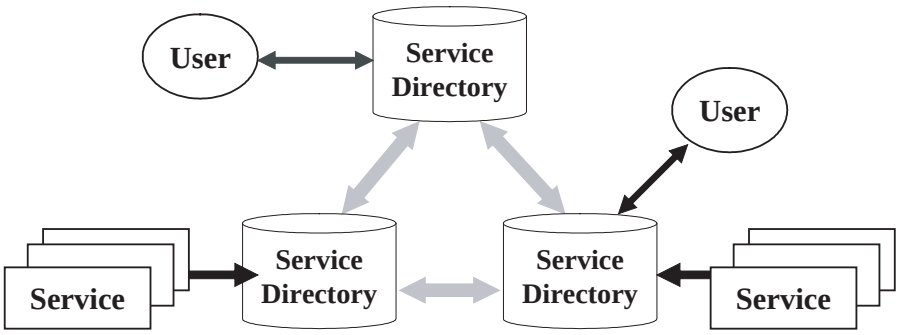
Adjie-Winoto, Schwartz, Balakrishnan, and Lilley, (1999) proposes a resource discovery system named *intentional naming system* (INS). The main idea is that resources or services are named using an ordered list of attribute-value pairs. Since service characteristics can be described by the service name itself, the service discovery procedure is equal to name resolving which is accompanied by the *intentional name resolver* (INR). The INR is actually a service directory that holds the global knowledge about names in the whole network. INS is different from other naming services (e.g., DNS), in that the name describes service attributes and values, rather than simple network locations of objects.

In conclusion, most centralized directory-based architectures have been designed for local networks or enterprise-wide networks which are under a common administration. The primary issue for these systems is scalability. As the number of services and clients increases, a centralized directory, even replicated, will not be feasible to accommodate a large number of registrations and lookups. In this context, the distributed repositories model has been suggested.

Distributed Service Directories Model

In the distributed directories model, the whole service domain is split into partitions, possibly according to organizational boundary, network topology, geographic locations etc. In each partition, there are one or more directories. The conceptual scheme of the distributed directories model is shown in Figure 3. The distributed directories model is different from the centralized directory model in that no directory has a complete global view of services available in the entire domain. Each directory holds only a collection of services in its partition, and is responsible for interaction with clients and services in the partition.

Figure 3. Distributed directories model



The service registration and query submission in the distributed model remain similar to that in the centralized directory model. But the service search operation becomes more complicated. If required services can be found by local directories, the discovery procedure is akin to that in the centralized directory model. But if not, the directories in other partitions should be asked, to ensure that a client can discover any service offers in the entire domain.

The directories in this model are organized in some way to achieve cooperation. As stated in (Wang, 2003), the directories can be organized in a hierarchy structure or in a mesh structure. While in the hierarchy structure there is a "belong to" relationship between directly connected directories, directories in the mesh architecture are organized in a flat interconnected form without hierarchy. The interconnection structure might have strong implications on query routing. In the hierarchy structure queries are passed along the hierarchy, either upward or downward, thus the routing path is inherently loop free. But the rigid hierarchy obstructs to shortcut the routing path in some cases. On the other hand, the mesh structure is advantageous for optimizing the routing path, but might rely on some mechanisms to avoid loop circles or repeated queries.

A typical example of distributed directories-based architecture is **service discovery service (SDS)**, developed in Berkeley (Hodes, Czerwinski, Zhao, Joseph, & Katz, 2002). The SDS is based on the hierarchy model which is maintained by periodic "heartbeat" messages between parent and child nodes. Each SDS server pushes service announcements to its parent. By this means, each SDS server gathers a complete view of all services present in its underlying tree. The significant feature of SDS is the hierarchical structure with lossy aggregation to achieve better scalability and reachability. The SDS server applies multiple hash functions (e.g., MD5), to various subsets of tags in the service description and uses the results to set bits in a fixed-size bit vector. The parent node ORs all bit vectors from its children to summarize available services in the underlying tree.

The hierarchical structure with lossy aggregation helps SDS to reach better scalability, while ensuring users to be able to discover all services on all servers. However, the SDS is more favorable

for applying in stationary network environments since it requires additional overheads to maintain the hierarchical structure and to propagate index updates. If services change attributes rapidly or join/leave frequently, it will generate too much communication burden. Moreover, the OR-operation during aggregation may cause "false positive" answers in query routing. Although it does not sacrifice correctness, it will lead to unneeded additional query forwarding.

The media gateway discovery protocol (MEGADIP) is developed especially for discovering media gateways that act as proxy for transforming or caching data between media source and end users (Xu, Nahrstedt, & Wichadakul, 2000). In MEGADIP the discovery procedure starts from the local directory, and forwards the query to directories along the routing path of the network layer between media source and destination. This idea is driven by the heuristics that a media gateway on or close to the end-to-end path is likely to find more bandwidth and/or to incur smaller end-to-end delay.

Other Issues in Service Discovery

The architectural models and various approaches presented above solved the service discovery problem to some extent. However, in order to let users comfortably and effectively locate mobile multimedia services and contents, there are still some issues to be addressed. From our point of view, interoperability, asynchronous service discovery, and semantic service discovery are the most important.

Interoperability

As previously stated, a number of service discovery approaches have been proposed. Despite that most of them provide similar functionality, namely automatically discovering services based on service characteristics, they have different features and are not compatible with each other. This incompatibility is one of the biggest obstacles for mobile users to really benefit from service discovery. From our point of view, it is more useful to make different approaches interoperable, than to design a new protocol to cover functionalities of existing protocols. So far, some solutions have been proposed to bridge service discovery mechanisms, but they are limited to pair-wise bridges, such as JINI to SLP (Guttman & Kempf, 1999). Authors in Friday, Davies, and Catterall (2001) proposed a general solution on a modified form of the Structured Query Language (SQL). However, no implementation details are presented in the paper. More generally, Wang and Seitz (2002) addressed this issue by providing an intermediary layer between mobile users and underlying service discovery protocols. The intermediary layer on the one hand provides clients with a general consistent view of service configuration and a universal means to formulate search requests, on the other hand is capable of talking with various types of service discovery protocols and handling service requests from users.

Asynchronous Service Discovery

Apart from the heterogeneous environments, most of the existing approaches rarely take the issues of thin client and poor wireless link into consideration. For example, synchronous operation is one of the intrinsic natures of most existing service discovery approaches, such as SLP, Jini, and SDS. Although synchronous operation simplifies protocol and application design, it is fastidious for mobile environments. The unexpected but frequent disconnections and possible long delay of wireless link greatly influence the usefulness and efficiency of synchronous calls. To relax the communication restrains in wireless environments, (Wang & Seitz, 2002) proposed in their CHAPLET system an approach to achieve asynchronous service discovery by adopting mo-

bile agents. The asynchronous service discovery allows mobile users to submit a service request, without having to wait for results, nor continuously keeping the permanently active connection in the process of service discovery.

Semantic Service Discovery

Most existing service discovery approaches support only syntactic-level searching (i.e., based on attribute comparison and exact value matching). However, it is often insufficient to represent a broad range of multimedia services in real world, and lacks of capability to apply inexact matching rules. Therefore, there is need to discover services in a semantic manner. Chakraborty, Perich, Avancha, and Joshi (2001) proposes in the DReggie project to use the features of DAML to reason about the capabilities and functionality of different services. They designed a DAML-based language to describe service functionality and capability, enhanced the Jini Lookup Service to enable semantic matching process, and provided a reasoning engine based on Prolog. Yang (2001) presents a centralized directory-based framework for semantic service discovery. However, the semantic-based service discovery is still in its infancy. To promote wide development of semantic service discovery, more research efforts should be devoted.

DISCOVERING MULTIMEDIA SERVICES AND CONTENTS IN AD HOC ENVIRONMENTS

Overview

There are two well-known basic variants of mobile communication networks: infrastructure-based networks and ad hoc networks. Mobility support described in the previous sections relies on the existence of some infrastructure. A mobile node in the infrastructure-based networks communicates with other nodes through the access points which act as bridge to other mobile nodes or wired networks. Normally, there is no direct communication between mobile nodes. Compared to infrastructure-based networks, ad hoc networks do not need any infrastructure to work. Nodes in ad hoc networks can communicate if they can reach each other directly or if intermediate nodes can forward the message. In recent years, mobile ad hoc networks are gaining more and more interest both in research and industry. In this section we will present some typical approaches that enable discover and locate mobile multimedia services and contents in ad hoc environments. First we present broadcast-based approaches, and then the geographic service location approach is discussed. Next, a cluster-based approach is introduced. Finally, we present a new service or content location solution that addresses the scalability problem in multi-hop ad hoc networks.

Broadcast-Based Approaches

Considering the fact that no infrastructure is available in ad hoc environments, service directory-based solutions are unusable for service discovery in ad hoc networks. Instead, assuming that network supports broadcasting, service discovery through broadcast is one of most widely adopted solutions. Two broadcast-based approaches are possible: (1) broadcasting client requests and (2) broadcasting service announcements. In the first approach, clients broadcast their requests to all the nodes in the ad hoc network. Servers hosting requested services reply back to the clients. In the second approach, servers broadcast their services to all the nodes in the network. Each client is thus informed about the location of every service in the ad hoc network. Since these both approaches are mainly based on broadcasting, their efficiency strongly depends on the broadcast efficiency. The service location problem in that context can be reduced to the broadcast problem in ad hoc networks. For this

reason, in the following, we present a summary of proposed approaches for broadcasting in ad hoc networks. These broadcast approaches are not designed specifically for service location but we believe that a broadcast-based service location protocol has to be informed about how broadcast is carried out. This will help in deploying a cross layer-based service location protocol.

The broadcast techniques can be categorized into four families: Williams and Camp (2002), simple flooding, Jetcheva, Hu, Maltz, and Johnson (2001), probabilistic broadcast, Tseng, Ni, Chen, and Sheu (1999), location-based broadcast, and neighbor information broadcast, Lim and Kim (2000) and Peng and Lu (2000). Flooding represents a simple mechanism that can be deployed in mobile ad hoc networks. Using flooding, a node having a packet to be broadcasted sends this packet to his neighbors who have to retransmit it to their own neighbors. Every node receiving the packet for the first time has to retransmit it. To reduce the number of transmissions used in broadcasting, other broadcast approaches are proposed. The probabilistic broadcast is similar to flooding except that nodes have to retransmit the broadcast packet with a predetermined probability. Randomly choosing the nodes that have to retransmit can improve the bandwidth use without influencing the reachability. In the case of location-based broadcast techniques, a node x retransmits the broadcast packet received from a node y only if the distance between x and y exceeds a specific threshold.

The information on the neighborhood can also be used to minimize the number of nodes participating in the broadcast packet retransmission. Lim and Kim (2000) uses the information about the one hop neighborhoods. Node A, receiving a broadcast packet from node B, compares its neighbors to those of B. It retransmits the broadcast packet only if there are new neighbors that will be covered and that will receive the broadcast packet. Other broadcast protocols are based on the 2 hop neighborhood information. The protocol used in Peng and Lu (2000) is similar to the one proposed in Lim and Kim (2000). The difference is that in Lim and Kim (2000) the neighborhood information is sent within HELLO packets, whereas in Peng and Lu (2000), the neighborhood information is enclosed within the broadcast packet.

The study carried out in Williams and Camp (2002) showed that the probabilistic and location broadcast protocols are not scalable in terms of the number of broadcast packet retransmissions. The neighborhood-based broadcast techniques perform better by minimizing the number of nodes participating to the broadcast packet retransmission. The most significant disadvantage of these protocols is that they are sensitive to mobility.

Geographic Service Location Approaches

A more interesting service location approach than broadcasting the whole network is to restrict broadcasting to certain regions. These regions can be delimited on the basis of predefined trajectories. In fact, recently, geometric trajectories are proposed to be used for routing (Nath & Niculeson, 2003) and content location in location-aware ad hoc networks (Aydin & Shen, 2002; Tchakarov & Vaidya, 2004). Aydin and Shen (2002) and Tchakarov and Vaidya (2004) are closely related where content advertisements and queries are propagated along four geographical directions based on the physical location information of the nodes. At the intersection point of the advertising and query trajectories the queries will be resolved. Moreover, Tchakarov and Vaidya (2004) improves the performance by suppressing update messages from duplicate resources. However, basically they still rely on propagating advertisements and queries through the network.

Cluster-Based Solutions

Besides enhancements in broadcast, clustering can also be used to improve the performance of

service discovery in mobile ad hoc networks. An interesting cluster-based service location approach designed for ad hoc networks is proposed in Koubaa and Fleury (2001) and Koubaa (2003). The proposed approach involves four phases: (1) the servers providing services are organized within clusters by using a clustering protocol. The cluster-heads, elected on the basis of an election protocol, have the role of registering the addresses of the servers in their neighborhoods (clusters). (2) A reactive multicast structure gathering the cluster-heads to which participate the cluster-heads of the created clusters is formed at the application layer. Each client or a server in the network is either a part of this structure or one hop away from at least one of the multicast structure members. (3) Clients send their request inside this multicast structure. (4) An aggregation protocol is used to send the replies of the cluster-heads within the multicast structure. The aim of the aggregation protocol is to avoid using different unicast paths for reply transmission by using the shared paths of the multicast structure.

A study comparing broadcast approaches to the cluster-based approach is carried out in Koubaa and Fleury (2002). This comparison study showed that clustering reduces the overhead needed for clients to send their requests and for servers to send back their replies. This reduction is noticeable when we increase the number of clients, the number of servers, and the number of nodes in the ad hoc network. The multicast structure used in Koubaa (2003) consists of a mesh structure which is more robust than a tree structure. The density of the mesh structure is dynamically adapted to the number of clients using it. The key idea of this dynamic density mesh structure is that the maintaining of the mesh is restricted to some clients called effective clients. Indeed, when the network is dense or the number of clients is high there is no need that all clients participate the multicast structure maintaining. This new mesh structuring approach is compared to ODMRP (Koubaa, 2003) where all the multicast users participate in the mesh maintaining. The comparison study showed that the proposed dynamic density mesh is more efficient than ODMRP. Compared to the tree-based multicast structure, the mesh-based multicast structure shows better server reply reachability performance but using more bandwidth.

Scalability Issue in Service Location

Currently it is well known that ad hoc networks are not scalable due to their limited capacity. The scalability problem is mainly related to the specific characteristics of the radio medium limiting the effective ad hoc network capacity. Even though, we think that designing specific solutions for scalable networks can help us at defining how much scalable is an ad hoc network. In the context of service location, authors in Koubaa and Wang (2004) state the problem of scalable service location in ad hoc networks and propose a new solution inspired by peer-to-peer networks called HCLP (hybrid content location protocol). The main technical highlights in approaching this goal include: (1) the hash function for relating content to zone, (2) recursive network decomposition and recomposition, and (3) content dissemination and location-based on geographical properties.

The hashing technique is used in HCLP both for disseminating and locating contents. But unlike the approaches in peer-to-peer systems where the content is mapped to a unique node, the hash function in HCLP maps the content to a certain zone of the network. A zone means in HCLP a certain geographical area in the network. The first reason for mapping content into zone, i.e. a subset of nodes, instead of an individual node, is mainly due to the fact that it could be expensive in radio mobile environments to maintain a predefined rigid structure between nodes for routing advertisements and queries. For example, in Stoica, Morris, Karger, Kaashoek, and Balakrishnan (2001), each joining and leaving of nodes has to lead to an adjustment of the Chord ring. More-

over, the fact that the routing in ad hoc networks is far less efficient and less robust than in fixed networks makes the adjustments more costly if there is node movement. The second reason for relating content to zone is that it is more robust to host a content within many nodes inside a zone than to host it within an individual node.

The underlying idea of network decomposition in HCLP is to achieve load distribution by maintaining the zone structure. It is well known that if the number of the nodes and contents in an unstructured and decentralized zone is beyond a certain limit, the network overhead related to content advertisement/location would become unsatisfactory. Therefore, to ensure a favorable performance and to achieve a better load distribution in HCLP, a zone could be divided into sub-zones recursively if the cost related to content advertisement/location using unstructured approaches in the zone exceeds a certain threshold.

To enable network decomposition in different zones a protocol is deployed to make it possible to nodes on the perimeter of the network exchanging their geographical locations. This will help estimating the position of the centre of the network. Knowing the locations of the nodes on the perimeter and the location of the network centre, a simple decomposition of the network into four zones is used. Each of these zones can also be decomposed again into four zones, etc.

In HCLP, for disseminating or locating a content in the network, a user first sends out its announcement or query request along one of four geographical directions (north, south, east, and west) based on geographic routing. In a dense network, the announcement or the request will then be caught on the routing path by a node that knows the central region of the network, in the worst case by a perimeter node on the network boundary. This node will then redirect the request into the direction of the central region, again by geographic routing. The node that belongs to the central region and receives this query message has the responsibility to decide whether to resolve the request directly within the zone or whether to redirect the request to the next level of the zone hierarchy, until the content is discovered.

Such a content dissemination and location scheme works completely decentralized. Moreover, only a small portion of nodes is involved in routing and resolving advertisement or query messages. Because not all nodes are necessary for maintaining routing information nor a global knowledge of the whole network is required, HCLP can be expected to be well scalable to large ad hoc networks.

CONCLUSION

The prevalence of portable devices and wide deployment of easily accessible mobile networks promote the usage of mobile multimedia services. In order to facilitate effectively and efficiently discovering desirable mobile multimedia services and contents, many research efforts have been done. In this chapter, we discussed existing and ongoing research work in the service discovery field both for infrastructure-based mobile networks and mobile ad hoc networks. We introduced three main architectural models and related approaches for service discovery in infrastructure networks, and pointed out some emerging trends. For discovering services and contents in ad hoc networks, we presented and compared proposed approaches based on either broadcast or cluster, and discussed the scalability issue in detail. We believe that service discovery will play an important role for successful development and deployment of mobile multimedia services.

REFERENCES

Adjie-Winoto, W., Schwartz, E., Balakrishnan, H., & Lilley, J. (1999). The design and implementation

of an intentional naming system. In *Proceedings of the 17th ACM Symposium on Operating Systems Principles (SOSP ´99).*

Aydin, I., & Shen, C. (2002, October). *Facilitating match-making service in ad hoc and sensor networks using pseudo quorum.* In the 11th IEEE International Conference on Computer Communications and Networks (ICCCN).

Chakraborty, D., Perich, F., Avancha, S., & Joshi, A. (2001, October). *DReggie: Semantic service discovery for m-commerce applications.* In the Workshop on Reliable and Secure Applications in Mobile Environment, in Conjunction with 20th Symposium on Reliable Distributed Systems (SRDS).

Friday, A., Davies, N., & Catterall, E. (2001, May). Supporting service discovery, querying, and interaction in ubiquitous computing environments. In *Proceedings of the 2nd ACM International Workshop on Data Engineering for Wireless and Mobile Access,* Santa Barbara, CA (pp. 7-13).

Goland, Y., Cai, T., Leach, P., Gu, Y. & Albright, S. (1999). *Simple service discovery protocol.* IETF Draft, draft-cai-ssdp-v1-03.txt.

Guttman, E., & Kempf, J. (1999). Automatic discovery of thin servers: SLP, Jini, and the SLP-Jini Bridge. In *Proceedings of the 25th Annual Conference of IEEE Industrial Electronics Society (IECON'99),* Piscataway, USA.

Guttman, E., Perkins, C., Veizades, J., & Day, M.(1999). *Service location protocol, version 2. IETF* (RFC 2608). Retrieved from http://www.ietf.org/rfc/rfc2608.txt

Hodes, T. D., Czerwinski, S. E., Zhao, B. Y., Joseph, A. D., & Katz, R. H. (2002, March/May). An architecture for secure wide-area service discovery. *ACM Wireless Networks Journal, 8*(2-3), 213-230.

Jetcheva, J., Hu, Y., Maltz, D., & Johnson, D. (2001, July). *A simple protocol for multicast and broadcast in mobile ad hoc networks.* Internet Draft draft-ietfmanet-simple-mbcast-01.txt, Internet Engineering Task Force.

Koubaa, H. (2003). *Localisation de services dans les réseaux ad hoc.* PhD thesis, Université Henri Poincaré Nancy,1, Mars 2003.

Koubaa, H., & Fleury, E. (2001, November). *A fully distributed mediator based service location protocol in ad hoc networks.* In IEEE Symposium on Ad hoc Wireless Networks, Globecom, San Antonio, TX.

Koubaa, H., & Fleury, E. (2002, July). *Service location protocol overhead in the random graph model for ad hoc networks.* In the IEEE Symposium on Computers and Communications, Taormina/Giardini Naxos, Italy.

Koubaa, H., & Wang, Z. (2004, June). *A hybrid content location approach between structured and unstructured topology.* In the 3rd Annual Mediterranean Ad hoc Networking Workshop, Bodrum, Turkey.

Lim, H., & Kim, C. (2000, August). *Multicast tree construction and flooding in wireless ad hoc networks.* In ACM MSWiM, Boston.

Nath, B., & Niculescu, D. (2003). Routing on a curve. *SIGCOMM Computer Communication Review, 33*(1), 155-160.

Peng, W., & Lu, X. (2000, August). *On the reduction of broadcast redundancy in mobile ad hoc networks.* In the 1st ACM International Symposium on Mobile Ad hoc Networking and Computing (MobiHoc), Boston.

Stoica, I., Morris, R., Karger, D., Kaashoek, M. F., & Balakrishnan H. (2001). Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of the 2001 Conference on*

Applications, Technologies, Architectures, and Protocols for Computer Communications (pp. 149-160). ACM Press.

Sun Microsystems Inc. (2003). *Jini technology core platform specification, version 2.0.* Retrieved June, 2003, from http://www.jini.org/nonav/standards/davis/doc/specs/html/core-title.html

Tchakarov, T., & Vaidya, N. (2004, January). *Efficient content location in wireless ad hoc networks.* In the IEEE International Conference on Mobile Data Management (MDM).

Tseng, Y., Ni, S., Chen, Y., & Sheu, J. (1999, August). The broadcast storm problem in a mobile ad hoc network. *5th Annual International Conference on Mobile Computing (MOBICOM),* Washington, DC, *31*(5), 78-91.

Wang, Z. (2003). *An agent-based integrated service platform for wireless and mobile environments.* Aachen, Germany: Shaker Verlag.

Wang, Z., & Seitz, J. (2002). An agent based service discovery architecture for mobile environments. In *Proceedings of the 1st Eurasian Conference on Advances in Information and Communication Technology,* Shiraz, Iran, October (LNCS 2510, pp. 350-357). Springer-Verlag.

Wang, Z., & Seitz, J. (2002, October). Mobile agents for discovering and accessing services in nomadic environments. In *Proceedings of the 4th International Workshop on Mobile Agents for Telecommunication Applications,* Barcelona, Spain (LNCS 2521, pp. 269-280). Springer-Verlag.

Williams, B., & Camp. (2002, June). *Comparison of broadcasting techniques for mobile ad hoc networks.* In the 3rd ACM International Symposium on Mobile Ad hoc Networking and Computing (MobiHoc), Lausanne, Switzerland.

Xu, D., Nahrstedt, D., & Wichadakul, D. (2000). *MeGaDiP: A wide-area media gateway discovery protocol.* In the 19th IEEE International Performance, Computing, and Communications Conference (IPCCC 2000).

Yang, X. W. (2001). A framework for semantic service discovery. In *Proceedings of the Student Oxygen Workshop, MIT Oxygen Alliance, MIT Computer Science and Artificial Intelligence Laboratory, 2001.* Retrieved from http://sow.csail.mit.edu/2001/proceedings/yxw.pdf

Zhao, W., & Guttman, E. (2000). *mSLP–Mesh enhanced service location protocol.* Internet Draft draft-zhao-slp-da-interaction-07.txt.

KEY TERMS

Aggregation: A process of grouping distinct data. Two different packets containing different data can be aggregated into a single packet holding the aggregated data.

Broadcast: A communication method that sends a packet to all other connected nodes on the network. With broadcast, data comes from one source and goes to all other connected sources at the same time.

Clustering: Identifying a subset of nodes within the network and vest them with the responsibility of being a cluster-head of certain nodes in their proximity.

Hash: Computing an address to look for an item by applying a mathematical function to a key for that item.

Mobile Ad Hoc Network: A kind of self-configuring mobile network connected by wireless links where stations or devices communicate directly and not via an access point. The nodes are free to move randomly and organize themselves arbitrarily, thus, the network's topology may change rapidly and unpredictably.

Multicast: A communication method that sends a packet to a specific group of hosts. With multicast, a message is sent to multiple destinations simultaneously using the most efficient strategy that delivers the messages over each link of the network only once and only creates copies when the links to the destinations split.

Scalability: The ability to expand a computing solution to support large numbers of components without impacting performance.

Service: An abstraction function unit with clearly defined interfaces that performs a specific functionality. Users, applications, or other services can use the service functionality through well-known service interfaces without having to know how it is implemented.

Service Directory: An entity in service discovery architecture that collects and stores information about a set of services within a certain scope, which is used for searching and/or comparing services during the service discovery procedure. Service directory is also known as service repository or directory agent. Service directory can be organized in central or distributed manner.

Service Discovery: The activity to automatically find out servers in the network based on the given service type and service attributes. The service discovery is, therefore, a mapping from service type and attributes to the set of servers.

This work was previously published in Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 165-178, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.30
DRM Technology for Mobile Multimedia

Sai Ho Kwok
California State University, Long Beach, USA

INTRODUCTION

Mobile multimedia has been promoted as a promising service and application in mobile e-commerce (m-commerce) by many mobile operators and mobile service providers, when high-speed mobile networks are expected to take off in the near future. However, at present, mobile multimedia is still in its infancy, accessed by relatively low-end mobile devices with limited bandwidth and resources. A typical example is Orange in Hong Kong which launched a low-grade multimedia service in 2000 to test the market with current mobile technologies. Due to the physical constraints of a 2.5G mobile network, audio broadcast is the best service that the network can offer up to date. However, in the near future, when advanced mobile networks and technologies become available, higher demands will be placed on the quality of mobile multimedia services. Such services support both audio and video data, for example, video conferencing, music video, video-on-demand and so on. Rights management deserves more serious concern because intellectual property of distributed multimedia content is as valuable as a company's physical assets (Doherty, 2002). This will become even more important when mobile multimedia services become marketable and an essential part of the business. The purpose of a digital rights management (DRM) system is to allow owners of digital assets (movies, songs) to distribute their products/services/contents electronically in a controlled way (Peinado, 2002). DRM technology makes various online payment schemes possible, such as pay-per-view, pay-per-download, pay-per-game and so on. Hence, mobile service providers are able to control end users' use of, and accessibility to, their products, and stand to gain huge profits from this capability with the DRM technology (Foroughi, Albin, & Gillard, 2002). A successful DRM system should address both business and technical issues (Grab, 2002), but this chapter only addresses and presents issues in the technical side due to the nature of this book. We present some critical issues of mobile DRM for mobile multimedia. A proposal of mobile DRM framework is presented to meet the urgent DRM needs with the existing 2.5G

mobile technology. This chapter is concluded by presenting future directions of mobile DRM for mobile multimedia.

BACKGROUND

Internet Commerce

In the Internet domain, Vidius Incorporated estimates 450,000 to 580,000 downloads of unprotected full-length films are transferred over the Internet daily (Grab, 2002). Protection of distributed multimedia has been a growing concern to creators, distributors, copyright owners, publishers, and governments. DRM is considered to be one of the desirable solutions to this problem, and it can protect distributed media contents delivered over the Internet.

Several international standard organizations have been developing DRM solutions for various distributed multimedia, for example, digital music and video. The Secure Digital Music Initiative (SDMI) (SDMI, 2003), backed by the Recording Industry Association of America (RIAA) and 200 music and technology companies (as of October 2003), has been proposed to provide a secure environment for music distribution over the Internet. Another standard being developed by the Moving Picture Experts Group (MPEG) is known as MPEG-21 (Bormans & Hill, 2002) dedicated to distributing digital multimedia content. MPEG-21 defines an interoperable framework for Intellectual Property Management and Protection (IPMP). The IPMP can be interoperable with other MPEG standards, for example, MPEG-4. Therefore, the property protection will be also applicable to most of the MPEG video standards in the future. In addition, there are commercial DRM systems especially for the wired Internet business. They include Windows Media Rights Manager by Microsoft, and MetaTrust by InterTrust Technologies (InterTrust, 2000).

The above DRM standards and systems can be classified into two groups, namely, cryptographic-based and watermark-based DRM solutions (Kwok, 2003). Cryptographic systems permit only valid key-holders to access the encrypted data after receiving it from the authenticated senders. However, once such data is decrypted, it is impossible to track its reproduction or retransmission. Therefore, cryptography only provides protection during data transmission. Digital watermarking technology seems to complement the cryptographic process and to protect copyright ownership (Kwok, 2002). Digital watermarks can be visible but they are preferably invisible identification codes that are permanently embedded in the data and present within the data after any decryption process (Doherty, 2002).

In order to manage digital rights effectively and efficiently, many commercial DRM solutions employ license management models (Kwok & Lui, 2002). A license management model consists of a digital license that keeps access and control rights. Corresponding rights enforcement DRM applications determine usage rights based on these digital licenses.

Mobile Commerce

The current 2.5G mobile technologies for mobile multimedia service are fundamentally different from those used for Internet commerce service, and they impose many limitations and constrains upon the sophistication of mobile multimedia service. This explains why existing DRM systems for Internet commerce cannot be applicable to DRM over the mobile environment in a straightforward way. Some of the most important technical and physical obstacles are summarized as follows:

1. *License management:* A mobile device usually has limited resources of both memory and processing power to handle and process license documents and rights-protected contents.

2. *Limited storage and processing power:* Due to the limited resources of the mobile device, it is not possible to download rights-protected contents to the mobile device and play it there.
3. *Rights insertion:* A sophisticated consumer's ID cannot be kept on the consumer's device due to the storage limitation, and it must be provided by another party or uploaded for rights insertion.
4. *Rights enforcement:* An active rights enforcement cannot take place at the mobile device because the device is not capable of intensive computation.
5. *Payment:* Mobile devices cannot support the elaborate computations required for the encryption and de-encryption process of electronic payment, and mobile networks may not be adequately secure to prevent the exposure of personal and credit card information.

CRITICAL ISSUES OF DRM FOR MOBILE MULTIMEDIA

For mobile multimedia, DRM involves specifying and associating rights with the distributed multimedia contents, placing controls on the content to enforce rights, enabling access checks, and tracking permissions usage and payment. For a general mobile service, the required capabilities include:

1. rights specification and rights label management;
2. rights authorization;
3. content protection, rights enforcement, and trusted rendering;
4. rights tracking; and
5. a security and commerce infrastructure.

Business transactions, such as payment, ordering, customer enquiry, and so forth, may occur between the concerned parties during content packaging, distribution, and usage. Managing rights in all these transactions is necessary. To support DRM operations in mobile multimedia, a DRM system needs to perform rights insertion and rights enforcement operations. In addition, a license management mechanism is also needed in managing license documents. A DRM solution for mobile multimedia should possess features stated in Table 1.

There are still many un-resolved technical problems and issues to be addressed before a successful DRM system for mobile multimedia can emerge. Some existing problems and issues are listed in Table 2.

Table 1. Summary of features of DRM solutions

• Media right protection and management
• Secure delivery and distribution of digital contents
• Processing authorization, data authentication and verification for content service
• Data security, integrity check, access control, and management for distributed systems and peer-to-peer (P2P) networks
• Multimedia watermarking for copyright protection, media authentication and integrity checking, finger-printing, and data annotation

Table 2. Problems and issues for mobile DRM

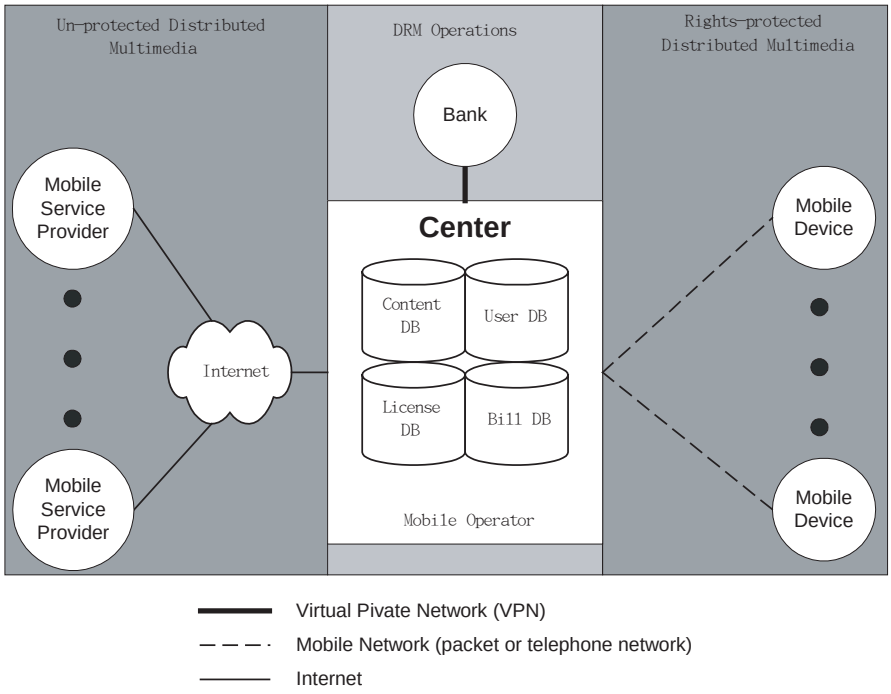
Mobile DRM standard: There is not yet a winner of mobile DRM standard. Open Mobile Alliance (OMA) DRM standard is one of the outstanding mobile DRM standards for mobile phones (Poropudas, 2003). However, other DRM standards, such as Windows Media DRM for Pocket PC (Microsoft, 2003), are highly competitive.
Trustful DRM protocol: Since DRM involves many parties, for example, technology service providers, mobile operators, service providers, creators, distributors and so forth, trust may not exist in all of these parties, for example, in a second-hand market. Hence, a trustful DRM protocol that can deal with DRM but without assuming mutual trust between involved parties is needed. A similar protocol was proposed by Cheung and Curreem (2002).
Robust and secure watermarking: A secure and robust watermarking algorithm is required to protect the distributed multimedia content. Such watermarking algorithm should resist attacks of any kinds. However, it cannot guarantee that a watermarking algorithm can resist all upcoming attacks (Tsang & Au, 2001).
Payment scheme: When trust does not exist, for example, in the second-hand market, a reliable payment scheme becomes an important issue.
Rights expression language: This is a need for a cross-platform rights expression language for all involved parties to specify and utilize their rights.

A MOBILE DRM FRAMEWORK FOR MOBILE MULTIMEDIA

This chapter proposes a mobile DRM framework for mobile multimedia, derived from Kwok (2002). The framework is a closed system that hosts all DRM operations within its center and provides a platform for mobile service providers to conduct business with their customers. Apart from rights insertion and enforcement operations, the center can handle transactions with a highly secure payment scheme. It is assumed that the central party is mutually trusted by both the businesses and consumers. A trustful center can be a mobile operator. The distributed multimedia contents between the center and the mobile users are rights-protected with a digital license and watermarking, while the multimedia contents transferring to the center do not require rights protection.

Figure 1 presents the mobile DRM framework for multimedia distribution in a mobile environment. The center of the framework is a mobile operator that manages information to and from mobile devices, mobile service providers and other concerned parties. The principal components include: (1) a mobile network infrastructure; (2) a DRM system; (3) a payment system; and (4) databases. There are three types of parties

Figure 1. DRM solution for mobile multimedia



involved in this framework: the mobile service providers (both official and unofficial sites), the bank, and the mobile users. The communication channels between different parties and the mobile operator are different from and independent of each other depending on the required security level. For example, a virtual private network (VPN) is used between the bank and the mobile operator, since highly confidential information is transferred through this channel, while the mobile operator relies on the packet network for multimedia content distribution, and the mobile service providers transfer multimedia contents to the mobile operator using the ordinary Internet.

The distinct features of the proposed framework include the following.

1. *DRM operation:* All DRM operations are performed by the center. It shifts all the processing and storage requirements to the center and relieves the burdens of the mobile devices and service providers.
2. *Independence of mobile devices:* The center can tailor the format of the distributed media for a specific mobile device. Besides, streaming technology is used in order to overcome the problems of processing power and storage requirement in the 2.5G mobile devices.
3. *Independence of mobile technology:* The framework can be applicable to 2.5G, 2.75G, 3G, 4G, and even higher because it does not depend on any specific mobile standard.

4. *Standardized rights expression language:* Since all DRM operations are managed by a single party—the mobile operator—the rights expression language can be standardized.
5. *Sharing and trading:* It facilitates media sharing and trading between users. Detail may be referred to Kwok (2002).
6. *Payment:* Transactions and payments are handled centrally through a secured channel.
7. *Ease of use and user satisfaction:* All DRM operations are completely transparent to mobile users and the mobile service providers.

FUTURE TRENDS

Mobile DRM for mobile multimedia is still at its infancy. The direction of mobile DRM is driven by the following factors.

1. *DRM standard:* One key player in mobile DRM standard, OMA mobile DRM has been proven successful in applying to music distribution. The standard is currently supported by some major labels, including Warner Music and BMG. However, the spectrum of mobile multimedia covers more than digital music, but also includes visual-audio data, such as movie, video conferencing, and so forth. It is still uncertain whether the market will accept OMA mobile DRM as the common standard for mobile multimedia.
2. *Mobile network:* Mobile multimedia demands a highly capable mobile network to support its services. An independent and constantly high transmission rate mobile network is the basic requirement for satisfactory mobile multimedia services. Unfortunately, the current mobile network, 2.5G or 2.75G, cannot provide a stable multimedia transmission. This problem will be overcome when 3G or 4G is launched.
3. *Mobile device:* The capabilities of mobile devices will be a major factor affecting the quality of mobile service. Pocket PCs usually perform better than smart phones when viewing mobile movie as their viewing screen and processing power are higher.

CONCLUSION

This chapter presents a mobile DRM framework for mobile multimedia. It is a practical and useful DRM framework when common mobile DRM standard and high bandwidth mobile channel are not available. This temporary but timely DRM solution could meet the urgent needs of DRM in mobile services. The primary objective of the framework is to impose DRM on mobile multimedia services without affecting the service providers and users. This is rather different from the emerging mobile DRM standards that require mobile users and service providers to adopt and apply their DRM technologies to mobile devices and distributed multimedia contents. However, privacy is the major problem of the proposed framework because the mobile operator possesses all of our transactions records. To respond to this problem, a possible solution may be encryption (Torrubia, Mora, & Marti, 2001) and an adapting system (Kenny & Korba, 2002).

ACKNOWLEDGEMENTS

The work described in this article was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKUST6256/03E).

REFERENCES

Bormans, J., & Hill, K. (2002). MPEG-21 overview v.5. Retrieved October 19, 2003, from *http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm*

Cheung, S.C., & Curreem, H. (2002). Rights protection for digital contents redistribution over the Internet. Paper presented at the *26th Annual International Computer Software and Applications.*

Doherty, S. (2002). Managing your digital rights. *Network Computing, 13*(19), 65-68.

Foroughi, A., Albin, M., & Gillard, S. (2002). Digital rights management: A delicate balance between protection and accessibility. *Journal of Information Science, 28*(5), 389-395.

Grab, E. (2002). Applying DRM techniques to video on the Internet: Characterizing problems and solutions. *SMPTE Journal-Society of Motion Picture & Television Engineers, 111*(3), 154-158.

InterTrust. (2000). InterTrust, The MetaTrust utility, announces OpenRights Initiative. Mountain View, CA: InterTrust Press.

Kenny, S., & Korba, L. (2002). Applying digital rights management systems to privacy rights management. *Computers & Security, 21*(7), 648-664.

Kwok, S.H. (2002). Chapter 5: Digital rights management for mobile multimedia. In E.P. Lim, Z. Shen, & K. Siau (Eds.), *Mobile commerce: Current states and future trends* (pp. 97-111). Hershey, PA: Idea Group Publishing.

Kwok, S.H. (2003). Digital watermarking for digital rights management. In L. Jain, H.C. Huang, & J.S. Pan (Eds.), *Intelligent watermarking techniques.* Hauppauge, NY: Nova Science Publishers.

Kwok, S.H., Cheung, S.C., Wong, K.C., Tsang, K.F., Lui, S.M., & Tam, K.Y. (2003). Integration of digital rights management into Internet open trading protocol (IOTP). *Decision Support Systems (DSS), 34*(4), 413-425.

Kwok, S.H., & Lui, S.M. (2002). A license management model for peer-to-peer music sharing. *International Journal of Information Technology and Decision Making (IJITDM), 1*(3), 541-558.

Kwok, S.H., Lui, S.M., Cheung, S.C., & Tam, K.Y. (2003). Digital rights management with Web services. *Electronic Markets, 13*(2), 133-140.

Microsoft. (2003). Windows Media DRM. Retrieved October 27, 2003, from *http://www.microsoft.com/windows/windowsmedia/drm.aspx*

Paskin, N. (2003). On making and identifying a "copy". *D-Lib Magazine, 9.*

Peinado, M. (2002). Digital rights management in a multimedia environment. *SMPTE Journal-Society of Motion Picture & Television Engineers, 111*(3), 159-163.

Poropudas, T. (2003). OMA digital rights arrive. Retrieved October 26, 2003, from *http://www.mobile.seitti.com/print.php?story_id=3136*

SDMI. (2003). Secure Digital Music Initiative. Retrieved October 20, 2003, from *www.sdmi.org*

Torrubia, A., Mora, F.J., & Marti, L. (2001). Cryptography regulations for e-commerce and digital rights management. *Computers & Security, 20*(8), 724-738.

Trowbridge, C. (2003, 1995 [October 13]). Image protection for archives, special collection libraries and museums in the WWW environment. Retrieved April 15, 2003, from *http://sunsite.berkeley.edu/Imaging/Databases/Fall95papers/trowbridge.html*

Tsang, K.F., & Au, O.C. (2001). A review on attacks, problems and weaknesses of digital wa-

termarking and the pixel reallocation attack. *Spie - the International Society for Optical Engineering, 4314*, 385-393.

KEY TERMS

Digital License: A digital license can be a separate file or message embedded in a media file. The license document states all of the terms and conditions concerning the use of the licensed media file. These terms and conditions can be static or dynamic depending on the payment scheme (Kwok, 2002).

Digital Rights Management (DRM): A set of technologies for content owners to protect their copyrights and stay in closer contact with their customers. In most instances, DRM is a system that encrypts digital media content and limits access to only those users who have acquired a proper license to play the content. That is, DRM is a technology that enables the secure distribution, promotion, and sale of digital media content on the Internet.

Identifiers and Metadata: Identifiers (unique labels for entities) and metadata (structured relationships between identified entities) are prerequisites for DRM. The essence of DRM is the control (licensing, etc.) of copies of entities; the identifiers and metadata are then essential to the management of this process, and to distinguishing and expressing relationships such as replicas and derivations (Paskin, 2003).

License Management: A mechanism to execute the terms and conditions stated in the license. This requires coordination among the media player, the media file, and other supporting modules; for example, the payment module. From the technical perspective, license management refers to issuing, hosting, and verifying the license (Kwok, 2002).

Rights Enforcement (or Verification): There are two types of rights enforcement: namely active enforcement and passive enforcement. The active enforcement takes place within the media player as a built-in function. The passive enforcement is an off-line ownership verification operation to check for the hidden owner identities (Kwok, 2002; Kwok, Cheung, Wong, Tsang, Lui & Tam, 2003; Kwok, Lui, Cheung, & Tam, 2003).

Rights Insertion: An operation to embed the identities of the concerned parties and assign business rules and conditions to the distributed multimedia content (Kwok, 2002; Kwok, Cheung, Wong, Tsang, Lui, & Tam, 2003; Kwok, Lui, Cheung, & Tam, 2003).

Watermarking: A technique for media authentication and forgery prevention. It is also viewed as an enabling technology to protect media from reuse without adequate credit or in an unauthorized way (Trowbridge, 2003). A watermarked media, *M'*, can be mathematically represented as $M'=M+W$ where M is the original media content and W is the embedded watermark. It is common that the extracted watermark, *W'*, could be different from the original watermark W because of the intentional or un-intentional attacks or post processing. To detect the watermark, a watermark detector is used to evaluate the similarity between W and W'.

This work was previously published in Encyclopedia of Information Science and Technology, Vol. 2, edited by M. Khosrow-Pour, pp. 918-923, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.31
V-Card:
Mobile Multimedia for Mobile Marketing

Holger Nösekabel
University of Passau, Germany

Wolfgang Röckelein
EMPRISE Consulting Düseldorf, Germany

ABSTRACT

This chapter presents the use of mobile multimedia for marketing purposes. Using V-Card, a service to create personalized multimedia messages, as an example, the advantages of sponsored messaging are illustrated. Benefits of employing multimedia technologies, such as mobile video streaming, include an increased perceived value of the message and the opportunity for companies to enhance their product presentation. Topics of discussion include related projects, as marketing campaigns utilizing SMS and MMS are becoming more popular, the technical infrastructure of the V-card system, and an outline of social and legal issues emerging from mobile marketing. As V-card has already been evaluated in a field test, these results can be implemented to outline future research and development aspects for this area.

INTRODUCTION

The chapter presents the use of mobile multimedia for marketing purposes, specifically focusing on the implementation of streaming technologies. Using V-card, a service for creating personalized multimedia messages, as an example, the advantages of sponsored messaging are illustrated. Topics of discussion include related projects, as marketing campaigns utilizing SMS and MMS are becoming more popular, the technical infrastructure of the V-card system, and an outline of social and legal issues emerging from mobile marketing. As V-card has already been evaluated in a field test, these results can be implemented to outline future research and development aspects for this area.

Euphoria regarding the introduction of the universal mobile telephony system (UMTS)

has evaporated. Expectations about new UMTS services are rather low. A "killer application" for 3rd generation networks is not in sight. Users are primarily interested in entertainment and news, but only few of them are actually willing to spend money on mobile services beyond telephony. However, for marketing campaigns the ability to address specific users with multimedia content holds an interesting perspective.

Advertisement-driven sponsoring models will spread in this area, as they provide benefits to consumers, network providers, and sponsors. Sponsoring encompasses not only a distribution of pre-produced multimedia content (e.g., by offering wallpapers), Java games, or ringtones based on a product, but also mobile multimedia services.

Mobile multimedia poses several problems for the user. First, how can multimedia content of high quality be produced with a mobile device. Cameras in mobile telephones are getting better with each device generation; still the achievable resolutions and framerates are behind the capabilities of current digital cameras. Second, how can multimedia content be stored on or transmitted from a mobile device. Multimedia data, sophisticated compression algorithms notwithstanding, is still large, especially when compared to simple text messages. External media, such as memory cards or the Universal Media Disk (UMD), can be used to a certain degree to archive and distribute data. They do not provide a solution for spreading this data via a wireless network to other users. Third, editing multimedia content on mobile devices is nearly impossible. Tools exist for basic image manipulation, but again their functionality is reduced and handling is complex.

Kindberg, Spasojevic, Fleck, and Sellen (2005) found in their study that camera phones are primarily used to capture still images for sentimental, personal reasons. These pictures are intended to be shared, and sharing mostly takes place in face-to-face meetings. Sending a picture via e-mail or MMS to a remote phone occurred only in 20% of all taken pictures. Therefore, one possible conclusion is that users have a desire to share personal moments with others, but current cost structures prohibit remote sharing and foster transmission of pictures via Bluetooth or infrared.

V-card sets out to address these problems by providing a message-hub for sublimated multimedia messaging. With V-card, users can create personalized, high-quality multimedia messages (MMS) and send those to their friends. Memory constraints are evaded by implementing streaming audio and video where applicable. V-cards can consist of pictures, audio, video, and MIDlets (Java 2 Micro-Edition applications). Experience with mobile greetingcards show that users are interested in high-quality content and tend to forward them to friends and relatives. This viral messaging effect increases utilisation of the V-card system and spreads the information of the sponsor. Haig (2002, p. 35) lists advice for successful viral marketing campaigns, among them:

- Create of a consumer-to-consumer environment
- Surprise the consumers
- Encourage interactivity

A V-card message is sponsored, but originates from one user and is sent to another user. Sponsoring companies therefore are actually not included in the communication process, as they are neither a sender nor a receiver. V-card is thus a true consumer-to-consumer environment. It also can be expected for the near future that high quality content contains an element of surprise, as it exceeds the current state of the art of text messaging. Interactivity is fostered by interesting content, which is passed on, but also by interactive elements like MIDlet games.

Additionally, Lippert (2002) presents a "4P strategy" for mobile advertising, listing four characteristics a marketing campaign must have:

- Permitted
- Polite

- Profiled
- Paid

"Permitted" means a user must agree to receive marketing messages. With V-card, the originator of the MMS is not a marketing company but another user, therefore the communication itself is emphasized, not the marketing proposition. Legal aspects regarding permissions are discussed detailed below. Marketing messages should also be "polite," and not intrusive. Again, the enhanced multimedia communication between the sender and the receiver is in the foreground, not the message from the sponsor.

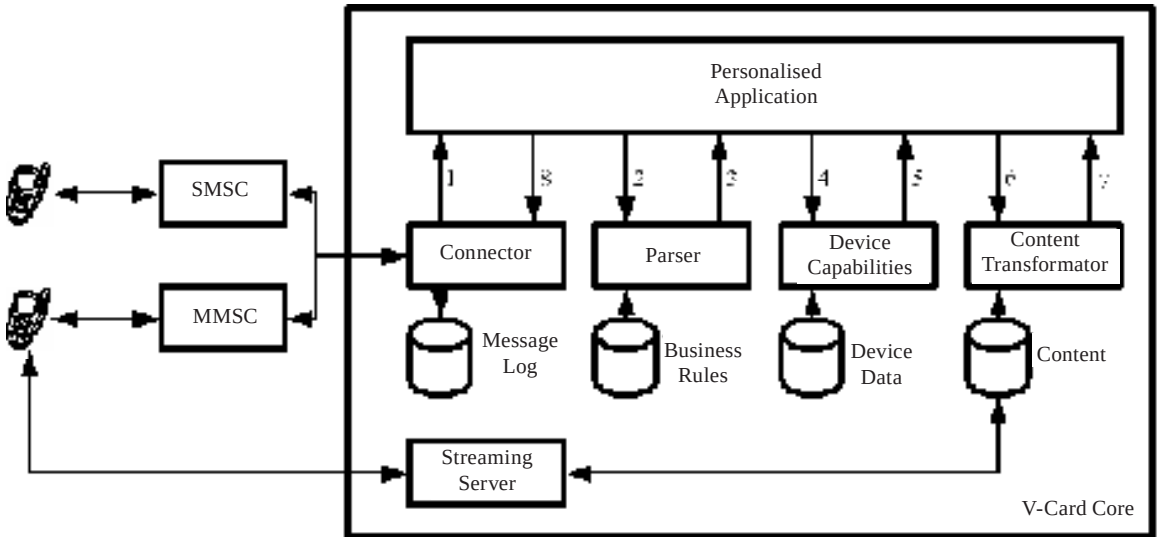
"Profiling" marketing tools enables targeted marketing and avoids losses due to non-selective advertising. Even if V-card itself is unable to match a sponsor to users, since users do not create a profile with detailed personal data, profiling is achieved by a selection process of the sender. As messages can be enhanced by V-card with media related to a specific sponsor, by choosing the desired theme the sender tailors a message to the interests of himself and the receiver. Usually, marketing messages should provide a target group with incentives to use the advertised service; the recipients need to get "paid." With V-card, sponsors "pay" both users by reducing the costs of a message and by providing high quality multimedia content.

V-CARD ARCHITECTURE

V-Card Core Architecture

Figure 1 shows the V-card core architecture and illustrates the workflow. First, the user with a mobile device requests a personalised application via the SMSC or Multimedia Messaging Service Centre (MMSC), which are part of the mobile network infrastructure. The message is passed on to the V-card core, where the connector decides which application has been called.

Figure 1. V-Card core architecture



After the request is passed on to the appropriate application (1), it is logged in the message log. A parser receives the message (2), extracts the relevant data for customisation, and returns this data (3)–this could include the receiver's phone number, the name of the sender or a message. Then, the capabilities of the receiving phone are queried from a database which holds all relevant data (4+5) like display size, number of colours, supported video and audio codecs.

Finally, the application transmits all the data gathered to the content transformator. Here, the pre-produced content is tailored with the input delivered by the user according to the capabilities of the device (6+7). The result is then sent via the connector (8) to the receiving user. Since the personalised applications and the data are separated, new applications can be easily created.

V-Card Streaming Technology

Since video content can not be stored directly on a mobile device due to memory limitations, a streaming server supplies video data to the device where the video is played, but not stored with the exception of buffered data, which is stored temporarily to compensate for varying network throughput. Streaming video and audio to mobile devices can be utilized for various services (e.g., for mobile education) (Lehner, Nösekabel, & Schäfer 2003). In the case of V-card, the MMS contains a link to adapted content stored on the content server. This link can be activated by the user and is valid for a certain amount of time. After the link has expired, the content is removed from the content server to conserve memory.

Currently, there are two streaming server solutions available for mobile devices. RealNetworks offers the HELIX server based on the ReadMedia format. RealPlayer, a client capable of playing back this format, is available for Symbian OS, Palm OS 5, and PocketPC for PDAs. Additionally, it is available on selected handsets, including the Nokia 9200 Series Communicators and Nokia Series 60 phones, including the Nokia 7650 and 3650. The other solution is using a standardized 3GPP-stream based on the MPEG4 format, which can be delivered using Apples Darwin server.

An advantage of implementing streaming technology for mobile multimedia is the fact that only a portion of the entire data needs to be transmitted to the client, and content can be played during transmission. Data buffers ensure smooth playback even during short network interruptions or fluctuations in the available bandwidth. As video and audio are time critical, it is necessary that the technologies used are able to handle loss of data segments, which do not arrive in time (latency) or which are not transmitted at all (network failure). GPRS and HSCSD connections allow about 10 frames per second at a resolution of 176 by 144 pixel (quarter common intermediate format QCIF resolution) when about 10 KBit per second are used for audio. Third generation networks provide a higher bandwidth, leading to a better quality and more stable connectivity.

A drawback of streaming is the bandwidth requirement. For one, the bandwidth should be constant; otherwise the buffered data is unable to compensate irregularities. Next, the available bandwidth directly influences the quality that can be achieved — the higher the bandwidth, the better the quality. Third, a transfer of mobile data can be expensive. A comparison of German network providers in 2003 showed that 10 minutes of data transfer at the speed of 28 KBit per second (a total amount of 19 Megabyte) resulted in costs ranging from 1 Euro (time-based HSCSD tariff) up to 60 Euro (packet-based GPRS by call tariff).

V-Card Examples

Figures 2 demonstrates a picture taken with the camera of a mobile device, rendered into a video clip by the V-card core. Figure 3 combines pictures and text from the user with video and audio content from the V-card hub. Figure 4 shows how simple text messages can be upgraded when a

Figure 2. V-Card with picture in video

Figure 3. V-Card with picture and text in video

Figure 4. V-Card with text in picture and audio

picture and an audio clip are added to create a multimedia message.

Since sponsoring models either influence the choice of media used to enhance a message, or can be included as short trailers before and after the actual message, users and sponsors can choose from a wide variety of options best suited for their needs.

LEGAL ASPECTS

It should be noted that the following discussion focuses on an implementation in Germany and today (2005Q1)–although several EU guidelines are applicable in this area there are differences in their national law implementations and new German and EU laws in relevant areas are pending.

Legal aspects affect V-card in several areas: consumer information laws and rights of withdrawal, protection of minors, spam law, liability, and privacy. A basic topic to those subjects is the classification of V-card among "Broadcast Service" ("Mediendienst"), "Tele Service" ("Teledienst"), and "Tele Communication Service" ("Telekommunikationsdienst"). According to § 2 Abs. 2 Nr. 1 and § 2 Abs. 4 Nr. 3 Teledienstegesetz (TDG) V-card is not a "Broadcast Service" and based on a functional distinction (see e.g., Moritz/Scheffelt in Hoeren/Sieber, 4, II, Rn. 10) V-card is presumed to be a "Tele Service."

Consumer information laws demand that the customer is informed on the identity of the vendor according to Art. 5 EU Council Decision 2000/31/EC, to § 6 TDG and to § 312c Bürgerliches Gesetzbuch (BGB) (e.g., on certain rights he has with regard to withdrawal). The fact that V-card might be free of charge for the consumer does not change applicable customer protection laws as there is still a (one-sided) contract between the costumer and the provider (see e.g., Bundesrat, 1996, p. 23). Some of these information duties have to be fulfilled before contract and some after. The post-contract information could be included

in the result MMS and the general provider information and the pre-contract information could be included in the initial advertisements and/or a referenced WWW- or WAP-site. Art. 6 EU Council Decision 2000/31/EC and § 7 TDG demand a distinction between information and adverts on Web sites and can be applicable, too. A solution could be to clearly communicate the fact that the V-card message contains advert (e.g., in the subject) (analogue to Art. 7(1) EU Council Decision 2000/31/EC, although this article is not relevant in Germany). The consumer might have a withdrawal right based on § 312d (1) BGB on which he has to be informed although the exceptions from § 312c (2) 2nd sentence BGB or § 312d (3) 2 BGB could be applicable. With newest legislation the consumer has to be informed on the status of the withdrawal rights according to § 1 (1) 10 BGB-Informationspflichtenverordnung (BGB-InfoV), whether he has withdrawal rights or not.

§ 6 Abs. 5 Jugendmedienschutzstaatsvertrag (JMStV) bans advertisements for alcohol or tobacco which addresses minors, § 22 Gesetz über den Verkehr mit Lebensmitteln, Tabakerzeugnissen, kosmetischen Mitteln und sonstigen Bedarfsgegenständen (LMBG) bans certain kinds of advertisements for tobacco, Art. 3(2) EU Council Decision 2003/33/EC (still pending German national law implementation) bans advertisements for tobacco in Tele Services. Therefore a sponsor with alcohol or tobacco products will be difficult for V-card. Sponsors with erotic or extreme political content will also be difficult according to § 4, 5 and 6(3) JMStV. § 12(2) 3rd sentence Jugendschutzgesetz (JuSchG) demands a labelling with age rating for content in Tele Services in case it is identical to content available on physical media. Since V-card content will most of the time special-made and therefore not available on physical media, this is not relevant.

The e-mail spam flood has led to several EU and national laws and court decisions trying to limit spam. Some of these laws might be applicable for mobile messaging and V-card, too. In Germany a new § 7 in the Gesetz gegen den unlauteren Wettbewerb (UWG) has been introduced. The question in this area is whether it can be assumed that the sent MMS is ok with the recipient (i.e., if an implied consent can be assumed). Besides the new § 7 UWG if the implied consent cannot be assumed a competitor or a consumer rights protection group could demand to stop the service because of a "Eingriff in den eingerichteten und ausgeübte Gewerbebetrieb" resp. a "Eingriff in das Allgemeine Persönlichkeitsrecht des Empfängers" according to §§ 1004 resp. 823 BGB.

Both the new § 7 UWG and previous court decisions focus on the term of an unacceptable annoyance or damnification which goes along with the reception of the MMS. The highest German civil court has ruled in a comparable case of advert sponsored telephone calls (BGH reference I ZR 227/99) that such an implied consent can be assumed under certain conditions e.g. that the communication starts with a private part (and not with the advertisement) and that the advertisement is not a direct sales pitch putting psychological pressure on the recipient (see e.g., Lange 2002, p. 786). Therefore if a V-card message consists of a private part together with attractive and entertaining content and a logo of the sponsor the implied consent can be assumed. The bigger the advertisement content part is the likelier it is that the level of a minor annoyance is crossed and the message is not allowed according to § 7 UWG (see e.g., Harte-Bavendamm & Henning-Bodewig, 2004, § 7, Rn. 171).

If users use the V-card service to send unwelcome messages to recipients V-card could be held responsible as an alternative to the user from whom the message originated. A Munich court (OLG München reference 8 U 4223/03) ruled in this direction in a similar case of an e-mail news letter service however focusing on the fact that the service allowed the user to stay anonymously. This is not the case with the mobile telephone numbers used in V-card, which are required to be associated with an identified person in Germany.

In addition to this the highest German court has in some recent decisions (BGH I ZR 304/01, p. 19 and I ZR 317/01, p. 10) narrowed the possibilities for a liability as an alternative by limiting the reasonable examination duties.

Manual filtering by the V-card service is a violation of communication secrecy and therefore not allowed (see e.g., Katernberg, 2003). Automatic filtering must not result in message suppression since this would be illegal according to German martial law § 206 (2) 2 Strafgesetzbuch.

The obligation to observe confidentiality has in Germany the primary rule that data recording is not allowed unless explicitly approved (§ 4 Bundesdatenschutzgesetz). Log files would therefore not be allowed with an exception for billing according to § 6 Gesetz über den Datenschutz bei Telediensten (TDDSG). These billing logs must not be handed over to third parties likely also including the sponsor.

As a conclusion, it can be noted that an innovative service like V-card faces numerous legal problems. During the project, however, it became clear that all these requirements can be met by an appropriate construction of the service.

EVALUATION OF V-CARD

Since V-card also has the ability to transmit personalised J2ME applications via MMS (see Figure 5 for an example), it surpasses the capabilities of pure MMS messages creating added value for the user, which normally do not have the possibility to create or modify Java programs. One example is a sliding puzzle where, after solving the puzzle, a user may use the digital camera of the mobile device to change the picture of the puzzle. After the modification, the new puzzle can then be send via V-card to other receivers.

Still, as previously mentioned, V-card requires a MMS client. It can therefore be regarded as an enhancement or improvement for MMS communication and is as such a competitor to the "normal" MMS. Hence, an evaluation framework should be usable to measure the acceptance of both "normal" MMS messaging and "enhanced" V-card messaging, creating results that can be compared with each other to determine the actual effect of the added value hoped to be achieved with V-card. While extensive research exists regarding PC-based software, mobile applications currently lack comprehensive methods for creating such evaluations. Therefore, one possible method was developed and applied in a fieldtest to evaluate V-card (Lehner, Sperger, & Nösekabel, 2004).

At the end of the project on June 3, 2004, a group of 27 students evaluated the developed V-card applications in a fieldtest. Even though the composition and size of the group does not permit to denote the results as representative, tendencies can be identified. The statistical overall probability of an error is 30%, as previously mentioned.

The questionnaire was implemented as an instrument to measure results. To verify the quality and reliability of the instrument, three values

Figure 5. V-card with MIDlet puzzle application

were calculated based on the statistical data. The questionnaire achieved a Cronbach alpha value of 0.89—values between 0.8 and 1.0 are regarded as acceptable (Cronbach, 1951). The split-half correlation, which measures the internal consistency of the items in the questionnaire, was calculated to be 0.77 with a theoretical maximum of 1.0. Using the Spearman-Brown formula to assess the reliability of the instrument, a value of 0.87 was achieved. Again, the theoretical maximum is 1.0. Therefore, the questionnaire can be regarded to be statistically valid and reliable.

One result of the fieldtest was that none of the students encountered difficulties in using any of the V-card applications, even though the usability of the mobile phone used in the fieldtest was regarded as less than optimal.

Overall 66% of the students thought that V-card was easy to use, 21% were undecided. It is very likely that the sample group leaned towards a negative or at least neutral rating as the usability of the end device was often criticised. This factor can not be compensated by the programmers of the mobile application. Another indicator for this rationale is the comparison with the results for the MMS client. Here, 75% of the group agreed to this statement, which is an increase of 9%. The similarity of the results suggests that also the rating for the usability of the MMS client was tainted by the usability of the device.

No uniform opinion exists regarding sponsored messages by incorporating advertising. Forty-two percent of the students would accept advertisements if that would lower the price of a message. Thirty-seven percent rejected such a method. The acceptable price for a V-card message was slightly lower compared to that of a non-sublimated MMS, which on the other hand did not contain content from a sponsor.

An important aspect for the acceptance of mobile marketing is the protection of privacy. In this area the students were rather critical. Sixty-three percent would reject to submit personal data to the provider of V-card. Since this information was not necessary to use V-card, only 17% of the sample group had privacy concerns while using V-card.

The mobile marketing component was perceived by all participants and was also accepted as a mean to reduce costs. This reduction should benefit the user, therefore a larger portion of the sample group rejected for V-card the idea for increased cost incurred by a longer or more intensive usage (88% rejected this for V-card, 67% for MMS).

As already addressed, the pre-produced content of V-card helped 50% of the users to achieve the desired results. The portion rejecting this statement for V-card was 25%, which is higher than the 8% who rejected this statement for MMS. This leads to the conclusion that if the pre-produced content is appropriate in topic and design for the intended message, it contributes to the desired message. However, it is not possible to add own content if the pre-produced content and the intention of the sender deviate. The user is therefore limited to the offered media of the service provider.

Overall, the ratings for V-card by the students were positive. Marketing messages, which were integrated into the communication during the fieldtest, were not deemed objectionable. The usability of V-card was also rated high. Main points that could be addressed during the actual implementation in the mobile market should include privacy and cost issues.

CONCLUSION

The new messaging service MMS has high potential and is being widely adopted today, although prices and availability are far from optimal. Mostly young people tend to use the fashionable messages which allow much richer content to be sent instantly to a friend's phone. This young

user group is especially vulnerable to debts due to their mobile phones though, or they have prepaid subscriptions letting them only send a very limited number of messages. By incorporating a sponsor model in V-card, this user group will be able to send a larger number of messages with no additional cost and thereby offering advertising firms a possibility to market their services and goods. For those users that are not as price sensitive, the large amount of professional media and the ease of the message-composition will be an incentive to use the service. The added value of the service should be a good enough reason to accept a small amount of marketing in the messages. Since V-card offers the sender and receiver an added value, the marketing message will be more acceptable than other forms of advertising where only the sender benefits from the advertisement.

Another advantage of V-card is the fact that the system takes care of the administration and storing of professional media and the complicated formatting of whole messages, thus taking these burdens from the subscriber. At the same time, V-card offers marketers a new way to reach potential customers and to keep in dialogue with existing ones. The ease of sending such rich content messages with a professional touch to a low price or even no cost at all will convince subscribers and help push 3G networks.

Overall, it can be expected that marketing campaigns will make further use of mobile multimedia streaming, aided by available data rates and the increasing computing power of mobile devices. Continuous media (video and audio), either delivered in real-time or on demand, will possibly become the next entertainment paradigm for a mobile community.

REFERENCES

Bundesrat. (1996). *Bundesrats-Drucksache* 966/96. Köln: Bundesanzeiger Verlagsgesellschaft mbH.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Haig, H. (2002). *Mobile marketing—The message revolution.* London: Kogan Page.

Harte-Bavendamm, H., & Henning-Bodewig, F. (2004). *UWG Kommentar.* München: Beck.

Hoeren, T., & Sieber, U. (2005). *Handbuch Multimedia-Recht.* München: Beck.

Katernberg, J. (2003). *Viren-Schutz/Spam-Schutz.* Retrieved from http://www.uni-muenster.de/ZIV/Hinweise/Rechtsgrundlage VirenSpamSchutz.html

Kindberg, T., Spasojevic, M., Fleck, R., & Sellen, A. (2005). The ubiquitous camera: An in-depth study of camera phone use. *IEEE Pervasive Computing, 4*(2), 42-50.

Lehner, F., Nösekabel, H., & Schäfer, K. J. (2003). *Szenarien und Beispiele für Mobiles Lernen.* Regensburg: Research Paper of the Chair of Business Computing III Nr. 67.

Lehner, F., Sperger, E. M., & Nösekabel, H. (2004). Evaluation framework for a mobile marketing application in 3rd generation networks. In K. Pousttchi, & K. Turowski (Eds.), *Mobile Economy—Transaktionen, Prozesse, Anwendungen und Dienste* (pp.114-126). Bonn: Köllen Druck+Verlag.

Lange, W. (2002). Werbefinanzierte Kommunikationsdienstleistungen. *Wettbewerb in Recht und Praxis, 48*(8), 786-788.

Lippert, I. (2002). Mobile marketing. In W. Gora, & S. Röttger-Gerigk (Eds.), *Handbuch Mobile-Commerce* (pp.135-146). Berlin: Springer.

KEY TERMS

MMS: Multimedia message service: Extension to SMS. A MMS may include multimedia content (videos, pictures, audio) and formatting instructions for the text.

Multimedia: Combination of multiple media, which can be continuous (e.g., video, audio) or discontinuous (e.g., text, pictures).

SMS (Short Message Service): text messages that are sent to a mobile device. A SMS may contain up to 160 characters with 7-bit length, longer messages can be split into multiple SMS.

Streaming: Continuous transmission of data primarily used to distribute large quantities of multimedia content.

UMTS (Universal Mobile Telecommunications System): 3rd generation network, providing higher bandwidth than earlier digital networks (e.g., GSM, GPRS, or HSCSD).

This work was previously published in Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 430-439, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.32
Acoustic Data Communication with Mobile Devices

Victor I. Khashchanskiy
First Hop Ltd., Finland

Andrei L. Kustov
First Hop Ltd., Finland

INTRODUCTION

One of the applications of m-commerce is mobile authorization, that is, rights distribution to mobile users by sending authorization data (a token) to the mobile devices. For example, a supermarket can distribute personalized discount coupon tokens to its customers via SMS. The token can be a symbol string that the customers will present while paying for the goods at the cash desk. The example can be elaborated further—using location information from the mobile operator, the coupons can only be sent to, for example, those customers who are in close vicinity of the mall on Saturday (this will of course require customers to allow disclosing their location).

In the example above, the token is used through its manual presentation. However, most interesting is the case when the service is released automatically, without a need for a human operator validating the token and releasing a service to the customer; for example, a vending machine at the automatic gas station must work automatically to be commercially viable.

To succeed, this approach requires a convenient and uniform way of delivering authorization information to the point of service—it is obvious that an average user will only have enough patience for very simple operations. And this presents a problem.

There are basically only three available local (i.e., short-range) wireless interfaces (LWI): WLAN, IR, and Bluetooth, which do not cover the whole range of mobile devices. WLAN has not gained popularity yet, while IR is gradually disappearing. Bluetooth is the most frequently used of them, but still it is not available in all phones.

For every particular device it is possible to send a token out using some combination of LWI and

presentation technology, but there is no common and easy-to-use combination. This is a threshold for the development of services.

Taking a deeper look at the mobile devices, we can find one more non-standard simplex LWI, which is present in all devices—acoustical, where the transmitter is a phone ringer. Token presentation through acoustic interface along with general solution of token delivery via SIM Toolkit technology (see 3GPP TS, 1999) was presented by Khashchanskiy and Kustov (2001). However, mobile operators have not taken SIM Toolkit into any serious use, and the only alternative way of delivering sound tokens into the phone-ringing tone customization technology was not available for a broad range of devices at the time the aforementioned paper was published.

Quite unexpectedly, recent development of mobile phone technologies gives a chance for sound tokens to become a better solution for the aforementioned problem, compared with other LWI. Namely, it can be stated that every contemporary mobile device supports either remote customization of ringing tones, or MMS, and in the majority of cases, even both, thus facilitating sound token receiving over the air.

Most phone models can playback a received token with only a few button-clicks. Thus, a sound token-based solution meets the set criteria better than any other LWI. Token delivery works the same way for virtually all phones, and token presentation is simple.

In this article we study the sound token solution practical implementation in detail. First, we select optimal modulation, encoding, and recognition algorithm, and we estimate data rate. Then we present results of experimental verification.

ACOUSTIC DATA CHANNEL

We consider the channel being as follows. The transmitter is a handset ringer; information is encoded as a sequence of sine wave pulses, each with specified frequency and amplitude.

Multimedia message sounds and most ringing tones are delivered as sequences of events in MIDI (musical instrument digital interface) format. A basic pair of MIDI events (note on and note off) defines amplitude, frequency, duration of a note, and the instrument that plays this note. MIDI events can be used to produce information-bearing sound pulses with specified frequency and amplitude.

Widely used support of polyphonic MIDI sequences allows playback of several notes simultaneously. Nonetheless, this has been proved worthless because in order to get distinguished, these notes have to belong to different non-overlapping frequency ranges. Then the bit rate that can be achieved would be the same as if wider frequency range was allocated for a single note.

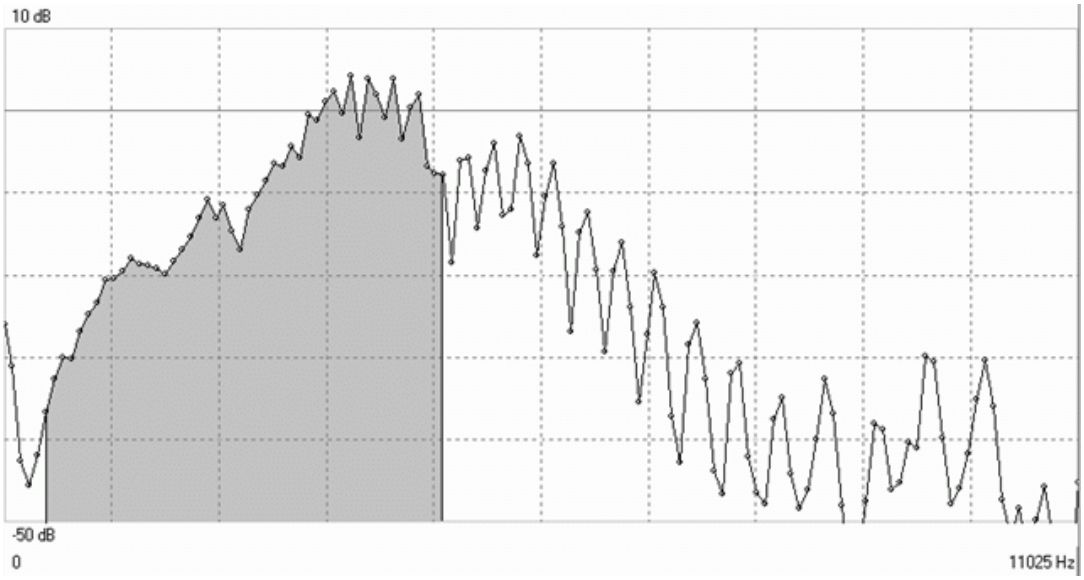
The receiver is a microphone; its analog sound signal is digitized and information is decoded from the digital signal by recognition algorithm, based on fast fourier transform (FFT) technique. FFT is, in our opinion, a reasonable trade-off between efficiency and simplicity.

We investigated acoustics properties of mobile devices. After preliminary comparison of a few mobile phone models, we found that ringer quality is of approximately the same level. All handsets have a high level of harmonic distortions and poor frequency response. The results shown in Figures 1 and 2 are obtained for a mid-class mobile phone SonyEricsson T630 and are close to average.

MIDI-based sound synthesis technology applies limitations on pulse magnitude, frequency, and duration. At the same time, ringer frequency response is not linear and the level of harmonic distortions is very high. Figure 1 shows frequency response measured with a sweeping tone or, to be precise, a tone leaping from one musical note to another. To obtain this, the phone played a MIDI sequence of non-overlapping in-time notes that covered a frequency range from 263 to 4200 Hz (gray area).

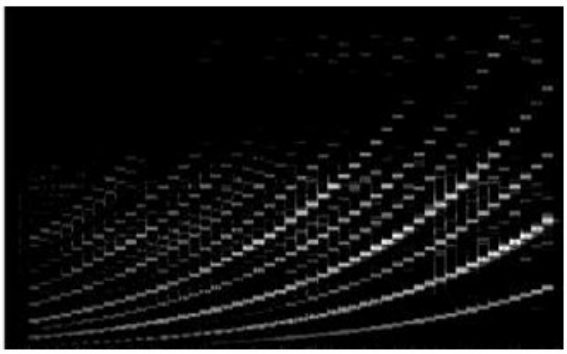
The frequency response varies over a 40 dB range, reaching its maximum for frequencies from approximately 2.5 to 4 kHz. Moreover, spectral

Figure 1. Frequency response measured with test MIDI sequence in hold-max mode



components stretch up to 11 KHz, which is caused by harmonic distortions. This is illustrated also by Figure 2.

Figure 2. A spectrogram of the test MIDI sequence



Horizontal axis is time; overall duration of the test sequence is 15 seconds. Vertical axis is sound frequency, which is in range from 0 to 11025 Hz. Brightness is proportional to sound relative spectral density; its dynamic range is 60 dB, from black to white.

We also found that frequency of the same note may differ in different handsets. Nevertheless, the ratio of note frequencies (musical intervals) remains correct, otherwise melodies would sound wrong.

For a simplex channel with such poor parameters, as reliable a data encoding method as possible is to be used. Frequency shift keying (FSK) is known as the most reliable method which finds its application in channels with poor signal-to-noise ratio (SNR) and non-linear frequency response.

It is not possible to negotiate transfer rate or clock frequency, as it is usually done in modem protocols because acoustic channel is simplex. To make the channel as adaptive as possible, we

have chosen to use differential FSK (DFSK), as it requires no predefined clock frequency. Instead, frequency leaps from one pulse to another provide the channel clocking. The difference between frequencies of consecutive pulses determines the encoded value.

Once encoding scheme is selected, let us estimate possible transfer rate before we can find the balance between data transfer rate and channel reliability. Suppose the transmitter generates a sequence of pulses of duration τ, which follow without gaps with repetition frequency f. If each frequency leap between two consequent pulses carries N bits of information, the overall bit rate p is obviously:

$$p = N \bullet f . \quad (1)$$

In DFSK, for each frequency leap to carry N bits, we must be able to choose pulse frequencies from a set of $2^N + 1$ values. If a pulse frequency can have n values, we will have

$$p = [\log_2(n-1)] \bullet f , \quad (2)$$

where by $[]$ we denote integer part. It follows from (1, 2), that to increase p, we must increase pulse repetition frequency f and the amount of possible values for pulse carrier frequencies n. However, if the recognition is based on spectral analysis, we cannot increase n and f independently. Let us show it. Assume for simplicity that pulse frequency can have any value within frequency range F. Then the number of available values of coding frequencies will be

$$n = [\log_2(F/\Delta f - 1)], \quad (3)$$

where Δf is the minimal shift of pulse frequency between two consecutive pulses. Maximum n is achieved with maximum F and minimum Δf. Both parameters have their own boundaries. Bandwidth is limited by the ringer capabilities, and frequency shift is dependant on pulse repetition frequency f, due to the fundamental rule of spectral analysis (Marple, 1987), which defines frequency resolution δf to be in reverse proportionality to observation time T:

$$\delta f = 1 / T. \quad (4)$$

How can (4) be understood in our case of a sequence of pulses? Having converted the signal into frequency domain, we will get the sequence of spectra. As information is encoded in the frequency pulses, we must determine the pulse frequency for every spectrum. This can only be done with certain accuracy δf called frequency resolution. The longer time T we observe the signal, the better frequency resolution is. So for given pulse duration τ, equation (4) sets the lower limit for frequency difference Δf between two consecutive pulses:

$$\Delta f \geq 1 / \tau = f. \quad (5)$$

This means, that if we increase pulse repetition rate f, then we have to correspondingly increase frequency separation Δf for the consecutive pulses; otherwise the spectral analysis-based recognizing device will not principally be able to detect signal.

Let us now try to estimate the data rate for the system we studied earlier. Figures 1 and 2 show that harmonic distortions are very high, and second and third harmonics often have higher magnitudes than the main tone. Consequently, the coding frequencies must belong to the same octave. Their frequency separation should be no less than defined by (5).

An octave contains 12 semitones, so possible frequency values f_i are defined by the following formula:

$$f_i = f_0 \bullet 2^{i/12}, \; i=0...11. \quad (6)$$

The minimum spacing between consecutive notes is for i=1; maximum for i=11.

In our case, we decided to use the fourth octave—as the closest to the peak area of phone ringer frequency response—in order to maximize SNR and thus make recognition easier. For it, f_0= 2093 Hz, and minimum spacing between notes is 125 Hz. Taking the maximum amount of N = 3 (9 coding frequencies), we can estimate transfer rate as:

$$p_{max} = 3 \bullet 125 = 375 \text{ bps}. \quad (7)$$

Recognition Algorithm (Demodulation)

The following algorithm was developed to decode information transferred through audio channel. Analog audio signal from the microphone is digitized with sampling frequency ***Fs*** satisfying Nyquist theorem (Marple, 1987). A signal of duration ***Ts*** is then represented as a sequence of ***Ts/Fs*** samples. FFT is performed on a sliding vector of ***M*** signal samples, where ***M*** is a power of 2.

- First, sequence of instant power spectra is obtained from the signal using discrete Fourier transform with sliding window (vector) of ***M*** samples. To get consecutive spectra overlapped by 50%, the time shift between them was taken ***M/2Fs***. Overlapping is needed to eliminate the probability of missing the proper position of a sliding window corresponding to the pulse existence interval, when the pulse duration is not much longer than analysis time significantly (at least twice).
- Second, the synchronization sound is found as sine wave with a constant, but not known in advance frequency, and a certain minimum duration.
- Third, the spectrum composed of maximum values over the spectra sequence (so-called hold-max spectrum) is used to find the pulse carrier frequencies. This step relies on the assumption that used frequency range does not exceed one octave. In other words, the highest frequency is less than twice the value of the lowest one.
- Forth, time cross-sections of spectra sequence at found carrier frequencies are used to recognize moments of sound pulse appearances.
- The last step is reconstruction of encoded bit sequence having the time-ordered set of frequency leaps.

Such an algorithm does not need feedback and can work with unknown carrier frequencies in unknown but limited frequency range. Recognizing the beginning of the transmission is critical for the correct work, so we added "synchronization header" in the beginning of the signal. The length of this header is constant, so the throughput of the system will rise with the message length.

Recognizer Parameters

Here we explain how the parameters of analyzer (***Fs, M***) are defined from that of signal (***f,*** Δf). After FFT, we have ***M/2*** of complex samples in frequency domain, corresponding to frequency range from 0 to ***Fs/2***. So for this particular case, frequency resolution obviously equals the difference between the consecutive samples in the frequency domain; namely,

$$df = Fs / M. \quad (7)$$

According to (4), minimum required time of analysis is

$$T = M / Fs. \quad (8)$$

It is obvious, that ***T*** must not exceed burst duration τ. Combining (8) and (5), we get:

$$M / Fs \le 1 / f \quad (9)$$

On the other hand, frequency resolution ***df*** must not exceed spacing Δf between carrier frequencies:

$$Fs / M \leq \Delta f \quad (10)$$

Combining (9) and (10), we will finally get:

$$f \leq Fs / M \leq \Delta f \quad (11)$$

which shows that values of analyzer parameters may be restricted when (5) is close to the equation. This imposes requirements on the sound recognition algorithm to work reliably nearby the "critical points," where the recognition becomes principally impossible.

EXPERIMENTAL RESULTS

We implemented a prototype of acoustic data channel with the mobile phone SonyEricsson T630, whose characteristics are seen in Figures 1 and 2.

For encoding, we developed software that encoded symbol strings in ASCII to melody played by an electric organ. The instrument was chosen from 127 instruments available in MIDI format, because its sound is the closest to the sine wave pulses model we used in calculations. It is maintained at approximately the same level over the whole note duration.

The recognizer consisted of a Sony ECM-MS907 studio microphone for signal recording, and a conventional PC with a sound card was used for signal analysis. FFT processing was done by our own software.

In the beginning of our experiments, we used the parameters described in the theoretical section. Later we found that at the highest possible transfer rate, data recognition is not reliable. So we gradually increased pulse duration until recognition became reliable. Eventually we selected the following modulation parameters: ***n=5*** (each frequency leap carries two data bits), notes were evenly distributed over the octave (C, D#, F, G, A in musical notation, and they correspond to frequencies 2093, 2489, 2794, 3136, and 3520 Hz), and pulse duration was 46 ms.

Figure 3 shows a spectrogram of recognizable signal from the microphone.

Horizontal axis is time, and overall signal duration is 2 seconds. Vertical axis is frequency, and one can see the leaps between consecutive pulses. Brightness is proportional to the signal intensity.

This example signal carries 88 bits of information (a string "hello world," coded as 11 ASCII characters), which makes the data transfer rate approximately 40 bps. Overhead from the synchronization header is ca. 25%; for longer messages the average transfer rate would be higher.

DISCUSSION

We have managed to implement a reliable data channel from the phone; the advantage of the proposed recognition algorithm is that it can work in the same way for every mobile device, independent on acoustic properties of different brands and models, although encoding frequencies are different.

The channel is principally one way: the handset cannot receive any feedback that can be used, for example, for error correction. Nevertheless, developed recognition algorithm provided good reliability. For a handset placed 30 cm from the microphone, in a room environment, recognition was 100% reliable. This condition corresponds to the output of the average phone in a "normal" room environment.

Ensuring reliability does not seem to be a very difficult task. First of all, SNR can be improved by increasing the number of receiving microphones. On the other hand, in practical systems simple shielding is very easy to implement. And finally, even one error in recognition is not fatal: the

user can always have another try. A recognizing device can easily identify cases of unsuccessful recognition and indicate the former case for the user to retry.

The recognition system can be implemented on any PC equipped with a sound card. The algorithm is so simple that the system can also be implemented as an embedded solution based on digital signal processors. Microphone requirements are not critical either: both the frequency response and SNR of entry level microphones are much better than those of mobile device ringers. This means that cheap stand-alone recognizers can be implemented and deployed at the points of service.

It is interesting to note that other devices capable of playing MIDI sequences (e.g., PDAs) can be used as well as mobile phones.

Measured transfer rate (40 bps) was considerably less than the estimation, obtained in our simple model—375 bps. We think that the reason for this was slow pulse decay rate in combination with non-linear frequency response. Amplitude of the note with frequency close to a local frequency response maximum might remain higher than amplitude of the consecutive note through the whole duration of the latter. Thus, the weaker sound of the second note might be not recognized.

However, we consider even such relatively slow transmission still suitable for the purposes of mobile authorization applications, because authorization data is usually small and its transmission time is not critical.

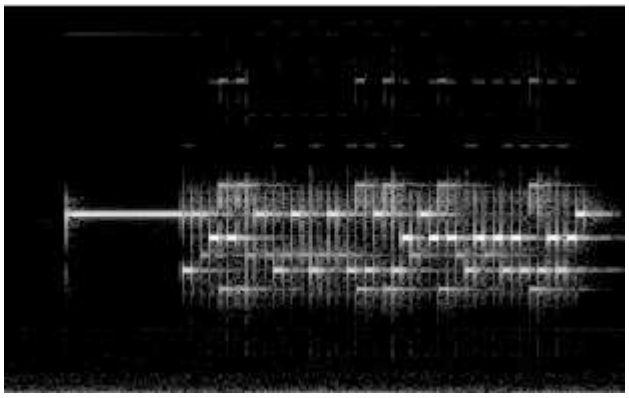
Our example (Figure 3) seems to be a quite practical situation—transmitting 11-symbol password during 2s is definitely not too long for a user. Typing the same token on the vending machine keyboard would easily take twice as long.

The acoustic presentation method might be an attractive feature for teenagers (e.g., mobile cinema tickets being one conceivable application).

ACKNOWLEDGMENTS

The authors would like to thank Petteri Koponen for the original idea.

Figure 3. Encoded "hello world"; note the leading synchronization header. Overall duration is approximately 2 seconds.



REFERENCES

Khashchanskiy, V., & Kustov, A. (2001). Universal SIM Toolkit-based client for mobile authorization system. *Proceedings of the 3rd International Conference on Information Integration and Web-Based Applications & Services (IIWAS 2001)* (pp. 337-344).

Marple, S. Lawrence Jr. (1987). *Digital spectral analysis with applications.* Englewood Cliffs, NJ: Prentice-Hall.

3GPP TS 11.14. (1999). *Specification of the SIM application toolkit for the Subscriber Identity Module-Mobile Equipment (SIM-ME) interface.* Retrieved from http://www.3gpp.org/ftp/Specs/html-info/1114.htm

KEY TERMS

Fast Fourier Transform (FFT): An optimized form of the algorithm that calculates a complex spectrum of digitized signals. It is most widely used to obtain a so-called power spectrum as a square of a complex spectrum module. Power spectrum represents energy distribution along frequency axis.

Frequency Resolution: The minimum difference in frequencies which can be distinguished in a signal spectrum.

Frequency Response: For a device, circuit, or system, the ratio between output and input signal spectra.

Frequency Shift Keying (FSK): The digital modulation scheme that assigns fixed frequencies to certain bit sequences. Differential FSK (DFSK) uses frequency differences to encode bit sequences.

Harmonic Distortions: Alteration of the original signal shape caused by the appearance of higher harmonics of input signal at the output.

IR: Short-range infrared communication channel.

Musical Instrument Digital Interface (MIDI): A standard communications protocol that transfers musical notes between electronic musical instruments as sequences of events, like 'Note On', 'Note Off', and many others.

Sampling Frequency: The rate at which analogue signal is digitized by an analogue-to-digital converter (ADC) in order to convert the signal into numeric format that can be stored and processed by a computer.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 15-19, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.33
The Design of Mobile Television in Europe

Pieter Ballon
Vrije Universiteit Brussel, Belgium

Olivier Braet
Vrije Universiteit Brussel, Belgium

ABSTRACT

Mobile television is potentially the most anticipated mass-market mobile application across Europe. This chapter examines the business model design of mobile TV by the various stakeholders currently piloting mobile broadcasting in the European national markets. It adapts a generic business model framework to systematically compare five recent pilots of the two mobile broadcasting technologies that are currently trialled most intensively in Europe, that is, digital video broadcasting-handheld (DVB-H) and digital audio broadcasting-Internet protocol (DAB-IP). The article illustrates the cross-impact of cooperation agreements between the various stakeholders with technological, service-related, and financial design choices. It also provides insights as to the likely business models in the upcoming commercialisation phase of mobile broadcasting in Europe.

INTRODUCTION

One of the most anticipated applications in Europe's mobile commerce and multimedia landscape is mobile TV. It is widely argued that mobile digital television has the potential of becoming one of the next high-growth consumer technologies (Kivirinta, Ali-Vehmas, Mutanen, Tuominen, & Vuorinen, 2004; Södergard, 2003), provided it is able to master its inherent complexities in terms of the various stakeholders required to cooperate (Shin, 2006). It has a clear and easily understandable value proposition towards the majority of end users: TV on a mobile device. Also, the technology lies at the crossroads of two powerful socio-technical trends: the ubiquity of mobile phones, and new forms of accessing media content.

In the European mobile market, digital TV on a mobile device is not a novelty. Initial TV services on mobile phones consist of streaming video over the cellular network. Third generation

(3G) cellular networks (i.e., Universal Mobile Telecommunications System [UMTS]) already allow for streaming video for a considerable time. In several European countries, a wide selection of rich video content is available over UMTS, with large markets such as Italy, the UK, and France as front-runners. The downside of this solution is that without network capacity investments the video images degrade in quality if there are too many simultaneous users, since content needs to be streamed to each user in a point-to-point fashion. Therefore, streaming content over cellular is a costly option for serving a mass audience. The Multimedia Broadcast Multicast Service (MBMS) standard could circumvent this by offering a multicast and a broadcast mode for existing cellular networks, but its implementation time path is currently unclear.

An alternative is offered by new point-to-multipoint digital TV standards such as DVB-H, DAB-based standards, and Media-FLO. These are able to offer high quality live broadcast TV, allowing mass-market service delivery in a more scalable way and at more attractive operational costs (but still considerable capital expenditures). However, since the current uptake of mobile video content over 3G is quite slow, some operators have expressed doubts as to whether investments in these new network technologies are necessary and are counting on the fact that their 3G property will be sufficient for the coming years.

Other major technology choices faced by prospective European mobile TV operators include whether or not to combine any new mobile broadcasting technologies with uplink technologies such as *global system for mobile communications* (GSM) and UMTS in order to ensure more flexibility and interactivity in the service offering, and whether new mobile broadcast standards should "piggy-back" on top of existing networks—*digital video broadcasting-terrestrial* (DVB-T) and DAB networks, respectively, - or whether they should be built as stand-alone networks.

The technological outlook on mobile broadcasting will be sketched briefly in the second section of this chapter. However, we aim to demonstrate that the main design choices to be addressed are not only, even not predominantly, techno-economic in nature. It is our assertion that the cross-impact of strategic cooperation and competition issues (e.g., related to control over this new market by broadcasters, content aggregators, or cellular network operators), market expectations (e.g., related to speed of uptake, service offerings, degree of interactivity), and legacy situations (e.g., related to existing networks and customer relations) will to a large extent determine the outcome of mobile digital TV in Europe (see Shin, 2006 for a similar argument on digital multimedia broadcasting [DMB] development in Korea). A four-level design framework, along with a detailed enumeration of mobile TV design issues, is presented in the third section.

To test both technological maturity and marketability of the new service, many mobile broadcast test and experimentation platforms (TEPs)[1] in the form of field trials and market pilots have been started in Europe since 2004. Two multicasting standards are being trialled quite intensively in the European area, that is, DVB-H and DAB-IP/DMB. Their commercialisation is expected to start in earnest from 2007 on, with small-scale commercialisation already available in 2006 in a few countries..

We selected and analysed five of the largest and most documented pilots (four DVB-H pilots and one DAB-IP pilot) using publicly available info, telephone interviews, and e-mail interactions with key pilot participants. The fourth section contains the results of the case analysis in terms of the design choices made, how these were interlinked, and which cooperation schemes were devised. Where possible, the consequences of the design choices for the commercialisation phase are indicated. Finally, the final section offers some concluding remarks in terms of the

models and strategies encountered in European mobile TV pilots.

TECHNOLOGY OUTLINE

The mobile broadcast landscape consists of three primarily non-proprietary standard families (Integrated Services Digital Broadcasting-Terrestrial [ISDB-T], DAB-based standards and DVB-H) developed by industry associations, and of the proprietary Media-FLO technology developed by Qualcomm. This section focuses on the DVB-H and DAB-based standards, as these are currently being piloted intensively throughout Europe. For a deeper analysis, see the rather extensive technological literature available on this subject (e.g., Curwen, 2006; Faria, Henriksson, Stare, & Talmola, 2006; Skiöld, 2006; Weck & Wilson, 2006).

The DVB-H Standard

DVB-H enjoys strong and organised support in Europe, as witnessed by the large amount of trials and pilots currently carried out, and by the forceful backing by European telecommunications giant Nokia, but it also has its supporters abroad. For instance, Intel Corporation, Modeo, Motorola, Nokia, and Texas Instruments created the Mobile DTV Alliance in January 2006 to promote the growth and evolution of DVB-H in the U.S.A.

As an extension of the DVB-T standard, DVB-H is relatively straightforward to implement, with several adjustments that make the standard more suitable for mobile communication. DVB-H uses significantly less bandwidth than DVB-T, approximately 300 kilobits versus 3 Megabytes per channel. Also, DVB-H saves on battery power by using the technique of time slicing, inserting the different video channels into the transmitted transport stream in bursts of data. The additional level of forward error correction (MPE-FEC) inserted in the DVB-H front end contributes to the robustness of the DVB-H signal.

DVB-H detractors regularly dispute the performance of DVB-H. They add that DVB-H channel switching is slow, unlikely to be able to deliver the stated data rates, and is susceptible to signal variations and problems with synchronisation. In fact, even DVB-H supporters acknowledge that the up to 6 seconds to switch channels is an issue, but claim that it is not insurmountable. DVB-H receiver manufacturers are confident they can drive down the channel switching to approximately 1.5 seconds as already achieved with DVB-T receivers.

Data on the amount of channels that DVB-H can carry as opposed to DAB-based standards varies. Currently, DVB-H seems to be able to offer considerably more channels, with between 10 and 20 channels per multiplex being offered in various trials, versus around 5 channels for DMB and DAB-IP.

DAB-IP, T-DMB, AND S-DMB

DAB-based standards include DAB-IP, terrestrial digital multimedia broadcasting (T-DMB), and S-DMB. DAB-IP can be described as a DMB I addition to DAB digital radio. More specifically, the network platform consists of DAB enhanced packet mode (EPM), in conjunction with an IP application. EPM was standardised by the WorldDAB Forum and enables video and other services—that are more sensitive to errors than the native audio services carried by DAB—to be carried.

DMB is an European Telecommunications Standards Institute (ETSI) standard developed in Europe that delivers mobile television services using the Eureka-147 DAB standard with additional error correction. Within the DMB sphere, a distinction is made between T-DMB and S-DMB. DAB-IP and T-DMB are both based on the DAB

transport layer, contrary to S-DMB. T-DMB uses the terrestrial network in Band III and/or Band L while S-DMB uses the satellite network in Band L. To complicate matters, S-DMB is actually not directly related to the DAB standard, but was developed in Korea using the System E International Telecommunication Union (ITU) standard based on code division multiple access (CDMA). S-DMB can deliver 13 video channels in a typical spectrum allocation.

T-DMB supporters argue that the scarcity of available spectrum will cripple the implementation and acceptance of DVB-H and MediaFLO, whereas T-DMB, due to its association with DAB, already has most of the required spectrum and infrastructure in place. T-DMB backers claim that it requires even less power than DVB-H or MediaFLO. Other reasons quoted by the T-DMB camp on why their standard is more suited than DVB-H for mobile digital TV are: lower channel switching time (around 1.5s), 30 frames per second (fps) versus just 15 fps on DVB-H (with traditional TV delivering between 25 and 30 fps), and the usage of 1.5 MHz channels requiring less power and circuit complexity (DVB-H uses 5 to 8 MHz channels).

However, as argued previously, DAB-based standards seem to be disadvantaged vis-à-vis DVB-H in terms of the number of channels per multiplex.

Principal Technical Components

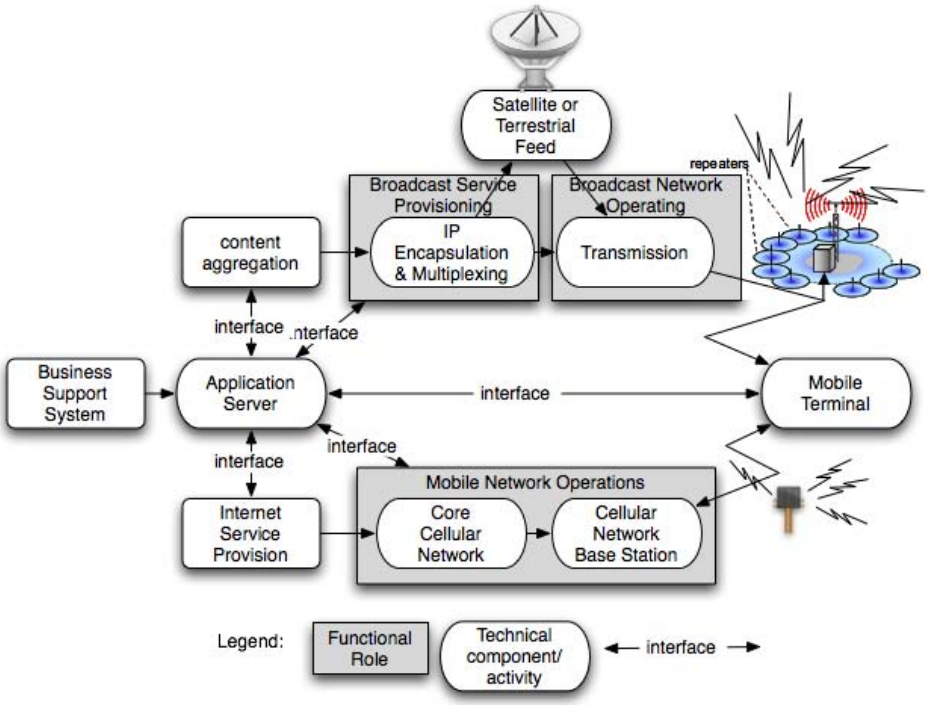
In a complete mobile broadcast system (including an uplink for interactive applications), the following functional roles and their constituting technical components can be distinguished:

- **Business support system:** Back office system for billing and help desk support.
- **Content aggregation:** Entails the collection of content in a content management system
- **Broadcast service provisioning:** Provides frequency allocation, IP encapsulation, and multiplexing
- **Broadcast network operating:** Entails network transmission of the content.
  - Note that, although the technical literature often groups IP-encapsulation, multiplexing, and the ownership + management of the transmitters into one actor—the broadcast network operator (BNO)—we will distinguish these roles, because in the business models these roles may be performed by different actors (see later on).
  - Therefore, the functions performed *before* the transmission will be referred to as broadcast service provisioning. All functions connected with the transmission of the DVB-H signal will be referred to as broadcast network operating.
- **Mobile network operations:** Provide the return channel through its cellular network.
- **The application layer:** provides for communication between the broadcast content and the mobile network operator (MNO) return channel.
- **The mobile terminal:** Has to be equipped with the suitable receiver in order to access the (free, subscribed or pay-per view) content.

METHODOLOGY AND DESIGN ISSUES

As illustrated by the technical architecture, cooperation between various stakeholders is necessary in order to bring mobile broadcasting to the market. Even though there might be significant differences between the pilot and the commercialisation phase, it may be assumed that the cooperation models currently employed in pilots to a certain extent

Figure 1. General technical architecture of mobile broadcasting (based a.o. on Digitag, 2005 and Pieck, 2005)



foreshadow the business models that will arise in the commercialisation phase (see also Dittrich & Van den Ende, 2006). In line with current thinking on strategic management and business model design (Ballon, 2007; Barney, 1991, 1997; Faber et al., 2003; Haaker, Bouwman, & Faber, 2004), our case analysis focuses on four business model design phases, which are equally relevant to the cooperation models used in the different pilots. These phases can be defined as follows:

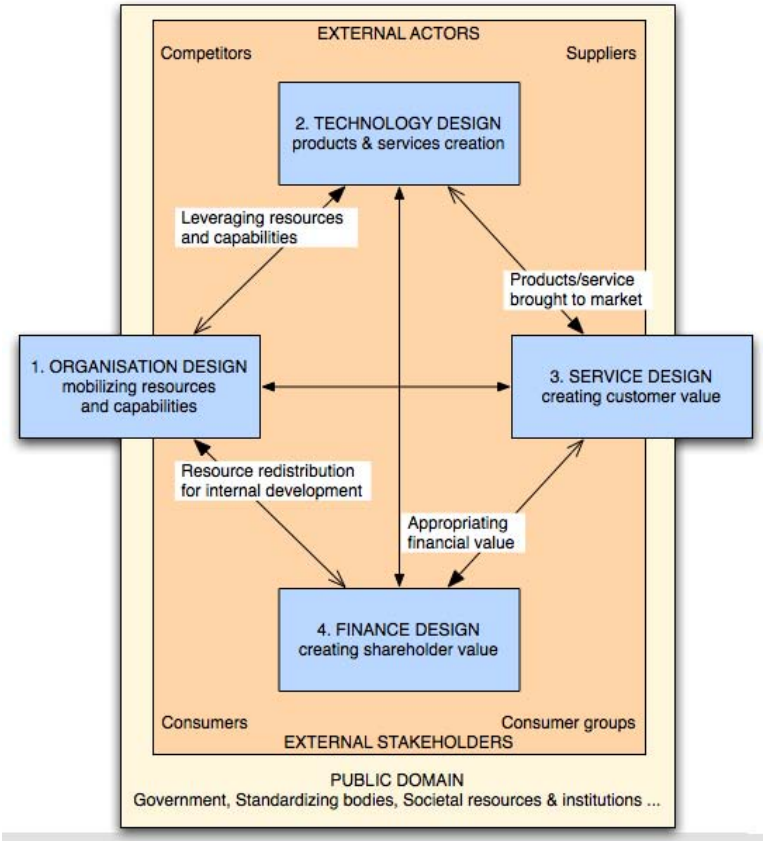
1. **Organisation design phase:** The organisation design involves defining a business scope (what customers will we try to reach and how), identifying distinctive competences, and making business governance decisions (make versus buy decisions).
2. **Technology design phase:** The technology design involved defining the technology scope (which technical design are we trying to develop and how), identifying the systemic competences that will contribute to the business strategies, and deciding on the IT governance (how will we develop or acquire the needed technical competences).
3. **Service design phase:** The service design involves choosing a specific value proposition towards the user, which implies choosing for a specific strategic scope.

4. **Financial design phase:** In a final phase, the financial modalities are formalised in binding contracts that clearly describe each partner's responsibilities, and the financial or other benefits they will receive in return.

Figure 2 illustrates how these design steps occur (chrono)logically, but cross-decisions are also possible, as illustrated by the horisontal and vertical arrows.

The framework presented previously emphasises organisation design as the starting point of any business modelling or cooperation modelling. This is especially relevant in cases such as mobile TV where convergence between various stakeholders and sectors increases the strategic importance of organisational design significantly. The focus of this chapter will be on cooperation models between stakeholders in the pilot phase of bringing mobile TV to the market. Therefore, a generic mobile TV value network model is constructed on which these models are subsequently mapped. They are described with the use of three main building blocks: business actors, business roles, and business relationships (see also Ballon et al., 2005). *Business actors* can be physical persons or corporations that participate in the creation of economic value, through the

Figure 2. Business modeling cycle



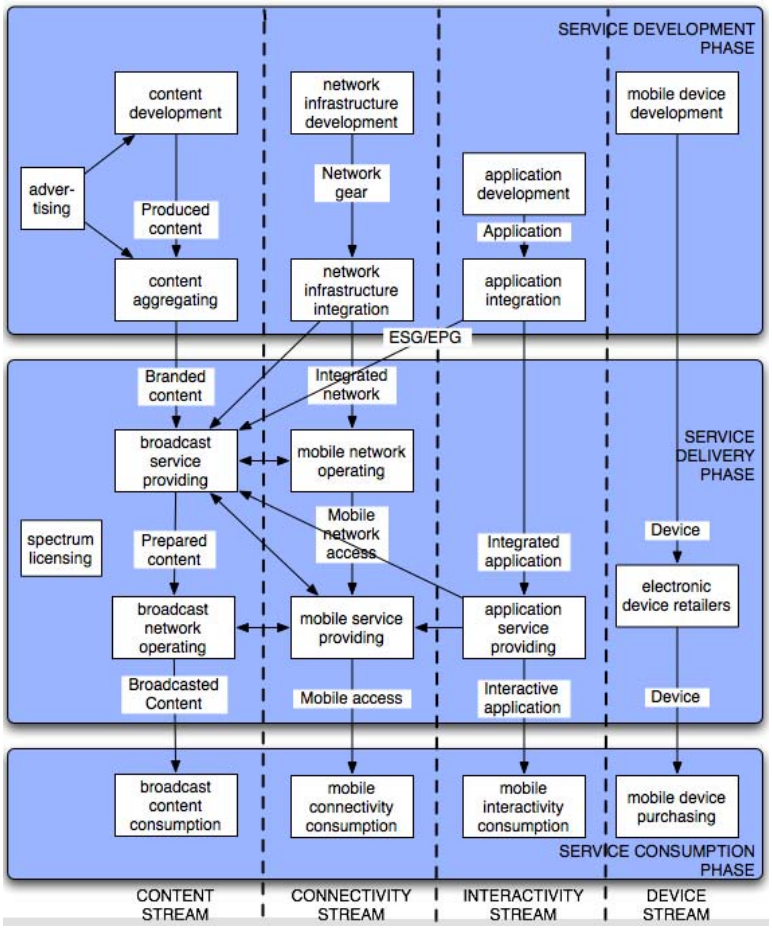
mobilisation of tangible or intangible resources within a business value network. *Business roles* are logical groups of business processes that are fulfilled by one or more actors. Business actors provide value to or derive value from the business roles they play. Finally, *business relationships* are the contractual exchanges of products or services for financial payments or other resources.

Organisation Design

The organisation design dimension concerns the relationships that are adopted between cooperating industrial partners to produce value for end customers. It entails which roles the partners take on, what resources each actor brings to the pilot and will bring during future commercialisation, and what kinds of cooperations arise during the delivery of specific mobile content services.

Figure 3 presents a generic value network for mobile digital TV delivery. The black arrows represent business relationships in the form of delivered services. The blue rectangles depict the different service deployment phases. The white rectangles are the business roles that actors can adopt. Each discrete role can be performed by a

Figure 3. Generic mobile digital TV value network



discrete actor, in which case the actor's name is inferred from the business role. Very often actors will perform more than one business role. Each configuration of roles, actors, and relationships constitutes a different cooperation model.

Absent from Figure 3—since this cannot be inferred from the mobile TV pilots—is the upwardly flow of revenues. Although the services flow more or less chronologically from the upper layers to the lower layers to the end users, the revenues may not. Most often, the infrastructure players receive their revenues in advance, that is, when the network operator purchases the networking gear. It is then up to the network operator to leverage this infrastructure into tempting service offerings towards the end users. To formulate this differently, one can say that there exists asynchronicity between the supply chain and the value chain. Because of the possible disconnect between a network operator's expenses and expected revenues, there are cases of vendor financing, where the infrastructure manufacturers supply credit lines to the network operators so they can foot the initial capex bill.

The following business roles were distinguished:

Service Development Phase

In this phase those business roles are situated that are necessary for the development and integration of the mobile digital content and of essential technical components such as the transmission infrastructure and the terminal devices.

- **Content development:** Development of the content that will be distributed. The distribution and branding of the content to the end user can be done directly by the content developer (which is rare) or via a content aggregator (which is common).
- **Content aggregation:** The acquisition, branding/marketing, and scheduling of the content that will be delivered to the broadcast service provider.
- **Network infrastructure development:** The designing and manufacturing of the network equipment, to be purchased and installed by the network operators (both mobile and broadcast network operators (BNOs).
- **Network infrastructure integration:** The installation of the network equipment acquired from network equipment manufacturing, at the request of the mobile and BNOs.
- **Application development:** The development of applications that will be used in the application layer.
- **Application integration:** The integration—when necessary—of application components into a platform or bundle.
- **Mobile device development:** The design and manufacturing of the end-user mobile devices with which the mobile content can be consumed.

Service Delivery Phase

In this phase the business roles are situated that transport and deliver the products and services to the end user, or prepare them to be sold through middlemen.

- **Broadcast service provision:** Preparation, encapsulation, encryption (with Digital Rights Management [DRM]), and multiplexing of the content so it can be delivered to the mobile devices via the BNO. An actor called the *datacast service operator* usually performs this role.
- **Broadcast network operation:** Operation (but not necessarily ownership) of the broadcast network. This can be combined with other networks. If a business actor, for example, already operates a DVB-T network,

they could be a more logical party to operate the DVB-H network.

- **Mobile network operating:** The operation and management of the mobile cellular network. This business role will be relevant if mobile broadcasting is integrated with a mobile cellular network.
- **Mobile service providing:** This role constitutes a layer between the MNO and the end user. Most often, this role together with the role of mobile network operation is performed by a single actor, which we call an MNO. A *mobile virtual network operator* (MVNO) stands for a special case of a business actor that provides network services to customers without owning the physical network infrastructure, but does sell mobile services to end users.
- **Application service providing:** The daily management and ownership of the (interactive) application service platform built by the application integrator.
- **Electronic device retailing:** The selling of handsets to end users. MNOs can also perform this role through the subsidisation or marketing of specific devices.

Service Consumption Phase

In this final phase the roles are situated that are related to consumption. Usually they are performed by a single actor (the end user), but they can also be unbundled, for example, in the case of a company buying a mobile device or mobile connectivity for its employees but not paying for the content or application consumption by the employees. One of the key questions is to what extent the different services will be combined into a single bill (and thus offered by a single customer owner). The roles are:

- **Broadcast content consumption:** The consumption of the broadcast content.
- **Mobile connectivity consumption:** The consumption of mobile connectivity services.
- **Interactivity consumption:** The return channel will usually but not necessarily run over a MNO's network. If broadcasters decide to sell mobile digital TV services to mobile TV-only devices, they could opt, for example, to use the fixed Internet as a (non-synchronised) return channel, instead of the mobile network.
- **Mobile device purchasing:** The act of purchasing the mobile device, be it a cellular plus DVB-H-enabled device or a stand-alone mobile TV terminal.

It should be noted that no single "service provider" role is included in Figure 3. Most business model literature assumes a unique service provider entity that ensures customer acquisition, billing, and customer care—in short, that possesses "customer ownership." However, in a potential unbundled market, it has to be envisaged that every provider, operator, retailer, or aggregator role can establish such a relationship with the end customer. Therefore, our design approach does not define customer ownership and the activities it entails as a specific role, but will rather treat it as an attribute that can be associated with several roles.

Technology Design

This section will describe the specific technology designs of each pilot, such as the network standard adopted for the pilot and what technological application choices were made. The following criteria were used to describe the technology design.

a. **Technical network architecture and device design:** This first criterion describes the network standard(s) that were adopted during the pilot, and that will possibly be adopted during commercialisation. For ex-

ample, the pilot participants could choose to build a standalone DVB-H network or a hybrid network such as a DVB-H network on top of a DVB-T network, and combinations of DVB-H or DAB-based standards with second generation (2G), 2.5G, or 3G networks as return channel. Also, the end user device used during the pilot is indicated.

b. **Interactivity:** This criterion refers to the kinds of interactive applications and functionalities that were developed during the pilot. This could include one button voting, voting via short message service (SMS), upload functionalities, or other forms of interactive applications.

c. Content protection: This entails the encryption or other security technology used, in order to protect the broadcasted content from being intercepted and/or re-used via other channels.

d. Electronic service guide (ESG): This criterion describes the technical standard chosen for the ESG. While the Electronic Program Guide (EPG) refers to the visual interface shown to the end user, the ESG is a structured document that contains information on all available services. With an ESG one can for instance describe whether a delivered service concerns a video game, home banking, or shopping.

Service Design

The service design dimension describes the specific characteristics of the developed end-user services, such as the degree of user interactivity allowed during the consumption of the services, and the different service bundles presented to the end user.

The potential services that can be delivered within a mobile digital TV value network to the end users can be situated along a continuum ranging from very low interactivity to high interactivity (see Table 2), which will influence the degree

Table 2. Potential mobile TV services

Scale of interactivity	Service
Low interactivity	• Standard broadcast TV channels • Special Interest TV: Niche content TV channels • Electronic Program Guide, personalised for the user • Pay-per-view TV • Scheduled content push or 'Near-Video-on-Demand' • Video on Demand • Mobile TV with integrated location-based services • Mobile TV with on-demand additional information services • Mobile TV with integrated e-commerce applications • Mobile TV with interactive entertainment services such as voting and gaming
High interactivity	• Mobile TV with video upload services • Social networking video applications

to which the mobile broadcast channel needs to cooperate with a mobile return channel.

Mobile phones have proven to be excellent conduits for interactivity using SMS. While SMS voting proves to be very popular while watching TV programs, the limitation of SMS lies in the simplicity of the interaction. For more sophisticated applications—such as allowing viewers to participate in game shows alongside the televised contestants—SMS is not convenient.

Local services can provide viewers with information on a city or region, such as weather forecasts, trailers of movies featured locally, and a teletext guide. With an interactive channel, viewers can request specific information. However, because a standard middleware interface is currently lacking, some further development is necessary before viewers will be able to trigger interactive services directly from the broadcast system.

As a rule, most TV content services listed previously will be delivered most efficiently over a mobile broadcast network. In contrast, interactive services will usually be delivered following a point-to-point distribution model over UMTS or, if the slower speeds are acceptable, over General Packet Radio Service (GPRS) or via SMS (Pilz, 2005). However, it is important to note that broadcast standards such as DVB-H can be used as stand-alone solutions for the delivery of low-interactivity services including near-video-on-demand.

Interactivity may be especially important when trying to reach the first adopter market segment. This segment is generally acquainted with on-demand content consumption such as personal video recorders (PVR), cable TV video-on-demand services, and the Internet and might consider a pure broadcast offering as a step back towards scheduled programming, with fixed viewing times of each show. Pre-downloaded content, which can be consumed when the end user has time to "snack" content, might prove to be at least equally popular. Nokia actually has a service called "Nokia Media Charger," that allows for push delivery of rich content.

The following criteria were used to describe the ways in which the service package was presented to the end users.

a. **User involvement:** This refers to the degree of interactivity experienced by users. User involvement can vary from low (no end-user involvement/interaction) to middle (user can give input, e.g., vote), to high (user can generate and post his/her own content). The degree of user involvement depends on the network characteristics, the chosen return channel, and the implementation of interactive technologies from the technology design.
b. **Product bundles:** This criterion describes the kind of product bundle that is offered to the end user. This can be a *package* (user takes a subscription on a collection of channels and does not have the authority to add or delete channels), *modules* (user can take a subscription on individual channels or theme packages), *individual views* (user can chose individual shows), or *hybrid* (mixes of the aforementioned bundles). Other product bundles that are possible are bundles with existing TV channels, bundles including new channels for specific mobile content, bundles with digital radio, bundles with interactive services, and so on.

Finance Design

Finally, the financial design criteria in this context relate to the costs of the network build-out, the revenue sharing agreements, and the business-to-consumer billing formulas.

In most pilots, no revenues were generated, and no revenue-sharing agreements were negotiated. However, during the interviews some executives provided information on how these financial matters might be resolved during commercialisation,

and this information is included in the pilots' descriptions.

The following criteria were used to describe the financial design decisions:

a. **Cost sharing agreements:** This first financial criterion describes how different actors carry the costs of the service roll-out. Three cost categories are taken into account. First, the **device cost** refers to the primary purchase cost of the handsets and to what degree the consumer has to pay the entire cost of the handset, or whether device subsidies are allowed. Second, the **network infrastructure costs** refer to the cost of building the transmission infrastructure. Third, the **content and application costs** refer to which partner carries what part of the content and/or application development cost. Besides the traditional approach, where content is aggregated by a traditional broadcaster, and applications are developed by or on behalf of a MNO, these efforts (and subsequent costs) could also be borne by other actors. For instance, a MNO could develop mobile TV content by purchasing and aggregating programs under his own brand, or even by building or acquiring a TV station of one's own.
b. **End-user billing:** This criterion describes the ways in which the user pays for the services provided. The billing formula will depend on the kinds of product bundles offered, but does not follow directly from that criterion. For example, being able to select individual shows does not necessarily imply pay-per-view pricing. Three basic end-user billing models can be distinguished: subscription based, pay-per-use, and free-to-air with advertisements. Between these three pure forms of revenue generation, any number of hybrid combinations can also arise.
c. **Revenue-sharing agreements.** The last criterion describes the ways in which the service supplier(s) agree on how the revenues generated through end-user billing are distributed throughout the value network, including the broadcasters, other content providers, and the MNOs.

CASE STUDY ANALYSIS

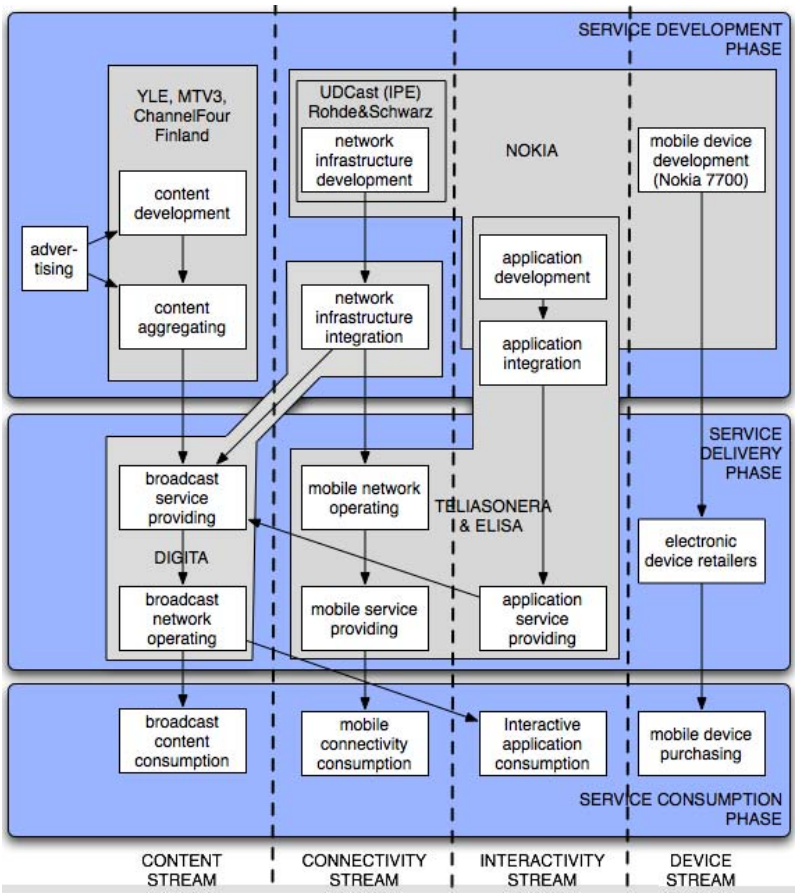
This section systematically compares a selection of European pilots using the framework defined previously. Five of the largest and most documented pilots (four DVB-H pilots and one DAB-IP pilot) were selected, that is, in Helsinki, Berlin, Paris, Oxford, and London. They were analysed using publicly available info, telephone interviews, and e-mail interactions with key pilot participants.[2] First, the organisational design of each pilot is addressed, highlighting the respective cooperation model. Next, the technical, service and financial design repercussions of these models are outlined.

Organisation Design

In this section the organisational cooperation that arose between the different pilot partners during the pilots is illustrated. Each figure is accompanied by a list of the roles performed by the business actor involved. When information could be obtained about the business roles and value networks during the future phase of commercialisation, the text expands on this issue. The information presented here is based on interviews with executives and publicly available pilot presentations.

Figure 4 illustrates the cooperation model adopted during the Helsinki trial. Note that the visual overlapping of one business actor by another, such as Nokia encapsulating UDCast and Rohde & Schwartz in Figure 4 does not imply that UDCast is a department of Nokia, but that different actors mutually performed different segments of the business role of "network infrastructure development" in the example below.

Figure 4. Helsinki pilot cooperation model

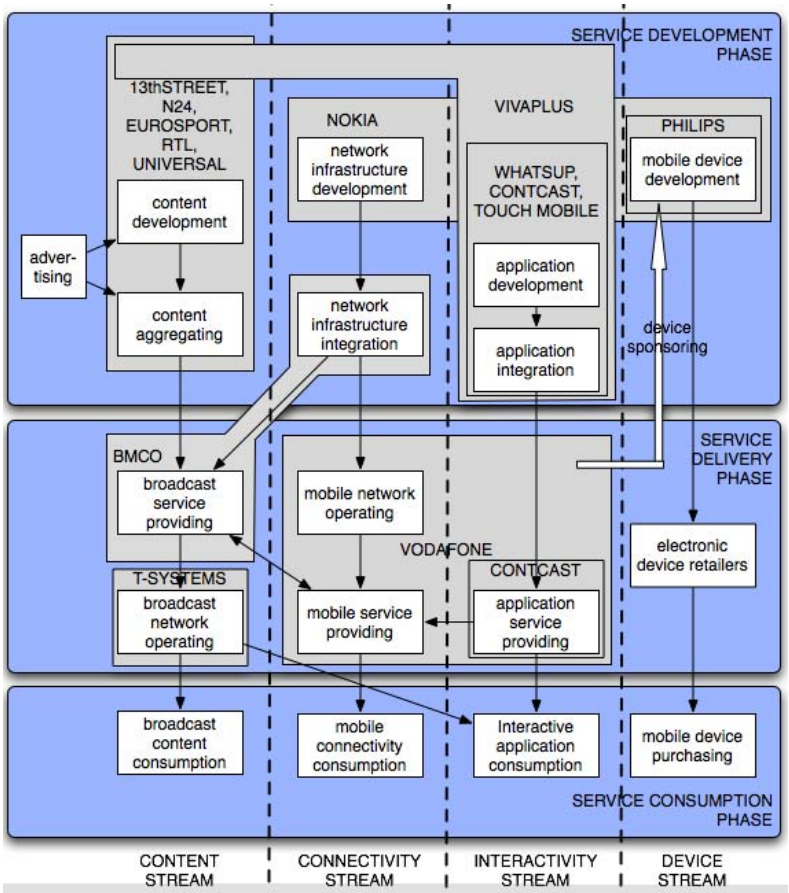


In Finland the MNOs Teliasonera and Elisa retained customer ownership during the pilot by offering the TV content service to the end users and offering help desk support. Nokia used the pilot to focus on the further development of its mobile service platform (later branded as Mobile Broadcast Solution or MBS 3.0). Digita (a unit of the French media group TDF) is the builder and owner of the DVB-T network in Finland. It used the pilot to learn about the pitfalls of rolling out a DVB-H transmission network. This practical experience proved fruitful when in March 2006 Digita won the Finnish DVB-H license beating Elisa, TeliaSonera, and Telemast Nordic. In May 2006 Digita announced that it had signed a contract with Nokia to use its DVB-H platform for the service. By the end of this year, Digita plans to reach 30% of the population by providing coverage in the Helsinki region as well as the cities of Turku, Tampere, and Oulu. Digita will hold the license for a period of 20 years.

During commercialisation, it is foreseen that Digita will adopt two business roles: broadcast service providing and broadcast network operating. Digita will be solely responsible for the management of the DVB-H digital multiplex and the transmission network, but will not offer mobile TV services directly to any end customers. The license includes a condition under which the license holder is obliged to sell network capacity to service operators. Digita will as BNO utilise an open network model for the DVB-H network, by offering access to the broadcast network to all service providers under equal, fair, and transparent terms. The role of *network infrastructure integration*, performed alone by Digita during the pilot given the lower complexity of the pilot context, will during commercialisation be jointly performed by the MNOs and Digita.

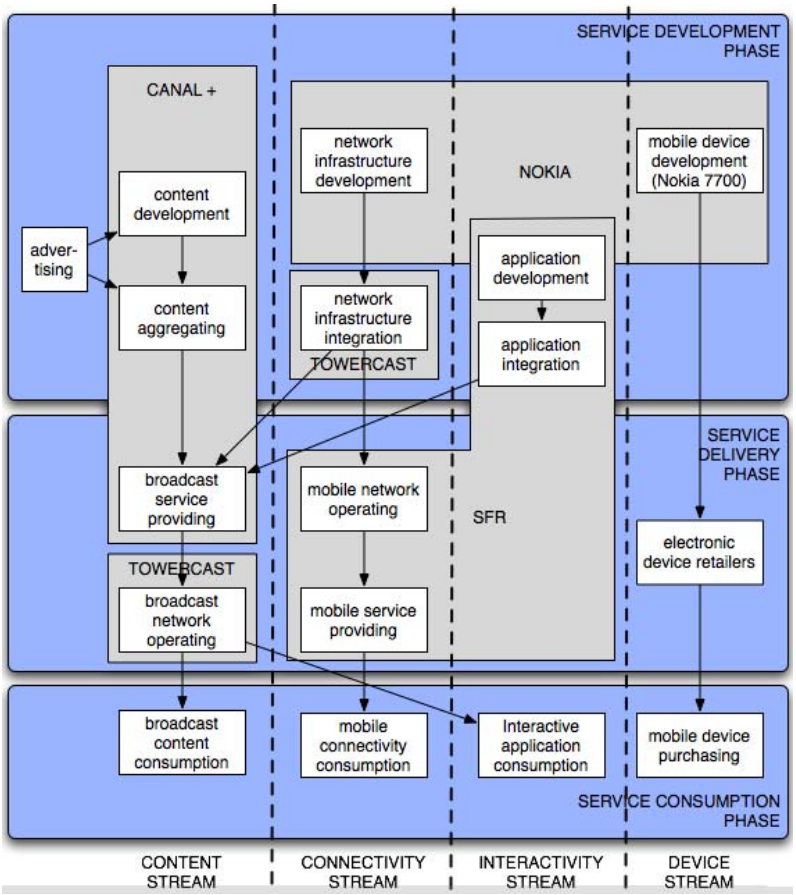
The basic cooperation model in the pilot, where Digita functioned as a common broadcast service provider for competing mobile operators, will nevertheless be replicated in the commercialisation phase, in which Digita will provide open access to its DVB-H platform to various service providers. Also, while functional roles related to broadcasting will remain in the hands of Digita,

Figure 5. Berlin pilot cooperation model



and will not be taken up by mobile operators, these mobile operators (and possibly other service providers) will most probably leverage their existing customer base to act as integrated service providers towards the customer (see also the sections on service and financial design). Figure 5 illustrates the cooperation model adopted during the Berlin trial. The role of broadcast network operator was taken by T-Systems, being already the DVB-T network operator in Berlin. During the pilot phase, half of a T-Systems DVB-T multiplex was used for DVB-H transmission, while the other half continued to provide commercial DVB-T programmes without any problems. This did cause the number of available TV programs to be restricted to four channels. BMCO provided network infrastructure integration and broadcast service provisioning. GPRS was used as the return channel (Sattler, 2005). Vodafone retained customer ownership and partially outsourced application service provision to Contcast. The development of the Berlin City Guide, ring tones downloading, and Get the Clip applications entailed cooperation between content aggregators and application developers.

Figure 6. Paris pilot cooperation model



During the commercialisation phase, it is expected that MNOs such as Vodafone will apply for a license to broadcast DVB-H. They will need to cooperate with an actor that has a media license, which is the local government's responsibility. While in Germany the spectrum license is handed out by the Bundesnetzagentur on a national level, the media licenses have to be applied for locally, namely each of the 15 regulators of the 16 federal states.

Already in the cooperation model used during the trial, and different from the Finnish model, there is a split between the BNO (T-Systems, the current DVB-T network owner) and the broadcast service provider (a consortium including Vodafone), which can be explained by the fact that two competing network operators are involved as network owner and network user, respectively. Figure 6 illustrates the cooperation model adopted during the Paris trial.

Until now, four multicasting pilots have been conducted in France, of which one was based on the DAB-IP standard, and three others based on the DVB-H standard. While the pilot under review here combined Towercast, Nokia, SFR, and Canal+Group, the other pilots used Sagem phones (instead of Nokia), and had TDF as BNO (instead of Towercast). The other pilots also involved more MNO partners (Orange, SFR, and Bouygues Telecom), and more content aggregators (TF1, TPS France Television, Radio France, and RTL, among others).

During the pilot described here, a DVB-H only pilot network was built and operated solely by Towercast. Also during the pilot, Nokia worked together with SFR for the development of the service platform. Canal+Group retained customer ownership.

During commercialisation, the DVB-H network is not expected to be built by the MNOs, but by TDF or Towercast, who are currently involved in constructing a DVB-T network. In France, no DVB-T operating licenses have been issued yet. Before analogue switch-off only one multiplex per French region is expected to be in place. In the commercialisation phase, it is expected that Canal+Group will come to market with an offering that will also target devices that only have mobile TV functionalities, without an integrated mobile phone. According to the interviewed executives, MNOs that come to market with a DVB-H brand will always choose a service that targets DVB-H + mobile phone terminal devices and will not introduce a second line of terminal devices that do not have integrated mobile telephony.

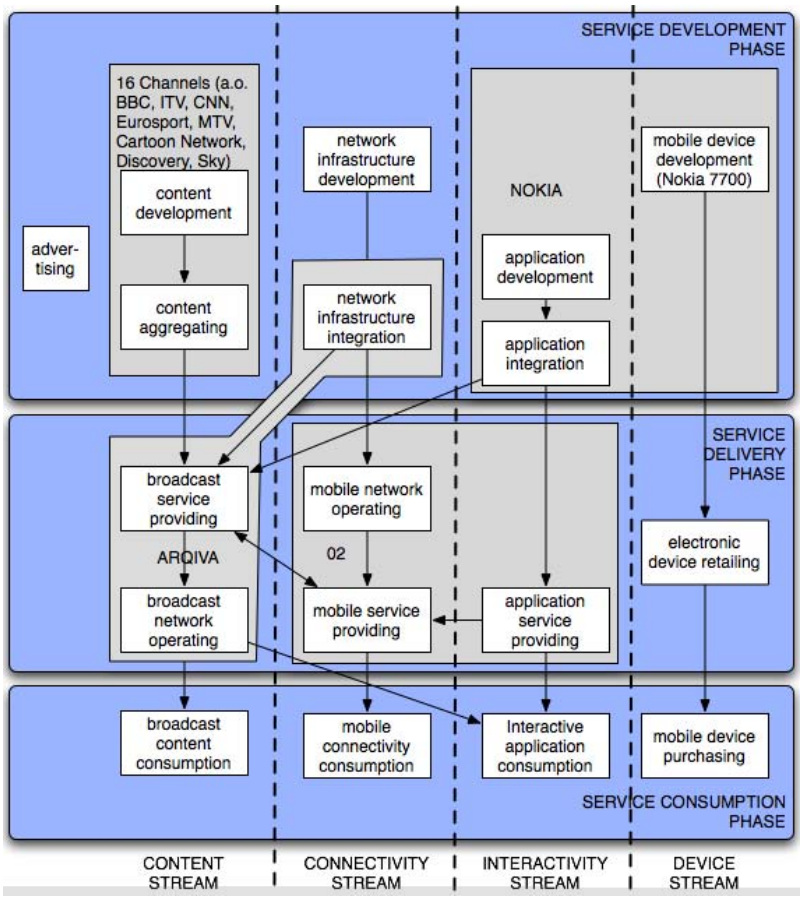
Again, it is interesting to see that the cooperation model in the pilot, in which broadcaster Canal+Group took an important role and combined the roles of content aggregation and broadcast service providing, will be mirrored in the commercialisation phase, in the sense that the broadcaster will probably come up with specific offers towards end customers, and thus will be competing with MNOs for the mobile TV market. Figure 7 illustrates the cooperation model adopted during the Oxford trial.

A DVB-H-only pilot network was built by Arqiva, while O2's network bandwidth was used as mobile telephony channel. Arqiva is also the company developing and implementing the digital terrestrial network for BBC, one of the digital terrestrial license holders in the UK.

During the pilot, all actors involved focused on their core competence. Nokia delivered the service platform software to Arqiva and mobile terminals to O2. The content suppliers provided content for the DVB-H broadcast, but did not interfere with the broadcast service provisioning side. The MNO O2 focused on mobile network operating, application service provisioning, and mobile service provisioning. The BNO Arqiva did have a more expanded role when compared to the Berlin or Paris pilot, in that it simultaneously performs the roles of broadcast provisioning and network operating. In the next pilot, Arqiva is restricted to broadcast network operation.

The Oxford pilot cooperation model is similar to the Helsinki pilot, in the sense that Arqiva

Figure 7. Oxford pilot cooperation model

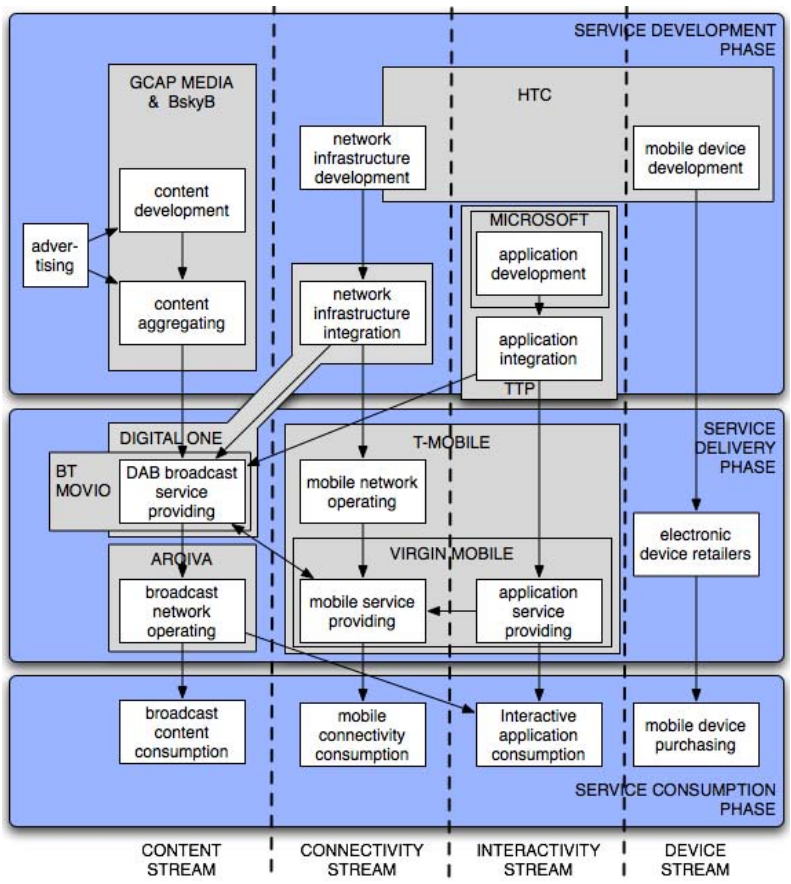


provided and operated the platform on which an MNO offered its services. The difference with the Helsinki trial is that no competing MNOs were involved in the Oxford pilot. There is little information on the commercialisation trajectory envisaged, but it may be expected that the MNO will want to retain full customer ownership. Figure 8 illustrates the cooperation model adopted during the London DAB-IP trial.

During the pilot, BT Movio, which leases 20% (later 30%) of the spectrum capacity on Digital One's DAB network, led the network build-out effort and acted as middleman between the content aggregators and the MNOs. The radio and TV content were delivered over the DAB-IP transmission network built by Arqiva, while T-Mobile's network was used as the return channel for additional interactivity and mobile telephony. T-Mobile offered Virgin Mobile, mobile network access, allowing Virgin Mobile to adopt the role of MVNO.

During commercialisation, BT Movio will probably act as a middleman between the owner of the broadcast network (which will be GCap

Figure 8. BT Movio London pilot cooperation model



Media, see next section on financial design) and the content aggregators, by offering a bundled wholesale package of broadcast access plus bundled content to the MNOs.

The pilot cooperation model already demonstrated the BT Movio wholesale model by leasing capacity from existing DAB network owner and granting access to MNOs and MVNOs under specific terms, as will be continued in the commercialisation phase.

Technology Design

The first technology design choices to be made concern the end-user devices and the network standards adopted during the pilots.

Three out of four of the DVB-H pilots built a DVB-H only pilot network, while in Berlin the network was built on top of DVB-T, which had an effect on the amount of TV content channels that could be used in Berlin. The Nokia 7710 was the dominant device during the four DVB-H

pilots. Only the Berlin pilot also conducted tests with another device, the Philips HoTMAN2. The London pilot used a smart phone developed jointly by BT, TTP, and HTC, and adopted the DAB-IP network standard.

Most users considered the Nokia 7710 as too big and unpractical. The Nokia N92 that is now arriving in the market is user friendlier, according to interviewed executives.

In most cases it is currently unknown what specific network technology choices will be made for the commercialisation phase. For instance, most pilots opted for DVB-H stand-alone configurations and many interviewees stressed the need for as many channels to be available as possible. On the other hand, the fact that in most cases DVB-T or DAB network owners were actively involved in the pilots seems to suggest that the mobile broadcast networks will be combined as much as possible with existing infrastructure.

Another design decision is whether the interactive applications that were developed during the pilots will have a direct effect on the degree of end-user interactivity in the service design (see section 5.2). Limited, "red button" interactivity was present in most pilots. The Paris pilot, in which broadcaster Canal+Group was the main actor, stands out as the only pilot where no extended forms of interactivity were developed, and where the focus remained primarily on downstream content delivery. The partners involved in the Berlin pilot developed the most interactive applications, that is, one-button voting for music videos, push of cinema trailers (users could then book cinema tickets directly through an interactive application), and the download of ring tones (a list of ring tones corresponding to currently played songs was broadcasted. Ring tones could be downloaded via the cellular network).

Concerning content protection, Berlin and Oxford did not implement content protection, considering the pilot context a controlled situation. While Helsinki adopted Nokia's content protection solution, Paris opted for the alternative conditional access through a SIM-card solution. Within the DAB-IP pilot, the industry participants opted for Windows Digital Rights Management solution. Interviewees stated that the choice of standards to be adopted during the commercialisation phase is an issue that will have to be cleared with the content owners beforehand. All observers agree that the choice of DRM is highly strategic and closely connected to the organisational design. If there is no lock-in by a SIM card, this might prove more interesting for actors who do not have a vested interest in mobile networks. Operators that do not have investments in a mobile network will not opt for a SIM-based solution, since this ties the customer to a MNO. Technical arguments can be given on why the SIM-based solution is still suitable in an age of convergence. SIM is a proven solution that provides a high level of security and reliability. But with the advent of converged services, using one service to tie the customer to a series of other services is experiencing pressures from non-MNO players.

Concerning the ESG, within the DVB-H field there are two camps with regard to the ESG standards (Yoshida, 2006). Nokia implemented its own version of the Open Mobile Alliance's OMA-BCAST specification on its DVB-H handsets, in a move against proponents of digital video broadcast-convergence of broadcast and mobile services (DVM-CBMS). Because of this rift, the two camps are promoting different ESGs. The Finnish and German Pilot adopted Nokia's non-standardised ESG, while in Paris the competing solution was chosen. Interviewed executives, even from pilots where the Nokia ESG was used, expressed doubts whether the Nokia non-standardised solution will be broadly adopted. In the DAB-IP pilot BT Movio developed a proprietary ESG.

Service Design

Given the aforementioned organisational and technological design decisions, the business actors then proceeded to offer the test users access

to a variety of service packages. The service bundles that were offered mostly consisted of TV content, but in some cases content was enriched with interactive applications.

The Helsinki pilot ran from March until July 2005. Five hundred test users received a basic package, which consisted of seven "free to air" television channels and three radio stations. In addition to the basic package users could subscribe to a supplementary package of seven premium service television channels. For some special events, like the Formula 1 Grand Prix in San Marino and Monaco, users had the possibility to buy one day's access in a pay-per-view model for $.50 a day.

The Berlin pilot started in July 2004 and took 8 weeks, during which the 20 test users had access to four television channels, one interactive channel and an interactive city guide of Berlin. The four television channels, with exception of the news channel, concentrated on the entertainment potential of mobile television.

In Paris, the pilot was conducted from September 2005 until June 2006, with 250 users. Access was provided to 10 television channels, four radio stations, and one channel that offered short programs to watch on mobile television (SFR TV). The user had also the possibility to subscribe individually to three additional channels or to choose the entire package of the three channels (Canal+, Sport+, and CineCinema Premier). Furthermore the user could watch additional content through a pay-per-view model.

In the Oxford pilot, which started in June 2005 and ran for 6 months, 400 users were offered 16 television channels among which 12 free-to-air channels, three pay-TV channels and one made-for-mobile channel: ShortsTV, a channel which offers short programmes developed for mobile television.

In the London pilot, BT Movio and Virgin Mobile let 1,000 users test their mobile digital TV service in the region of London, inside the M25 highway area. The users of this pilot were able to access three television channels and 52 radio stations.

It was already discussed shortly whether forms of interactivity were offered to the end user. Usage of an EPG is considered as the most basic form of interactivity. Additional types of interactivity were found in Berlin, Oxford and London.

In the Finnish pilot the focus was on TV content delivery, and interactivity was limited to the use of the EPG and some on-demand downloads (Sandell, 2005). In Berlin users were able to consult movie trailers and book cinema tickets for Berlin movie theatres through the What's Up application. The Vivaplus application was an interactive music channel where users could vote for music clips. Finally, users could download ring tones that were delivered over the cellular network. In Paris, besides the interactions with the EPG some on-demand downloads were possible. In Oxford users were able to record short content to their mobile device. In the London trial, the degree of interactivity was not very high. Although the user could use the red button functionality, a proportion of them were afraid to use it because of lack of good communication towards the user, concerning the price of each interaction.

The section on financial design will detail the customer ownership models in the pilots. It can already be stated here that most interviewees agreed that cooperation between business actors on the service design level is necessary in order to offer the consumer an integrated package.[3] It would be too confusing for consumers if they have to buy access from a separate firm, and their content from another firm. The consumer will expect that the purchase of network access will come together with a reasonable amount of basic content. Therefore, the MNOs (supposing most do retain full customer ownership) will have to negotiate content deals in order to be able to offer attractive packages.

The youth market (ages 18-35) is considered as the most important market segment for mobile TV (Page, Watt, & Menon, 2005), so it is expected that

content aggregators geared towards the youngest demographic segment such as music TV stations, will have important bargaining power as part of entering the bouquet. Jason Hirschhorn, MTV's chief digital officer, has stated that MTV would enter into discussions with operators over an advertising-based business model (Best, 2006).

While the chances are slim that telecom operators will massively start developing or commissioning the development of content on their own, it is very probable that the content aggregators will also develop interactive applications or services alongside their TV content. But as a rule it is foreseen that existing content aggregators will provide most of the video content, while MNOs will mainly develop the interactive services. Therefore, cooperations between the MNOs and the broadcast service platform will have to be guaranteed for interactive applications that can run over DVB-H.

Concerning time-shifting services, design choices between near-video on demand (near-VOD), VOD, and PVR have to be made. MNOs appear to favour VOD most, since this offers an opportunity to utilise their 3G property. Near-VOD, where content is downloaded at an earlier point in time (e.g., overnight) ranks second, but PVR creates tensions with the content aggregators. In reality, it is feasible that a hybrid solution will be implemented to circumvent the content industry's doubts about copyright protection. In this solution, content can be downloaded at an earlier time, but a one-time activation over a mobile network is then necessary to unlock the downloaded content. Content providers do not prefer lock-in by a SIM-based solution, since this hampers the amount of platforms they can offer their content on. Ideally, only one encryption scheme is used across several platforms.

Financial Design

In this final section, the financial design decisions taken during the pilots and the possible repercussions on the commercial financial design are described. Three design criteria are considered: (1) the sharing of the infrastructural cost, (2) the pricing of the product bundle offered to the end users, and (3) the revenue sharing arrangements (if any) among the different partners involved. It needs to be noted that during the pilots little to no revenues were generated, except in the Helsinki pilot and the London DAB-IP pilot. Therefore both revenue and cost-sharing agreements were rare to nonexistent during the pilots. Contacts with executives from the different pilots did however offer some insight into the financial arrangements that could arise during commercialisation. If not mentioned otherwise, the pilot costs incurred were carried by each individual pilot participant individually.

Concerning who will bear the cost of network roll-out during commercialisation, the discussions are still ongoing within most pilot consortia. In most countries there seems to be a movement away from cooperative models that were considered in a number of pilots, where various partners jointly funded the DVB-H roll-out, towards a wholesale model, where a single entity deploys and funds the roll-out of the network, and then gets the right to resell it. Some variation can arise on who will be able to resell access to the platform to interested parties. The wholesalers are not necessarily the actors with existing DVB-T or DAB networks (see BT Movio). In Germany, where T-Systems is currently building out the DVB-T network, the situation is still unclear on who will apply for the service licenses, and who will fulfil the role of reseller.

Only during the Finnish, French, and the London BT Movio pilot the end users were billed (and only some in the French case). No specific total amounts were made publicly available of the revenues collected, though. The pricing plans of the partners does show that all the MNOs are planning on retaining customer ownership. Only in France Canal+Group is counteracting this logic by also planning to include a mobile

digital TV subscription in its pay-TV bundle. It is expected that some pay-TV business actors in other European countries (e.g., Sky in the UK) will follow this example.

Concerning customer ownership, most MNOs are aiming to leverage their intimate customer relations during the DVB-H roll-out. But in each country any actor that has a customer base such as pay-TV broadcasters (France) or MVNOs (UK) may wish to be able to extend their service offering towards their customers.

Given the informal character of the pilots, no revenue-sharing arrangements were negotiated. Interviewed executives did express numerous hypotheses on what they thought would probably happen during the commercialisation phase. However, it was clearly expressed that the revenue split would be primarily between the end user service provider ("the customer owner") and the broadcast service provider ("the spectrum owner"). Although only a limited amount of data were available on the financial design from the pilots and the subsequent commercial roll-outs. However, at least some information could be obtained on the commercialisation phase. It appears that the license holder (or the party that leases spectrum from the license holder) on the one hand, and the customer owner (the actor that sells the subscription/service to the end customers) on other hand will divide the lion's share of revenues between them. Some anxiety among the content aggregators that occupy neither of these roles about revenue-sharing agreements that might be suboptimal for them was already reported in a few pilots.

CONCLUSION

This chapter aimed to provide a detailed and systematic analysis of the issues for bringing mobile broadcasting to market, and of the solutions found in five major pilots throughout Europe. We focused on organisation design, and in particular on the cooperation models employed by various stakeholders during the pilot phase.

Various cooperation models were found, specifically regarding the roles of broadcast service provider and BNO. It became clear that pilot cooperation models in this sense already foreshadow envisaged business models in commercialisation phase. However, some significant differences were found and reported in this chapter, particularly regarding the (types of) actors assuming the new and important role of broadcast service provider, and on the (types of) actors assuming customer ownership.

In Helsinki and Oxford, the role of BNO and the role of broadcast service provisioning were performed by one actor. In the other cases different actors performed these two roles. This split of responsibilities is expected to persist in many cases during commercialisation, with BNOs, content aggregators, MNOs, and intermediaries all taking an interest in broadcast service provisioning and/or spectrum ownership.

The business responsibility of customer ownership includes customer acquisition, end-user billing (including handling bad debt), and customer care (help desk support). The BNOs did not retain customer ownership in any pilot, nor did the broadcast service providers or the content aggregators (such as broadcasters), except in France where a content aggregator (pay-TV channel owner Canal+Group) retained customer ownership alongside one MNO (SFR). This scenario is expected to repeat itself during French commercialisation.

Concerning technical design, a majority of European pilots has chosen the DVB-H standard, although a sizeable minority opts for DAB-IP/DMB. It is not entirely clear to what extent the new mobile technologies will be tied in with the existing DVB-T and DAB networks, but given the cost advantages and the involvement of current DVB-T and DAB network owners, coupling, at least to a certain degree, seems probable in most countries. It also became clear that organisational

arrangements (e.g., whether MNOs or broadcasters respectively took the lead in the project) affected technical design decisions such as the selected standard for the ESG, or the degree of interactive applications.

In terms of service design, the pilots experimented with a wide variety of product bundles, subscription and pay-per-view schemes, and usually—but not necessarily—including some forms of interactivity. The combination of basic packages with premium packages was widespread throughout the pilots.

Regarding network roll-out, there seems to be a move away from the cooperative models (i.e., various partners jointly funding the roll-out) that were at one time considered in various pilots, towards a wholesale model, where a single entity deploys and funds the entire roll-out and then resells access to various service providers. These wholesalers are often actors currently owning or building out DVB-T or DAB networks, although not necessarily so, as the BT Movio example demonstrates. In Germany, where telco T-Systems is currently building out the DVB-T network, the situation is still unclear. In most cases, the MNOs seem to limit their involvement in the infrastructural effort to opening their network for the placement of repeaters necessary for full coverage.

Most of the MNOs aimed to leverage their current customer base into the mobile broadcasting arena. But other types of actors with an established customer base such as pay-TV broadcasters (France) or MVNOs (UK) also showed interest in acquiring access to the platform, or acquiring a service license, in order to offer mobile TV services directly to customers themselves.

Finally, from the information available, it appears that the spectrum owner (i.e., the license holder or the actor that leases spectrum from the license holder) on the one hand, and the customer owner (the actor that sells and guarantees the service to the end customer) on the other hand will divide the lion's share of revenues among them. Several content aggregators that neither own spectrum nor have direct customer ownership have expressed fears that revenue share agreements may turn out to be suboptimal for them. The evidence gathered here suggests that these fears may well materialise.

ACKNOWLEDGMENT

This article is based on results from the MADUF project (IBBT project 0052), which is funded by the IBBT (Interdisciplinary Institute for BroadBand Technology) of Flanders, Belgium, as well as by various partner companies. The authors gratefully acknowledge their interview partners for the information provided on various mobile TV pilots, as well as Dr. Jo Pierson, Katrien Dreessen (both IBBT-SMIT), and the other MADUF partners involved for their insightful comments and suggestions.

REFERENCES

Ballon, P. (2007, August). Business modelling revisited: The configuration of control and value. *The Journal of Policy, Regulation and Strategy for Telecommunications, Information and Media*.

Ballon, P., Pierson, J., & Delaere, S. (2005, September 4-6). *Open innovation platforms for broadband services: Benchmarking European practices.* Paper presented at ITS (International Telecommunications Society) 16th European Regional Conference, Porto, Portugal.

Ballon, P., Pierson, J., & Delaere, S. (2007) Fostering Innovation in Networked Communications: Test and Experimentation Platforms for Broadband Systems. In S. Heilesen & S. S. Jensen (Eds.) *Designing for Networked Communications: Strategies and Development.* Hershey: Idea Group Publishing, pp. 137-167.

Barney, J. B. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99-120.

Barney, J. B. (1997). *Gaining and sustaining competitive advantage*. Reading, MA: Addison-Wesley.

Best, J. (2006, February 16). Is free the way forward for mobile TV? *Silicon.com*. Retrieved from http://networks.silicon.com/mobile/0,39024665,39156508,00.htm

Braet, O., Ballon, P., & Dreessen, K. (2006). *Cooperation models for DVB-H rollout.* Final Report for IBBT-project MADUF ("Maximize DVB-H Usage in Flanders").

Curwen, P. (2006). Mobile television. *Communications & Strategies, 62,* 183-195.

Digitag (2005) *Television on a handheld receiver: Broadcasting with DVB-H.* Geneva: DigitAG :The Digital Terrestrial Television Action Group.

Dittrich, K., & Van den Ende, J. (2006, June 18-20). *Organizational forms for the development of new broadband services: A dynamic model for the degree of integration between collaborating firms*. Paper presented at the DRUID Summer Conference 2006, Copenhagen, Denmark.

Faber, E., Ballon, P., Bouwman, H., Haaker, T., Rietkerk, O., & Steen, M. (2003, June 9-11). *Designing business models for mobile ICT services.* Positioning paper for workshop on concepts, metrics & visualization. In *Proceedings of the Bled E-commerce conference*, Bled, Slovenia.

Faria, G., Henriksson, J., Stare, E., & Talmola, P. (2006). DVB-H: Digital broadcast services to handheld devices. *Proceedings of the IEEE, 94*(1), 194-209.

Haaker, T., Bouwman, H., & Faber, E. (2004). Customer and network value of mobile services: Balancing requirements and strategic interests. In R. Agarwal, L. Kirsch, & J. I. DeGross (Eds.), *Proceedings of the 25th international conference on Information systems (ICIS 2004)* (pp. 1-14).

Kivirinta, T., Ali-Vehmas, T., Mutanen, T., Tuominen, T., & Vuorinen, M. (2004). *Forecasting market demand for mobile broadcast services in Finland* (Rep. No. 51530C). Finland: Helsinki University of Technology.

Page, M., Watt, M., & Menon, N. (2005). *Mobinet 2005—Raising the stakes.* Retrieved from http://www.atkearney.com/main.taf?p=5,3,1,121,1

Pieck, R. (2005, September 14). *DVB-H broadcast to mobile devices.* Retrieved from http://www.newtec.be/fileadmin/webfolder/whitepaper/DVB-H_White_Paper.pdf

Pilz, K. (2005). *TV goes mobile with DVB-H—Swisscom's approach developing a market entry scenario with DVB-H based products.* Retrieved from http://www.ipdc-forum.org/resources/documents/6-Swisscom.pdf

Sandell, L. (2005). *FinnishMobileTV: Analysis on logfile data, April-June2005.* Retrieved from www.mobiletv.nokia.com/download_counter.php?file=/pilots/finland/files/Finnpanel_press_all_channels.pdf

Sattler, C. (2005, November 8). *BMCO newsletter.* Retrieved from http://www.bmco-forum.org/

Shin, D. H. (2006). Prospectus of mobile TV: Another bubble or killer application? *Telematics and Informatics, 23,* 253-270.

Skiöld, D. (2006). An economic analysis of DAB and DVB-H. *EBU Technical Review.* Retrieved from http://www.ebu.ch/en/technical/trev/trev_305-skiold.pdf

Södergard, C. (Ed.). (2003). *Mobile television: Technology and user experiences: Report on the Mobile TV project* (VTT publications 506).

Weck, C., & Wilson, E. (2006, January). Broadcasting to handhelds: An overview of systems and

services. *EBU Technological Review.* Retrieved from http://www.ebu.ch/en/technical/trev/trev_305-wilson.pdf

Yoshida, J. (2006, March 2). Protocol spat threatens to fragment DVB-H market. *EE Times.* Retrieved from http://www.eetimes.com/news/latest/business/showArticle.jhtml?articleID=181500546

ENDNOTES

1 For a conceptualization and overview of European TEPs, see Ballon, Pierson, and Delaere (2007).

2 Detailed references are in the original research report Braet, Ballon, and Dreessen (2006).

3 It was not within the scope of this chapter to assess the way the new service was experienced by the end users. Most pilots reported favourably on the way their service was received by the end users. Nevertheless, some critical results can be quoted. Users were not happy with some of the first generation devices such as the Nokia 7710, which were considered too heavy and clumsy. Also, users complained when there were not enough channels available, a result recorded during BT Movio's DAB-IP pilot (three TV channels) and the German pilot (four TV channels). Finally, the BT Movio pilot proved that users were reluctant to use interactive applications if the pricing model was unclear.

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, and J. Day, pp. 150-173, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.34
The MP3 Player as a Mobile Digital Music Collection Portal

David Beer
University of York, UK

INTRODUCTION

MP3 players are often described as *music collections in our pockets* or the *pocket jukebox*. Indeed, it would seem that MP3 players have significantly transformed music collections, music collecting practices, and contemporary understandings of the music collection. The MP3 player may be used to store, retrieve, and reproduce digital music files, and, therefore, it can be described as a portal—if we define the term portal as an entrance, doorway, or gateway—into these simulated (Baudrillard, 1983) mobile music collections. It is an interface between the human body and archives of digitally compressed music. This can perhaps be understood as constituting a kind of *musical cyborg*, a cybernetic organism, a hybrid of human and machine (Haraway, 1991). The MP3 player, in this hybridised sense, is a gateway into the digital, virtual, or simulated (Baudrillard, 1983) material cultural realm of music, a mobilised cyber-collection. The question then is what becomes of the music collection and the music collector when music shifts from the objectified disc and spool to the digital compression format and MP3 player portal? And, what are the social and cultural implications of the MP3 player portal's increasing pervasiveness and embeddedness in the flows of everyday life? The purpose of this article is to briefly introduce and discuss these questions alongside some of the technical details of the MP3 player. This article aims to use the material and technical details and definitions of the MP3 player to open up a range of possible questions that may be pursued in future research in this area. I will begin by defining the MP3 and the MP3 player.

BACKGROUND: MP3

The MP3 player, such as those manufactured by Sony, Creative, and Apple, can perhaps best be understood as a music retrieval interface that provides a portal for its appropriator to access an archive of digitally stored music files. These

may be selected and reproduced or illuminating the increasingly inert user, the device may select the tracks on behalf of the listener. An example of this is the *Shuffle* function on the Apple iPod (see next). This extension of the random play function of the compact disk (CD) player can perhaps be offered as an example of the increasing intelligence of the machine and the increasing inertia of the appropriator (Gane, 2005; Kittler, 1999).

According to Duncan and Fox (2005):

One of the oldest—and probably best known—compression/decompression formats (codecs) is MP3. It is popular with users for its near-CD quality and relative high speed of encoding and decoding. It is less popular with the music industry because it lacks controls to prevent copying. (Duncan et al., 2005, p. 9)

MP3, an abbreviation of *Motion Picture Experts Group One Audio Layer Three*, originated in 1991 as a system for broadcasting media files. MP3 is a file compression format that has the capacity to reduce music files to around one-twelfth of their original size (Mewton, 2001, p. 25), thus making the transfer across the Internet far more rapid and the space required to store the music much smaller. However, and contrary to the utopian rhetoric of the information or digital age, these are not perfect reproductions. The process of compression removes elements from music files so as to reduce them in size effectively; this leads to some of the subtleties of the music being removed. This then is a somewhat alternative vision to the perfect and infinite reproducibility that digitalisation has come to represent.

The MP3 format can be understood to have mobilised the music collection by compressing it, or miniaturizing it (Haraway, 1991), to fit into these pocket sized retrieval and reproduction devices.

THE MP3 PLAYER

The MP3 player, then, is a device that may be networked with the Internet (usually) through a connection with a computer, provided that the relevant software is installed upon it. A CD containing the required software usually comes with a newly purchased MP3 player. This connection made via the USB (Universal Serial Bus), USB2, or Firewire port or connector on the back of the computer enables music files stored on the computer's hard drive or accessed directly through the Internet to be downloaded onto the MP3 player where they are stored. The MP3 player then enables the appropriator to retrieve their music and reproduce the music file, often through headphones, although a variety of technologies are now available through which MP3 players may be docked (amplifying the music through speakers around open spaces).

MP3 players vary somewhat in size but, to give an idea of dimensions, are usually somewhere between the size of a box of matches and a pack of playing cards (more exact dimensions are included in the following discussion of the iPod). However, contrary to the image this suggests, the MP3 player is not a discrete, standardised, or self-contained device that takes on a single form or design. The current trend is for the combination of MP3 players with other technologies to create hybrid devices, the most significant of which is the combination of MP3 and mobile telephone technologies. This creates always-already networked MP3 players that may access networked archives of music files and therefore, exceed the storage capabilities of an isolated MP3 player and the collecting practices of its owner. Recently, highlighting their dynamic form, MP3 players have also been hybridised with camcorders, sunglasses, and even confectionary packaging to create novelty devices.

MP3 players are highly mobile portal technologies upon which anything between around 120

and 15,000 songs may be stored, dependent on the device. The music collection is then entirely mobile and may be comfortably carried around; weight is bypassed as an inhibiting problematic. It is now a common site in the street to see people interfacing with MP3 players and other mobile music devices (mobile CD, tape, and MiniDisk players). Indeed the scale of use and the details of the practices of these cyborgs (Haraway, 1991) may well represent one of the biggest challenges facing studies of contemporary music collecting practices. This is not to mention the implications that these devices have for the human body and the everyday spaces, which they populate (Bull, 2000, Thibaud, 2003). Before developing these future research questions, and to crystallize the material dimensions of the MP3 player, I will first focus briefly on a specific example of the MP3 player, the Apple iPod.

THE iPOD

The Apple iPod (see www.apple.com) has come to dominate the emerging MP3 player market. Due to a series of high profile advertising campaigns and innumerable editorial pieces, it has obtained a high international profile. Possibly the most interesting of these advertising campaigns came in 2003. This incorporated a two-page advert, which juxtaposed images of what had become the conventional record collection, records, tapes, and CD on the left hand page, and the image of the iPod on the right hand page. This attempt to redefine or "recraft" (Haraway, 1991) the music collection had some success, although it is not clear what part, or to what extent, this advertising campaign had in this shift in musical consciousness. Yet from purely anecdotal evidence, and the sales figures available for the iPod, it appears that music collecting practices have indeed shifted to momentarily rely on the outdated dualism from the actual or physical to the virtual and non-physical.

We now find the iPod dominates contemporary music discourse; the non-capitalised "i" prefix appears frequently in media discourse to evoke the downloading phenomenon and issues related to it. Furthermore, the descendant term Podcasting (Crofts, Dilley, Fox, Retsema, & William, 2005) is now becoming increasingly widely used to describe a practice of downloading pockets of music from the Internet onto the hard drive of computers and MP3 players. A practice that numerous companies such as British Telecom and the BBC (Radio 4) are buying into, as well as musician community sites such as www.garageband.com, in addition to the vast numbers of private podcasters.

In terms of its form, there are now five distinct models of iPod on the market, these are the original iPod, the iPod Mini, the iPod Shuffle, the iPod Nano, and the new iPod with video screen. Although the iPod Mini has now been discontinued to be replaced, it seems, by the iPod Nano. These iPod's come in various sizes and have the capability to hold various numbers of songs. To highlight this, and to give some sense of scale, I will look at the iPod, with the largest memory, and the iPod Shuffle, with the smallest memory.

The new video screen iPod, which has replaced the original iPod, is available (at the time of writing) in two forms or models; these are the 30GB memory model, which holds up to 7,500 songs, weighs 136g, and measures 103.5 x 61.8 x 11mm, or the 60GB memory model, which holds up to 15,000 songs, weighs 157g, and measures 103.5 x 61.8 x 14mm. The iPod Shuffle, the smallest of the iPods, also comes in two forms, a 512MB memory model, which holds up to 120 songs, and weighs 22g, or the 1GB memory model, which holds up to 240 songs, and weighs 22g (www.apple.com/uk).

These iPod's, despite the fact that they have come to be described as an MP3 player, in fact, like the connected iTunes Internet site (www.itunes.com), use the advanced audio coding

(AAC) format. MP3 is one of a number of digital compression formats; there are innumerable other similar formats that are available such as AAC, WMA, some of which are encrypted like liquid audio for example, yet it is the dominance of the MP3 that has caused it to become the representative label for an entire series of music compression technologies.

RECONTEXTUALISATIONS AND SIMULATIONS

To return to the broader question of the implications of the MP3 player, we find that the collection is recontextualised in two senses. First, it has moved from discs to digital files. Second, it has moved the collection on mass from private domestic spaces to public spaces—thereby extending the work of the personal stereo or car stereo by providing instant access to entire music collections rather than being restricted to a tape, CD, or MINIDisk's worth.

In light of these recontextualisations, the iPod and other similar digital technologies have created the possibility for a reconsideration of the music collection. And as such, along with other digital technologies, have generated a vast series of questions around ownership and the way in which we approach material cultural artefacts. The spaces taken up by racks, boxes, stands, rooms, shelves, piles, holders, wallets, sleeves units, and record bags have been transposed onto the hard-drive. The digital music file collection takes up space on a hard drive, a kind of virtual space.

On the issue of collecting, Walter Benjamin has suggested that:

One has only to watch a collector handle the objects in his glass case. As he holds them in his hands, he seems to be seeing through them into their distant past as though inspired. (Benjamin, 1999, p. 62)

If we cannot hold and feel these collections, admire them, have them populate the spaces of our everyday lives, or present them as a concretised representation of aspects of identity, what are the consequences (Sterne, 2003)? What becomes of Benjamin's book collector and the experiences of collecting when music collections are no longer rows or piles of discs or tapes but are merely lists of artists and songs on a screen, a collection that cannot be held in the hand, touched, and smelt. Indeed the MP3 music collection never grows in a physical sense (used here in a conventional form). Rather it is a kind of simulated (Baudrillard, 1983) music collection, a collection in hyperspace, or perhaps, a hyperreal (Baudrillard, 1983) music collection that is neither real nor illusion, virtual nor actual, but rather it moves freely between these interlocked spheres, or to use Haraway's terminology, this music collection, as it is reproduced from the virtual music file into actual material sounds that reverberate around the spaces and organisms, or as it is "burnt" or inscribed from the MP3 file onto a CD, permeates the boundary between these dualisms (Haraway, 1991). This then opens up vast sets of complex and problematic questions concerning the understanding of music. One consequence of this recontextualising and redefinition of the music collection is the recent explosion in music theft in the form of music file sharing, which has lead to a number of ongoing legal battles. It would seem that the MP3 file has far exceeded the music theft possibilities of bootlegging, piracy, and shoplifting. Perhaps the removal of the object form, the physical disk, or spool, has radically transformed the notion of ownership and has created the possibility for large-scale music theft. This again is a question that requires further examination as the numerous legal conflicts ensue and conclude. These questions concern the issues of ownership and theft in the digital age, and the related issues of copyright, security, access, and encryption.

FUTURE ISSUES

The important issue from the point of view of this brief exploratory article is what future issues relating to the MP3 player require examination. These future research questions can perhaps be understood to fall into three interrelated categories: *mobility*, *ownership*, and *collecting*. The central question that informs these three categories is that of transformation and the implications of the MP3 player. These perceived transformations require rigorous empirical examination in the form of close-up analyses of the MP3 player in praxis (Beer, 2005a), the MP3 player in the mundane flows of everyday life (Beer, 2005b), in short, studies of the MP3 in/and the "richness of the ordinary" (Sandywell, 2004). To obtain even a tentative notion of transformation these studies must be historically (Sandywell, forthcoming) and culturally embedded.

Existing approaches in this area present a number of opportunities for extended study. Take for example, the empirically grounded approaches to music and music technologies in everyday life found in the work of Bull (2000, 2004), DeNora (2000, 2003), and Shuker (2004), the theoretically informed radical posthumanism of Kittler (1999) (Gane, 2005), the historically and culturally embedded descriptions of Sterne (2003), or, even, the critical or dialectical materialist approach to music technologies of Adorno (2002a, 2002b, 2002c, 2002d). We also find now an emerging and varied (practical, instructional, legal, and analytical) body of literature on music and the Internet (see for example Beer, 2005c; Jones, 2000; Mewton, 2001; Waugh, 1998), which, over the coming years, as the implications of networked communications technologies and music production and reproduction proliferate, is certain to escalate rapidly.

It is perhaps now time to consider the MP3 player as a deeply embedded everyday technology around which individualised yet networked everyday practices are structured and defined. This then requires a system of analysis that accesses these everyday practices and uncovers the complex appropriations of MP3 technologies within the broader context of the digital or information age. This is the challenge for a sociology or social psychology of music technologies, or a technologically focused cultural studies, as the MP3 player portal mobilises, re-contextualises, and networks the digital music collection.

REFERENCES

Adorno, T. W. (2002a). The radio symphony: An experiment in theory. In R. Leppert (Ed.), *Essays on music* (pp.251-269). California: University of California Press.

Adorno, T. W. (2002b). The curves of the needle. In R. Leppert (Ed.), *Essays on music* (pp.271-276). California: University of California Press.

Adorno, T. W. (2002c). Opera and the long-playing record. In R. Leppert (Ed.), *Essays on music* (pp.283-286). California: University of California Press.

Adorno, T. W. (2002d). The form of the phonograph record. In R. Leppert (Ed.), *Essays on music* (pp.277-282). California: University of California Press.

Baudrillard, J. (1983). *Simulations*. New York: Semiotext[e].

Beer, D. (2005a). Reflecting on the digit(al)isation of music. *First Monday, 10*(2). Retrieved from http://www.firstmonday.org/issues/issue10_2/beer/index.html

Beer, D. (2005b). Sooner or later we will melt together: Framing the digital in the everyday. *First Monday, 10*(8). Retrieved from http://www.firstmonday.org/issues/issue10_8/beer/index.html

Beer, D. (2005c) Music and the Internet, Special Issue No.1. *First Monday, 10*(7). Retrieved from http://firstmonday.org/issues/special10_7/

Benjamin, W. (1999). Unpacking my library. In H. Arendt (Ed.), *Illuminations* (pp. 61-69). London: Pimlico.

Bull, M. (2000). *Sounding out the city: Personal stereos and the management of everyday life*. Oxford: Berg.

Bull, M. (2004). Automobility and the power of sound. *Theory, Culture, & Society, 21*(4/5), 243-259.

Crofts, S., Dilley, J., Fox, M., Retsema, A., & William, B. (2005). Podcasting: A new technology in search of viable business models. *First Monday, 10*(9). Retrieved from http://www.firstmonday.org/issues/issue10_9/crofts/index.html

Denora, T. (2000). *Music in everyday life*. Cambridge: Cambridge University Press.

Denora, T. (2003). *After Adorno: Rethinking music sociology*. Cambridge: Cambridge University Press.

Duncan, N. B., & Fox, M. A. (2005). Computer-aided music distribution: The future of selection, retrieval, and transmission. *First Monday, 10*(4). Retrieved from http://www.firstmonday.org/issues/issue10_4/duncan/index.html

Gane, N. (2005). Radical post-humanism: Friedrich Kittler and the primacy of technology. *Theory, Culture, & Society, 22*(3), 25-41.

Haraway, D. (1991). A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century. In *Simians, cyborgs, and women: The reinvention of nature* (pp. 149-181). London: Free Association Books.

Jones, S. (2000). Music and the Internet. *Popular Music, 19*(2), 217-230.

Kittler, F. A. (1999). *Gramophone, film, typewriter*. California: Stanford University Press.

Mewton, C. (2001). *All you need to know about music and the Internet revolution*. London: Sanctuary.

Sandywell, B. (2004). The myth of everyday life: Toward a heterology of the ordinary. *Cultural Studies. 18*(2/3), 160-180.

Sandywell, B. (Forthcoming) Monsters in cyberspace: Cyberphobia and cultural panic in the information age. *Information, Communication & Society*. (forthcoming, 2006)

Shuker, R. (2004). Beyond the "high fidelity" stereotype: Defining the (contemporary) record collector. *Popular Music, 23*(3), 311-330.

Sterne, J. (2003). *The audible past: Cultural origins of sound reproduction*. London: Duke University Press.

Thibaud, J. P. (2003). The sonic composition of the city. In M. Bull, & L. Back (Eds), *The auditory culture reader* (pp. 329-341). Oxford: Berg.

Waugh, I. (1998). Music on the Internet (and where to find it). Kent: PC Publishing.

KEY TERMS

CD: An abbreviation of compact disk. CD is a digital storage and reproduction technology commonly associated with music.

Compression Format: A technology (or software) for reducing the size of files to enable storage and transfer, some are encrypted some are not, for example MP3, AAC, and Liquid Audio.

Cyborg: A cybernetic-organism, a hybrid of human and machine, organic and inorganic. Most famously appropriated from cyberpunk literature in the social theory of Haraway and other socialist-feminist writers.

IPod: Perceived as the dominant "MP3 player" (also plays AAC format) on the market. A product of Apple (see www.apple.com).

Jukebox: A device through which selections of records may be chosen and played back, usually activated by the insertion of a coin and the depression of a series of numbered buttons corresponding to the demarcated number of the chosen record. These are predominantly found in public spaces such as bars, restaurants, cafeterias, and public houses.

MP3: A file compression format capable of reducing the size of music files to facilitate transfer and storage.

Music Collection: The practice of accumulating and storing objects on which music is inscribed. Such as vinyl records, tapes, CDs, MiniDisks, and, more recently, MP3 and other digital compression files.

Podcasting: This term is a combination of "iPod" and "Broadcasting." Podcasting is often described as musical blogging (Web Logging), by which selections of music may be accessed and downloaded in relation to chosen genres, types, and styles.

Posthuman: An emergent theory of technologies that places technologies at the forefront of the analysis. It is based centrally on the premise that technologies are increasingly intelligent and that human experience is centred around technological interfaces and interfacing. See for example the work of McLuhan, Haraway, Kittler, and Hayles.

Simulation: A concept of the French philosopher Jean Baudrillard that deals directly with the inseparability of the real and the non-real in the contemporary media age. See Jean Baudrillard's 1983 text *Simulations* (New York: Semiotext[e]).

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 637-641, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.35
Wireless Technologies for Mobile Computing and Commerce

David Wright
University of Ottawa, Canada

INTRODUCTION

At the time of writing (1Q06) most countries have a small number (2-6) of major cellular operators offering competing 2.5G and 3G cellular services. In addition, there is a much larger number of operators of WiFi networks. In some cases, a major cellular operator, for example, Deutsche Telekomm and British Telecom, also offers a WiFi service. In other cases, WiFi services are provided by a proliferation of smaller network operators, such as restaurants, laundromats, airports, railways, community associations and municipal governments. Many organizations offer WiFi free of charge as a hospitality service, for example, restaurants. Cellular services offer ubiquitous, low data rate communications for mobile computing and commerce, whereas WiFi offers higher data rates, but less ubiquitous coverage, with limitations on mobility due to business as opposed to technology reasons.

Emerging networks for mobile computing and commerce include WiMAX and WiMobile (Wright, 2006), which offer higher data rates, lower costs and city-wide coverage with handoff of calls among multiple base stations. These new technologies may be deployed by the organizations that currently deploy cellular and WiFi networks, and also may give rise to a new group of competitive wireless network operators.

This article identifies the capabilities needed for mobile computing and commerce and assesses their technology and business implications. It identifies developments in the wireless networks that can be used for mobile computing and commerce, together with the services that can be provided over such networks. It provides a business analysis indicating which network operators can profitably deploy new networks, and which network operators need to establish business and technology links with each other so as to better serve their customers. The resulting

range of next generation service, technologies and network operators available for mobile computing and commerce is identified.

WIRELESS NETWORK ARCHITECTURES

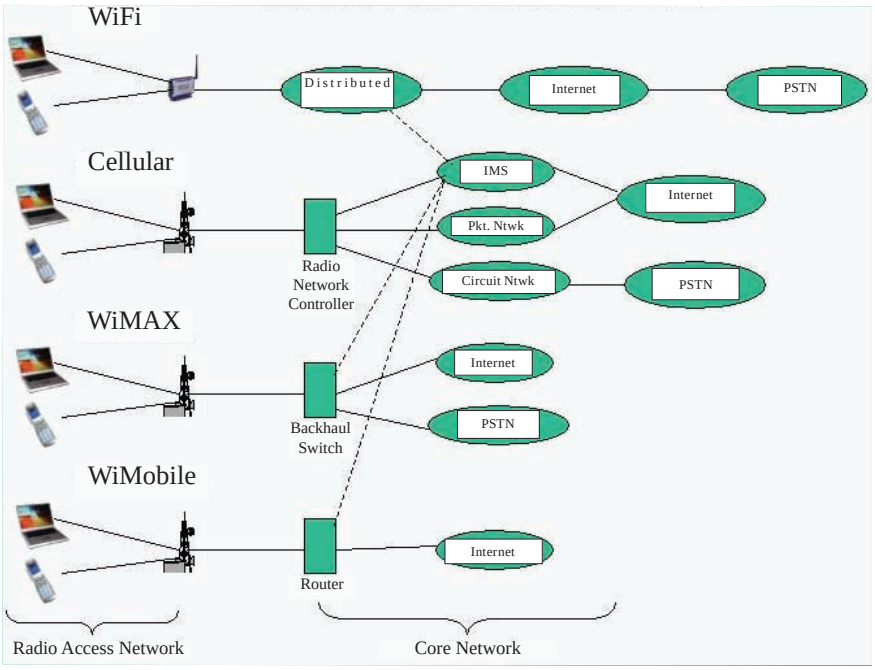
Figure 1 illustrates the network architectures for WiFi, Cellular, WiMAX and WiMobile, including the radio access network on the left and the wired core network on the right.

The cellular architecture is the most sophisticated in that the core network includes a circuit network (for legacy circuit switched voice calls), a packet network (for data calls) and an IP Multimedia Subsystem, IMS (for migration of all traffic onto the Internet).

These three networks essentially allow the cellular operator to maintain control over all calls to and from the mobile device, and hence derive revenue from them. In particular the IMS network contains servers for establishing voice and video calls over IP, authenticating users, maintaining records of the current location of a mobile user, accounting, and security. Cellular operators are migrating traffic from their circuit and packet networks onto the IMS.

By contrast, WiFi (IEEE, 1999a, 1999b, 1999c, 2003), WiMAX (IEEE, 2006; Ghosh et al., 2005), and WiMobile (IEEE, 2006; Lawton, 2005) are simply radio access technologies and do not specify a core network. They therefore allow more direct access from a mobile device to the Internet. In particular, the WiMobile specification, which is under development at the time of writing, emphasizes that its design is being optimized for operation with IP. This more open access to the Internet allows a mobile user to set up, for instance, a VoIP call using a third party

Figure 1. Wireless network architectures



service without the involvement of the wireless network operator. As the user moves from one access point to another, the call can be maintained using Mobile IP, involving servers maintained by the user's ISP, not by the wireless network operator. Mobile IP can operate over diverse wireless access technologies as described by Benzaid et al. (2004).

If the operator of a WiFi, WiMAX or WiMobile network wishes to maintain more control over the traffic passing through their network and hence participate more in the revenue generated by that traffic, they can build an IMS network. Alternatively if they already operate a cellular network, they can provide access to their existing IMS network, as shown by the dashed lines in Figure 1.

REQUIREMENTS FOR MOBILE COMPUTING AND COMMERCE

Any wireless transmitter/receiver has a limited range in order to comply with government regulations regarding maximum power output. A mobile user therefore may move out of the range of its current wireless access point, and it is necessary to handoff the communication to another access point using either the same or a different wireless technology. Handing off the communication means that the current IP session is maintained, for example, the user continues to browse a Web site as a registered user, a VoIP call is not interrupted, and an enterprise user with a laptop-based secure VPN to an enterprise network continues to use the same VPN. There are four requirements in order to achieve handoff suited to mobile computing and commerce:

1. It must be possible to switch the call from one access point to another
2. If the user is receiving quality of service, QoS, for example, a guaranteed low latency, that QoS is maintained after the handoff, and an acceptable number of packets are lost during handoff.
3. If the access points are operated by different network operators, there must be a business arrangement between them regarding mediation of the billing for the call.
4. The organization deploying the wireless access network must be able to make a profit or to have a business model that focuses on hospitality service.

Requirements 1 and 2 are technology related and are discussed next, followed by the business requirements 3 and 4.

TECHNOLOGY ISSUES

A mobile device that is capable of using multiple wireless access technologies, such as those described above, can continuously scan its radio environment to search for access points that it could potentially use. Some of them may not be available, if, for instance, they are operated by companies with which the user does not have a subscription. In order to choose among the available access points within range the mobile device can apply criteria including: data rate, cost, ability to handoff seamlessly, and QoS; delay (important for voice) and packet loss rate (important for data). For instance, a mobile device with an interactive voice/video call in progress could choose the lowest cost network that provides acceptable delay. A device downloading a large data file could choose the network with the highest data rate given limitations on cost and packet loss rate. Once the network is selected, handoff is initiated.

Handoff among WiFi, WiMAX and WiMobile is handled by IEEE (2006). Handoff between cellular and one of these three technologies is complicated by the need to interwork with the cellular circuit, packet and IMS networks.

- In the case of WiFi, this interworking is provided by a specification from the industry consortium UMA, Unlicensed Mobile Access (2006), which is incorporated as part of the GSM cellular network specifications, release 6.
- In the case of WiMAX, similar issues are involved and are being resolved by the WiMAX Forum (2006).
- WiMobile is at an early stage of development and interworking with cellular is not a priority at this stage. A specification may be developed later, or alternatively, WiMobile may differentiate itself from the other technologies by becoming a "native-IP" access mechanism, similar to DSL and cable modem in which customers have direct access to the Internet.

This discussion addresses requirements 1, 2 above. We now move on to requirements 3, 4.

BUSINESS ISSUES

This section presents business strategies for wireless access network operators that take into account sources of revenue related to mobile computing and commerce, plus the need to compete with other technologies and network operators. Earlier work in this area (du Preez & Pistorius, 2003) dates from a time when 3G and wireless data services were emerging technologies. The present section incorporates developments in technology and services to date.

The sources of revenue are given in Table 1 and are classified in two ways:

1. Whether the service is provided by a content provider or a network operator, which may be the wireless access network operator or another network operator. For instance, a VoIP service could be provided by the wireless operator or by a third party such

Table 1. Wireless access network operators revenue sources

Revenue Source	**Service provided by:** **(N) Network Operator** **(C) Content Provider**	**Revenue accrues to:** **(W) Wireless Network Operator** **(3) Third Party Service Provider** **(S) Shared with Content Provider**
Voice/video calls	N	W 3
Audio content	C	S
Video content	C	S
Gateway to PSTN	N	W 3
Geographic info (e.g., travel directions, highway safety)	N C	W S
Location enabled advertising	N C	W S
Location enabled buddy lists	N	W
Multimedia Messaging Service	C	W 3 S
Gaming	C	S
QoS	N	W
VPN	N	W 3

as Vonage. Either way it is provided by a network operator.

2. Who receives the revenue for the service: the wireless access network operator, a third party or a sharing arrangement with a content provider.

It can be seen from Table 1 that there is a large number of mobile computing and commerce services that can be provided by a mix of wireless network operators, content providers and third parties. In addition there are non-revenue generating services such as e-mail and Web browsing. A clear business strategy is needed to operate successfully in competition with the other players. Strategies suited to the different types of wireless network operators are given in Table 2.

Table 2 divides wireless access network operators into three groups: incumbent cellular operators, hospitality providers such as restaurants and municipalities, and new competitors, who are starting operations based on the availability of new technology. The incumbent cellular operators have complex core networks as shown in Figure 1 and incur costs of operating legacy technologies. They seek to deploy all possible wireless technologies in order to accommodate the needs of all customers. By contrast the new competitors seek to reduce their costs by only operating the most recent technologies. Both these groups are operating commercial services and therefore use licensed spectrum so that their customers do not experience interference from other users. The hospitality providers, however, are providing a free service. Their customers accept that the performance may vary according to the demands of other users and therefore the operators reduce their costs by using unlicensed spectrum.

Both the incumbents and the new competitors aim to deliver the full range of services listed in Table 1 to their customers, typically from the IMS, so as to maintain control over the revenue. The hospitality providers, however, are typically providing access only, allowing their customers to get services from any third party they wish, since they do not seek to generate revenue from

Table 2. Strategy for wireless access network operators

	Cellular Operators	**Hospitality providers**	**Competitive Wireless Network Operators**
Technologies	2.5G, 3G, WiFi, WiMax, WiMobile	WiFi, unlicensed WiMAX	WiMAX, WiMobile
Revenue sources	Generate revenue from the full range of services	Provide Internet access for the full range of services.	Generate revenue from the full range of services
IMS strategy	Lock customers into IMS-based services.	Establish partnerships and interfaces to the IMS of other operators	Build IMS. Establish partnerships and interfaces to the IMS of other operators
Competitive strategy	Buy up competitors.	Avoid competing with other operators by a competitive bid process.	Differentiate from incumbents by offering low cost services, focusing on IP, developing next generation services, for example, presence, location, QoS.

their networks. For location-based services, the hospitality provider can provide the third party with information about the customer's current location.

The cellular incumbents typically already have an IMS in place and aim to lock customers into service provided by that IMS. The new competitors need to build an IMS and then establish partnerships with other wireless operators so that calls originating on one IMS can be handed off to another operator. These partnerships are also important to the hospitality providers since they typically have no interest in developing their own IMS.

The competitive strategy of incumbent cellular operators towards WiFi operators historically has been to buy them up, and this strategy is also appropriate for WiMAX and WiMobile operators. The strategy of hospitality operators is to avoid competition, and this is particularly important for municipalities, who should not be seen to use tax dollars to compete against private industry. In order to avoid this perception, they can use a competitive bid process allowing any operator the opportunity to bid on the contract to build and operate their network. The strategy of the new competitors is to compete on three fronts. First, they can offer low cost services, since they do not have the cost of operating legacy networks. Second, they can offer a full range of next generation services, such as presence and location-based services, thus positioning themselves as state-of-the-art suppliers. Third, they can sell QoS guarantees to their customers, since new technologies such as WiMAX and WiMobile are particularly suited to providing such guarantees.

CONCLUSION

The enabling technologies for mobile computing and commerce are developing rapidly. New wireless technologies such as WiMAX and WiMobile offer extended coverage and improved QoS compared to WiFi; and higher data rates and lower costs compared to 2.5G and 3G cellular. A wide range of services is available over these technologies including services that generate revenue (a) for the wireless operator, such as location-based services, (b) for a third party, such as VoIP and (c) for a content provider, such as entertainment. Wireless network operators, including incumbent cellular operators, hospitality providers and new competitive wireless network operators, need to develop strategies that allow handoff of calls among the different technologies and operators. Strategies include locking customers into an IMS, interworking with other operators' IMSs, buying out competitors and developing a broad range of state-of-the-art services such as location and presence services.

The mobile computing and commerce user can therefore expect a proliferation of services (Table 1), a number of different network operators (Table 2), an array of different wireless technologies, WiFi, 3G, WiMAX and WiMobile, and a mobile device that can make the best choice among these alternatives at any point in time and space.

REFERENCES

Benzaid, M., Minet, P., Al Agha, Kh., Adjih, C., & Allard. G. (2004). Integration of mobile-IP for universal mobility. *Wireless Networks, 10*(4), 377-388.

du Preez, G. T., & Pistorius, C. W. I. (2003). Analyzing technological threats and opportunities in wireless data services. *Technological Forecasting and Social Change, 70*(1), 1-20.

Ghosh, A., Wolter, D. R., Andrews, J. G., & Chen, R. (2005, February). Broadband wireless access with WiMax/802.16: Current performance benchmarks and future potential. *IEEE Communications, 43*(2), 129-136.

IEEE. (1999a). *802.11 Wireless LAN: Medium access control (MAC) and physical layer (PHY) specifications*. New York: IEEE Publications.

IEEE. (1999b). *802.11a high-speed physical layer in the 5 GHz band*. New York: IEEE Publications.

IEEE. (1999c). *802.11b higher-speed physical layer (PHY) extension in the 2.4 GHz band*. New York: IEEE Publications.

IEEE. (2003). *802.11g further higher-speed physical layer extension in the 2.4 GHz band*. New York: IEEE Publications.

IEEE. (2006a). *802.16e air interface for fixed and mobile broadband wireless access systems: Amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands*. New York: IEEE Publications.

IEEE. (2006b). *802.20 mobile broadband wireless access* (In Progress). Retrieved March 2006, from http://grouper.ieee.org/groups/802/20

IEEE. (2006c). *802.21 media independent handover services* (In Progress). Retrieved March 2006, from http://grouper.ieee.org/groups/802/21/

Lawton, G. (2005). What lies ahead for cellular technology? *IEEE Computer, 38*(6), 14-17.

UMA. (2006). *Unlicensed mobile access*. Retrieved from http://www.umatechnology.org/specifications/index.htm

WiMAX Forum. (2006). Retrieved March 2006, from www.wimaxforum.org

Wright, D. (2006). Wireless technologies for mobile computing and commerce. In D. Taniar (Ed.), *Encyclopedia of mobile computing and commerce*. Hershey, PA: Idea Group Reference.

KEY TERMS

IMS, IP Multimedia Subsystem: Part of the wired core network containing servers for establishing voice and video calls over IP, authenticating users, maintaining records of the current location of a mobile user, accounting, and security.

Location-Based Services: Services that take into account the users current geographical location, for example, advertising locally available products and services, providing directions and alerting drivers to traffic congestion and road accidents.

Mobile IP: An Internet standard that allows a mobile user to move from one point of attachment of the network to another while maintaining an existing TCP/IP session. Incoming packets to the user are forwarded to a server in the user's new access IP subnetwork.

Presence: The ability of a user device to specify characteristics, such as whether the user is online, whether the user is willing to receive calls, whether the user is willing to receive calls of a given type (e.g., voice, video, data, MMS) from specified other users and what is the user's current location to a specified degree of accuracy.

Quality of Service (QoS): Features related to a communication, such as delay, variability of delay, bit error rate and packet loss rate. Additional parameters may also be included, for example, peak data rate, average data rate, percentage of time that the service is available, mean time to repair faults and how the customer is compensated if QoS guarantees are not met by a service provider.

WiFi: A commercial implementation of the IEEE 802.11 standard in which the equipment has been certified by the WiFi Alliance, an industry consortium.

WiMAX: A commercial implementation of the IEEE 802.16 standard in which the equipment has been certified by the WiMAX Forum, an industry consortium.

WiMobile: Another name for the IEEE 802.20 standard, which is in course of development at the time of writing (1Q06).

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 1038-1042,

Chapter 3.36
Mobile Handheld Devices for Mobile Commerce

Wen-Chen Hu
University of North Dakota, USA

Jyh-haw Yeh
Boise State University, USA

Hung-Jen Yang
National Kaohsiung Normal University, Taiwan

Chung-wei Lee
Auburn University, USA

INTRODUCTION

With the introduction of the World Wide Web, electronic commerce has revolutionized traditional commerce and boosted sales and exchanges of merchandise and information. Recently, the emergence of wireless and mobile networks has made possible the extension of electronic commerce to a new application and research area: mobile commerce (MC), which is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of mobile handheld devices. In just a few years, mobile commerce has emerged from nowhere to become the hottest new trend in business transactions. Despite a weak economy, the future of mobile commerce is bright according to the latest predictions (Juniper Research Ltd., 2004). Internet-enabled mobile handheld devices are one of the core components of a mobile commerce system, making it possible for mobile users to directly interact with mobile commerce applications. Much of a mobile user's first impression of the application will be formed by his or her interaction with the device, therefore the success of mobile commerce applications is greatly dependent on how easy they are to use. This article first explains the role of handheld devices in mobile commerce systems and then discusses the devices in detail. A mobile handheld device includes

six major components: (a) a mobile operating system (OS), (b) a mobile central processor unit (CPU), (c) a microbrowser, (d) input and output (I/O) devices, (e) memory, and (f) batteries. Each component is described, and technologies for the components are given.

BACKGROUND

Internet-enabled mobile handheld devices play a crucial role in mobile commerce as they are the devices with which mobile users interact directly with mobile commerce applications. This section first introduces a mobile commerce system and then illustrates how it is used to carry out a mobile transaction. A mobile commerce system is inherently interdisciplinary and could be implemented in various ways. Figure 1 shows the structure of a mobile commerce system and a typical example of such a system (Hu, Lee, & Yeh, 2004). The system structure includes six components: (a) mobile commerce applications, (b) mobile handheld devices, (c) mobile middleware, (d) wireless networks, (e) wired networks, and (f) host computers.

To explain how the mobile commerce components work together, Figure 2 shows a flowchart of how a user request is processed by the components in a mobile commerce system.

Figure 1. A mobile commerce system structure

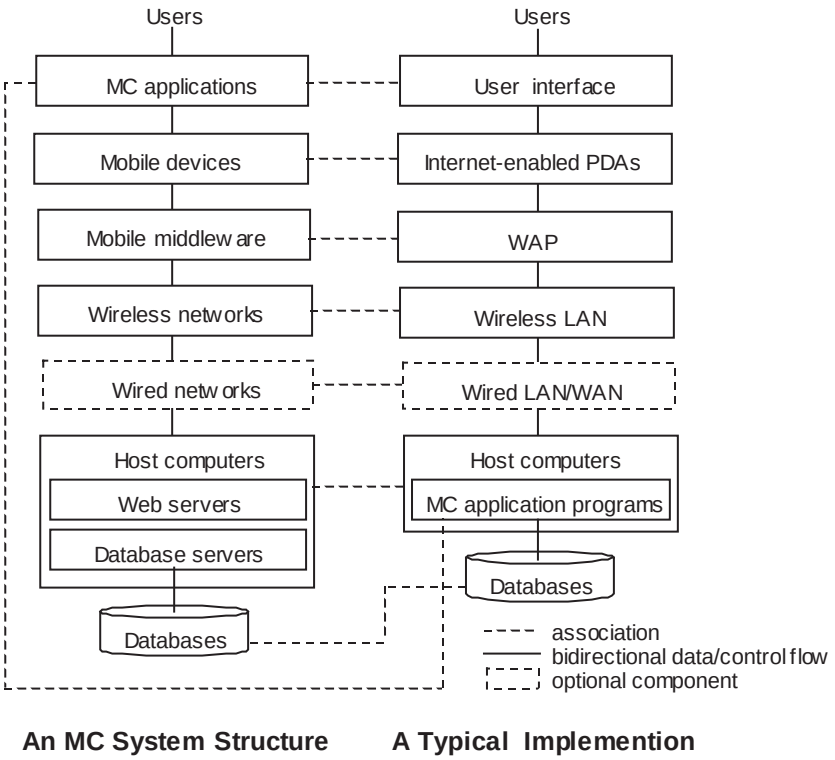
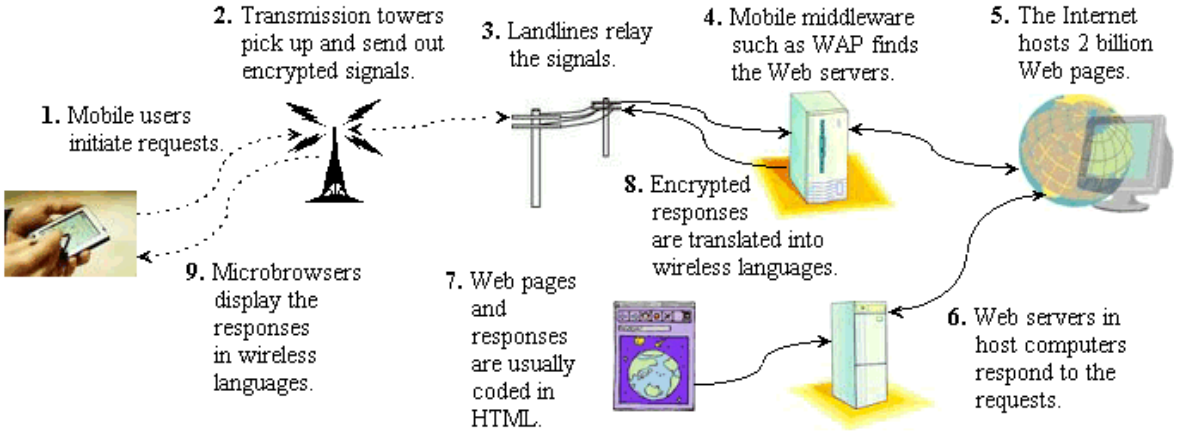


Figure 2. A flowchart of a user request processed in a mobile commerce system



MOBILE HANDHELD DEVICES

Mobile users interact with mobile commerce applications by using small wireless Internet-enabled devices, which come with several aliases such as handhelds, palms, PDAs (personal digital assistants), pocket PCs (personal computers), and smart phones. To avoid any ambiguity, a general term, mobile handheld devices, is used in this article. Mobile handheld devices are small general-purpose, programmable, battery-powered computers, but they are different from desktop PCs or notebooks due to the following special features.

- Mobility
- Low communication bandwidth
- Limited computing power and resources such as memory and batteries

Figure 3 shows a typical system structure for handheld devices, which includes the following six major components: (a) a mobile operating system, (b) a mobile central processing unit, (c) a microbrowser, (d) input and output devices, (e) memory, and (f) batteries. Brief descriptions of all the components are given in the coming sections.

Mobile Operating Systems

Simply adapting desktop operating systems for mobile handheld devices has proved to be a futile endeavor; an example of this effort is Microsoft Windows CE. A mobile operating system needs a new architecture and different features in order to provide adequate services for handheld devices. Several mobile operating systems are already available and each employs a different architecture and implementation. Figure 4 shows a generalized mobile operating system structure, which can be visualized as a six-layer stack.

Although a wide range of mobile handheld devices are available in the market, the operating

Figure 3. System structure of mobile handheld devices

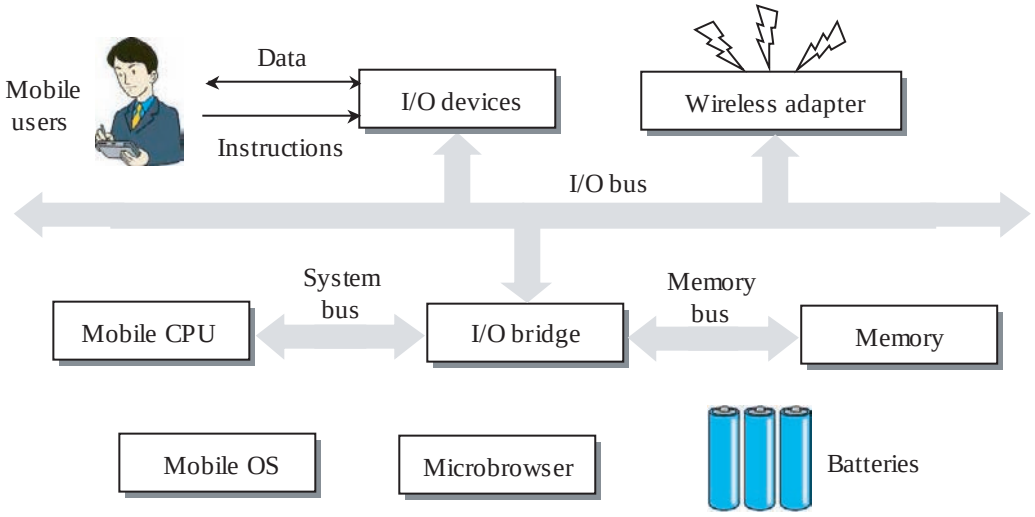
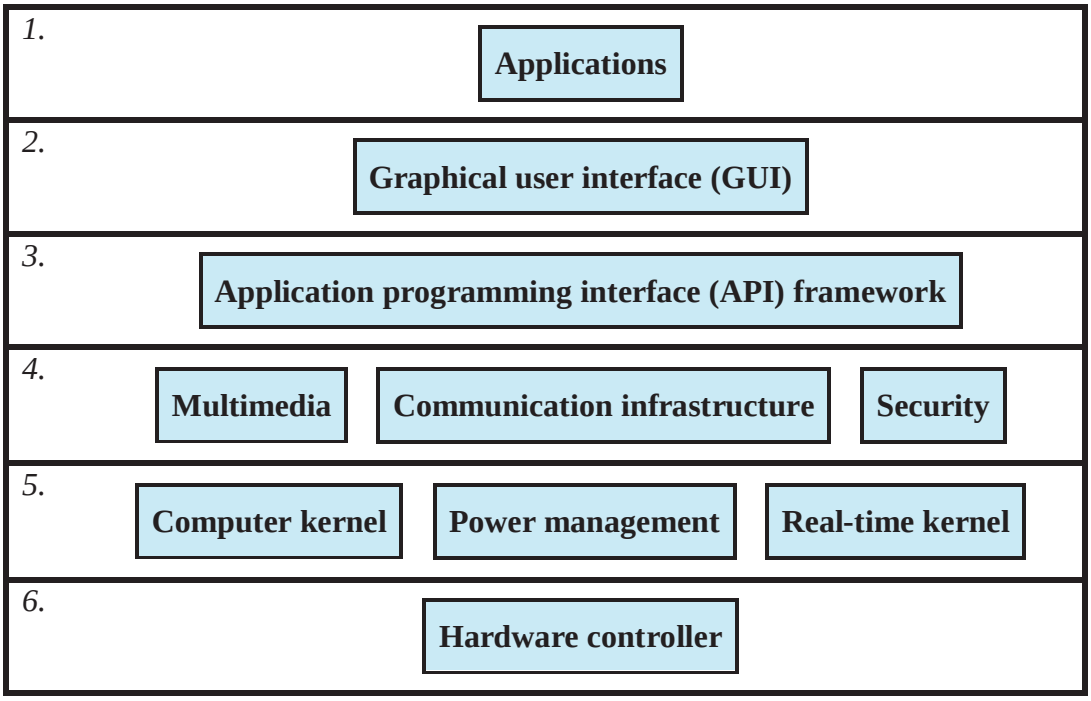


Figure 4. A generalized mobile operating system structure



systems, the hubs of the devices, are dominated by just three major organizations. The following two lists show the operating systems used in the top three brands of smart cellular phones and PDAs in descending order of market share.

- **Smart Cellular Phones:** Microsoft Smartphone 2002, Palm OS 5, and Symbian OS 7 (Vaughan-Nichols, 2003)
- **PDAs:** Palm OS 5, Microsoft Pocket PC 2002, and Symbian OS 7 ("Mobile Computing," 2003)

The market share is changing frequently, and claims concerning the share vary enormously. It is almost impossible to predict which will be the ultimate winner in the battle of mobile operating systems.

Mobile Central Processing Units

The core hardware in mobile handheld devices are the mobile processors, and the performance and functionality of the devices are largely dependent on the capabilities of the processors. There used to be several brands available, but recently mobile processors designed by ARM Ltd. have begun to dominate the market. Handheld devices are becoming more sophisticated and efficient everyday, and mobile users are demanding more functionality from the devices. For example, in 2002, *In-Stat/MDR* predicted that worldwide mobile Internet-access device unit shipments would increase from approximately 430 million that year to approximately 760 million in 2006 ("Demand Increasing," 2002). To achieve this advanced functionality, in addition to the obvious feature, low cost, today's mobile processors must have the following features.

- **High Performance:** The clock rate must be higher than the typical 30 MHz for Palm OS PDAs, 80 MHz for cellular phones, and 200 MHz for devices that run Microsoft's Pocket PC.
- **Low Power Consumption:** This prolongs battery life and prevents heat buildup in handheld devices that lack the space for fans or other cooling mechanisms.
- **Multimedia Capability:** Audio, image, and video applications are recurring themes in mobile commerce.
- **Real-time Capability:** This feature is particularly important for time-critical applications such as voice communication.

Microbrowsers

Microbrowsers are miniaturized versions of desktop browsers such as Netscape Navigator and Microsoft Internet Explorer. They provide graphical user interfaces that enable mobile users to interact with mobile commerce applications. Due to the limited resources of handheld devices, microbrowsers differ from traditional desktop browsers in the following ways.

- smaller windows
- smaller footprints
- fewer functions and multimedia features

Several microbrowsers, such as Microsoft Mobile Explorer and Wapaka Java Micro-Browser, are already available. America Online (AOL), Inc., the parent company of the Netscape Network, and Nokia are developing and marketing a Netscape-branded version of Nokia's WAP microbrowser, with AOL-enhanced features, for use across a wide variety of mobile handheld devices. Figure 5 shows a typical microbrowser, Mobile Browser version 7.0 from Openwave Systems, which includes the following features: compatibility with WAP (Open Mobile Alliance Ltd., n.d.) or i-mode (NTT DoCoMo, n.d.), multimedia support, color images and animation, and dual network-stack, HTTP (hypertext transfer protocol) and WSP, support (Openwave Systems Inc., n.d.).

Figure 5. Openwave® Mobile Browser version 7

Input and Output Devices

Various I/O devices have been adopted by mobile handheld devices. The only major output device is the screen, whereas there are several popular input devices, including the following:

- **Keyboards:** There are two kinds of keyboards: built-in keyboards and external, plug-in keyboards. The problem with the former is that they are too small for touch-typing, whereas the latter suffers from inconvenience. Fabric keyboards that can be rolled up or folded around the handheld devices are being developed to relieve the problem of external keyboards.
- **Touch Screens and Writing Areas with Styli:** A touch screen is a display that is sensitive to human touch, allowing a user to interact with the applications by touching pictures or words on the screen. A stylus is an input device used to write text or draw lines on a surface as input to a handheld device. A handheld device equipped with a writing area and a stylus needs a handwriting-recognition function, but existing systems do not yet have a satisfactory recognition rate. Graffiti, employed by many handheld devices, is the most popular writing software.

Some mobile handheld devices can also react to voice input by using voice-recognition technology.

Memory

Desktop PCs or notebooks usually have between 64 to 256 MB of memory available for users, whereas handheld devices typically have only 4 to 64 MB. PDAs normally have more storage space than smart cellular phones. The former commonly have 16 MB, and the latter may have a memory size as low as a few kilobytes. Three types of memory are usually employed by handheld devices.

- **Random Access Memory (RAM):** There are two basic types of RAM: dynamic RAM (DRAM) and static RAM (SRAM). Dynamic RAM, the more common type, needs to be refreshed thousands of times per second in order to hold data, whereas static RAM does not need to be refreshed, making it faster but also more expensive than dynamic RAM.
- **Read-Only Memory (ROM):** ROM is manufactured with fixed contents, and it is usually used to store the programs that boot the device and perform diagnostics. It is inherently nonvolatile storage, in contrast to RAM.
- **Flash Memory:** This is a kind of nonvolatile storage similar to EEPROM (electrically erasable, programmable read-only memory), but updating can only be done either in blocks or for the entire chip, making it easy to update. Flash memory is not as useful as random access memory because RAM can be addressable down to the byte (rather than the block) level.

It is expected that hard disks, which provide much more storage capacity, will be adopted by handheld devices in the near future. A comprehensive survey of storage options can be found in Scheible (2002).

Batteries

Rechargeable lithium ion batteries are the batteries most commonly used by handheld devices. The life of this kind of battery is short, generally only a few hours of operating time. Battery technology will not significantly improve unless and until manufacturers begin to switch to fuel cells, which is unlikely in the near future. A fuel cell operates like a battery, but unlike a battery, a fuel cell does not run down or require recharging and will continue to produce energy in the form of electricity and heat as long as fuel is supplied. Since the fuel cell relies on chemical energy rather than combustion, emissions would be much lower than emissions from the cleanest existing fuel-combustion processes.

Synchronization

Synchronization connects handheld devices to desktop computers, notebooks, and peripherals in order to transfer or synchronize data. The traditional method of synchronization uses serial cables to connect handheld devices and other computing equipment. Now, however, many handheld devices use either an infrared (IR) port or Bluetooth technology to send information to other devices without needing to use cables.

- IrDA Data, a standard formulated by the Infrared Data Association (n.d.) to ensure the quality and interoperability of infrared hardware, is designed for data transfer over distances of up to 1 meter, acting as a point-to-point cable replacement.
- Bluetooth wireless technology is a specification aiming at simplifying communications among handheld devices, printers, computers, and other devices based on short-range radio technology. The Bluetooth 1.1 specification (Bluetooth SIG, Inc., n.d.) consists of two documents: the core, which provides design specifications, and the profile, which provides interoperability guidelines.

FUTURE TRENDS

Mobile handheld devices are usually divided into two types: smart cellular phones and Internet-enabled PDAs. These two kinds of devices started out as very different products, but they have gradually blended into each other. In the near future, it will be difficult to tell the differ-

ence between these two types of devices. The newest products such as tablet PCs belong to the category of PDAs because both have similar functionality. There are numerous mobile devices available in the market today. Table 1 lists some major mobile-device specifications, although several table entries are incomplete as some of the information is classified as confidential due to business considerations.

From Table 1 and previous discussions, the future trends of mobile handheld device components are observed.

1. **Operating Systems:** There are several popular operating systems available; the big three are (a) Palm OS, (b) MS Pocket PC/Smartphone, and (c) Symbian OS. It is hard to tell the eventual winner at this moment.
2. **CPU:** The ARM processors (Cormie, 2002) have already dominated and will dominate the market.
3. **Microbrowsers:** Most HTML pages cannot be displayed on microbrowsers, which

Table 1. Specifications of some major mobile handheld devices

Vendor & Device	Operating System	Processor	Installed RAM/ROM	Input Methods	Key Features
Compaq iPAQ H3870	MS Pocket PC 2002	206 MHz Intel StrongARM 32-bit RISC	64 MB/32 MB	Touchscreen	Wireless email/ Internet
Handspring Treo 300	Palm OS 3.5.2H	33 MHz Motorola Dragonball VZ	16 MB/8 MB	Keyboard/ Stylus	CDMA network
Motorola Accompli 009	Wisdom OS 5.0	33 MHz Motorola Dragonball VZ	8 MB/4 MB	Keyboard	GPRS network
Nokia 9290 Communicator	Symbian OS	32-bit ARM9 RISC	16 MB/8 MB	Keyboard	WAP
Nokia 6800	Series 40			Keyboard	Innovative keyboard integration
Palm i705	Palm OS 4.1	33 MHz Motorola Dragonball VZ	8 MB/4 MB	Stylus	Wireless Email/ Internet
Samsung SPH-i330	Palm OS 4.1	66MHz Motorola Dragonball Super VZ	16 MB/8 MB	Touchscreen/ Stylus	Color screen
Sony Clie PEG-NR70V	Palm OS 4.1	66 MHz Motorola Dragonball Super VZ	16 MB/8 MB	Keyboard/ Stylus/ Touchscreen	Multimedia
Sony Ericsson T68i			800KB	Keyboard	Multimedia Messaging Service
Toshiba E740	MS Pocket PC 2002	400 MHz Intel PXA250	64 MB/32 MB	Stylus/ Touchscreen	Wireless Internet

will be gradually improved to adopt more HTML pages.

4. **Input Methods:** The two major input methods are and will be touch screens and styli, and keyboards.
5. **Memory:** 64 MB or even 128 MB memory for a handheld device will be common.
6. **Batteries:** Fuel cells are likely the most promising method for extending battery life. However, they will not be available in the near future.

CONCLUSION

The emerging wireless and mobile networks have extended electronic commerce to another research and application area: mobile commerce. Internet-enabled mobile handheld devices are one of the core components of mobile commerce systems as they are needed for mobile users to directly interact with mobile commerce applications. Understanding the devices and knowing their functions and capabilities is vital for the success of mobile commerce applications. A handheld device relies on a wide range of disciplines and technologies for its success. To facilitate understanding, this article broke down the functions of a handheld device into six major components, which can be summarized as follows.

1. **Mobile Operating Systems:** Simply adapting desktop operating systems for handheld devices has proved to be futile. A mobile operating system needs a completely new architecture and different features to provide adequate services for handheld devices. A generalized mobile operating system structure can be visualized as a six-layer stack: (a) applications, (b) a GUI, (c) an API framework, (d) multimedia, a communication infrastructure, and security, (e) a computer kernel, power management, and a real-time kernel, and (f) a hardware controller.
2. **Mobile Central Processing Units:** Handheld devices are becoming more sophisticated and efficient everyday, and mobile users are demanding more functionality from their devices. To achieve this advanced functionality, in addition to the obvious feature, low cost, today's mobile processors must have the following features: (a) high performance, (b) low power consumption, (c) multimedia capability, and (d) real-time capability. The cores and architectures designed by Cambridge-based ARM Holdings Ltd. have begun to dominate the mobile CPU market.
3. **Microbrowsers:** Microbrowsers are miniaturized versions of desktop browsers such as Netscape Navigator and Microsoft Internet Explorer. They provide graphical user interfaces that allow mobile users to interact with mobile commerce applications. Microbrowsers usually use one of the following four approaches to return results to the mobile user: (a) wireless language direct access, (b) HTML direct access, (c) HTML to wireless-language conversion, and (d) error.
4. **Input and Output Devices:** Various I/O devices have been adopted by mobile handheld devices. The only major output device is the screen, but there are several popular input devices; among them are (a) keyboards and (b) touch screens and writing areas that need styli.
5. **Memory:** Three types of memory are usually employed by handheld devices: (a) RAM, (b) ROM, and (c) flash memory. Hard disks, which provide much more storage capacity, are likely to be adopted by handheld devices in the near future.
6. **Batteries:** At present, rechargeable lithium ion batteries are the most common batteries used by handheld devices. However, the life of this kind of battery is short, and the technology will not significantly improve

unless and until manufacturers begin to switch to fuel cells, which may not happen for at least several years.

Synchronization connects handheld devices to desktop computers, notebooks, or peripherals to transfer or synchronize data. Without needing serial cables, many handheld devices now use either an infrared port or Bluetooth technology to send information to other devices.

REFERENCES

Bluetooth SIG, Inc. (n.d.). *Bluetooth specifications.* Retrieved August 12, 2004, from https://www.bluetooth.org/foundry/specification/document/specification

Cormie, D. (2002). *The ARM11 microarchitecture.* Retrieved July 21, 2004, from http://www.arm.com/support/59XGYS/$File/ARM11+Microarchitecture+White+ Paper.pdf

Demand increasing for mobile Internet access devices: Handsets represent primary growth driver. (2002). *In-Stat/MDR*. Retrieved July 8, 2004, from http://www. instat.com/press.asp?ID=250&sku=IN020280MD

Hu, W., Lee, C., & Yeh, J. (2003). Mobile commerce systems. In N. Shi (Ed.), *Mobile commerce applications* (pp. 1-23). Hershey: Idea Group Publishing.

Infrared Data Association. (n.d.). *Technical summary of "IrDA DATA" and "IrDA CONTROL."* Retrieved July 15, 2004, from http://www.irda.org/standards/standards.asp

Juniper Research Ltd. (2004). *Mobile commerce & micropayment strategies.* Retrieved February 3, 2004, from http://www.juniperresearch.com/reports/17_ MCommerce/main.htm

Microsoft Corp. (2003a). *Pocket PC*. Retrieved June 25, 2003, from http://www.microsoft.com/windowsmobile/products/pocketpc/default.mspx

Microsoft Corp. (2003b). *Smartphone*. Retrieved June 23, 2003, from http://www.microsoft.com/windowsmobile/products/smartphone/default.mspx

Mobile computing. (n.d.). *PCTechGuide*. Retrieved July 2, 2004, from http://www.pctechguide.com/25mobile.htm

NTT DoCoMo. (n.d.). *i-mode*. Retrieved June 28, 2004, from http://www.nttdocomo.com/

Open Mobile Alliance Ltd. (n.d.). *WAP (wireless application protocol).* Retrieved July 21, 2004, from http://www.openmobilealliance.org/tech/affiliates/wap/wap index.html

Openwave Systems Inc. (n.d.). *Mobile Browser V7.* Retrieved July 15, 2004, from http://www.openwave.com/products/device_products/phone_tools/mobile_ browser_7.html

Palm Source, Inc. (n.d.). *Palm OS*. Retrieved December 22, 2003, from http://www.palmsource.com/palmos/

Scheible, J. P. (2002). A survey of storage options. *IEEE Computer, 35*(12), 42-46.

Vaughan-Nichols, S. J. (2003). OSs battle in the smart-phone market. *IEEE Computer, 36*(6), 10-12.

KEY TERMS

Electronic Commerce: It is the exchange or buying and selling of commodities, services, or information, or the transfer of funds on the Internet through the use of desktop computers.

Flash Memory: This is a kind of nonvolatile storage similar to EEPROM, but updating can only be done either in blocks or for the entire chip, making it easy to update.

Microbrowsers: They are miniaturized versions of desktop browsers such as Netscape Navigator and Internet Explorer. Microbrowsers, due to the limited resources of handheld devices, are different from the traditional desktop browsers in the following features: (a) smaller windows, (b) smaller footprints, and (c) less functions and multimedia features.

Mobile Commerce: It is the exchange or buying and selling of commodities, services, or information, or the transfer of funds on the Internet (wired or wireless) through the use of Internet-enabled mobile handheld devices.

Mobile Handheld Device: It is a small general-purpose, programmable, battery-powered computer that can be held in one hand by a mobile user. It is different from a desktop or notebook computer due to the following features: (a) mobility, (b) low communication bandwidth, and (c) limited computing power and resources such as memory and batteries. There are two major kinds of handheld devices: (a) smart cellular phones and (b) PDAs.

Stylus: A stylus is an input device used to write text or draw lines on a surface as input to a handheld device.

Synchronization: Synchronization connects handheld devices to desktop computers, notebooks, and peripherals in order to transfer or synchronize data. Other than using serial cables to connect handheld devices and other computing equipment, many handheld devices use either an infrared port or Bluetooth technology to send information to other devices.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 792-798, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.37
Mobile Commerce Multimedia Messaging Peer

Kin Choong Yow
Nanyang Technological University, Singapore

Nitin Mittal
Nokia Pte Ltd, Singapore

INTRODUCTION

In a mobile-commerce world, shops could provide product brochures, cards, sounds, songs and so forth in the form of multimedia messaging presentations, which could be used by a customer to send to friends. Shopping malls will have information kiosks equipped with wireless access capabilities, and could perform searches across the mall's network to update its multimedia message repository. Customers can download and distribute to their friends such multimedia content via mobile messaging, leading to increased revenue for the shops.

Over the years, mobile messaging has become an essential means of communication, and it is going to be even more so with the merging of the Internet and Mobile Networks. The ability to message from a phone to a computer on the Internet and vice versa is making messaging a powerful means of communication (Yeo, Hui, Soon, & Lau, 2001).

This article discusses the development of a multimedia messaging client for a personal digital assistant (PDA) and a Kiosk providing multimedia messages composition, search, share and send capabilities. Various messaging technologies, enabling wireless technologies and the peer-to-peer model, are also discussed and evaluated in this article. We substantiate the ideas discussed in this article with a description of an MMS PDA client application using JXTA with specific references to a shopping mall scenario.

BACKGROUND

Short Messaging Service

Text messaging uses the short messaging service (SMS, 100-200 characters in length), and involves

sending text messages between phones. Examples include "C U L8ER" and "OK. AT FLAT OR OFFICE." It is quick and dirty, hard to use the keypad, abrupt, punctuation challenged and incredibly useful and popular. Text messaging also has a lot of advantages, such as its convenience, availability on all phones and discreteness.

Text messaging is most prevalent in the youth market (Tan, Hui, & Lau, 2001), and especially teenagers, who are able to manipulate the difficulty of entering text with the mobile phone keypad. In fact, it is suspected that this steep learning curve and the necessary insider knowledge are two of the things that appeal to the youngsters (Bennett & Weill, 1997).

Multimedia Messaging Service

The multimedia messaging service (MMS), as its name suggests, is the ability to send and receive messages comprising of a combination of text, sounds, images and video to MMS-capable handsets (MMS Architecture, 2002). The trends for the growth in MMS are taking place at all levels within GSM (Patel & Gaffney, 1997), enabling technologies such as GPRS, EDGE, 3G, Bluetooth and Wireless Access Protocol (WAP).

MMS, according to the 3GPP standards, is "a new service, which has no direct equivalent in the previous ETSI/GSM world or in the fixed network world." Here is an introduction to the features of this innovative new service:

- MMS is a service environment that allows different kinds of services to be offered, especially those that can exploit different media, multimedia and multiple media.
- MMS will enable messages to be sent and received using lots of different media, including text, images, audio and video.
- As more advanced media become available, more content-rich applications and services can be offered using the MMS service environment without any changes.

Table 1. SMS vs. MMS

Feature	**SMS**	**MMS**
Store and Forward (non real time)	Yes	Yes
Confirmation of message delivery	Yes	Yes
Communications Type	Person to person	Person to person
Media Suppored	Text plus binary	Multiple- Text, voice, video
Delivery mechanism	Signalling channel	Data traffic channel
Protocols	SMS specific e.g. SMPP	General Internet e.g. MIME SMTP
Platforms	SMS Center	MMS Relay plus others
Applications	Simple person to person	Still images

- The MMS introduces new messaging platforms to mobile networks in order to enable MMS. These platforms are the MMS Relay, MMS Server, MMS User Databases and new WAP Gateways.
- MMS will require not only new network infrastructure but also new MMS-compliant terminals. MMS will not be compatible with old terminals, which means that before it can be widely used, MMS terminals must reach a certain penetration.

Implications of SMS on MMS

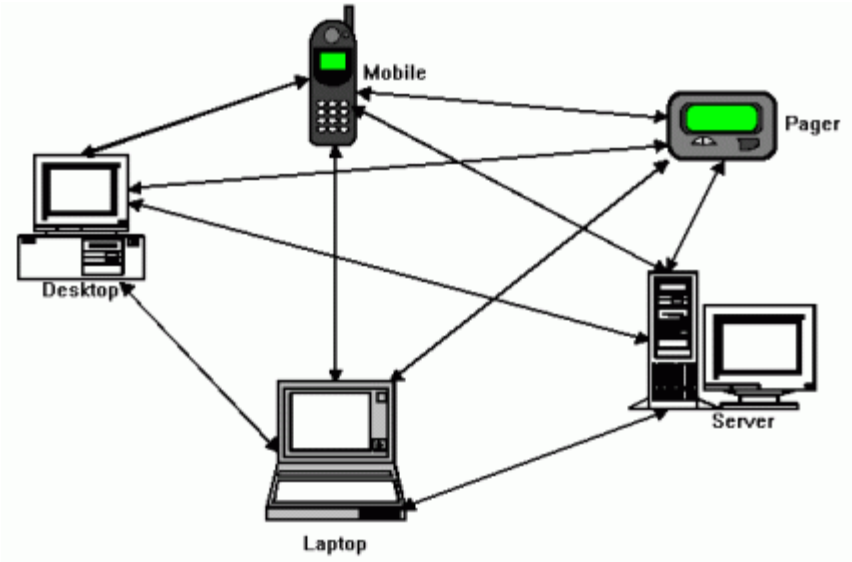
The current SMS has some unique advantages that other non-voice services do not have, such as store and forward and confirmation of message delivery. However, SMS also has some disadvantages, such as limited message length, inflexible message addressing structures and signalling channel slowness.

Person-to-Person (P2P) Model

Today, the most common distributed computing model is the client-server model (Chambers, Duce, & Jones, 1984). In the client-server architecture, clients request services and servers provide those services. A variety of servers exist in today's Internet: Web servers, mail servers, FTP servers and so forth. The client-server architecture is an example of a centralized architecture, where the whole network depends on central points to provide services. Regardless of the number of clients, the network can exist only if a server exists (Berson, 1992).

Like the client-server architecture, P2P is also a distributed computing model (Yemini, 1987). However, the P2P architecture is a decentralized architecture where neither client nor server status exists in a network (Madron, 1993). Every entity in the network, referred to as a peer, has equal status, meaning that an entity can either request a

Figure 1. P2P model



service (a client trait) or provide a service (a server trait). Figure 1 illustrates a P2P network.

Though peers all have equal status in the network, they do not all necessarily have equal physical capabilities. A P2P network might consist of peers with varying capabilities, from mobile devices to mainframes (Budiarto & Masahiko, 2002). A mobile peer might not be able to act as a server due to its intrinsic limitations, even though the network does not restrict it in any way.

JXTA

Jxta strives to provide a base P2P infrastructure over which other P2P applications can be built (Project Jxta, 2002). This base consists of a set of protocols that are language independent, platform independent and network agnostic. These protocols address the bare necessities for building generic P2P applications (Jxta Technology Overview, 2002). Designed to be simple with low overheads, the protocol's target is to build, to quote the Jxta vision statement, "every device with a digital heartbeat."

JXTA vs. .NET and JINI

Jxta's XML-based messaging is similar to Microsoft's .Net and SOAP technologies. But that is a very thin foundation for comparison. As more and more third-party protocols make use of XML, it will become obvious that just using XML as a message format says nothing at all about an actual networking technology. Comparing the high-level, policy-rich, Web-services-based infrastructure that is .Net to the low-level, fundamental, policy-neutral nature of Jxta is a futile exercise.

Project Jxta and the Jini project are also fundamentally different. Both of them have some similarity in higher-level interaction, enabling true distributed computing over a network. However, the similarity abruptly ends there. Strategic differences between the two are: Jxta started life as a completely interoperable technology (any platform, any programming language, any vendor). Sun is a mere contributor in the community. Jini is a Java-centered technology that Sun will integrate and deploy strategically in future product offerings. Sun will maintain a degree of control over Jini's evolution.

MULTIMEDIA MESSAGING PEER FOR MOBILE COMMERCE

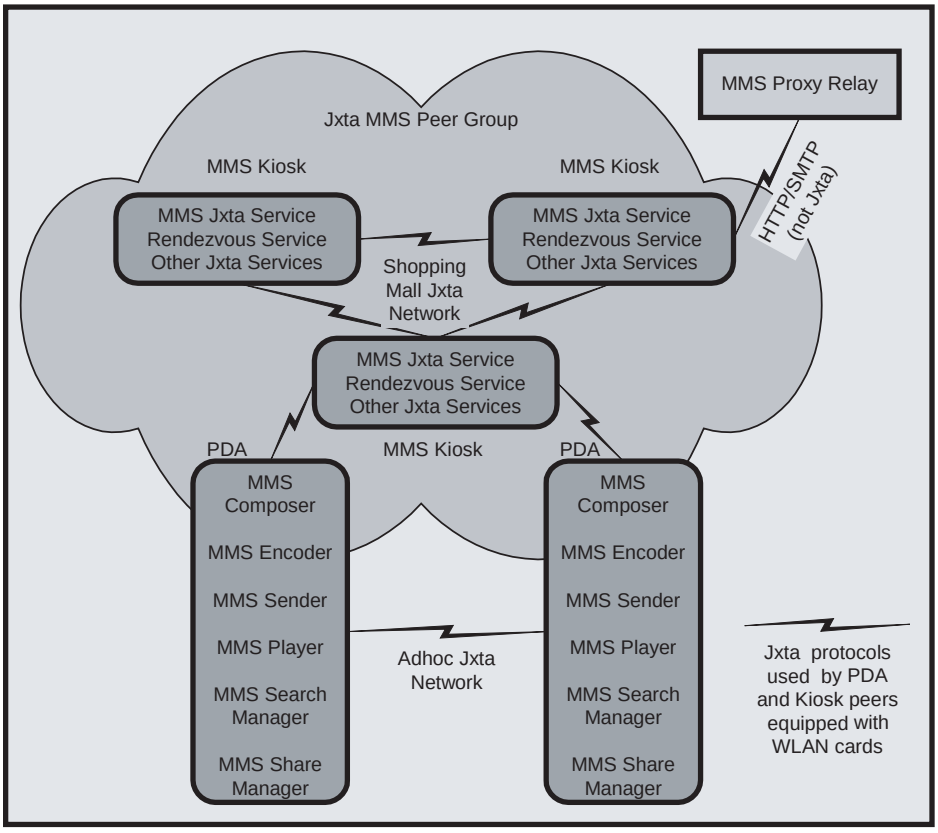
Most shopping malls have information kiosks, which could be equipped with a network point and wireless service access capabilities using technologies like Bluetooth and WiFi. These kiosks could perform searches across the mall's network to update its multimedia content repository and provide a common contact point for all the shops in the mall.

The kiosks could provide product brochures, cards, postcards, pictures, comic strips, sounds, songs and so forth in the form of MMS presentations, which could be used by a customer to send to another person. The customer need not visit all the shops and need not verbally describe a product to another person before making a decision to buy something. This leads to increased revenue for the shops. These kiosks could also provide multimedia messages intended for fun and entertainment purposes and charge for them.

MMS Peer and Kiosk Architecture

An MMS Peer would not only be a multimedia messaging client like the one on a mobile phone today but also would provide the capabilities to search for content and share content with other mobile devices in the vicinity. It would also not require a WAP stack, as it would send the messages directly to the MMS Proxy using either its HTTP or SMTP interface, which are expected to be accessible through the Jxta-based MMS service

Figure 2. MMS Peers and kiosk architecture



provided by a peer like the kiosk or another more powerful mobile device like a laptop, which is connected to the Internet and in the vicinity.

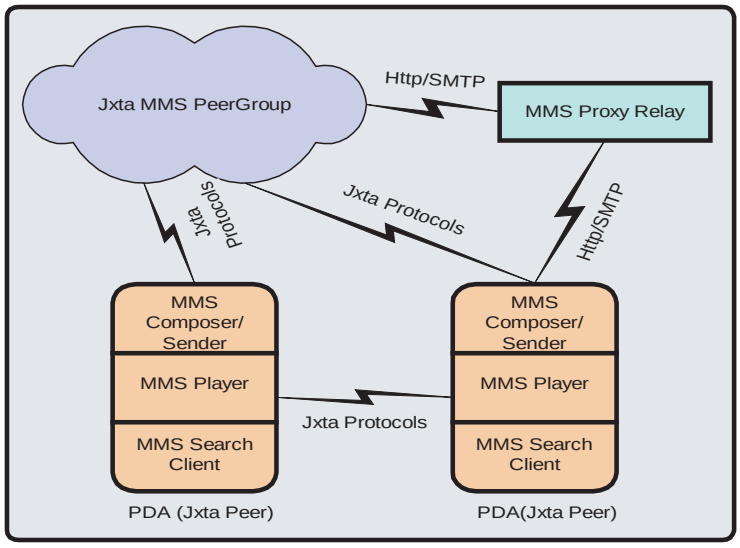
MMS Kiosk and Jxta

In a P2P environment like Jxta, commonly accessed information gets replicated (the peers have a choice to keep a copy of content passing through them) and becomes available at peers fewer hops away. This avoids "hotspots" and is ideal for content sharing, where the content can be of any type. For an MMS Kiosk searching for multimedia messages, the situation is no different and it would thus be ideal to use a P2P framework to advertise and search for multimedia messages and media content.

Design and Implementation of MMS Peers

Figure 3 shows the architecture due to the availability of the Jxta platform for a PDA (Jxta Platform, 2002). This architecture is almost fully P2P except the interaction between the Jxta Peer and the MMS Proxy-Relay. This offers the advantage

Figure 3. MMS kiosk environment architecture



that a customer can become part of a Peer Group due to other customers around him. This opens opportunities for customers to exchange MMS messages they have on their devices.

The PDA has been shown to have only an MMS Composer, MMS Player and MMS Sender. The MMS Composer composes a message by aggregating all the media and presentation information provided by the user. The MMS Sender performs an HTTP Post to the MMS Proxy-Relay to send the message to its destination. An MMS Player is also provided to the PDA client to view an MMS message before sending. The Kiosk/Shop is what provides the service to allow a customer to search for MMS messages and send them. The kiosk and the shops are part of a Jxta MMS Peer Group.

The protocols that the PDA can use to directly send to the MMS Proxy are either HTTP or SMTP (if the MMS Proxy-Relay provides an SMTP interface). The communication between the kiosk/shop and the PDA can be over Bluetooth, IEEE 802.11b or Infrared. Infrared is not a good choice due to its very limited range.

PDA MMS Peer Design

The MMS Peer on the PDA consists of four modules:

- MMS Composer
- MMS Encoder and Sender
- MMS Player
- MMS Jxta Search and Share

MMS Composer

This module allows a user to compose an MMS on the move. It allows the user to select the media content and provide layout details and timing information for the slides of the MMS presentation.

The process results in the generation of an SMIL file, which contains the presentation details of the media. Subsequently, a Jar file (JAR Documentation, 2002) is created with all the media files and the SMIL file. The MMS Sender (in the next section) takes the Jar as its input, encodes it into an MMS and sends it.

MMS Encoder and Sender

MMS can be sent either using HTTP Post or SMTP if the MMS Proxy-Relay provides both interfaces. The two modes of sending the message could be chosen based on the priority of the message. Using SMTP takes longer to send, as the message has to be ultimately encoded according to MMS standards (MMS Encapsulation Specification, 2002). Hence, SMTP could be used to send low-priority messages.

MMS Player

The MMS Player uses takes a Jar as input, extracts all the media and SMIL parts, and uses a SMIL parser to parse the SMIL and play it. The slides in the SMIL presentation are rendered using *double buffering.* The AMR audio is first converted to the WAV format and then played. Imelody (.imy files) files cannot be played by the application at this stage.

MMS Jxta Search and Share

The MMS application, once started, represents a Jxta peer. The peer becomes a member of a universal group called the NetPeerGroup. The peer then starts to discover its peers and available peer groups. The Search and Share module relies on two main modules called the Peer Group Manager

Figure 4. MMS Player and Sender

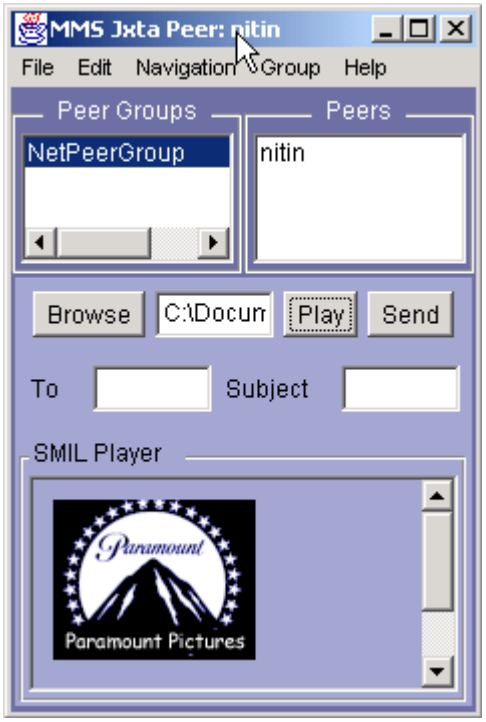
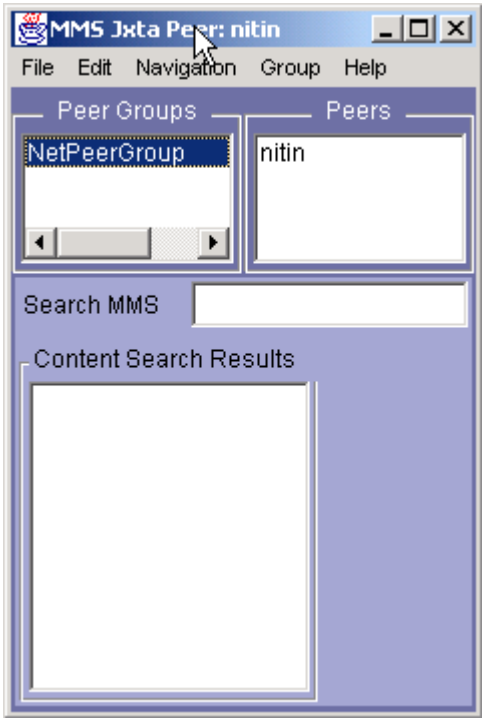


Figure 5. MMS Search



and the Search Manager. The Search and Share use the Content Management Service.

Graphical User Interface (GUI) Design

The GUI was designed keeping the PDA in mind. The GUI uses as many components that can be either easily clicked or tapped with a stylus. The following things were taken into consideration for the GUI design:

- A user would always want to have the list of peers and peer groups in front of him or her because of constant interaction with these entities.
- The limited screen size of the PDA requires that every function be provided without cluttering the screen. Thus, every function is provided on a new screen.
- This sort of layout would be ideal if service clients are to be loaded dynamically upon discovery of a service.

Figure 4 shows the MMS player and sender GUI. It also shows the list of peer groups and peers currently visible. The peer groups and peers list keeps getting updated automatically. Figure 5 shows the MMS Search GUI. A user can enter the keywords and press enter to search. A button will be added also to allow easy use on the PDA.

Comparison with Other MMS Solutions

There are some other MMS clients for the PDA that exist now. The one from Anny Way (MMS—Opportunities, migration and profits, 2003) is specifically for Pocket PC. EasyMessenger (EasyMessenger, 2003) from Oksijen Technologies is the only other Personal Java-based MMS client but without the additional P2P features provided by us. Electric Pocket's Pixer (Pixer MMS, 2003) is another MMS Client that only supports Pocket PC and Palm OS and can be used to send images only. There seems to be no work done on using MMS and Jxta together or, for that matter, not even EMS or SMS and Jxta.

All the solutions above are MMS clients with a view to sending multimedia messages, a progress from SMS or EMS. The MMS Peer was developed with a view to making not only messaging a more pleasant and easier experience but also to provide features that would facilitate access to a variety of content. The searching and sharing of content from peers in the vicinity (shoppers) as well as content stores (kiosks) make it a compelling multimedia messaging solution.

FUTURE TRENDS

MMS Message Receiver Module for the PDA

The next step is to implement a receive module for the PDA so the MMS Peer is able to achieve two-way messaging. When a message is retrieved directly from the MMS Proxy-Relay using the HTTP GET method, the MMS Proxy-Relay will return with a message along with the HTTP headers. The HTTP headers can be easily skipped by looking for two consecutive carriage returns and line feed pairs. After this, the encoded MMS header are read byte by byte until the byte of number of body parts is reached. In this way, the MMS Peer will be able to both send and receive messages with other peers.

Service Client Plug-In Feature

The Service Client Plug-In feature refers to the client download option. The current implementation assumes the client for a service to be there on the peer. As the peer already has core Jxta functionalities, it is a good idea to use them to provide this feature. The advertisements of a service could specify the location of a client,

which could be transferred over to the peer and dynamically loaded. This is possible in Java using the API for loading classes. To enable this feature, one could create a Jxta Service that has clients registered with it.

PDA to PDA Messaging

With the existing application framework, PDA to PDA MMS messaging can be easily enabled using the Jxta messaging layer. As PDAs are more capable than mobile phones, even video could be enabled for PDA to PDA messaging. All it would mean is using another media type in the SMIL or encoded message.

To account for different PDAs communicating, the User Agent Profile Specification (UAProf) could be used for capability negotiation. The UAProf schema for MMS characteristics (client transactions) could be adapted to the PDA situation. The XML messaging layer for Jxta would enable the use of this XML scheme effectively.

CONCLUSION

The Jxta platform Personal Java port came out very recently and the application was designed and implemented with it in mind. If the basic platform functionalities have been ported correctly, then it should not take long to port this whole application to the PDA. The application conforms to the Personal Java standard when checked with the compliance tool. This implies it should work on the PDA without requiring any changes.

Currently, the MMS client by itself works on the PDA. An MMS Player and sharing of content were developed. The former was implemented while searching for a reasonably priced Bluetooth SDK for WinCE and trying various Bluetooth PCMCIA cards with the freely available Bluetooth stack called CStack. This project will have a commercial value when shopping malls in Singapore install wireless networks or have wireless kiosks.

In conclusion, with the increase in memory and processing power of a plethora of mobile devices found in the market, and the ongoing improvements in available bandwidth to the user, MMS is a service to look forward to, and more so with peer-to-peer technologies like Jxta, which will make it truly ubiquitous.

REFERENCES

Bennett, M., & Weill P. (1997). Exploring the use of electronic messaging infrastructure: The case of a telecommunications firm. *The Journal of Strategic Information Systems, 6*(1), 7-34.

Berson, A. (1992). *Client/server architecture.* New York: McGraw-Hill.

Budiarto, S. N., & Masahiko, T. (2002). Data management issues in mobile and peer-to-peer environments. *Data & Knowledge Engineering, 41*(2-3), 183-204.

Chambers, F. B., Duce, D. A., & Jones, G. P. (Eds.). (1984). *Distributed computing.* London; Orlando: Academic Press.

EasyMessenger. (2003). Retrieved from www.o2.com.tr/easymessenger.htm

JAR Documentation. (2002). Retrieved from http://java.sun.com/products/jdk/1.1/docs/guide/jar/

Jxta Platform. (2002). Retrieved from http://platform.jxta.org

Jxta Technology Overview. (2002). Retrieved from www.jxta.org/project/www/docs/TechOverview.pdf

Madron, T. W. (1993). *Peer-to-peer LANs: Networking two to ten PCs.* New York: Wiley.

MMS—Opportunities, migration and profits. (2003). Retrieved from www.annyway.com/annyway-com2.htm

Multimedia Messaging Service (MMS) Architecture Overview. (2002). Retrieved from www.wapforum.org/what/technical.htm

Patel A., & Gaffney K. (1997). A technique for multi-network access to multimedia messages. *Computer Communications, 20*(5), 324-337.

Pixer MMS. (2003). Retrieved from http://electricpocket.com/products/carriers.html

Project Jxta: Getting Started. (2002). Retrieved from http://www.jxta.org/project/www/docs/GettingStarted.pdf

Tan, D. H. M, Hui, S. C., & Lau, C. T. (2001). Wireless messaging services for mobile users. *Journal of Network and Computer Applications, 24*(2), 151-166.

Yemini, Y. (Ed.). (1987). *Current advances in distributed computing and communications*. Rockville: Computer Science Press.

Yeo, C. K., Hui, S. C., Soon, I. Y., & Lau, C. T. (2001). A unified messaging system on the Internet. *Microprocessors and Microsystems, 24*(10), 523-530.

KEY TERMS

Application Programming Interfaces (APIs): Programming tools that provide developers with a simple, consistent mechanism for extending the functionality of an application and for accessing existing computing systems.

Distributed System: A system made up of components that may be obtained from a number of different sources that together work as a single distributed system, providing the run-time infrastructure supporting today's networked computer applications.

Multimedia: Involving or encompassing more than one concurrent presentation medium, such as text, sound and/or motion video.

Peer-to-Peer: A communications model in which each party has the same capabilities and either party can initiate a communication session.

Plug-In: Programs that can easily be installed and used as part of a Web browser. A plug-in application is recognized automatically by the browser and its function is integrated into the main HTML file that is being presented.

Protocol: A special set of rules that end points in a telecommunication connection use when they communicate with each other.

WAP Stack: A set of protocols that covers the whole process of wireless content delivery, from the definition of WML and WMLScript for creating and layout of the actual content and the specification of security measures in the WTLS to the lowest parts of the stack dealing with the actual transport of content.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 779-785, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.38
Mobile and Electronic Commerce Systems and Technologies

Wen-Chen Hu
University of North Dakota, USA

Chyuan-Huei Thomas Yang
Hsuan-Chuang University, Taiwan

Jyh-haw Yeh
Boise State University, USA

Weihong Hu
Auburn University, USA

ABSTRACT

The emergence of wireless and mobile networks has made possible the introduction of electronic commerce to a new application and research subject: mobile commerce. Understanding or constructing a mobile or an electronic commerce system is an arduous task because the system involves a wide variety of disciplines and technologies and the technologies are constantly changing. To facilitate understanding and constructing such a system, this article divides the system into six components: (i) applications, (ii) client computers or devices, (iii) mobile middleware, (iv) wireless networks, (v) wired networks, and (vi) host computers. Elements in these components specifically related to the subject are described in detail and lists of current technologies for component construction are discussed. Another important and complicated issue related to the subject is the mobile or electronic commerce application programming. It includes two types of programming: client-side and server-side programming, which will be introduced too.

INTRODUCTION

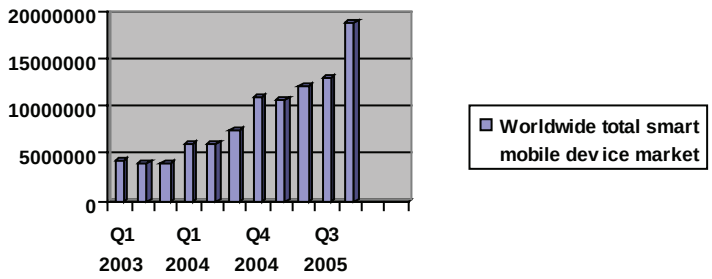
With the introduction of the World Wide Web, electronic commerce has revolutionized traditional commerce and boosted sales and exchanges of merchandise and information. Recently, the emergence of wireless and mobile networks has made possible the extension of electronic commerce to a new application and research area: mobile commerce, which is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of mobile handheld devices. In just a few years, mobile commerce has emerged from nowhere to become the hottest new trend in business transactions. The future of mobile commerce is bright according to the following predictions:

- Figure 1 shows the growth in demand for smart mobile devices including handhelds, wireless handhelds, and smart cellular phones through 2006, as estimated by the research firm Canalys (2004a, 2004b, 2004c, 2005a, 2005b, 2005c, 2005d, & 2006).
- Cumulative sales of smartphones will reach 1 billion units by the first quarter of 2011 according to IDC, a market research company (Symbian Limited, 2006).
- According to various reports, the estimated worldwide shipments of the following three equipments in 2006 were:
  - *PDAs and smartphones*: 84 million (Gartner, Inc., 2006);
  - *Celular phones*: 986 million (cellular-news, 2006); and
  - *PCs*: 250 million (Silicon Valley Daily, 2006).

The worldwide shipments of PDAs and smartphones in 2006 had a 57% increase from the same period last year, according to Gartner, Inc. Smartphone shipments bolstered the market growing 75.5% to reach 34.7 million units, more than four times the size of the PDA market. PDA shipments increased by 5.7% totaling 7.4 million units. Though the unit sales were less than one tenth of the worldwide mobile phone sales in 2006, they were not too far away from the worldwide PC sales in 2006.

- Juniper Research has published a report forecasting that the global mobile commerce market will be an $88 billion industry by 2009 (Glenbrook Partners, LLC., 2004) compared to $8.5 trillion of business-to-business electronic commerce in 2005 (Gartner, Inc. 2001).

Figure 1. Worldwide total smart mobile device market



- Jupiter Research says ringtone revenues, which have been doubling in recent years, will reach $724 million in 2009. Mobile gaming revenues should reach $430 million the same year. The market research firm forecasts that total m-commerce sales globally could reach $3.6 billion in 2006 (Brad, 2006).

Mobile commerce is an effective and convenient way of delivering electronic commerce to consumers from anywhere and at any time. Realizing the advantages to be gained from mobile commerce, companies have begun to offer mobile commerce options for their customers in addition to the electronic commerce they already provide (Yankee Group, 2001). However, it requires a tremendous effort to understand or construct a mobile or an electronic commerce system because it involves such a wide range of disciplines and technologies. To lessen the difficulty, this article will divide the system into six components: (i) applications, (ii) client computers or devices, (iii) mobile middleware, (iv) wireless networks, (v) wired networks, and (vi) host computers. Since each component is large enough to be a research area by itself, only elements in components that are specifically related to mobile or electronic commerce are explained in detail. Lists of the technologies used for component construction are also discussed. Related research on mobile commerce systems can be found in the article by Varshney, Vetter, and Kalakota (2000).

System Requirements

A wide variety of technologies are used to build mobile or electronic commerce systems. No matter what kinds of technologies are used, the requirements for both mobile and electronic commerce systems include:

- The system uses the state-of-the-art technologies.
- The system is easy to deploy and adapt by content providers, telecommunication companies, and computer/device manufacturers.
- The system allows end users to perform transactions easily.
- The system allows products to be personalized or customized upon request. For example, Web content can be viewed via either browsers or micro browsers.
- Maximum interoperability is desirable because so many technologies are now available and new techniques are constantly being invented for the use of mobile or electronic commerce systems.
- Program/data independence is held, that is, changing the system components will not affect the existing programs/data.
- End-to-end security and user privacy are rigorously enforced.

Requirements solely for mobile commerce systems include:

- The system allows end users to perform transactions easily, in a timely manner, and ubiquitously.
- The system provides supports for a wide variety of mobile commerce applications such as location finding to content providers.
- The applications can be accessed from a wide range of handheld devices.

SYSTEM STRUCTURES

This section illustrates the system structures of electronic and mobile commerce and explains the procedures of mobile commerce transactions. A modular approach will be used to study the systems.

An Electronic Commerce System Structure

Electronic commerce describes the manner in which transactions take place over networks, mostly the Internet. It is the process of electronically buying and selling goods, services, and information. An electronic commerce system is inherently interdisciplinary and there are many different ways to implement it. Figure 2 shows the structure of a traditional electronic commerce system and a typical example of such a system. The system structure includes four components, some of which are at least partly shared by mobile commerce systems: (i) electronic commerce applications, (ii) client computers, (iii) wired networks, and (iv) host computers.

Figure 2. An electronic commerce system structure

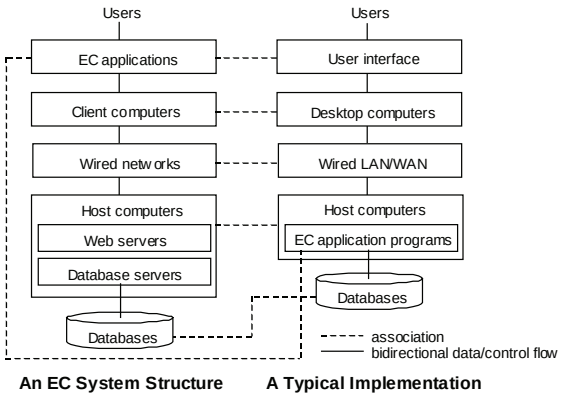
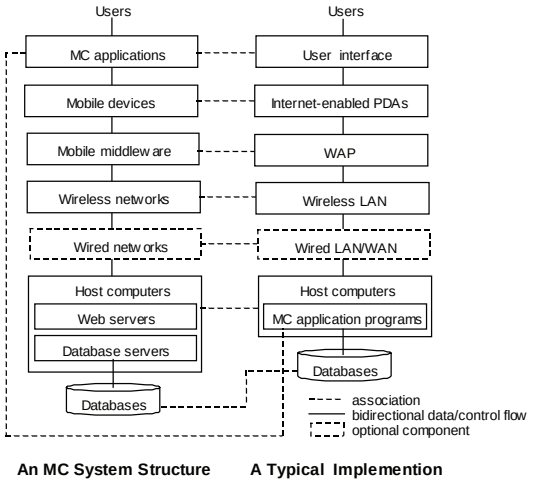


Figure 3. A mobile commerce system structure



A Mobile Commerce System Structure

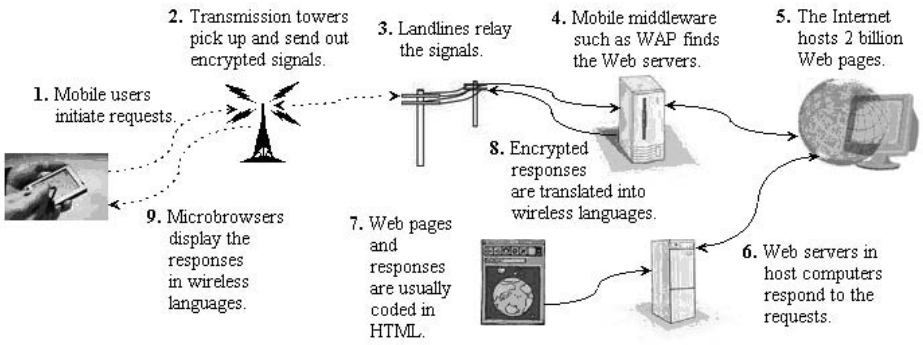
Much like the concept of e-commerce, mobile commerce is a type of business conducted 100% electronically through computer networks; m-commerce is a concept of buying and selling goods and services via wireless networks with a mobile device. Compared to an electronic commerce system, a mobile commerce system is much more complicated because components related to mobile computing have to be included. Figure 3 shows the structure of a mobile commerce system and an example of such a system that is currently possible based on the existing technologies (Hu, Lee, & Yeh, 2004). The system structure includes six components: (i) mobile commerce applications, (ii) mobile handheld devices, (iii) mobile middleware, (iv) wireless networks, (v) wired networks, and (vi) host computers. The network infrastructure for mobile commerce systems consists of mobile middleware and wired & wireless networks. The wired networks component has the same structure and implementation as that needed by an electronic commerce system.

Mobile Commerce Transaction Processing

Mobile commerce transaction processing is complicated. To explain how the mobile commerce components work together for a transaction, Figure 4 shows a flowchart of how a user request is processed by the components in a mobile commerce system, along with brief descriptions of how each component processes the request:

1. *Mobile commerce applications*: A content provider implements an application by providing two sets of programs: client-side programs, such as user interfaces on micro browsers, and server-side programs, such as database access and updating.
2. *Mobile handheld devices*: Handheld devices present user interfaces to the mobile end users, who specify their requests on the interfaces. The devices then relay the user requests to the other components and later display the processing results using the interfaces.

Figure 4. A flowchart of a user request processed in a mobile commerce system



3. *Mobile middleware*: The major purpose of mobile middleware is to seamlessly and transparently map Internet contents to mobile stations that support a wide variety of operating systems, markup languages, micro browsers, and protocols. Most mobile middleware also encrypts the communication in order to provide some level of security for transactions.
4. *Wireless and mobile networks*: Mobile commerce is possible mainly because of the availability of wireless networks. User requests are delivered to either the closest wireless access point (in a wireless local area network environment) or a base station (in a cellular network environment).
5. *Wired networks*: This component is optional for a mobile commerce system. However, most computers (servers) usually reside on wired networks such as the Internet, so user requests are routed to these servers using transport and/or security mechanisms provided by wired networks.
6. *Host computers*: Host computers process and store all the information needed for mobile commerce applications, and most application programs can be found here. They include three major components: Web servers, database servers, and application programs and support software.

APPLICATIONS

The emergence of electronic and mobile commerce creates numerous business opportunities and applications. Electronic commerce, defined as the buying and selling of goods and services and the transfer of funds through digital communications, includes a wide variety of applications, such as auctions, banking, marketplaces and exchanges, recruiting, and retailing, to name but a few. Mobile commerce applications not only cover the electronic commerce applications, but also include new applications, for example, mobile inventory tracking, which can be performed at any time and from anywhere by using mobile computing technology.

Electronic Commerce Applications

This sub-section discusses some new business models, which were not seen before, created by electronic commerce. Other than the "buy-and-sell" model, the following list gives some other common models created by e-commerce (Turban, et al, 2004):

- *Affiliate marketing:* Affiliate marketing is a marketing method, which allows other Websites to receive a commission by selling your products or services. For the example of Amazon.com's Associates Program, the associates drive Internet traffic to Amazon through specially formatted links that allow Amazon to track sales and other activities. The partners can receive up to 10% in referral fees on all qualifying revenue made through their links to Amazon's products and services. Amazon sends monthly payments to those associates.
- *Comparing prices:* This method presents a list of services or products based on a consumer's specifications. mySimon.com is a comparison shopping site for apparel, computers, electronics, jewelry, video games, and more. It gathers prices on millions of products from thousands of stores, so customers can compare products and find the best price before he or she buys.
- *Customization and personalization:* Customization or personalization is to design and creation of content that meets a customer's specific needs. For example, Dell is based on a simple concept: by selling computer systems directly to customers. This direct business

model eliminates retailers that add unnecessary time and cost. Instead of picking one from few standard models, Dell customers can specify their requirements such as memory sizes and CPU models and Dell will build the systems based on their specifications.

- *Electronic marketplaces and exchanges:* Electronic marketplaces are Internet Websites acting as a meeting point between supply and demand and electronic exchanges are a central marketplace with established rules and regulations where buyers and sellers meet to trade futures and options contracts or securities. Electronic marketplaces and exchanges provide benefits to both buyers and sellers because they are more efficient than traditional ones.
- *Electronic tendering systems:* Tendering is potential suppliers bid competitively for a contract, quoting a price to the buyer. Large buyers usually make their purchases through a tendering (bidding) system, which is more effective and efficient with the help of electronic commerce.
- *Group purchasing:* Large-quantity purchasing usually receives lower prices than small-quantity purchasing does. Electronic commerce allows a group of customers or organizations to place their orders together and negotiate for a better deal. For example, Amerinet members saved more than $300 million in 2003 through group purchasing health care equipments and products.
- *Name your price:* With this model, the product or service prices are set by customers instead of sellers. Priceline.com is the first company applying this method. The following example shows how the "Name Your Price" of Priceline.com works. With Priceline's "Name Your Own Price" hotel reservation service, customers choose the star level of hotel they want, along with the desired neighborhood, dates and price they want to pay. Priceline then works to find a hotel room at the customer's desired price. There is no guarantee that any offer will be accepted due to the changeability of room availability and pricing. Customers learn the specific hotel name and location after the purchase is completed.
- *Online auctions:* Traditional auctions usually require bidders to attend the auctions, whose items are limited. Online auctions allow bidders from everywhere to bid products or services provided by various sellers without needing to show up. eBay.com is the world's largest online auction site. It offers an online platform where millions of items are traded each day.

Mobile Commerce Applications

Mobile commerce applications cover almost everything in our daily lives such as traveling and foods. Table 1 lists some major mobile commerce applications along with explanations of three applications related to traveling (Sadeh, 2002):

- *Map services*: Map services provide various useful functions to mobile users. Some of the functions include:
  - *Directions*, which are driving/walking directions from the starting location to destination;
  - *Maps*, which include traditional clear maps;
  - *Local hangouts and businesses recommendations*, which provide suggestions for restaurant/gas-station/grocery-store/movie-theater; and
  - *Satellite imagery*, which includes real images from satellites.

  A few mobile map services are available. Google Maps for Mobile (n.d.) lets users find local hangouts and businesses across

Table 1. Major mobile commerce applications

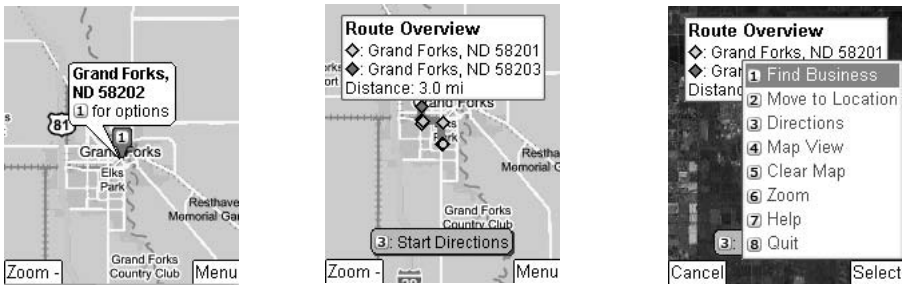
Mobile Category	Major Applications	Clients
Advertising	Targeted ads, location-based ads	Business
Commerce	Mobile transactions and payments	Business
Education	Mobile classrooms and labs	Schools and training centers
Enterprise resource planning	Resource management, managing a mobile workforce	All
Entertainment	Games/images/music/video downloads and on-line gaming	Entertainment industry
Health care	Accessing and updating patient records	Hospitals and nursing homes
Inventory tracking and dispatching	Product tracking and dispatching	Delivery services and transportation
Traffic	Global positioning, routing services, toll paying, traffic advisories	Transportation and auto industries
Travel and weather	Reservation services	Airlines, hotels, travel agencies

town or across the country—right from their phones. Figure 5 shows three screenshots from the Google's map services where

a. a clear map of the location with a postal code 58202;
b. directions from the postal code 58201 to 58203; and
c. a satellite map of (b) and a menu.

- *Travel*: Travel expenses can be costly for a business or an individual. Mobile commerce could help reduce operational costs by providing mobile travel management services to travelers. It can be used to provide assistance to customers by using the mobile channels to locate a desired hotel nearby, purchase tickets, make transportation arrangements, and so on. The travel section of Yahoo! Mobile (n.d.) includes the following services:
  - *Travel guides*: Allow mobile users to research 500,000 places to stay and things to do in over 40,000 cities worldwide with user reviews, photos, and maps, save favorite places into a custom trip plan, and get good travel deals.
  - *Trip planner*: It is a tool that lets mobile users save hotels, attractions, restaurants, maps, and more to a customized travel guide. Users can add travel dates, their own comments, even bookmarks for other sites to their trip.
  - *FareChase*: Yahoo! FareChase is a travel search engine that helps travelers scour the Web for the best flights and hotels that meet their budget and travel schedules.
  - *Deals*: This service provides various top deals from hotels to car rentals.

Figure 5. Screenshots of the Google's map services



(a) a clear map *(b) directions* *(c) a satellite image of (b) and a menu*

Figure 6. Screenshots of Yahoo! Mobile: Travel

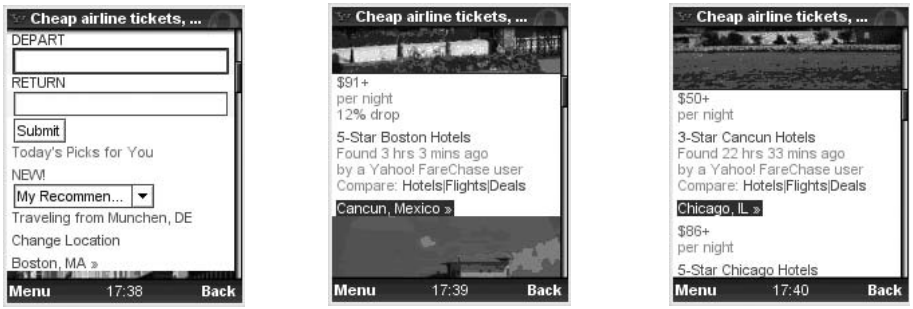


Figure 7. Screenshots of Google weather reports

Of course, mobile users can search travel-related information on Yahoo! Travel. Figure 6 shows screenshots of Yahoo! Mobile: Travel.

- *Weather*: Weather reports are important for travelers, who may pack their bags or plan their trips based on the weather predictions. Most mobile service providers offer weather reports. For example, the Google provides the following local weather information: (i) 3-day weather forecasts including lowest and highest temperatures, (ii) current temperature, (iii) humidity, (iv) weather conditions with pictures, and (v) wind including direction and speed. Figure 7 shows screenshots of Google weather reports.

CLIENT COMPUTERS OR DEVICES

Desktop and laptop computers are on the client-side of electronic commerce systems, whereas mobile handheld devices are for mobile commerce systems. An Internet-enabled mobile handheld device is a small general-purpose, programmable, battery-powered computer that is capable of handling the front end of mobile commerce applications and can be operated comfortably while being held in one hand. It is the device via which mobile users interact directly with mobile commerce applications. The differences between these two client machines are given in Table 2. There are other kinds of computers such as tablet computers, which are a special kind of PCs.

Client-Side Programming

Mobile or electronic commerce application programming involves a variety of technologies and languages. It consists of two kinds of programming:

- *Client-side programming*, which is to develop software running on client computers or devices. It is mostly related to Web interface

Table 2. Differences between desktop & laptop computers and handheld devices

	Desktop and Laptop Computers	**Mobile Handheld Devices**
Browser	Desktop browsers	Micro browsers
Functions	Full	Limited
Major Input Methods	Keyboards and mice	Stylus and soft keyboards
Major Output Methods	Screens and printers	Screens
Mobility	Low	High
Networking	Wired	Wireless and mobile
Transmission Bandwidth	High	Low
Power Supply	Electrical outlets	Batteries
Screen	Normal	Small
Size	Desktop	Handheld
Weight	Normal	Light

construction. Popular languages for Web interface construction include CSS, DOM, (X)HTML, JavaScript, WML, WMLScript, XML, XSL(T), and so forth. Other than Web interface construction, client-side programming can be used to build client-side applications such as address and schedule books. The tools and languages used for client-side application development are based on the client-side operating systems, for example, Visual Studio for Windows and C/C++ for Palm OS.

- *Server-side programming*, which is to develop software running on servers. The software normally receives requests from browsers and sends the results from databases/files/programs back to the browsers for display. Popular server-side languages include C/C++, Java, Perl, PHP, and so forth. Other than Web applications, it can be used to implement numerous applications such as instant messaging and telephony. However, this kind of applications is normally related to network programming such as TCP/IP programming and will not be covered in this book.

This sub-section discusses Web interface construction. The server-side programming will be covered in the section of host computers. Other than building a Web system from scratch by using various languages and tools, some common software packages are available for developing Web applications easily and quickly. Those packages can be divided into three categories: (i) multimedia editors, (ii) HTML editors, and (iii) integrated development environments (IDEs):

- *Multimedia editors*, which are used to create, edit, and post animation, audio, images, and videos on Web pages. Adobe Systems, Inc. provides two popular multimedia editors:
  - *Flash*, which is an authoring environment for creating animations, advertisements, and various Web page components, to integrate video into Web pages, and more recently, to develop rich Internet applications. Flash Professional is an IDE while Flash Player is a virtual machine used to run, or parse, the Flash files.
  - *Photoshop*, which is image-editing and graphics creation software.
- *HTML editors*, which are used to create static Web pages. Three popular HTML editors are:
  - *Adobe Dreamweaver*, which is WYSIWYG (what you see is what you got) authoring software that allows Web developers to generate HTML and JavaScript source code while viewing the site as they work.
  - *Microsoft Expression Web*, which is a design tool to create sophisticated standards-based Web sites. It combines both FrontPage and Visual Studio technologies in a new user interface for creating XHTML, CSS, XML, XSLT, and ASP.NET 2.0. Where appropriate, the user interface and features of Expression Web and Visual Studio are identical.
  - *Microsoft SharePoint Designer*, which will enable information workers to develop applications and solutions on top of the SharePoint platform to enable organizational agility, business process automation, and get the value of Microsoft Office applications on the SharePoint platform.

The category of integrated development environments (IDEs) will be covered in the section of Host Computers.

MOBILE MIDDLEWARE AND WIRELESS NETWORKS

Mobile middleware and wireless networks are for mobile commerce systems only. The mobile middleware is optional, but the system will be greatly simplified with it. A mobile commerce system is already complicated enough. Without mobile middleware, the mobile system becomes even more complicated.

Mobile Middleware

The term middleware refers to the software layer between the operating system and the distributed applications that interact via the networks. The primary mission of a middleware layer is to hide the underlying networked environment's complexity by insulating applications from explicit protocols that handle disjoint memories, data replication, network faults, and parallelism (Geihs, 2001). The major task of mobile middleware is to seamlessly and transparently map Internet contents to mobile handheld devices that support a wide variety of operating systems, markup languages, micro browsers, and protocols. WAP and i-mode are the two major kinds of mobile middleware:

- *WAP (wireless application protocol)*, which is a secure specification that allows users to access information instantly via mobile handheld devices such as smart phones and PDAs (Open Mobile Alliance Ltd., n.d.). WAP supports most wireless networks including CDPD, CDMA, GSM, PDC, PHS, TDMA, FLEX, ReFLEX, iDEN, TETRA, DECT, DataTAC, and Mobitex. WAP is supported by all operating systems. Ones specifically engineered for handheld devices include PalmOS, EPOC, Windows CE, FLEXOS, OS/9, and JavaOS. Although WAP supports HTML and XML, the WML language is specifically designed for small screens and one-hand navigation without a keyboard.
- *i-mode*, which is a mobile Internet service that has caused a revolution in both business and private lifestyles in Japan (NTT DoCoMo, Inc., 2007). Forty-six million subscribers have been attracted to this service since its debut in February 1999 and currently more than 95,000 Internet sites are providing a variety of contents. The use of packet transmissions offers continuous access, while the use of a subset of HTML makes content creation easy and provides simple conversion of existing websites.

Table 3 compares i-mode to WAP.

Table 3. A comparison between the two major types of mobile middleware

	WAP	**i-mode**
Developer	Open Mobile Alliance	NTT DoCoMo
Implementation	A protocol	A complete mobile Internet service
Web Language	WML (wireless markup language)	CHTML (compact HTML)
Major Technology	WAP gateway	TCP/IP development
Key Features	Widely adopted and flexible	Highest number of users and easy to use

Wireless Networks

Wireless communication capability supports mobility for end users in mobile commerce systems. Wireless LAN, MAN, and WAN are the major components used to provide radio communication channels so that mobile service is possible. In the WLAN category, the Wi-Fi standard with 11 Mbps throughput dominates the current market. However, it is expected that standards with much higher transmission speeds, such as IEEE 802.11a and 802.11g, will replace Wi-Fi in the near future. Compared to WLANs, cellular systems can provide longer transmission distances and greater radio coverage, but suffer from the drawback of much lower bandwidth (less than 1 Mbps). In the latest trend for cellular systems, 3G standards supporting wireless multimedia and high-bandwidth services are beginning to be deployed. The wireless telephone technology includes several generations as follows:

- *0G (1945-1973),* which refers to mobile radio telephone systems.
- *1G (1980s),* which is analog cell phone standards including NMT and AMPS.
- *2G (1990s),* which is digital cell phone standards divided into TDMA-based and CDMA-based standards depending on the type of multiplexing used.
- *2.5G (late 1990s),* which is implemented a packet switched domain in addition to the circuit switched domain.
- *3G (early 2000s),* which includes wide-area wireless voice telephony and broadband wireless data, all in a mobile environment.
- *4G (2000s),* which provides end-to-end IP solution where voice, data, and multimedia streaming can be served at higher data rates with anytime-anywhere concept.

A wide variety of technologies and standards for wireless telephones are available. Some of the major ones include:

- *CDMA (code division multiple access),* which is based on a spread spectrum method. The method transmits a signal by "spreading" it over a broad range of frequencies. This provides reduced interference and can increase the number of simultaneous users within a radio frequency band. With CDMA, each conversation is digitized and then tagged with a code.
- *GSM (global system for mobile communications),* which is one of the most popular standards for mobile phones and is specifically developed to provide system compatibility across country boundaries, especially the Europe. It is based on TDMA (time division multiple access) technology, which works by dividing a radio frequency into time slots and then allocating slots to multiple calls. Therefore, GSM allows eight simultaneous calls on the same radio frequency.
- *IEEE 802.11,* which includes an encryption method, the wired equivalent privacy algorithm. WLAN (wireless local area network), based on 802.11, allows a mobile user connecting to a local area network (LAN) through a wireless (radio) connection. This wireless data transmission speed of WLAN is up to 54 Mbps.
- *IEEE 802.16,* which ensures compatibility and interoperability between broadband wireless access equipment. WiMAX (worldwide interoperability for microwave access), based on 802.16, provides wireless data over long distances, in a variety of different ways, from point to point links to full mobile cellular type access. In practical terms this enables a user, for example, to browse the Internet on a laptop computer without

Table 4. Wireless telephone technology evolution

	2G (10 Kbps – 40 Kbps)	**2.5G** (20 Kbps – 171 Kbps)	**3G** (60 KBps – 54 Mbps)	**4G** (50 Mbps – 1 Gbps)
CDMA track	IS-95	CDMA 2000	W-CDMA	UMTS Revision 8 (LTE)
GSM track	GSM	GPRS	EDGE	
IEEE 802.11 track			Wi-Fi	
IEEE 802.16 track				WiMAX

physically connecting the laptop to a wall jack.

Table 4 shows major technologies and standards used in the wireless telephone generations.

Wired Networks

Wired networks are used to transmit data for electronic and mobile commerce. This component is a requirement for electronic commerce, but not necessary for mobile commerce, though mobile commerce would be greatly benefited by applying wired networks to its data communication because data transmission using wireless networks is more expensive than using wired networks. Among several types of wired networks, three major types are:

- *Local area network (LAN),* which spans a relatively small space of only a few square kilometers or less such as an office building. It generally offers a throughput of 10 Mbps or 100 Mbps and is usually based on ethernet technology, which is a network protocol using a bus topology and defining a specific implementation of the physical and data link layers in the OSI model (IEEE 802.3).
- *Metropolitan area network (MAN),* which spans a geographical area greater than an LAN but less than a WAN, such as few city blocks or a whole city. MAN typically uses wireless infrastructure or optical fiber connections to link its sites and it may connect multiple LANs together. Its maximum throughput is no less than 44 Mbps and it uses the distributed queue dual bus technology based on the IEEE 802.6 standard.
- *Wide area network (WAN),* which spans a wide geographic area, such as state or country, and uses specialized computers to connect smaller networks, such as LANs. It generally offers a throughput of 1.5 Mbps or more. WANs typically use wide area network services from telecommunications carriers, whose technologies include standard phone lines, ISDN (integrated services digital

network), or other high-speed services. Two examples of WAN are the Internet, the largest network in the world, and an airline corporation using WAN to connect its offices around the world.

HOST COMPUTERS

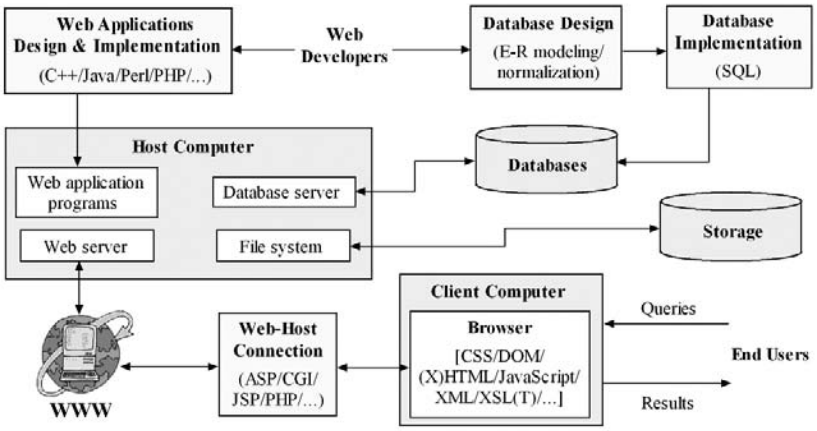
This component is similar for both electronic and mobile commerce systems because host computers are usually not aware of the differences among the targets, browsers, or micro browsers they serve. The application programs are responsible for apprehending their clients and responding to them accordingly. Most of the electronic/mobile commerce application programs reside in this component, except for client-side programs such as cookies or user interface using markup languages. A user request such as checking out or adding items to the shopping cart is actually processed at a host computer, which contains three major kinds of software specifically for electronic or mobile commerce transactions: (i) Web servers, (ii) databases and database servers, and (iii) application programs and support software. Figure 8 shows a structure of three-tiered client-server Web systems. Before examining the three kinds of software in more detail, the following sub-section studies the LAMP stack, which is one of the popular Web technologies used these days.

The LAMP Stack

One of the popular technologies for Web content implementation is the LAMP stack, which includes four components (Lawton, 2005):

1. *Linux*, which is a free open-source operating system based on Unix.
2. *Apache Web server*, which is an open-source HTTP server for modern operating systems including UNIX and Windows NT.
3. *MySQL*, which is an open source relational database management system (RDBMS) that uses structured query language (SQL), the most popular language for adding, accessing, and processing data in a database.
4. *Script languages*, such as Perl, PHP, and Python:

Figure 8. Three-tiered client-server Web system structure



- *Perl (practical extraction and report language),* which is often used for scanning text and printing formatted reports. It provides extensive support for regular expression matching, dynamically scoped variables and functions, extensible run-time libraries, exception handling and packages, and provide/require. The use of Perl has grown significantly since its adoption as the language of choice of many World Wide Web developers. CGI interfaces and libraries for Perl exist for several platforms and Perl's speed and flexibility make it well suited for form processing and on-the-fly page creation. Perl5 is a major rewrite and enhancement of Perl version 4. It adds nested data structures and object-oriented features.
- *PHP (HyperText preprocessor),* which is a general-purpose server-side scripting language that is especially suited for dynamic and interactive Web sites and can be embedded into HTML. It is the widely-used, free, and efficient alternative to competitors such as Microsoft's ASP. The PHP syntax is very similar to Perl and C. PHP is often used together with MySQL (DBMS) and Apache (Web server) on various operating systems. A PHP file may contain text, HTML tags, and scripts. Scripts in a PHP file are executed on the server.
- *Python*, which is a dynamic object-oriented programming language that can be used for many kinds of software development. It offers strong support for integration with other languages and tools, comes with extensive standard libraries, and can be learned in a few days. Python runs on Windows, Linux/Unix, Mac OS X, OS/2, Amiga, Palm Handhelds, and Nokia mobile phones. Python has also been ported to the Java and .NET virtual machines. Python is distributed under an OSI-approved open source license that makes it free to use, even for commercial products.

LAMP has several key advantages over proprietary software development approaches. Two of them are:

- *Cost*: Open source software is either free or low cost compared to proprietary software.
- *Open source*: Anyone can review, modify, and work with open source software; developers can improve and fix the technology faster.

Web Servers

A Web server is a server-side application program that runs on a host computer and manages the Web pages stored on the Web site's databases or files. There are several kinds of Web server software including public domain software from NCSA and Apache, and commercial packages from Microsoft, Netscape, and others. Three popular Web servers are:

- *Apache HTTP servers*, which are a collaborative software development effort aimed at creating a freely-available source code implementation of an HTTP (Web) server. They are jointly managed by a group of volunteers located around the world. Since April 1996, Apache has been the most popular HTTP server on the Internet. It was developed in early 1995 based on code and ideas found in the most popular HTTP server of the time, NCSA httpd 1.3. It has since evolved to rival almost any other Unix

based HTTP server in terms of functionality and speed. It features highly configurable error messages, DBM-based authentication databases, and content negotiation (Apache Software Foundation, n.d.).

- *Microsoft's Internet information services (IIS)*, which provide a Web application infrastructure for all versions of Windows servers (Microsoft, Corp., n.d.a). It is the world's second most popular Web server after Apache.
- *Sun Java system Web servers*, which provide organizations with a single deployment platform for Web services, JavaServer Pages (JSP), Java Servlet technologies, NSAPI, and CGI (Sun Microsystems, Inc., n.d.). They also offer built-in HTTP reverse-proxy capabilities to provide a highly scalable HTTP front-end to application servers or other HTTP origin servers. Its bundled FastCGI interface provides a stable and scalable platform for third party scripting technologies such as PHP, Ruby On Rails, Perl, Python, and more.

Database Servers

A database server manages database access functions, such as locating the actual record being requested or updating the data in databases. Some popular databases include:

- *IBM DB2*: DB2 9 is a hybrid data server with management of both XML and relational data. It includes the following major features:
  - XML data store;
  - Integration with relational data;
  - Eclipse-base developer workbench; and
  - Integration with leading application infrastructures like PHP, Java, and .NET
- *Microsoft*: Microsoft provides two kinds of databases: (i) *Access* for desktop computers and (ii) *SQL Server* for the server engines in client-server solutions:
  - *Access*: The Microsoft Access is a full-featured multi-user relational database management system that designed for the Microsoft Windows operating systems. It makes extensive use of drag-and-drop and visual design for queries, forms, and reports. Access comes with an integrated development environment, including incremental compilation, a fully interactive visual debugger, breakpoints, and single step-through. These capabilities combine to make Microsoft Access a powerful platform for developing client-server database solutions.
  - *SQL Server*: The SQL Server is a comprehensive database software platform providing enterprise-class data management and integrated business intelligence (BI) tools (Microsoft, Corp., n.d.b). The SQL Server data engine lies at the core of this enterprise data management solution. In addition to providing support for relational databases or XML, SQL Server combines the functions in analysis, reporting, integration, and notification. Close integration with Microsoft Visual Studio, the Microsoft Office System, and a suite of new development tools, including the Business Intelligence Development Studio, sets SQL Server apart.
- *MySQL*: MySQL is an open-source, multi-threaded, multi-user SQL relational database management system. It is used in more than 11 million installations ranging from large corporations to specialized embedded applications. Not only is MySQL the

world's most popular open source database, it is a key part of LAMP (Linux, Apache, MySQL, PHP/Perl/Python), a fast growing open source enterprise software stack. More and more companies are using LAMP as an alternative to expensive proprietary software stacks because of its lower cost and freedom from lock-in. MySQL is flexible and runs on more than 20 platforms including Linux, Windows, OS/X, HP-UX, AIX, and Netware.

- *Oracle databases*, whose newest version is Oracle10g. The following list shows the Oracle database migration (Oracle, n.d.):
  - *Oracle7.2*, which is a client-server based relational database management system (RDBMS). The query language is based on SQL.
  - *Oracle8i*, which is an RDBMS with object capabilities included. Java has been added to the database capabilities.
  - *Oracle9i*, which features full XML database functionality with the new Oracle XML DB feature, and other improvements.
  - *Oracle 10g*, which is the first database designed for enterprise grid computing. Grid computing provides an environment in which individual users can access computers, databases, and experimental facilities simply and transparently, without having to consider where those facilities are located.

Other than the server-side database servers, a growing trend is to provide a client-side mobile database or an embedded database to a handheld device with a wide range of data-processing functionality. The functionality is frequently very sophisticated, and the flat file system that comes with these devices may not be able to adequately handle and manipulate data. Embedded databases have very small footprints, and must be able to run without the services of a database administrator and accommodate the low-bandwidth constraints of a wireless network. Some leading embedded-databases are Progress Software databases, Sybase's Anywhere products, and Ardent Software's DataStage (Ortiz, 2000).

Application Programs and Support Software

Application programs and support software are responsible for handling server-side processing. Three generations of programming languages and environments are used for server-side Web application development:

1. *1st generation*: Traditionally, conventional programming languages such as C/C++ and Java are used for Web development.
2. *2nd generation*: Dynamic programming languages such as Perl and PHP gradually replace conventional languages for Web development. A dynamic language basically enables programs that can change their code and logical structures at runtime, adding variable types, module names, classes, and functions as they are running. These languages frequently are interpreted and generally check typing at runtime.
3. *3rd generation*: Recently, a couple of IDEs (integrated development environments) are used for Web development:
   - *Adobe ColdFusion*, which is an application server and software development framework used for the development of computer software in general, and dynamic Web sites in particular.
   - *Microsoft ASP.NET*, which is part of Microsoft's .NET platform and is the successor to ASP technology. ASP.

NET is a free technology that allows programmers to create dynamic web applications.

- *Microsoft Visual Studio,* which is Microsoft's flagship software development product for computer programmers. It lets programmers create standalone applications, Web sites, Web applications, and Web services that run on any platforms supported by Microsoft's .NET Framework.
- *NetBeans IDE*, which is an open-source IDE for software developers. It is used to create professional cross-platform desktop, enterprise, Web, and mobile applications.
- *Ruby On Rails (ROR),* which is a full-stack framework for developing database-backed Web applications according to the model-view-control pattern.
- *Sun Java Studio IDE,* which is a development platform with features such as UML modeling, instant collaboration, and application profiling. It is used to develop, debug, tune, and deploy enterprise applications, Web services, and portal components based on the Java EE platform.
- *Zend Core*, which is the production PHP 5 stack that provides the certified, enhanced capabilities with support and services that professionals need for PHP development and production.

SUMMARY

The emerging wireless and mobile networks have extended electronic commerce to another research and application subject: mobile commerce. A mobile or an electronic commerce system involves a range of disciplines and technologies. This level of complexity makes understanding and constructing such a system an arduous task. To facilitate this process, this article divided a mobile or an electronic commerce system into six components, which can be summarized as follows:

1. *Applications*: Electronic commerce applications are already broad. Mobile commerce applications not only cover those applications, but also include new applications, which can be performed at any time and from anywhere by using mobile computing technology.
2. *Client computers or devices*: Desktop and notebook computers are for electronic commerce and mobile handheld devices, including smart cellular phones and PDAs, are used to perform mobile transactions. Handheld devices are convenient and have many advantages over desktop computers, but they are limited by their tiny screens, small memory, low processing power, and short battery life, and suffer from wireless network transmission problems. Numerous mobile devices are available in the market, but most use one of three major operating systems: Palm OS, Microsoft Windows Mobile, and Symbian OS. At this moment, Symbian OS leads the market, although it faces a serious challenge from Windows Mobile.
3. *Mobile middleware (mobile commerce systems only)*: Mobile middleware is used to facilitate mobile communication. It is not required for mobile commerce systems, but it can greatly reduce the complication of mobile communication. WAP and i-mode are the two major kinds of mobile middleware. WAP is widely adopted and flexible, while i-mode has the highest number of users and is easy to use. It is difficult to predict which middleware will be the eventual winner in the end; it is more likely that the two will

be blended somehow at some point in the future.

4. *Wireless networks (mobile commerce systems only)*: Wireless communication capability supports mobility for end users in mobile commerce systems. Wireless LAN, MAN, and WAN are major components used to provide radio communication channels so that mobile service is possible. In the WLAN category, the Wi-Fi standard with 11 Mbps throughput dominates the current market. It is expected that standards with much higher transmission speeds, such as IEEE 802.11a and 802.11g, will replace Wi-Fi in the near future. Compared to WLANs, cellular systems can provide longer transmission distances and greater radio coverage, but suffer from the drawback of much lower bandwidth (less than 1 Mbps). In the latest trend for cellular systems, 3G standards supporting wireless multimedia and high-bandwidth services are beginning to be deployed. WCDMA and CDMA2000 are likely to dominate the market in the future.
5. *Wired networks*: This component is a requirement for electronic commerce systems, but not necessary for mobile commerce systems, though mobile commerce systems will be greatly benefited by applying wired networks to its data communication because data transmission using wireless networks is more expensive than using wired networks. Among several types of wired networks, three major types are (i) LAN (local area network), (ii) MAN (metropolitan area network), and (iii) WAN (wide area network) based on the sizes of their covering areas.
6. *Host computers*: Host computers process and store all the information needed for mobile and electronic commerce applications, and most application programs can be found here. They include three major components: (i) Web servers, (ii) database servers, and (iii) application programs and support software.

Another important issue about mobile and electronic commerce systems is application programming. Electronic and mobile commerce programming, involving a wide variety of technologies and languages, consists of two kinds of programming:

- *Client-side programming*, which is to develop software running on client computers or devices. It is mostly related to Web interface construction. The popular languages for Web interface construction include CSS, DOM, (X)HTML, JavaScript, WML, WMLScript, XML, XSL(T), and so forth.
- *Server-side programming*, which is to develop software running on servers. The software normally receives requests from browsers and sends the results from databases/files/programs back to the browsers for display. The popular server-side languages include C/C++, Java, Perl, PHP, and so forth.

REFERENCES

Apache Software Foundation. (n.d.). *Apache HTTP server project.* Retrieved June 21, 2007, from http://httpd.apache.org/

Brad, S. (2006). *Mobile commerce hits the big time*. Retrieved November 13, 2006, from http://www.wirelessweek.com/article/CA6311136.html?text=qpass

Canalys. (2004a). *A world of difference*. Retrieved April 14, 2006, from http://www.canalys.com/pr/2004/r2004061.pdf

Canalys. (2004b). *Global mobile device market shows tremendous growth*. Retrieved March 22,

2006, from http://www.canalys.com/pr/2004/r2004081.pdf

Canalys. (2004c). *Global smart phone shipments treble in Q3.* Retrieved December 3, 2006, from http://www.canalys.com/pr/2004/r2004102.pdf

Canalys. (2005a). *Global smart mobile device sale surge past 10 million in quarter.* Retrieved April 25, 2006, from http://www.canalys.com/pr/2005/r2005041.pdf

Canalys. (2005b). *Smart phones up, handhelds down globally in Q2.* Retrieved January 15, 2006, from http://www.canalys.com/pr/2005/r2005071.pdf

Canalys. (2005c). *Global mobile device shipments hit new peak in Q4 2004.* Retrieved May 02, 2006, from http://www.canalys.com/pr/2005/r2005012.pdf

Canalys. (2005d). *Worldewide smart phone market soars in Q3.* Retrieved December 3, 2006, from http://www.canalys.com/pr/2005/r2005102.pdf

Canalys. (2006). *Smart mobile device market growth remains steady at 55%*. Retrieved December 3, 2006, from http://www.canalys.com/pr/2006/r2006071.pdf

cellular-news. (2006). *Worldwide mobile phone sales up—except in Japan.* Retrieved May 3, 2006, from http://www.cellular-news.com/story/20573.php

Gartner, Inc. (2001). *Worldwide business-to-business Internet commerce to reach $8.5 trillion in 2005.* Retrieved February 26, 2006, from http://www.gartner.com/5_about/press_room/pr20010313a.html

Gartner, Inc. (2006). *Gartner says worldwide combined PDA and smartphone shipments market grew 57 percent in the first half of 2006.* Retrieved October 30, 2006, from http://www.gartner.com/it/page.jsp?id=496997

Geihs, K. (2001). Middleware challenges ahead. *IEEE computer, 34*(6), 24-31.

Glenbrook Partners, LLC. (2004). *Mobile commerce market forecast.* Retrieved June 2, 2006, from http://www.paymentsnews.com/2004/08/mobile_commerce.html

Google. (n.d.). *Google maps for mobile.* Retrieved March 12, 2007, from http://www.google.com/gmm/

Hu, W.-C., Lee, C.-W., & Yeh, J.-H. (2004). Mobile commerce systems. In S. Nansi, (Ed, *Mobile commerce applications* (pp. 1-23). Idea Group Publishing.

Lawton, G. (2005). LAMP lights enterprise development efforts. *IEEE Computers, 38*(9), 18-20.

Microsoft, Corp. (n.d.). *Internet information services.* Retrieved June 15, 2007, from http://www.microsoft.com/WindowsServer2003/iis/default.mspx

Microsoft, Corp. (n.d.). *SQL server 2005.* Retrieved May 6, 2007, from http://www.microsoft.com/sql/default.mspx

NTT DoCoMo, Inc. (2007). *i-mode*. Retrieved June 12, 2007, from http://www.nttdocomo.com/services/imode/index.html

Open Mobile Alliance Ltd. (n.d.). *WAP forum.* Retrieved June 13, 2007, from http://www.openmobilealliance.org/tech/affiliates/wap/wapindex.html

Oracle. (n.d.). *Oracle databases.* Retrieved May 25, 2007, from http://www.oracle.com/database/index.html

Ortiz, S., Jr. (2000). Embedded databases come out of hiding. *IEEE Computer, 33*(3), 16-19.

Sadeh, N. (2002). *M-commerce: Technologies, services, and business models.* New York: John Wiley & Sons.

Silicon Valley Daily. (2006). *HP regains lead in global PC sales.* Retrieved June 11, 2007, from http://www.svdaily.com/gartner1.html

Sun Microsystems, Inc. (n.d.). *Sun Java system Web server.* Retrieved June 19, 2007, from http://www.sun.com/software/products/web_srvr/home_web_srvr.xml

Symbian Limited. (2006). *Fast facts.* Retrieved December 10, 2006, from http://www.symbian.com/about/fastfacts/fastfacts.html

Turban, E., King, D., Lee, J., & Viehland, D. (2004). *Electronic commerce 2004: A managerial perspective*. Prentice Hall.

Varshney, U., Vetter, R. J., & Kalakota, R. (2000). Mobile commerce: A new frontier. *IEEE Computer*, *33*(10), 32-38.

Yahoo! (n.d.). *Yahoo! Mobile*. Retrieved June 21, 2007 from http://mobile.yahoo.com/

Yankee Group. (2001). *Over 50% of large U.S. enterprises plan to implement a wireless/mobile solution by 2003*. Retrieved December 10, 2002 from http://www.yankeegroup.com/public/news_releases/news_release_detail.jsp?ID=PressReleases/news_09102002_wmec.htm

This work was previously published in the Journal of Electronic Commerce in Organizations, edited by M. Khosrow-Pour, Volume 6, Issue 3, pp. 54-73, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 3.39
E-Commerce Services Based on Mobile Agents

Giancarlo Fortino
DEIS, University of Calabria, Italy

Alfredo Garro
DEIS, University of Calabria, Italy

Wilma Russo
DEIS, University of Calabria, Italy

INTRODUCTION

The Internet offers a unique opportunity for e-commerce to take central stage in the rapidly growing online economy. With the advent of the Web, the first generation of business-to-consumer (B2C) applications was developed and deployed. Classical examples include virtual shops, on-demand delivery of contents, and e-travel agency. Another facet of e-commerce is represented by business-to-business (B2B), which can have even more dramatic economic implications since it far exceeds B2C in both the volume of transactions and rate of growth. Examples of B2B applications include procurement, customer relationship management (CRM), billing, accounting, human resources, supply chain, and manufacturing (Medjahed, Benatallah, Bouguettaya, Ngu, & Elmagarmid, 2003).

Although the currently available Web-based and object-oriented technologies are well-suited for developing and supporting e-commerce services, new infrastructures are needed to achieve a higher degree of intelligence and automation of e-commerce services. Such a new generation of e-commerce services can be effectively developed and provided by combining the emerging agent paradigm and technology with new Web-based standards such as ebXML (2005).

Agents have already been demonstrated to retain the potential for fully supporting the development lifecycle of large-scale software systems which require complex interactions between autonomous distributed components

(Luck, McBurney, & Preist, 2004). In particular, e-commerce has been one of the traditional arenas for agent technology (Sierra & Dignum, 2001). Agent-mediated e-commerce (AMEC) is concerned with providing agent-based solutions which support different stages of the trading processes in e-commerce, including needs identification, product brokering, merchant brokering, contract negotiation and agreement, payment and delivery, and service and evaluation. In addition, the mobility characteristic of peculiar agents (a.k.a. *mobile agents*), which allows them to move across the nodes of a networked environment, can further extend the support offered by the agents by featuring advanced e-commerce solutions such as location-aware shopping, mobile and networked comparison shopping, mobile auction bidding, and mobile contract negotiation (Kowalczyk, Ulieru, & Unland, 2003; Maes, Guttman, & Moukas, 1999).

To date, several agent- and mobile agent-based e-commerce applications and systems have been developed which allow for the creation of complex e-marketplaces—that is, e-commerce environments which offer buyers and sellers new channels and business models for trading goods and services over the Internet.

However, the growing complexity of agent-based marketplaces demands for proper methodologies and tools supporting the validation, evaluation, and comparison of: (1) models, mechanisms, policies, and protocols of the agents involved in such e-marketplaces; and (2) aspects concerned with the overall complex dynamics of the e-marketplaces.

The use of such methodologies and tools can actually provide the twofold advantage of:

1. analyzing existing e-marketplaces to identify the best reusable solutions and/or identify hidden pitfalls for reverse engineering purposes; and
2. analyzing new models of e-marketplaces before their actual implementation and deployment to identify, *a priori*, the best solutions, thus saving reverse engineering efforts.

This article presents an overview of an approach to the modeling and analysis of agent-based e-marketplaces (Fortino, Garro, & Russo, 2004a, 2005). The approach centers on a Statecharts-based development process for agent-based applications and systems (Fortino, Russo, & Zimeo, 2004b) and on a discrete event simulation framework for mobile and multi-agent systems (MAS) (Fortino et al, 2004a). A case study modeling and analyzing a real consumer-driven e-commerce service system based on mobile agents within an agent-based e-marketplace on the Internet (Bredin, Kotz, & Rus, 1998; Wang, Tan, & Ren, 2002) is also described to demonstrate the effectiveness of the proposed approach.

BACKGROUND

In a broad sense, an agent is any program that acts on behalf of a (human) user (Karnik & Triphati, 1998). An agent can just sit there and interact with its environment and with other agents through conventional means, such as local/remote procedure calls and asynchronous messaging, or through more advanced coordination infrastructures such as *tuple* spaces and *event*-based systems. Agents that do not or cannot move are called "stationary agents." Conversely, a mobile agent is a program that represents a user in a computer network and can migrate autonomously from node to node to perform some computation on behalf of the user. Thus mobility is an orthogonal property of agents—that is, not all agents are mobile. Also mobile agents can interact with their environment and, notably, with other agents through mobility-aware and mobility-unaware infrastructures (Fortino & Russo, 2005). Indeed, the emergence of mobile agents was motivated by the benefits they provide for creating distributed systems. In

fact, as Lange and Oshima (1999) pointed out in their seminal paper, there are at least seven good reasons to employ mobile agents: reduction of network load, overcoming of network latency, encapsulation of protocols, asynchronous and autonomous execution ("dispatch your agents, shut off your machine"), dynamic adaptation, seamless system integration, and robustness and fault-tolerance.

An agent-based e-marketplace (AEM) is a distributed multi-agent system formed by both stationary and mobile agents which provide e-commerce services to end-users within a business context. AEMs are, as previously pointed out, distributed large-scale complex systems which require tools which are able to analyze not only the AEM at the *micro* level (i.e., behaviors and interactions of their constituting agents), but also the AEM at the *macro* level (i.e., the overall AEM dynamics).

In Griss and Letsinger (2000), an agent-based framework for e-commerce simulation games has been developed by using *Zeus*, a Java-based multi-agent system developed at the British Telecom Lab. Its goal is to evaluate the potential consequences of novel combinations of market models, business strategies, and new e-services through multi-player shopping games, in which agents represent various typologies of sellers, buyers, brokers, and services.

In Wang et al. (2002), an infrastructure for Internet e-marketplaces based on the *Aglets* mobile agents that enables real commercial activities by consumers, agents, and merchants, has been proposed. Its goal is not only to provide an advanced e-commerce service, but also to evaluate several dispatching models for mobile agents.

Bredin et al. (1998) describe a simulated environment for mobile agents which allows analyzing the market-based resource control system of the *D'Agents* mobile agent system and, in particular, the resource allocation mechanism of its resource manager using a sealed-bid, second-price auction policy.

Although useful insights about AEM micro and macro levels can be acquired by playing e-commerce simulation games and, then, analyzing the obtained results, or by evaluating real e-commerce applications, discrete event simulators are essential for evaluating how AEMs work on scales much larger than that achievable in games or in applications which involve humans. In fact, discrete event simulation is currently extensively exploited as a strategic tool in most research and application areas which are directly or indirectly related to computer science. In this context, the article proposes an approach based on discrete event simulation and shows its application to the analysis of micro-level issues of a consumer-driven AEM: validation and evaluation of services based on mobile agents for product searching and buying.

MODELING AND ANALYSIS OF MOBILE AGENT-BASED SYSTEMS

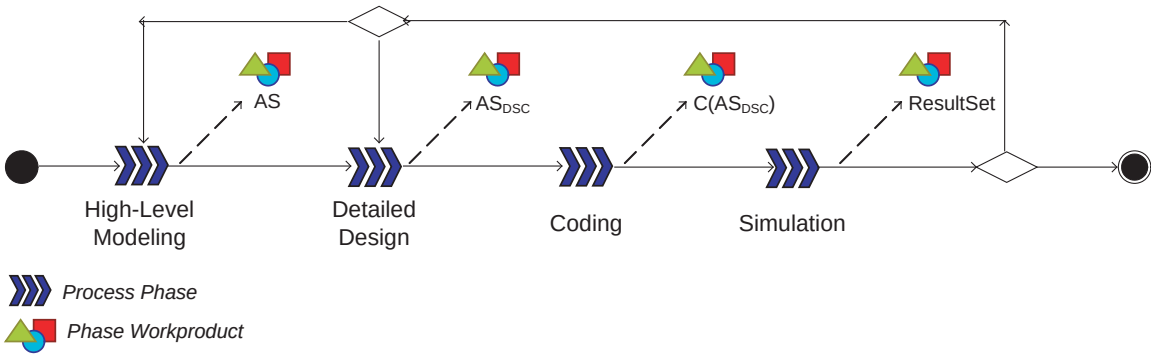
The StateCharts-Based Approach for Modeling and Analysis

The proposed approach (Fortino, Garro & Russo, 2005) consists of the following phases: high-level modeling, detailed design, and coding and simulation (see Figure 1).

The *High-Level Modeling* of an agent-based system can be supported through well-established agent-oriented methodologies (such as the Gaia methodology; Wooldridge, Jennings, & Kinny, 2000) which cover the phases of requirements capture, analysis, and high-level design. An agent-based system (AS) can be modeled as follows:

AS = <AT, LCL, act, ser, pro>,
where:
AT (Agent Types) is the set of types of agents embodying activity, offering services, and interacting with each other.

Figure 1. Phases and work products of the proposed approach



LCL (Logical CommunicationLinks) is the set of logical communication channels among agent types which embody interaction protocols.

act: $AT \rightarrow$ *activity description* is the activity relation which associates one or more activities to an agent type.

ser: $AT \rightarrow$ *service description* is the service relation which associates one or more services to an agent type.

pro: $LCL \rightarrow$ *interaction description* is the protocol relation which associates an interaction protocol to a logical communication channel.

The *Detailed Design* of an AS is achieved through a Statecharts-based formalism, namely Distilled StateCharts (DSC) (Fortino et al., 2004b), which allows for the specification of the behavior of the agent types and the interaction protocols among the agent types. In fact, a Statecharts-based specification of an entity describes both internal behavior and coordination through the reception and generation of events (Harel & Gery, 1997). DSC allow for the specification of the behavior of lightweight agents which have the following features: event-driven, single-threaded, capable of transparent migration, and executing chains of atomic actions. The DSC-based specification of an AS (AS_{DSC}) can be expressed as follows:

$AS_{DSC} = \{Beh(AT_1), \ldots, Beh(AT_n)\}$,
where:

$Beh(AT_i)$ is the DSC-based specification of the behavior of the i-th agent type.

$Beh(AT_i) = <S_{Beh}(AT_i), E_{Beh}(AT_i)>$, where $S_{Beh}(AT_i)$ is a hierarchical state machine incorporating the activity and interaction handling of the i-th agent type and $E_{Beh}(AT_i)$ is the related set of events to be handled triggering state transitions in $S_{Beh}(AT_i)$.

The *Coding* of an AS_{DSC}, $C(AS_{DSC})$, is carried out through the Java-based Mobile Active Object (MAO) Framework (Fortino et al., 2004b). In particular, $Beh(AT_i)$ can be seamlessly translated into a composite object (called MAOBehavior object), which is the object-based representation

of $S_{Beh}(AT_i)$, and into a set of related event objects representing $E_{Beh}(AT_i)$.

Finally, the *Simulation* phase of AS_{DSC} is supported by a Java-based discrete event simulation framework for distributed agent systems. The framework provides:

1. *Basic Simulation Objects*:
   - **Agent (Ag):** Represents a stationary or a mobile agent and includes a pair of objects: <MAOId, MAOBehavior>, where MAOId is the unique agent identifier and MAOBehavior is an agent behavior object.
   - **Event (Evt):** Represents the event for intra- and inter-Ags interactions.
   - **AgentServer (AgS):** Represents the agent server hosting **Ag**s.
   - **VirtualNetwork (VN):** Represents the logical network of hosts on which **AgS** are mapped.
   - **UserAgent (UA):** Represents a user, directly connected to an **AgS**, who can create, launch, and interact with **Ag**s.
2. A *Simulation Engine* Enabling:
   - execution of **Ag**s by interleaving their **Evt**s processing;
   - transmission of **Evts** among **Ag**s; and
   - migration of **Ag**s.

On the basis of the framework, a simulator program can be implemented and executed to obtain a ResultSet containing validation traces and performance parameter values. While the validation of agent behaviors and interactions is carried out on execution traces automatically generated, the performance evaluation relies on the specific agent-based system to be analyzed; the performance evaluation parameters are therefore set ad-hoc. The ResultSet can also be used to feed back the high-level modeling and detailed design phases.

A Consumer-Driven Agent-Based E-Marketplace

A consumer-driven e-marketplace is an e-marketplace in which the exchange of goods is driven by the consumers that wish to buy a product. The modeled AEM, inspired by the systems presented in Bredin et al. (1998) and Wang et al. (2002) consists of a set of stationary and mobile agents (see Figure 2) which provides basic services for the searching, buying, selling, and payment of goods.

The identified types of agents are:

- **User Assistant Agent (UAA):** Associated with users and assists them in: (1) looking for a specific product that meets their needs; and (2) buying the product according to a specific buying policy.
- **Access Point Agent (APA):** Represents the entry point of the e-marketplace. It accepts requests for buying a product from a registered UUA.
- **Mobile Consumer Agent (MCA):** Represents an autonomous mobile agent dealing with the searching, contracting, evaluation, and payment of goods.
- **Vendor Agent (VA):** Represents the vendor of specific goods.
- **Yellow Pages Agent (YPA):** Represents the contact point of the distributed Yellow Pages Service (YPS) providing the location of agents selling a given product. The organization of the YPS can be: (1) *Centralized* (C), where each YPA stores a complete list of Vendor Agents; (2) *One Neighbor Federated* (1NF), where each YPA stores a list of VAs and keeps a reference to only another YPA; or (3) *M-Neighbors Federated* (MNF), where each YPA stores a list of VAs and keeps a list of at most M YPAs.
- **Bank Agent (BA):** Represents a reference bank supervising money transactions between MCAs and VAs

The identified types of interactions between the agent types are described below by relating them to the system workflow triggered by a user's request (see Figure 2):

1. **Request Input (UAA → APA):** The UAA sends a request to the APA containing a set of parameters selected by the user for searching and buying the desired product—that is, the product description (*Prod_Desc*), the maximum product price (P_{MAX}) the user is willing to pay, and the type of buying policy (*BP*).
2. **Service Instantiation (APA→ MCA):** The APA creates a specific MCA and provides it with the set of user parameters, the type of searching policy (*SP*), and the location of the initial YPA to be contacted. Upon creation, the MCA moves to the initial YPA location.
3. **Searching (MCA→ YPA):** The MCA requests a list of locations of VAs selling the desired product to the YPA. The YPA replies with a list of VA locations and, possibly, with a list of linked YPA locations.
4. **Contracting & Evaluation (MCA → VA):** The MCA interacts with the found VAs to request an offer (P_{offer}) for the desired product, evaluates the received offers, and selects an offer, if any, for which the price is acceptable (i.e., $P_{offer} <= P_{MAX}$) according to the type of *BP*.
5. **Buying (MCA → VA → BA):** The MCA moves to the location of the selected VA and pays for the desired product using a given amount of e-cash (or bills) triggering the following money transaction: (a) the MCA gives the bills to the VA; (b) the VA sends the bills to a BA; (c) the BA validates the authenticity of the bills, disables them for re-use, and, finally, issues an amount of bills equal to that previously received to the VA; and (d) the VA notifies the MCA.
6. **Result Report (MCA → UAA):** The MCA reports the buying result to the UUA.

Analysis of Mobile Agent-Based Services

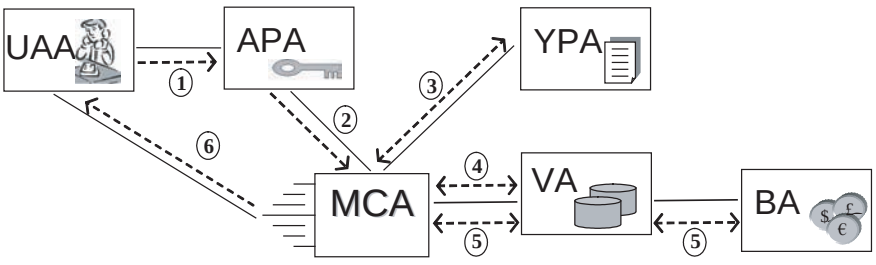
A model of MCA is defined on the basis of the tuple:

<SP, BP, TEM>,

where:

- *SP* is a searching policy in {ALL, PA, OS}:

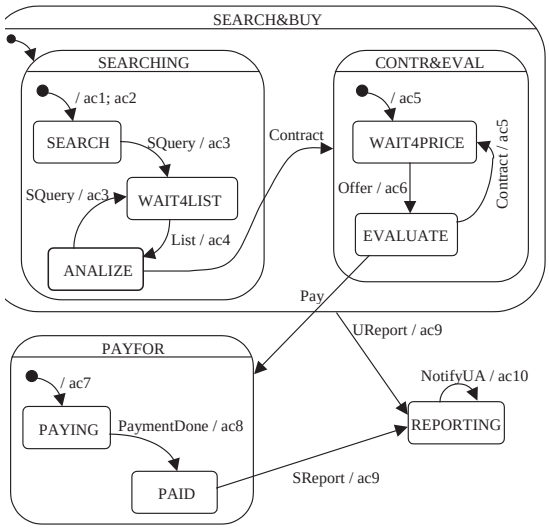
Figure 2. The reference consumer-driven agent-based e-marketplace model: the types of agents, the logical communication links among them, and the sequence of agent interactions



a. ALL: All YPAs are contacted.
b. *Partial* (PA): A subset of YPAs are contacted.
c. *One-Shot* (OS): Only one YPA is contacted.

- *BP* is a buying policy in {MP, FS, FT, RT}:
  a. *Minimum Price* (MP): The MCA first interacts with all the VAs to look for the best price of the desired product; then, it buys the product from the VA offering the best acceptable price.
  b. *First Shot* (FS): The MCA interacts with the VAs until it obtains an offer for the product at an acceptable price; then, it buys the product.
  c. *Fixed Trials* (FT): The MCA interacts with a given number of VAs and buys the product from the VA which offers the best acceptable price.
  d. *Random Trials* (RT): The MCA interacts with a random number of VAs and buys the product from the VA which offers the best acceptable price.
- *TEM* is a task execution model in {ITIN, PAR}:
  a. *Itinerary* (ITIN): The *Searching* and *Contracting & Evaluation* phases are performed by a single MCA which fulfils its task by sequentially moving from one location to another.
  b. *Parallel* (PAR): The *Searching* and *Contracting & Evaluation* phases are performed by a set of mobile agents in a parallel mode. In particular, the MCA is able to generate a set of children (generically called workers) and to dispatch them to different locations; the workers can, in turn, spawn other workers.

Figure 3. DSC-based behavior of the MCA models <,*,ITIN>*



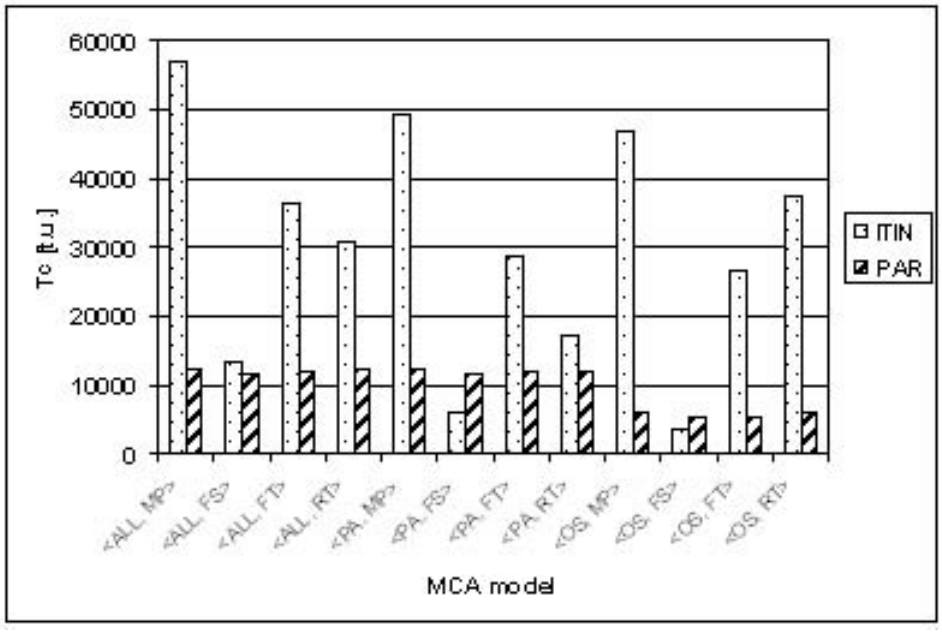

```
ac1 :nextYP=0;
ac2 :YPATarget=(MAOId)yp.elementAt(nextYP++);
     generate(new Move(self(), YPATarget.getCurrLocation()));
     generate(new SQuery(self()));
ac3 :generate(new VAListQuery(self(), YPATarget, vaListQuery));
ac4 :List reply = (List)mevent;
     Proc proc = processYPAReply(reply);
     if (proc.continueSearching()) ac2();
     else if (proc.noVendors()) sa2();
     else { nextVA=0; sa1(); }
sa1 :VATarget=(MAOId)va.elementAt(nextVA++);
     generate(new Move(self(), VATarget.getCurrLocation()));
     generate(new Contract(self()));
sa2 :generate(new UReport(self()));
ac5 :generate(new PriceQuery (self(), VATarget, priceQuery));
ac6 :Offer offer = (Offer)mevent;
     Eval eval = evaluateVAOffer(offer);
     if (eval.buySoon()) generate(new Pay(self()));
     else if (eval.moveAndBuy()){
      generate(new MAOMove(self(), VATarget.getCurrLocation()));
      generate(new Pay(self()));
     }
     else if (eval.noBuy()) sa2();
     else sa1();
ac7 :bills = prepareBills(price);
     generate(new PayFor(self(), VATarget, bills);
ac8 :nbills = nbills - eval.price();
     generate(new SReport(self()));
ac9 :generate(new Move(self(), self().getHomeLocation()));
     generate(new NotifyUA(self()));
ac10:reportTR();
```

Thus, each one of the defined models implements the product buying service differently. Figure 3 shows the generic DSC-based behavior of the MCA models <*,*,ITIN>. For the sake of brevity, the explanation is not given here, but readers unfamiliar with DSC-based programming can refer to Fortino et al. (2004b).

In order to analyze and compare the MCA models, the Task Completion Time (T_{TC}) parameter was defined as follows: $T_{TC}=T_{CREATION}-T_{REPORT}$ where, $T_{CREATION}$ is the creation time of the MCA and T_{REPORT} is the reception time of the MCA report. Accordingly, a simulator program was implemented which allows for computation of T_{TC} for each MCA model by varying the *Yellow Pages* organization, the number of YPAs (N_{YPA}), and the number of VAs (N_{VA}). In particular, the simulation scenario was set up as follows:

- Each stationary agent (UAA, APA, YPA, VA, BA) executes in a different agent server.
- Agent servers are mapped onto different network nodes which are completely connected through links having the same characteristics. The communication delay (δ) on a network link is modeled as a lognormally distributed random variable with a mean, μ, and a standard deviation, σ (Floyd & Paxson, 2001).
- Each UAA is connected to only one APA.
- The price of a product, which is uniformly distributed between a minimum (PP_{MIN}) and a maximum (PP_{MAX}) price, is set in each VA at initialization time and is never changed; thus the VAs adopt a fixed-pricing policy to sell products.

Figure 4. Task completion time of the MCA models in an e-marketplace with $N_{YPA}=10$, $N_{VA}=80$, and YPS=2NFBT



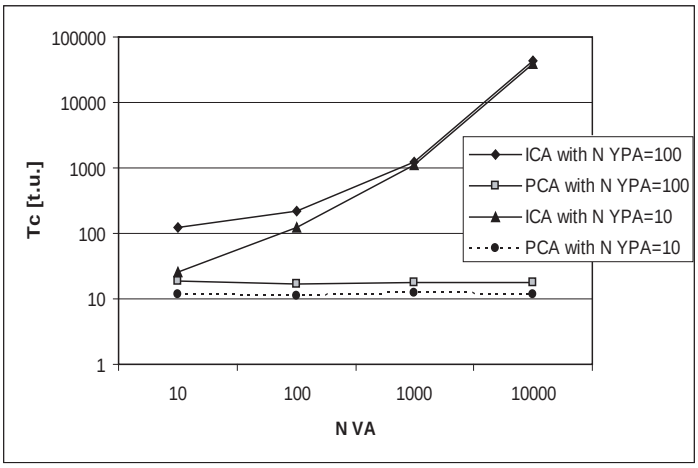
- Each YPA manages a list of locations of VAs selling available products.
- An UAA searches for a desired product, which always exists in the e-marketplace, and is willing to pay a price P_{MAX} for the desired product which can be any value uniformly distributed between PP_{MAX} and $(PP_{MAX}+PP_{MIN})/2$.

Simulations were run by varying (1) the organization of the Yellow Pages (C, 1NF and 2NF organized as a binary tree or 2NFBT), (2) the number of YPA agents in the range [10..1000], and (3) the number of VA agents in the range [10..10000]. These ranges were chosen for accommodating small as well as large e-marketplaces. The duration of the performed simulations were set so to allow for the completion of the buying task carried out by the MCA.

Figure 4 shows the T_{TC} of the <*,*,ITIN> and <*,*,PAR> models in a *medium-sized* e-marketplace with N_{YPA}=10, N_{VA}=80, and YPS=2NFBT. The lowest-performance model is the <ALL,MP,ITIN> model. The <ALL,MP,*> models are the only models guaranteeing both a successful purchase and the best purchase since they are able to find the VA offering the minimum price. The <*,*,PAR> models always outperform the <*,*,ITIN> models, but the <*,FS,*> models where the <*,FS,ITIN> models perform similarly or slightly better than the <*,FS,PAR> models. However, in the latter case, purchase of the desired product at the best price is not guaranteed.

In order to compare the performances of PCA (Parallel Consumer Agent) and ICA (Itinerary Consumer Agent) models, the results obtained for the <ALL, MP, *> MCA models adopting a YPA organization of the 2NFBT type are reported in Figure 5, where results were obtained setting N_{YPA} ={10, 100} and varying N_{VA}. In agreement with the analytical model reported in Wang et al. (2002), the PCA, due to its parallel dispatching mechanism, outperforms the ICA when N_{VA} and N_{YPA} are increased.

*Figure 5. Performance evaluation of the <ALL, MP, * > models for an e-marketplace with YPS=2NFBT, N_{YPA} ={10, 100}, and variable N_{VA}*



FUTURE TRENDS

To date, *Agents* have been employed primarily for product and merchant discovery and brokering (Sierra & Dignum, 2001). The next stage will involve moving into real trading, which will require considerable research and development efforts, including the definition, implementation, and notably, analysis of new products and services such as market-specific shells, payment and contracting methods, risk assessment and coverage, quality and performance certification, security, and trust management.

Moreover, in the very near future, a rapid growth in agent-mediated auctions is expected. Auction is a long-established and well-understood trading mechanism, and the agent technology can be used to develop and support agent-mediated auction houses (Luck et al., 2004).

In order to test these new trading and auction services within large-scale MAS, discrete-event simulation seems to be the most appropriate and reliable tool. Therefore, flexible and robust agent-oriented, discrete-event simulation frameworks must be carefully designed and developed to support analysis of MAS at different levels of granularity: from agent behaviors, protocols, and services (*micro-level*) to global MAS behavior (*macro-level*).

CONCLUSION

Using *agents* to support e-commerce (both B2C and B2B) is considered a key challenge for the agent community. This article has presented an integrated approach which effectively models and analyzes e-commerce services based on *agents*. In particular, a consumer-driven AEM was modeled, and the product searching and buying strategies carried out by the mobile consumer agents in this AEM were analyzed. The consumer-driven AEM model used here was derived from real systems using agent-based e-marketplaces on the Internet (Bredin et al., 1998; Wang et al., 2002). In line with the *future trends* which have been delineated, the proposed approach is being applied to the modeling and analysis of more complex AEMs and related services, and also enhanced by exploiting game theory and economics models.

REFERENCES

Bredin, J., Kotz, D., & Rus, D. (1998, May). Market-based resource control for mobile agents. *Proceedings of ACM Autonomous Agents.*

ebXML. (2005). *Specifications documents.* Retrieved from http://www.ebxml.org

Floyd, S., & Paxson, V. (2001). Difficulties in simulating the Internet. *IEEE/ACM Transactions on Networking, 9*(4), 392-403.

Fortino, G., Garro, A., & Russo, W. (2004a). From modelling to simulation of multi-agent systems: An integrated approach and a case study. *Multiagent System Technology (MATES).* Berlin: Springer-Verlag (LNAI 3187).

Fortino, G., Garro, A., & Russo, W. (2005). Modelling and analysis of agent-based electronic marketplaces. *IPSI Transactions on Internet Research, 1*(1), 24-33.

Fortino, G., & Russo, W. (2005, March 13-17). Multi-coordination of mobile agents: A model and a component-based architecture. *Proceedings of the ACM Symposium on Applied Computing, Special Track on Coordination Models, Languages and Applications,* Santa Fe, NM.

Fortino, G., Russo, W., & Zimeo, E. (2004b). A statecharts-based software development process for mobile agents. *Information and Software Technology, 46*(13), 907-921.

Griss, M., & Letsinger, R. (2000, June). Games at work—Agent-mediated e-commerce simulation. *Proceedings of the ACM Conference on Autonomous Agents,* Barcelona, Spain.

Harel, D., & Gery, E. (1997). Executable object modelling with statecharts. *IEEE Computer, 30*(7), 31-42.

Karnik, N. M., & Triphati, A. R. (1998). Design issues in mobile-agent programming systems. *IEEE Concurrency, 6*(3), 52-61.

Kowalczyk, R., Ulieru, M., & Unland, R. (2003, October 7-10). Integrating mobile and intelligent agents in advanced e-commerce: A survey. In R. Kowalczyk, J. P. Müller, H. Tianfield, & R. Unland (Eds.), *Proceedings of the Agent Technologies, Infrastructures, Tools, and Applications for E-Services, NODe 2002 Agent-Related Workshops*, Erfurt, Germany. Berlin: Springer-Verlag (LNCS 2592).

Lange, D. B., & Oshima, M. (1999). Seven good reasons for mobile agents. *Communications of the ACM, 42*(3), 88-89.

Luck, M., McBurney, P., & Preist, C. (2004). A manifesto for agent technology: Towards next generation computing. *Autonomous Agents and Multi-Agent Systems, 9*(3), 203-252.

Maes, P., Guttman, R. H., & Moukas, A. (1999). Agents that buy and sell: Transforming commerce as we know it. *Communications of the ACM, 42*(3), 81-91.

Medjahed, B., Benatallah, B., Bouguettaya, A., Ngu, A. H. H., & Elmagarmid, A. K. (2003). Business-to-business interactions: Issues and enabling technologies. *The VLDB Journal, 12,* 59-85.

Sierra, C., & Dignum, F. (2001). *Agent-mediated electronic commerce: The European AgentLink perspective*. Berlin: Springer-Verlag (LNAI 1991).

Wang, Y., Tan, K. L., & Ren, J. (2002). A study of building Internet marketplaces on the basis of mobile agents for parallel processing. *World Wide Web: Internet and Web Information Systems, 5,* 41-66.

Wooldridge, M., Jennings, N. R., & Kinny, D. (2000). The Gaia methodology for agent-oriented analysis and design. *Journal of Autonomous Agents and Multi-Agent Systems, 3*(3), 285-312.

KEY TERMS

AEM: An agent-based e-marketplace is a distributed multi-agent system formed by stationary and mobile agents which provide e-commerce services to end-users within a business context.

Agent: In a broad sense, any program that acts on behalf of a (human) user.

AMEC: Agent-mediated e-commerce is concerned with providing agent-based solutions which support different stages of the trading processes in e-commerce, including needs identification, product brokering, merchant brokering, contract negotiation and agreement, payment and delivery, and service and evaluation.

Distilled StateCharts: A statecharts-based formalism for lightweight mobile agents.

ebXML: Electronic business XML is an XML-based language and infrastructure which aims at enabling B2B interactions among companies of any size.

E-Marketplace: An electronic marketplace is an e-commerce environment which offers new channels and business models for buyers and sellers to trade goods and services over the Internet.

Mobile Agent: A program that represents a user in a computer network and can migrate autonomously from node to node, to perform a computation on behalf of the user.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 319-326, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.40
B-POS Secure Mobile Payment System

Antonio Grillo
Universita di Roma "Tor Vergata", Italy

Alessandro Lentini
Universita di Roma "Tor Vergata", Italy

Gianluigi Me
Universita di Roma "Tor Vergata", Italy

INTRODUCTION

The B-POS (Bluetooth Point of Sale) is the prototype of a secure, mobile macropayment system. Since heterogeneous wireless network technologies such as PANs, LANs, and WANs have well-known security weaknesses, it is mandatory to enforce security services, such as authentication, confidentiality, integrity, and non-repudiation. This article describes a Java-based macropayment system prototype featuring security and independence from an e-money third party acting as an intermediary. This system can rely on the existing financial network infrastructure (e.g., credit card, ATM networks).

BACKGROUND

Currently, most of m-payment (payment performed with a mobile device) systems rely upon mobile WAN (wide area network, e.g., GSM/GPRS/UMTS), enabling the customer to buy contents (by mobile carrier) or goods billed to the mobile phone contract account or to a prepaid card (in a business model called "walled garden"). The widespread diffusion of Bluetooth-enabled mobile phones, however, can possibly boost the deployment of new application paradigms based on personal area networks (PANs) and NFC (near field communications), enlarging the payment paradigm to financial and banking systems and circuits (e.g., EFC —electronic financial circuits),

so achieving two major benefits: (1) to escape from the "walled garden," so acquiring the capability to buy every good and every service, not only those from the mobile carrier; and (2) collecting the payment capabilities in a personal trusted device (PTD, e.g., the smartphone) without dealing with the ATM/credit cards in the wallets, supporting both micropayments and macropayments (Me, 2003).

There has been a considerable amount of research focusing on the adoption of mobile payments using a POS (Me & Schuster, 2005). Most of the research effort on usability led to description of the adoption factors influencing the consumer in the adoption of the payment solution (Dahlberg, Mallat, & Oorni, 2003; Mallat, 2004; Pousttchi, 2003; Zmijewska, Lawrence, & Steele, 2004). Other research has focused on finding the most critical factor of success and the different requirements of mobile payment systems (Hort, Gross, & Fleisch, 2002; Muller, Lampe, & Fleisch, 2004). Many more studies focused on the adoption intentions of the consumers and the merchants toward a new electronic payment system (Plouffe & Vandenbosch, 2001). Early local mobile payment systems (cash like, micropayments) were pioneered by Chaum CAFÉ-IR (www.chaum.com/CAFE_Project.htm) based on public-key encryption and the blind signature scheme of Ecash: this system uses a smartcard and an electronic prepaid "wallet" to complete transactions via the InfraRed technology. The work of Blaze, Ioannidis, and Keromytis (2001) on microchecks (over the InfraRed links) is (somewhat) similar to ours, except for (at least) its minor concerns with fraudulent transactions (due to their small amount). Another important difference is that the payer is not required to authenticate the merchant during a transaction. Several other e-check systems have been implemented during recent years, but they were never customized for mobile/local transactions (e.g., SET, echeque, First Virtual). The foremost e-check system, Kerberos based, was NetCheque (http://gost.isi.edu/info/NetCheque/).

Currently, several countries have adopted mobile payments: in the Asian market, Singapore, South Korea, and Japan reached an advanced market stage, for example, the FeliCa contactless payment system, currently the de-facto standard method in Japan with over 20 million users, counts over six million FeliCa-enabled handsets and POS installed in all the major shop chains (www.sony.net/Products/felica). Europe is following close behind with successful m-payment services already launched in Austria, Croatia, and Norway, and in Italy, Telecom Italia Lab unveiled in December 2005 a contactless system called Z-SIM, where mobile phones can communicate with any terminal or object by very simple interaction. As a rule, a mobile payment system can be operator independent, where billing is based on an association between a credit card or bank account to the mobile phone (e.g., the Italian major credit card distributor CartaSi has recently launched its own mobile payment system).

MAIN FOCUS OF THE ARTICLE

B-POS aims to be a secure mobile macropayment system for local, contactless, and operator-independent payment systems, involving three different entities—bank, shop, and customer (smartphone)—communicating via secure channels, as shown in Figure 1. Due to well-known mobile vulnerabilities, especially regarding Bluetooth (Nichols & Lekkas, 2001, 402-415; Jacobson & Wetzel, 2001), it is mandatory to enforce security, firstly on wireless links. For this reason, the requirement of macropayment system security is met at the application layer, avoiding various communication layer vulnerabilities, (e.g. E3—electromagnetic environmental effects) or new, unpredictable vulnerabilities (e.g., wireless transport layer security (WTLS) gap in versions

prior to WAP 2.0). This task is performed using an asymmetric keys schema for the authentication with a derived symmetric session key. A mutual authentication is implemented among B-POS's communication parties, so the entities involved can trust each other.

The Bluetooth link connects the smartphone to the shop (1). Information exchanged between the shop and the bank happens on a secure channel. The customer refers to his own bank through a virtual private network (VPN), taking advantage of both these channels (3).

Since the end user trust should be placed into a customer-reliable authority, we centralized all the responsibilities into the bank, whose former task was to release and setup the BPOS mobile application on the customer device.

In order to achieve a widespread diffusion between customers, the prototype is suited for devices as PDAs and smartphones equipped with Kilo Virtual Machine (KVM). Former benefits in adoption of J2ME reside in advantages to use cross-platform code and embedded security mechanisms (safe box). Several further considerations suggest the J2ME platform as appropriate to support m-payments (Sun Microsystems, 2000a; Cervera, 2002):

- **Broad user experience:** The J2ME API provides enhanced possibilities to present GUI, for example, event handling and rich graphics.
- **Comprehensiveness:** The details of the machine architecture, operating system, and display environment are all handled transparently by the Java virtual machine (JVM). The same Mobile Information Device Profile (MIDP) m-payment client can run on all the MIDP-compliant devices (Sun Microsystems, 2000a, 2000b; Cervera, 2002). This allows the m-payment system providers to target a wider range of end-users.
- **Reduced network and server load:** The J2ME-based applications can operate when disconnected, and they only interact with a server when necessary. J2ME has its own runtime environment and the capability to store data in the mobile device.

The ease of use and the cheapness of further requirements led to a wireless point-to-point communication between the shop and the customer. For this reason, Bluetooth technology represents the most suitable alternative due to its limited range and the mobile phone market penetration (in Europe).

Money or virtual checks are not exchanged between entities, thus reducing fraud and stealing due to eavesdroppers. BPOS payments are performed in the bank context, where a customer authorizes the bank to commit a fund transfer from his or her personal bank account to the shop's.

Figure 1. BPOS architecture

Application's Schema Description

We divide the whole transaction in different phases (see Figure 2). Firstly, a former procedure is required, where the user physically interacts with the bank/CA, in order to install on his smartphone the BPOS application. He could request for this service from his trusted bank where he owns an account.

Setup Procedure

The BPOS Bank's portion code creates a symmetric key associated to a specific user that had requested the BPOS service. The key, stored into the BPOS application, will be installed on the user's personal device. The bank asks the user to turn on the application, forcing the user to select an application PIN, used to protect the system from unauthorized spawn of the application. Furthermore the customization of the application, needed for the individual payment, includes the storage of a specific certificate and an asymmetric key pair on the user device: this operation means that bank plays the role of CA. For further transactions, mutual authentication between the user personal device and the bank is no longer needed. Moreover the setup of the VPN guarantees the confidentiality of the exchanged information, avoiding spoofing attacks and repudiability, while a control acknowledge message prevents replay attacks.

Figure 2. BPOS application's schema

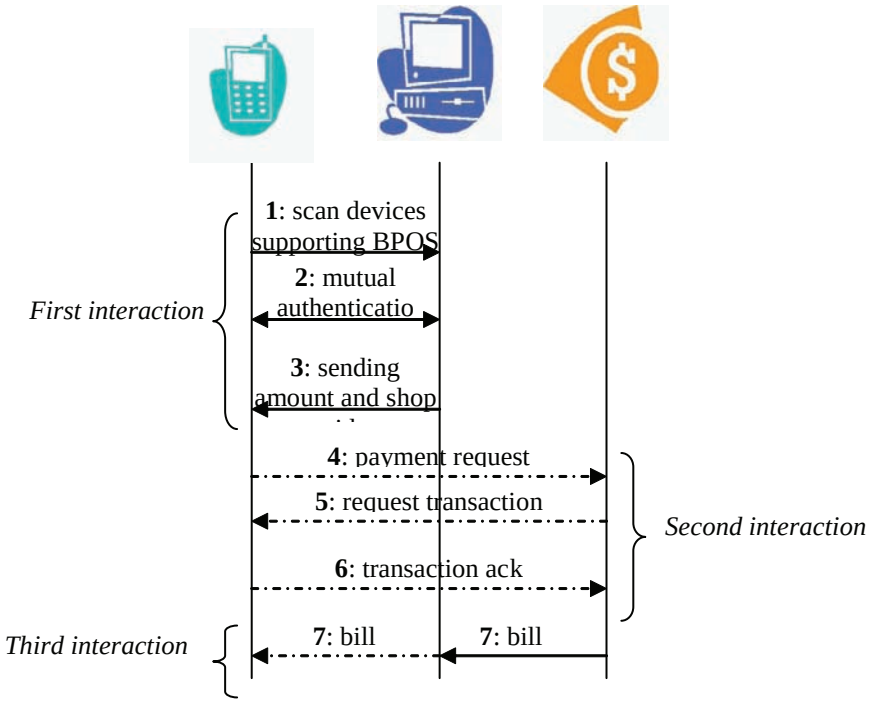


Figure 2 shows the application schema, enlightening three different interactions between the involved entities:

1. smartphone-shop (via Bluetooth),
2. smartphone-bank (via VPN), and
3. bank-shop-smartphone.

First interaction takes place on a Bluetooth connection and consists of three sub-steps: device discovery, (1) mutual authentication (2), and the payment amount dispatch (3). In the former phase, the user runs the BPOS application on his device, inserts his PIN, and starts the scan for devices. The scanning phase is filtered by the service presence in order to identify only the BPOS partner, then the user selects the shop where he is purchasing in a proposed list of shops.

The second sub-step of the user interaction with *shop* is the mutual authentication, achieving two different goals:

1. shop-customer, ensuring to the customer that the shop is the place where he is purchasing; and
2. customer-shop, ensuring to the shop that the ticket is addressed to the customer required to pay.

Table 1. Notation

Mutual Authentication Protocol:

i. $A \rightarrow B: Cert_A$
ii. $A \leftarrow B: Cert_B , PK_A(C_B , R_B)$
iii. $A \rightarrow B: SK_A(C_B) , PK_B(C_A , R_A)$
iv. $A \leftarrow B: SK_B(C_A)$
v. $A \rightarrow B: AS$

Symbol	**Description**
A	*Smartphone*
B	*Shop*
$Cert_A$	*Certificate of A*
$Cert_B$	*Certificate of B*
$PK_A(m)$	*Message "m" encrypted with the Public Key of A*
$PK_B(m)$	*Message "m" encrypted with the Public Key of B*
$SK_A(m)$	*Message "m" encrypted with the Private Key of A*
$SK_B(m)$	*Message "m" encrypted with the Private Key of B*
C_A	*Challenge of A for B*
C_B	*Challenge of B for A*
R_A	*Random number of A*
R_B	*Random number of B*
AS	*Arrangement of session key with HMAC-SHA256(R_A, R_B)*

The latter authentication side is important because the ticket is the only method to identify the current paying customer.

After the successful mutual authentication phase, the shop (payer) and the smartphone (payee) agree on a session key, used by the shop to send to the client:

- the amount to be paid, and
- its unique ID registered to the bank.

Second interaction involves the user and the bank over a VPN (see Figure 2), via a three-step communication: smartphone sending request for payment (4), bank sending request for the transaction confirm (5), and smartphone transaction acknowledge (6).

The smartphone needs to send a request to the bank for payment, so it creates a packet containing:

- customer identification number, registered to the bank (not encrypted); and
- an embedded packet encrypted with the symmetric key shared with the bank containing:
  - shop ID registered to the bank, received from the shop in the previous step (Figure 2, step 3);
  - amount to be paid received from the shop in the previous step (Figure 2, step 3);
  - a nonce X.

Moreover, the bank, after processing the information received, accesses its database in order to find information about the entities involved in the transaction, and consequently sends a request for transaction confirm to the smartphone via the established VPN. Hence, the bank translates information recovered in a new packet and sends it back to the smartphone. Finally, the smartphone/customer sends a transaction acknowledge to the bank. The customer views all the previous information on the device's display, so he can decide whether to confirm or not the requested transaction. This step is crucial to unmask a malicious subject pretending to be the genuine shop. In order to avoid replay attacks, the protocol uses a transaction Ack to keep the session state. Ack is computed using random X combined by cod.op to random Y (these are produced in steps 4 and 5).

Third interaction involved the bank sending the transaction status (a virtual ticket) to other entities, which protects the latter one by the use of the VPN.

Technical Features

Since B-POS fulfills the security as central requirement, the system is based on a large, careful use of cryptographic technique. Due to the hardware device restriction, the crypto-primitives implementation relied on the BouncyCastle APIs (http://www.bouncycastle.org, http://java.sun.com/products/javacomm/index.jsp). The Elliptic Curve Integrated Encryption Scheme (ECIES) algorithm, with a 192-bit elliptic curve (EC), provides support for asymmetric key needs and symmetric encryption/decryption, and ensures packet integrity between the shop and the customer. The ECIES combines the EC asymmetric encryption with the Advanced Encryption Standard (AES, 128-bit key) and adds a SHA-1 hash for message authentication. In comparison to a 1024-bit RSA key, the ECC (Elliptic Curve Cryptography) provides shorter keys, shorter encrypted messages, and faster private key operations. The system relies on the Elliptic Curve Digital Signature Algorithm (ECDSA) to sign the transmitted data with additional 48-byte signature value, thus avoiding its unnoticed modification. In this PKI, the bank entity produces the certificate and the asymmetric keys for other system entities. The X.509 certificate stored on the mobile devices and

shop is signed by a 1024-bit RSA public key of the bank, embedded in a PKCS12 envelop, which protects this information with a password. The mutual authentication phase between the shop and the smartphone could be performed in a secure way, exchanging the certificates guaranteed from the CA.

In a second phase of the application schema, the communications between the authenticated entities are protected by a session key. This key is derived from two random numbers R, derived from the challenge response authentication phase relying on a HMAC-SHA256(R1,R2) hash function. The message integrity is ensured via AES and a SHA-1 hash.

FUTURE TRENDS

The BPOS is suitable for the end user because of its ease of use, security, and independence from third parties; the overall time to complete the transaction is upper bounded at an acceptable two minutes. Further studies and field trials have to be carried on in environments with multiple Bluetooth devices running, since Bluetooth inquiry performances are currently not acceptable in terms of reliability of service.

CONCLUSION

In this article we proposed a local mobile payment system with respect to:

- **Security:** Communication sessions between principals are based on well-understood cryptographic techniques.
- **Traceability:** Transactions are not kept anonymous.
- **Usability and convenience:** Costs of deployment and management for all principals involved (consumers, merchants, and banks) are acceptable.
- **Portability and interoperability:** System design is based on "de facto" standards (hardware equipment and software modules).

In particular, the system does not rely on a third party because it makes use of the existing bank network infrastructure. This assumption, together with use of open software on smartphones, minimizes the impact (and costs) on merchants, customers, and banks.

Since current trends in mobile phone technology move towards a direction of miniaturization and higher computational and graphical performance, allowing the completion of the whole payment procedure in less than a minute, we believe that the BPOS prototype can help the development of new forms of payments, local payments, with very low impact on customers (payers), due to widespread diffusion of mobile phone and shops, allowing their reuse in the same financial network.

The final advantage consists of the avoidance of all the fraudulent activities dealing with ATM/credit cards.

REFERENCES

Blaze, M., Ioannidis, J., & Keromytis, A.D. (2001). Offline micropayments without trusted hardware. *Proceedings of Financial Cryptography 2001.* Retrieved from www.crypto.com/papers/knpay.pdf

Cervera, A. (2002). *Analysis of J2ME™ for developing mobile payment systems.* Retrieved from http://www.microjava.com/articles/Final01092002.pdf

Dahlberg, T., Mallat, N., & Oorni, A. (2003). Trust enhanced technology acceptance model: Consumer acceptance of mobile payment solutions. *Proceedings of the Stockholm Mobility Roundtable 2003.*

Hort, C., Gross, S., & Fleisch, E. (2002). *Critical success factors of mobile payment.* Technical Report, M-Lab.

Jacobson, M., & Wetzel, M. (2001). Security weaknesses in Bluetooth. *Topics in Cryptology-CT-RSA* (pp. 176-191). Retrieved from http://www.belllabs.com/user/markusj/bluetooth.pdf

Mallat, N. (2004). Theoretical constructs of mobile payment adoption. *IRIS27.*

Me, G., & Schuster, A. (2005, Fall). A secure and reliable local payment system. *Proceedings of IEEE VTC.*

Me, G. (2003). Security overview for m-payed virtual ticketing. *Proceedings of the 14th IEEE PIMRC* (pp. 844-848).

Muller, G.S., Lampe, M., & Fleisch, E. (2004). Requirements and technologies for ubiquitous payment. *Proceedings of Multikonferenz Wirtschaftsinformatik, Techniques and Applications for Mobile Commerce.*

Nichols, R.K., & Lekkas, P.C. (2001). *Wireless security: Model, threats and solutions* (pp. 402-415). New York: McGraw-Hill Telecom.

Plouffe, C.R., & Vandenbosch, M. (2001). Intermediating technologies and multi-group adoption: A comparison of consumer and merchant adoption intentions toward a new electronic payment system. *The Journal of Product Innovation Management, 18*(2), 65-81.

Pousttchi, K. (2003). Conditions for acceptance and usage of mobile payment procedures. *Proceedings of the 2nd International Conference on Mobile Business* (pp. 20-210).

Sun Microsystems. (2000a). *Connected, Limited Device Configuration (CLDC) specification, version 1.*

Sun Microsystems. (2000b). *Mobile Information Device Profile (MIDP) specification, version 1* (pp. 551-561).

Zmijewska, A., Lawrence, E., & Steele, R. (2004). Towards understanding of factors influencing user acceptance of mobile payment systems. *Proceedings of IADIS WWW/Internet 2004,* Madrid.

KEY TERMS

Bluetooth: A standard for cable-free, short-range connectivity between mobile phones, mobile PCs, handheld computers, and other peripherals developed by Bluetooth Special Interest Group (Ericsson, IBM, Intel, Nokia, and Toshiba Consortium, http://www.bluetooth.com). It uses short-range, low-power radio links in the 2.4 GHz Instrumentation Scientific and Medical (ISM) band.

Elliptic Curve Integrated Encryption Scheme (ECIES): One of the most popular ECC (Elliptic Curve Cryptography) schemes, defined in *ANSI X9.63-2002, Public Key Cryptography for the Financial Services Industry: Key Agreement and Key Transport Using Elliptic Curve Cryptography* and *IEEE P1363a: Standard Specifications for Public Key Cryptography: Additional Techniques (draft).* ECC is based on arithmetic using elliptic curves, providing shorter key lengths, and under some conditions, improved performance over systems based on integer factorization and discrete logarithms.

Java2 Micro Edition (J2ME): A version of Java used to develop applications running on a consumer wireless device platform (like a smartphone or PDA). A MIDlet is a J2ME™ application, and the KVM Kilo Virtual Machine is the java virtual machine for mobile devices.

Macro/Micro-Payment: An electronic payment can be loosely categorized according to the amount of money conveyed from the payer (customer) to the payee (merchant): *macropayments* for transaction volumes exceeding $10 (10€) , *micropayments* for amounts less than $10 (10€).

Mobile Payment: An electronic payment can be defined as "the transfer of electronic means of payment from the payer to the payee through the use of an electronic payment instrument," where the electronic payment instrument is viewed as "a payment instrument where the forms are represented electronically and the processes to change the ownership of the means of payment are electronic" (Mobile Payment Forum, 2002). Mobile payments are a subset of electronic payments, where the payee device performs at least one payment step via a wireless link.

Point of Sale (POS): An electronic device used by a retail business (payer) to process credit/debit card transactions. The payee swipes or slides his credit/debit card through the machine, or in mobile POS, connects through radio network instead of swiping the card.

Public Key Cryptography Standards (PKCS12): A series of documents, published by RSA Data Security, defining standard protocols to enable compatibility among public key cryptography implementations. PKCS12 defines a file format commonly used to store private keys with accompanying public key certificates protected with a password-based symmetric key.

Smartphone: An electronic handheld device that integrates mobile phone capabilities, personal information management, short-medium range network capability (e.g., RFid, Bluetooth Wi-Fi), and resources to run ad-hoc application in the same device.

Virtual Private Network (VPN): A way to provide a confidential network through a public network. In IP-based networks, VPNs are defined by a network of secure links over a public IP infrastructure, for example, PPTP (Point-to-Point Tunneling Protocol), L2TP (Layer 2 tunneling protocol), and IP Security (IPSec).

X.509: An ITU-T standard for public key infrastructure (PKI). X.509 specifies standard formats for public key certificates and a certification path validation algorithm.

This work was previously published in Encyclopedia of Information Ethics and Security, edited by M. Quigley, pp. 55-61, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.41
Mobile Banking Systems and Technologies

Cheon-Pyo Lee
Mississippi State University, USA

Merrill Warkentin
Mississippi State University, USA

INTRODUCTION

The last decade has witnessed the rapid growth of mobile communication devices and wireless technologies across the globe. The convergence of mobile devices and wireless technologies has not only changed the way many activities are conducted, but has also provided a foundation for a new type of technology-aided commerce called mobile commerce (m-commerce). As e-commerce's next evolutionary stage, m-commerce opens up new business opportunities in business-to-consumer (B2C) markets in addition to extending current operations in e-commerce and traditional brick-and-mortar businesses (Varshney & Vetter, 2002). The significant power of m-commerce is primarily a result of the anytime-anywhere connectivity of wireless devices, which provides unique experiences and services (Figge, 2004; Zwass, 2003).

One of the most promising and value-added m-commerce services is mobile banking (Lee, McGoldrick, Keeling, & Doherty, 2003; Mallat, Rossi, & Tuunainen, 2004). Mobile banking is the newest electronic delivery channel to be offered by banks in which technology has become an increasingly vital element, and it provides convenience and enhanced value to both banks and customers. With its clear benefits, mobile banking is now gaining rapid popularity in European and Asian countries with the significant market penetration of mobile handsets and the optimally designed marketing tactics of service providers (Suoranta & Mattila, 2004). However, mobile banking is still marginally adopted across the globe, and especially in the U.S., the growth appears much slower than anticipated (Mallat et al., 2004). In the United States, there are only a small number of banks that have actually introduced mobile banking services, and most other

mobile banking efforts are in small-scale trials (Charny, 2001). Therefore, the technology which will be employed in the United States market has been of interest not only to financial institutions, but also to mobile technology developers and future users.

BACKGROUND

M-commerce is defined as any transaction with a monetary value—either direct or indirect—that is conducted over a wireless telecommunication network (Barnes, 2002). However, there is no clear definition for mobile banking services, so often the traditional banking services using mobile handsets (i.e., making transaction by calling a call center using a mobile phone) are considered as mobile banking services. Thus, it is very important to define a clear boundary of mobile banking service to avoid confusion. In this article, mobile banking refers to a client-server system that is specifically designed for mobile devices, allowing banking customers to use handheld devices to access their accounts, pay bills, authorize fund transfers, or perform other activities. Table 1 shows various mobile banking services currently provided.

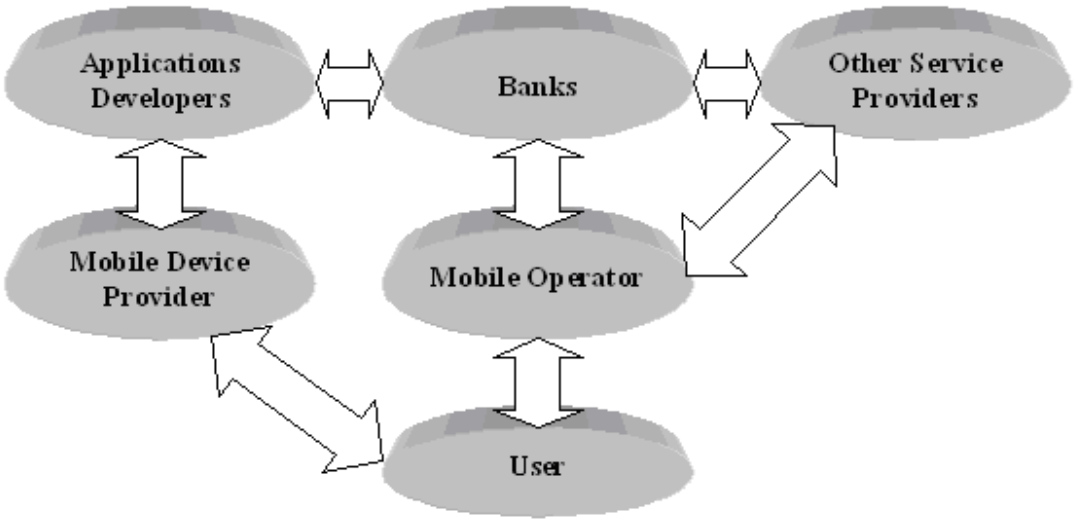
Mobile banking has two big advantages over the narrow sense of e-banking: security and convenience (Herzberg, 2003). E-banking is based on account-holder authentication by the payment system which can fail in multiple ways but do not distinguish the source of fraud. However, mobile devices, usually with a built-in display and keyboard, are well positioned to provide a technical solution for reducing fraud and allowing the fair allocation of responsibility for damages from fraud. In addition, unlike e-banking, the transactions through mobile banking can be made anywhere whether on foot or in cars, planes, or trains.

Mobile banking services began in 1999 in European and Asian countries, and have gained rapid popularity with the significant market penetration of mobile phones, the optimally designed marketing tactics of service providers, and the increased exposure to mobile technology (Suoranta & Mattila, 2004). Like many other mobile commerce services, mobile banking services are provided by several different entities, with which the customers of mobile banking services must interact to complete a successful mobile banking transaction, especially the mobile device provider, mobile operator, and content provider (Varshney & Vetter, 2002; see Figure 1). Thus, each entity in the mobile banking cycle must assist the others to attract more customers to mobile banking. The fastest way to promote the growth of mobile

Table 1. Mobile banking services

• Check the balances of checking and savings accounts, investment account, business banking accounts, lines of credit, credit card accounts, and loan and mortgage accounts
• Electronic funds transfer (EFT)
• Pay bills and taxes
• Request a check book
• Inquire about check status
• Customize the statements according to the user's specific needs and requirements

Figure 1. Mobile banking life cycle (Adapted from Varshney & Vetter, 2002)



banking services is mutual cooperation among the entities (Datta, Pasa, & Schnitker, 2001).

Mobile Devices

The significant potential of mobile banking derives mainly from the fact that the mobile device is a familiar device which is always with the user (Mattila, 2003). Thus, satisfaction with the mobile device is a very significant factor of the mobile banking adoption decision. Mobile device in this article refers to those devices that are used to connect to mobile services (Tarasewich, Nickerson, & Warkentin, 2002). Various mobile devices are available, including mobile phones, personal digital assistants (PDAs), wireless-enabled handheld computers, laptop computers, vehicle-mounted technologies, and personal message pager devices. Among them, wireless-enabled laptops, PDAs, and handsets are currently highly preferred mobile devices for mobile banking (M-Commerce Insider, 2001). Since the use of mobile banking depends on the capabilities of mobile devices, users' satisfaction with various factors of mobile devices significantly influences their adoption of mobile banking. Some of the features of mobile devices which prevent the adoption of mobile banking are small multifunction keypads, less computation power, and limited memory and disk capacity (Siau, Lim, & Shen, 2001).

Mobile Operator

Like many other m-commerce services, mobile banking services are so new that no single company has all the expertise required to develop and deliver compelling services on its own, but many studies point out that mobile operators play a significant role since customers access their networks to perform all transactions (Donegan,

2000; Varshney, 2003). Due to the significant power of mobile operators in mobile banking services, banks often see mobile operators trying to control the financial transaction, and the relationship between banks and the mobile operator is often described as one of mutual distrust (DeZoysa, 2001). The significant power of mobile operators in the m-commerce cycle also imposes important duties to perform in support of m-commerce, such as content provider relationship management, content billing, settlement, and customer care (Buellingen & Woerter, 2004). Thus, the m-commerce users utilize the service of the mobile operator more frequently than that of any other entity, and the service attributes of mobile operators, such as call quality and tariff level, not only influence the satisfaction with the mobile operator, but also influence the adoption of mobile banking services. For instance, current pricing strategies, mainly based on time-usage, of many mobile operators are considered to prevent the mobile users from adopting mobile banking (Yeo & Huang, 2003).

Banks as Content Provider

In mobile banking services, banks are the entity providing content, which is considered one of the most important factors regardless of whether a site is Web based or wireless. Poor quality of content is considered to be significant barrier to m-commerce (Venkatesh, Ramesh, & Massey, 2003). Therefore, users increasingly choose their mobile operators on the basis of the content available, and a majority of mobile network service subscribers are willing to switch mobile operators to get better mobile content (Barnett, Hodges, & Wilshire, 2000). However, the important feature of m-commerce content is that a successful Web interface does not simply translate into a successful mobile interface (Lee & Benbasat, 2004; Venkatesh et al., 2003). Therefore, in addition to accurate and stable content, proper fonts and colors that fit into mobile device screens are also considered significant factors to attract customers to mobile banking services.

MOBILE BANKING TECHNOLOGIES

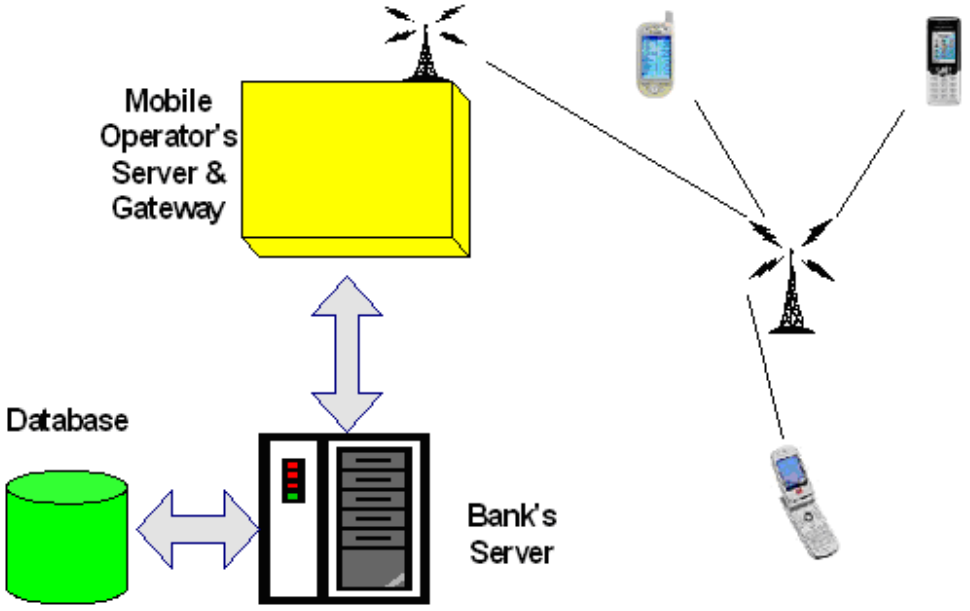
Mobile banking services are conducted by communication between the user's handset and the bank's server (see Figure 2). When the user requests a service, it is transmitted to mobile operator and forwarded to its gateway server. The server later interprets the request and forwards it to the bank's server. Then, the bank's server validates the customer for the mobile number and other information based on its database, and processes a valid request. The result is then passed on to the mobile operator's server which converts the format and transmits it to the user's mobile handset for display.

Like other entities in the mobile banking service cycle, the technologies employed for mobile banking services also play a significant role in mobile banking adoption. Various technologies have been tested and implemented in mobile banking systems, including short message service (SMS), Java, wireless application protocol (WAP), iMode, XHTML, and integrated-circuit (IC) chip. Among them, SMS and WAP have been the most popular technologies for mobile banking systems in Europe and some Asian countries (Marenzi, 2004).

Short Message Service (SMS)

An SMS mobile banking system uses the popular text-messaging service. When the customer requests information by sending an SMS containing a service command to a pre-specified number, the bank responds with a reply SMS containing the specific information. The most important feature of SMS mobile banking systems which attracts many customers is that it can be used conveniently with low cost. In addition, mobile banking with SMS enables every owner of a mobile phone to

Figure 2. How mobile banking services work



use mobile banking services. Therefore, in spite of its significant disadvantages, such as low data throughput, relatively long transaction time, limited service, and difficulty of use, SMS mobile banking services have been a dominant mobile banking technology (Kwon, 2004).

Wireless Application Protocol (WAP)

WAP has been implemented as a mobile banking technology to deliver fast and accurate banking services to customers (Lomax, 2002). WAP supports simple pages with 'menu' structures which are needed for most financial transactions (Economist, 2000). Therefore, it was popularly implemented as a major technology for mobile banking systems. However, even though WAP applications offer a greater graphical look and feel on current mobile phones, only a small percentage of the mobile users have WAP-enabled mobile devices, and only a small proportion of them actually have the knowledge of how to use them (Telephonyworld, 2004). In addition, WAP runs over circuit-switched networks in which the user is charged based on the time they spend online or how long they occupy the circuit (Yan, 2003). Since the initial connection takes a long time and there is a data call charge for browsing the service, WAP-based mobile banking service costs are usually high. Not surprisingly, WAP mobile banking service has proved too complex to use and has been replaced with less advanced technology, SMS, in many places (Zeddies, 2003).

iMode

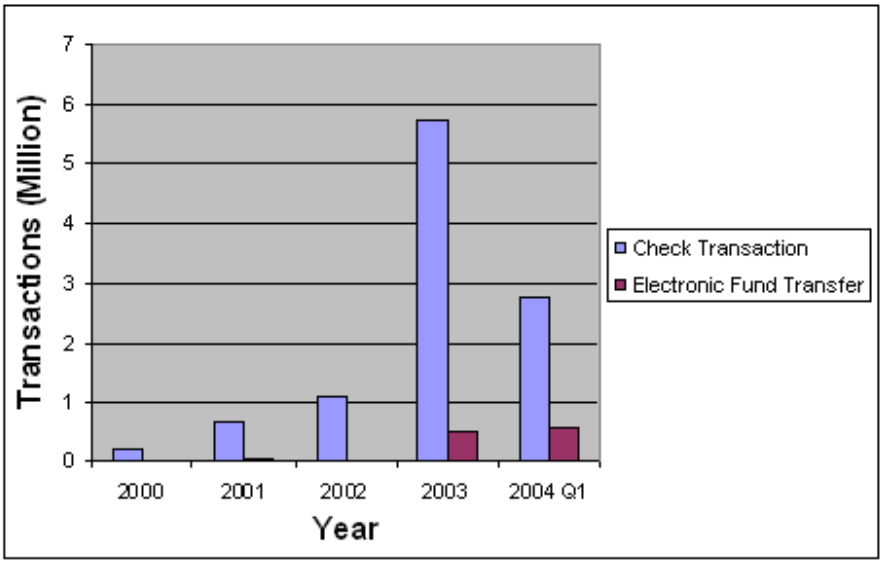
Recently some Western European banks started providing new mobile banking services using

the Japanese iMode technology in the hope of revolutionizing mobile banking (Marenzi, 2004). iMode created a successful business model and gained popularity in Japan. The packet-switching iMode uses the Packet Data Cellular Protocol to speed transmission, so users pay for each packet of data rather than their time (Yan, 2003). Therefore, the cost of using iMode banking is much cheaper than WAP-based mobile banking. However, this popular interactive service neither appeals to European customers nor changes the dominant trend of SMS in Europe and Asian mobile banking markets (Marenzi, 2004).

Integrated-Circuit (IC) Chip

Like in most other countries in Asia and Europe, SMS was also the first technology employed as a mobile banking technology in South Korea, where more than 70% of its 48 million citizens carry one or more mobile handsets and one-third of all mobile phone subscribers use their handsets for m-commerce activities (Kim, 2004). However, unlike most other countries, SMS mobile banking services in South Korea, where well-advanced e-banking services are available, did not attract customers at all and was considered as a failure since not much interest was shown (Kwon, 2004). Since then, various technologies including WAP and virtual machine (VM) have been tested and used to attract customers (Korea, 2004), but customers did not display much interest. However, a new integrated-circuit (IC) chip technology provided by the nation's smallest mobile operator changed the history of mobile banking in South Korea (Kim, 2004). The new chip-based offerings were a big hit, as a total 280,000 people signed up for the new services during the first four months after its introduction in September 2003. Since then, the number of mobile banking users has

Figure 3. The number of mobile banking transactions in South Korea (Source: www.bok.or.kr/contents_admin/info_admin/main/home/financial/payment/material/info/mobile.pdf



dramatically increased and reached 1.1 million as of September 2004 (Korea, 2004; see Figure 3).

The most important feature of IC chip-based mobile banking is that bank account data is encrypted on a smart-card chip, so it enables customers to connect to their account quickly and securely by pressing a single button on their mobile phone. In addition, to prevent someone else from using a mobile phone or IC chip, a special password or code given by banks is required for the user to make each transaction. Another important feature of IC chip-based mobile banking systems in South Korea is that they were introduced in cooperation between banks and mobile operators (Herald, 2004). Therefore, better services, such as relatively low-cost fixed data service (around $7 per month), are offered to users (Kim, 2004). Korea's mobile banking service history clearly shows the significant role of technology implemented in mobile banking services.

FUTURE TRENDS

In the United States, differing technical standards and the 'called-party-pays' system somewhat inhibit the use of mobile phones and mobile commerce (Economist, 2000). However, the market has been dramatically expanded over the last two-and-a-half years. The mobile data services market in the United States has grown from nothing to a market worth an estimated $1.5 billion, and various services have been added in the m-commerce marketplace (Media, 2004). However, the majority of mobile data usage in the United States to date revolves around text messaging and low-value downloads such as ringtones (Media, 2004).

In 2004, the first for individual subscribers in the country, AT&T Wireless Services Inc., an affiliate of Japan's NTT DoCoMo Inc., started offering a mobile phone-based fast data service in the United States, enabling users to browse the Web, e-mail, and share video clips (Wireless, 2005). Sprint, another mobile operator, introduced a new mobile banking service to small and independent business owners with spring mobile loan officer (Sprint, 2004). Thus, though mobile banking services have suffered from customer disinterest, many experts insist that mobile banking will eventually become a significant channel for personal banking. One Atlanta-based bank already reported recent success in mobile banking (RCR, 2004), and a major U.S. bank found that 93% of its customers are very interested in mobile banking services (O'Connell, 2004). All the evidence suggests that if mobile banking services are provided in the proper way, they will soon be a major transaction channel for personal banking.

CONCLUSION

There are three important factors which will decide the success of mobile banking services. First is the technology, *per se*, which when implemented for mobile banking systems will significantly influence the use of mobile banking. The success of South Korea's IC chip-based mobile banking service clearly shows that a good technology ultimately appeals to customers. Secondly, it is important that the technology should be approved before it is implemented by future users. The success of SMS and the failure of WAP demonstrates that the ability to offer simple services done well has proved far more attractive than sophisticated services done poorly (Lomax, 2002). Finally, the service of all entities in a mobile commerce cycle, especially mobile operators and mobile device providers, will significantly influence the mobile banking adoption. Therefore, all the entities in a mobile banking cycle should cooperate and assist each other in order to promote the rapid growth of mobile banking.

REFERENCES

Barnes, S. J. (2002). The mobile commerce value chain: Analysis and future developments. *International Journal of Information Management, 22*(2), 91-108.

Barnett, N., Hodges, S., & Wilshire, M. J. (2000). M-commerce: An operator's manual. *The McKinsey Quarterly, 1*(3), 163-173.

Buellingen, F., & Woerter, M. (2004). Development perspectives, firm strategies and applications in mobile commerce. *Journal of Business Research, 57*(12), 1402-1408.

Charny, B. (2001). *Nokia banks on mobile banking.* Retrieved March 10, 2005, from http://news.zdnet.com/2100-9595_22-276400.html

Datta, A., Pasa, M., & Schnitker, T. (2001). Could mobile banking go global? *The McKinsey Quarterly, 1*(4), 71-80.

DeZoysa, S. (2001). *Who do you trust? Should banks or mobile operators be entrusted with m-commerce security?* Retrieved January 20, 2005, from http://www.telecommagazine.com/default.asp?journalid =2&func=articles&page=0112i13&year=2001&month=12

Donegan, M. P. (2000). *Whose kingdom is it?* Retrieved January 20, 2005, from http://www.telecommagazine.com/default.asp?journalid=2&func=articles&page=0007i10&year=2000&month=7

Economist. (2000). Survey online finance: The other sort of channel conflict. *The Economist, 355*(8171), SO6.

Figge, S. (2004). Situation-dependent services—A challenge for mobile network operators. *Journal of Business Research, 57*(12), 1416-1422.

Herald, K. (2004). *South Korea's mobile banking race accelerates.* Retrieved January 15, 2005, from http://www.epaynews.com

Herzberg, A. (2003). Payments and banking with mobile personal devices. *Communications of the ACM, 46*(5), 53-58.

M-Commerce Insider. (2001). Telephia: Affinity for mobile e-commerce depends on device. *M-Commerce Insider, 2*(6), 1.

Kim, T. (2004). Korea sets trends in global mobile banking. *The Korea Times.* Retrieved February 15, from http://times.hankooki.com/lpage/special/200405/kt200405161 4041511440.htm

Korea, T. B. (2004). *The mobile payment systems in Korea*. Retrieved January 10, 2005, from http://www.bok.or.kr/contents_admin/info_admin/main/home/financial/payment/material/info/mobile.pdf

Kwon, S. H. (2004). *New technology in mobile banking.* Retrieved January 15, 2005, from http://www.etnews.co.kr/news/detail.html?id=200402040104

Lee, M. S. Y., McGoldrick, P. J., Keeling, K. A., & Doherty, J. (2003). Using zmet to explore barriers to the adoption of 3G mobile banking services. *International Journal of Retail & Distribution Management, 31*(6/7), 340-348.

Lee, Y. E., & Benbasat, I. (2004). A framework for the study of customer interface design for mobile commerce. *International Journal of Electronic Commerce, 8*(3), 79-102.

Lomax, V. (2002, December). WAP lash. *Financial World,* 44-49.

Mallat, N., Rossi, M., & Tuunainen, V.K. (2004). Mobile banking services. *Association for Computing Machinery. Communications of the ACM, 47*(5), 42-46.

Marenzi, O. (2004). *Will iMode save mobile banking in western Europe?* Retrieved February 10, 2005, from http://www.celent.com/PressReleases/20031023/MobileEurope.htm

Mattila, M. (2003). Factors affecting the adoption of mobile banking services. *Journal of Internet Banking and Commerce, 8*(1).

Media. (2004). Wireless: Playing catch-up. *New Media Age, 23.*

O'Connell, B. (2004). Wireless banking. *Bank Technology News, 17*(8), 43.

RCR. (2004). Text messaging enables new wireless banking service. *RCR Wireless News, 23*(2), 13.

Siau, K., Lim, E. P., & Shen, Z. (2001). Mobile commerce: Promises, challenges, and research agenda. *Journal of Database Management, 12*(3), 4-13.

Sprint. (2004). Sprint unveils mobile banking services solution. *Telephone IP News, 16*(7).

Suoranta, M., & Mattila, M. (2004). Mobile banking and consumer behavior: New insights into the diffusion pattern. *Journal of Financial Services Marketing, 8*(4), 354-366.

Tarasewich, P., Nickerson, R. C., & Warkentin, M. (2002). Issues in mobile-commerce. *Communications of the Association for Information Systems, 8*, 41-64.

Telephonyworld. (2004). *Zinek's mobile banking system components lead the financial industry.* Retrieved December 15, 2004, from http://www.telephonyworld.com/cgi-bin/news/viewnews.cgi?category=all&id= 1088725646

Varshney, U. (2003). Wireless I: Mobile and wireless information systems: Applications, networks, and research problems. *Communications of the Association for Information Systems, 12*(11), 1-23.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7*(3), 185-198.

Venkatesh, V., Ramesh, V., & Massey, A.P. (2003). Understanding usability in mobile commerce. *Communications of the ACM, 46*(12), 53-56.

Wireless. (2005). Booming next-generation mobile data services market: Anyone's game. *Wireless News, 1.*

Yan, X. (2003). Mobile data communications in China. *Communications of the ACM, 46*(12), 80-85.

Yeo, J., & Huang, W. (2003). Mobile e-commerce outlook. *International Journal of Information & Decision Making, 2*(2), 313-332.

Zeddies, R. (2003). *Secure mobile banking by SMS is starting up in Germany: Replacing WAP banking.* Retrieved Feb 10, 2005, from http://www.intercomms.net/AUG03/content/tecways.php

Zwass, V. (2003). Electronic commerce and organizational innovation: Aspects and opportunities. *International Journal of Electronic Commerce, 7*(3), 7-37.

KEY TERMS

Circuit Switching: A type of communications in which a dedicated channel (or circuit) is established for the duration of a transmission.

Integrated Circuit (IC): A small electronic device made out of a semiconductor material. Integrated circuits are used for a variety of devices, including microprocessors, audio and video equipment, and automobiles.

Mobile Banking: A client-server system that enables banking customers to use handheld devices to access their accounts, pay bills, authorize funds transfers, or perform other activities.

Mobile Commerce: Any transaction with a monetary value—either direct or indirect—that

is conducted over a wireless telecommunication network.

Mobile Services: Services provided by a mobile operator that enable individual customers to access information and applications anytime-anywhere.

Packet Switching: A type of communication in which packets are individually routed between nodes, without a previously established communication path.

Wireless Application Protocol (WAP): A secure specification that allows users to access information instantly via handheld wireless devices such as mobile phones, pagers, two-way radios, smartphones, and communicators.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 754-759, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.42
Mobile Clinical Learning Tools Using Networked Personal Digital Assistants (PDAs)

Bernard Mark Garrett
University of British Columbia, Canada

INTRODUCTION

The School of Nursing at the University of British Columbia has more than 300 nursing students engaged in supervised clinical practice in hospital and community settings around Vancouver. Likewise, the Faculty of Medicine has more than 200 medical students undertaking supervised clinical experience locally and remotely in the Prince George and Vancouver Island regions. The management of these clinical experiences and the promotion of learning while in an active clinical setting is a complex process.

BACKGROUND

Supporting the students at a distance while undertaking their clinical experience is particularly resource-intensive. It requires the creation and maintenance of good communication links with the clinical and administrative staff, active management, clinical visits from faculty, and the provision and management of remotely based resources. However, there were few existing resources that helped to contextualize and embed clinical knowledge in the workplace in the practice setting (Landers, 2000). A technological solution was developed and implemented using several clinical applications designed for use on personal digital assistants (PDAs).

MOBILE CLINICAL LEARNING TOOLS

A suite of PDA-based tools were created for a pilot study with the involvement of nursing and medical students during the academic year of 2004-2005 to achieve the following objectives:

- To demonstrate the potential use of mobile networked technologies to support and improve clinical learning.

- To develop and evaluate a range of mobile PDA tools to promote reflective learning in practice and to engage students in the process of knowledge translation.
- To develop and evaluate a suite of pedagogic tools that help contextualize and embed clinical knowledge while in the workplace.
- To evaluate the value of networked PDA resources to help prevent the isolation of students while engaged in clinical practicum.

The tools developed provide a mobile clinical learning environment incorporating an e-portfolio interface for the Pocket PC/Windows Mobile (Microsoft, 2004) operating system. They were implemented on i-mate PDAs equipped with GSM/GPRS (Global System for Mobile Communications/General Packet Radio Service; GSM World, 2002). This platform offered considerable flexibility for the project. It supported the use of cellular telephone connectivity and Pocket Internet Explorer Web browser (which has a full Internet browser with support for HTML, XML/XSL, WML,cHTML, and SSL); the i-mate device had sufficient memory for the storage of text, audio, image, and video data, with a large screen and a user-friendly interface with an integrated digital camera.

The tools included a mobile e-portfolio (with a multimedia interface) designed to promote professional reflection (Chasin, 2001; Fischer et al., 2003; Hochschuler, 2001; Johns, 1995; Kolb, 1984). These mobile learning tools were designed to promote the skills of documentation of clinical learning, active reflection, and also to enable students to immediately access clinical expertise and resources remotely. Community clinical placements are being used for the testing domain, as there are currently no restrictions on using cellular network technology in these areas, whereas this is currently restricted in acute hospital settings in British Columbia and many other parts of the world.

THE PDA INTERFACE DESIGN

The main interface to the clinical tools was based on a clinical e-tools folder on the Pocket PC containing icon-based shortcuts to a number of specific applications (Figure 1).

The clinical e-portfolio tool represented the major focus for the project, allowing the student to access clinical placement information; log clinical hours; achieve clinical competencies; record portfolio entries in the form of text, pictures, or video clips; and record audio memos. This provides the user with a very adaptable interface, allowing them to choose how they input data. For example, a text-based entry describing a clinical procedure may be accompanied by a picture or audio memo.

The e-portfolio tool also incorporates a reflective practice wizard promoting the students to work through the stages of the Gibbs reflective

Figure 1. Screenshot of the clinical e-tools folder

cycle (Gibbs, 1988) when recording their experiences. This wizard also allows students to record their experiences with multimedia, including text, audio, digital images, or video input. Once the data have been recorded in the e-portfolio, they can be synchronized wirelessly (using the built-in GSM/GPRS or Bluetooth connectivity) with a Web-based portfolio. The data then can be reviewed and edited by the student or by clinical tutors.

The other icons represent the following applications:

- The synch portfolio icon initiates synchronization of the content of the student's e-portfolio on the PDA with that of a remote server.
- The University of British Columbia (UBC) library icon presents a shortcut to a Pocket Internet Explorer Web access to the UBC library bibliographic health care database search (CINAHL, Medline, etc.).
- The Pocket Explorer icon presents a shortcut to Pocket Internet Explorer for mobile Web access.
- The e-mail icon presents a shortcut to the Pocket PC mobile e-mail application.

The other icons on the screen (Diagnosaurus, ePocrates, etc.) represent third-party clinical software that was purchased and loaded onto the PDAs in order to support the students learning in the clinical area (e.g. a drug reference guide).

FUTURE TRENDS

In the future, the PDA will provide a one-stop resource to support clinical learning. Students also will be able to examine their learning objectives, record their achievements, and record notes/memos attached to specific clinical records for later review. Where students have particular concerns or questions that cannot be answered immediately in the clinical area, they will be able to contact their supervisors or faculty for support using e-mail, cell phone, or multimedia messaging service (MMS) communications.

The use of multimedia in PDA interfaces is likely to become much more widespread as the cost of these devices reduces and they become more accessible to a wider spectrum of the population. This already is occurring with the merging of cell phone and PDA technologies and the uptake of MMS and use of audio and video data entry on mobile devices (deHerra, 2003).

In the long term, multimedia mobile learning tools will encourage a more structured process of professional reflection among students in supervised clinical practice (Conway, 1994; Copa et al., 1999; Palmer et al., 1994; Reid, 1993; Sobral, 2000). When unexpected learning opportunities arise, students will be able to quickly review online materials in a variety of formats and prepare for their experience, record notes, record audio memos or images during their practice, and review materials following their experience.

An expansion in the use of such mobile clinical learning tools is envisaged, and there is considerable scope for the widespread application of such tools into areas where students are engaged in work-based learning. We are likely to see the integration of these technologies into mainstream educational practice in a wide variety of learning environments outside of the classroom.

CONCLUSION

The value of these new tools to students in clinical practice remains to be demonstrated, as the evaluation stage of the project has yet to be completed. The project also has highlighted the necessity of addressing some of the weaknesses of current PDA design, such as the small display screen and the need for more built-in data security. However, initial feedback appears promising, and the interface design appears to promote reflective

learning in practice and engage students in the process of knowledge translation.

REFERENCES

Chasin, M.S. (2001). Computers: How a palm-top computer can help you at the point of care. *Family Practice Management, 8*(6), 50-51.

Conway, J. (1994). Profiling and nursing practice. *British Journal of Nursing, 3*(18), 941-946.

Copa, A., Lucinski, L., Olsen, E., & Wollenberg, K. (1999). Promoting professional and organizational development: A reflective practice model. *Zero to Three, 20*(1), 3-9.

deHerra, C. (2003). *What is in the future of Windows mobile pocket PCs.* Retrieved August 12, 2004, from http://www.cewindows.net/commentary/future-wm.htm

Fischer, S., Stewart, T.E., Mehta, S., Wax, R., & Lapinsky, S.E. (2003). Handheld computing in medicine. *Journal of the American Medical Informatics Association, 10*(2), 139-149.

Gibbs, G. (1988). *Learning by doing. A guide to teaching and learning methods.* Oxford: Oxford Polytechnic.

GSM World. (2002). GSM technology: GPRS platform? Retrieved June 24, 2004, from *http://www.gsmworld.com/technology/sms/intro.shtml*

Hochschuler, S.H. (2001). Handheld computers can give practitioners an edge. *Orthopedics Today, 21*(6), 56.

Johns, C. (1995). The value of reflective practice for nursing. *J. Clinical Nurs, 4,* 23-60.

Kolb, D.A. (1984). *Experiential learning.* Englewood Cliffs, NJ: Prentice Hall.

Landers, M.G. (2000). The theory-practice gap in nursing: The role of the nurse teacher. *Journal of Advanced Nursing, 32*(6), 1550-1556.

Microsoft. (2004). *Windows mobile based pocket PCs.* Retrieved August 10, 2004, from http://www.microsoft.com/windowsmobile/pocketpc/ppc/default.mspx

Palmer, A., Buns, S., & Bulman, C. (1994). *Reflective practice in nursing: The growth of the professional practitioner.* Oxford: Blackwell Science.

Reid, B. (1993). "But we're doing it already": Exploring a response to the concept of reflective practice in order to improve its facilitation. *Nurse Ed Today, 13,* 305-309.

Sobral, D.T. (2000). An appraisal of medical student reflection-in-learning. *Medical Education, 34,* 182-187.

KEY TERMS

Bluetooth: A short-range wireless radio standard aimed at enabling communications between digital devices. The technology supports data transfer at up to 2Mbps in the 2.45GHz band over a 10m range. It is used primarily for connecting PDAs, cell phones, PCs, and peripherals over short distances.

Digital Camera: A camera that stores images in a digital format rather than recording them on light-sensitive film. Pictures then may be downloaded to a computer system as digital files, where they can be stored, displayed, printed, or further manipulated.

e-Portfolio: An electronic (often Web-based) personal collection of selected evidence from coursework or work experience and reflective

commentary related to those experiences. The e-portfolio is focused on personal (and often professional) learning and development and may include artefacts from curricular and extra-curricular activities.

General Packet Radio Service (GPRS): A standard for wireless communications that operates at speeds up to 115 kilobits per second. It is designed for efficiently sending and receiving small packets of data. Therefore, it is suited for wireless Internet connectivity and such applications as e-mail and Web browsing.

Global System for Mobile Communications (GSM): A digital cellular telephone system introduced in 1991 that is the major system in Europe and Asia and is increasing in its use in North America. GSM uses Time Division Multiple Access (TDMA) technology, which allows up to eight simultaneous calls on the same radio frequency.

i-Mate: A PDA device manufactured by Carrier Devices with an integrated GSM cellular phone and digital camera**.** The device also incorporates a built-in microphone and speaker, a Secure Digital (SD) expansion card slot, and Bluetooth wireless connectivity.

Personal Digital Assistant (PDA): A small handheld computing device with data input and display facilities and a range of software applications. Small keyboards and pen-based input systems are commonly used for user input.

Pocket PC: A Microsoft Windows-based operating system (OS) for PDAs and handheld digital devices. Versions have included Windows CE, Pocket PC, Pocket PC Phone Edition, and Windows Mobile. The system itself is not a cut-down version of the Windows PC OS but is a separately coded product designed to give a similar interface.

Wizard: A program within an application that helps the user perform a particular task within the application. For example, a setup wizard helps guide the user through the steps of installing software on his or her PC.

This work was previously published in Encyclopedia of Human Computer Interaction, edited by C. Ghaoui, pp. 404-407, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 3.43
3G Mobile Medical Image Viewing

Eric T. T. Wong
The Hong Kong Polytechnic University, Hong Kong

Carrison K. S. Tong
Pamela Youde Nethersole Eastern Hospital, Hong Kong

ABSTRACT

Teleradiology is the technology of remote medical consultation using X-ray, Computed Tomographic or Magnetic Resonance images. It was commonly accepted by clinicians for its effectiveness of making diagnosis for patients at critical situations. Since the huge size of data volume involved in teleradiology [American College of Radiology et al., 2003], clinicians are not satisfied with the relatively slow data transfer rate. It limits the technology to fixed-line communication between the doctor's home and his office. In this project, a mobile high speed wireless medical image viewing system using 3G Wireless Network [Collins et al., 2001], Virtual Private Network and One-Time Two-Factor Authentication (OTTFA) technologies is presented. Using this system, teleradiology can be achieved by using a 3G PDA phone to query, retrieve and review the patient's record at anytime and anywhere in a secure environment. Using this technology, the patient-data availability can be improved significantly, which is crucial to timely diagnosis of patients at critical situations.

INTRODUCTION

Teleradiology is the technology of remote medical consultation using X-ray, Computed Tomographic (CT), or Magnetic Resonance (MR) images. This technique is commonly accepted by clinicians for its effectiveness of making diagnosis for patients at critical situations. For effective implementation of teleradiology, many technical problems including data integrity, accessibility, size of data volume, compression method and bandwidth of linkage should be considered. Hitherto, due to the huge size of data volume involved, clinicians are not satisfied with the slow data transfer rate. It limits the use of the technology to fixed line communication between a doctor's office and

his/her home. In this project, a mobile high speed wireless medical image viewing system using Third Generation (3G) mobile, Virtual Private Network (VPN), Common Gateway Interfacing (CGI), dynamic JPEG compression, WEB, Structural Query Language (SQL) [DuBois et al., 2002], Digital Imaging and Communication in Medicine (DICOM) [NEMA et al., 2004], and One-Time Two-Factors Authentication (OTTFA) technologies was developed. Using this system, teleradiology has been enhanced to a large extent – image data query and retrieval can be transferred from a hospital data centre to any notebook Personal Computer (PC), or to any 3G Personal Digital Assistant (PDA) phone at anytime and anywhere in a secure environment. Hence, the patient-data availability can be improved significantly, which is quite important for patients at critical situations.

BACKGROUND

Teleradiology involves the process of sending radiographic images from one point to another through digital, standard telephone lines, wide area network (WAN), or over a local area network (LAN). The radiographic images can be acquired either by a video capture board such as a frame grabber or the console of a medical imaging modality. After acquisition, the images were digitally stored in a teleradiology workstation in which the images were ready to be sent to a remote site over a network such as ethernet.

In the field of medical imaging, most of the images were stored in Digital Imaging and Communications in Medicine (DICOM) formation. DICOM is a standard that is a framework for medical-imaging communication. It was developed by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) [Bidgood et al., 1992] with input from various vendors, academia, and industry groups. It is referred to as "version 3.0" because it replaces versions 1.0 and 2.0 of the standard previously issued by ACR and NEMA, which was called the "ACR-NEMA" standard. It provides standardized formats for images, a common information model, application service definitions, and protocols for communication. Based upon the Open System Interconnect (OSI) reference model, which defines a 7-layer protocol, DICOM is an application-level standard, which means it exists inside layer 7 (the uppermost layer).

Today most teleradiology systems run over standard telephone [Oguchi et al., 2001] and ISDN (Integrated Services Digital Network) lines which are available in most parts of the world. Other high-speed lines, including T1 line and SMDS (Shared Multi-megabit Data Services) will also become more popular as their prices continue to drop. Over the next couple of years, we should see a substantial migration to wireless network such as IEEE 802.11x wireless and 3G (Third Generation) [Collins et al., 2001] networks, which offer higher flexibility than fixed line networks.

Digital images, whether viewed on a computer monitor, transmitted over a phone line [Reponen et al., 2000], or stored on a hard disk or archival medium, are pictures that have a certain spatial resolution. The spatial resolution, or size, of a digital image is defined as a matrix with a certain number of pixels (information dots) across the width of the image and down the length of the image. The more the number of pixels, the better is the image resolution. This matrix also has depth. This depth is usually measured in bits and is commonly known as shades of grey: a 6-bit image contains 64 shades of grey; 7-bit, 128 shades; 8-bit, 256 shades; and 12-bit, 4096 shades. The size of a particular image is referenced by the number of horizontal pixels "by" (or "times") the number of vertical pixels, and then by indicating the number of bits in the shades of grey as the depth. For example, an image might have a resolution of 640 x 480 and 256 shades of grey, or 8 bits deep. The number of bits in the data set can be calculated by the product of 640 x

480 and 8, which equals to 2,457,600 bits. Since there are 8 bits in a byte, the 640 x 480 image with 256 shades of grey is 307,200 bytes or .3072 megabytes of information.

Although images should be permanently archived as raw data or with only lossless data compression (i.e. no data is destroyed), hardware and software technology exists that allows teleradiology systems to compress digital images into smaller file sizes so that the images can be transmitted faster. Compression [Nelson et al., 1995] is usually expressed as a ratio: 3:1, 10:1, or 15:1. A 10:1 compression factor means that for each piece of information in the original image's matrix, ten are compressed. Certain images can withstand a substantial amount of compression without a visual difference: CT and MR images have large areas of black surrounding the actual patient image information in virtually every slice. The loss of some of those pixels has no impact on the perceived quality of the image nor does it significantly change reader-interpretive performance.

Transmission time has to follow the laws of size. The only way to reduce the transmission time is either to increase the speed of the modem or reduce the number of bits (compress the image) being sent.

The following formula is used to calculate the time to transmit an image:

(Matrix Size) x (Matrix Depth + X bits) x (Percentage of Compression) / (Network Speed)
= **Transmission Time (Seconds)**

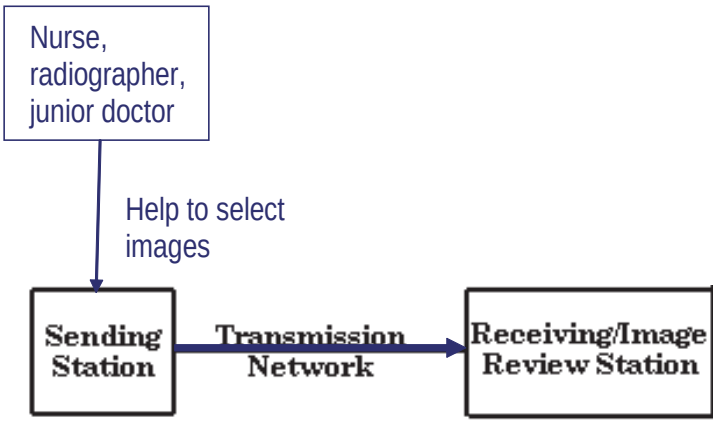
Where Matrix Depth is expressed as shades of grey: 256 shades of grey equal 8 bits; 128 shades of grey equal 7 bits; 64 shades of grey equal 6 bits. For transmission control, all devices add X bits as overhead when transmitting data.

Client/server computing developed from the need to move systems used for application development and operations from expensive mainframes to more efficient, less expensive--yet just as powerful--workstations. Client/server architecture involves the use of two types of computers: a client computer, which runs applications and makes requests for data and other resources, and a server, which processes the client's requests by distributing the requested resources. Client/server computing is known as a cooperative distribution system because both the client and the server cooperate in performing a task. For example, if a client requests a record from the server, the server uses its resources to process the entire file, while the client computer uses its resources to run an application that reads and writes individual records on the file. The server does not need to send the entire file to the client, thus diminishing network traffic or traffic over communication lines.

Client/server computing has numerous advantages for the medical world and for the field of radiology in particular. It permits workstations to achieve computing power previously only available from mainframes--at a fraction of the mainframe costs. By efficiently dividing resources, client/server computing reduces network traffic and improves response time. This efficiency offers a significant advantage to physicians who need to receive images quickly and who require real-time image navigation and manipulation to perform diagnostic tasks effectively. Client/server computing facilitates the use of graphical-user interfaces, making teleradiology and Picture Archiving and Communication System (PACS) applications [Tong & Wong, 2005] easier to use and more responsive. Additionally, clients and servers can be run on different platforms, allowing end users to free themselves from particular proprietary architectures. Software applications designed for client-server computing can interface seamlessly with most Hospital Information System (HIS) [RCR et al., 1999] or Radiology Information System (RIS) systems, while providing rapid soft-copy image distribution.

While most teleradiology systems used over the last decade were intended for on-call purposes,

Figure 1. Schematic diagram of traditional tele-radiology



the past two years have seen a rapid increase in the use of teleradiology to link hospitals and affiliated satellite facilities, other primary hospitals and imaging centres. As teleradiology allows radiologists using their time more efficiently, volume of images transmitted increases without substantially increasing the costs. Although a number of enabling technologies have developed for effective over-read networks, such as the more affordable high-speed telecommunications networks and improved data compression techniques [Nelson et al., 1995] in recent years, the traditional teleradiology technology is only available within fixed line communication system between the doctor's Home and his office and this requires manual image selection from the hospital-side. Both factors have limited the flexibility of the teleradiology technology.

SOLUTION

3G [Collins et al., 2001] is a generic name for a set of mobile technologies launched at the end of 2001 using a host of high-tech infrastructure networks, handsets, base stations, switches and other equipment to allow mobiles to offer high-speed Internet access, data, video and CD-quality music services. Data speeds in 3G networks, being up to 2 Megabits per second, show an improvement on the current technology. Today, various standards of 3G techniques are found in the commercial market including WCDMA, CDMA2000, UMTS and EDGE technologies.

Types of 3G

WCDMA – Wideband Code Division Multiple Access

A technology for wideband digital radio communications of Internet, multimedia, video and other capacity-demanding applications. WCDMA has been selected for the third generation of mobile telephone systems in Europe, Japan and the United States. Voice, images, data, and video are first converted to a narrowband digital radio signal. The signal is assigned a marker (spreading

code) to distinguish it from the signal of other users. WCDMA uses variable rate techniques in digital processing and it can achieve multi-rate transmissions. WCDMA has been adopted as a standard by the International Telecommunication Union (ITU) under the name International Mobile Telecommunications-2000 (IMT-2000) direct spread.

CDMA 2000 – Code Division Multiple Access 2000

Commercially introduced in 1995, CDMA quickly became one of the world's fastest-growing wireless technologies. In 1999, the International Telecommunications Union (ITU) selected CDMA as the industry standard for new "third-generation" (3G) wireless systems. Many leading wireless carriers are now building or upgrading to 3G CDMA networks in order to provide more capacity for voice traffic, along with high-speed data capabilities. Today, over 100 million consumers worldwide rely on CDMA for clear, reliable voice communications and leading-edge data services.

UMTS – Universal Mobile Telecommunication

The name for the third generation mobile telephone standard in Europe, standardized by European Telecommunications Standards Institute (ETSI). It uses WCDMA as the underlying standard. To differentiate UMTS from competing network technologies, UMTS is sometimes marketed as 3GSM, emphasizing the combination of the 3G nature of the technology and the GSM standard which it was designed to succeed. At the air interface level, UMTS itself is incompatible with GSM. UMTS phones sold in Europe(as of 2004) are UMTS/GSM dual-mode phones, hence they can also make and receive calls on regular GSM networks. If a UMTS customer travels to an area without UMTS coverage, a UMTS phone will automatically switch to GSM (roaming charges may apply). If the customer travels outside of UMTS coverage during a call, the call will be transparently handed off to available GSM coverage. However, regular GSM phones cannot be used on the UMTS networks.

EDGE – Enhanced Data for Global Evolution

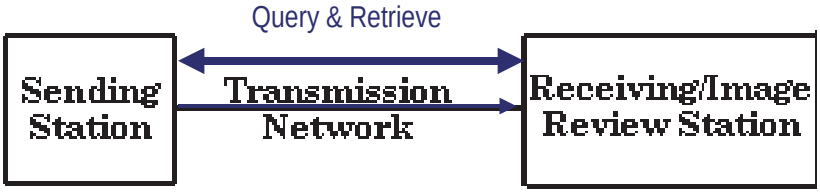
A technology that gives GSM the capacity to handle services for the third generation of mobile telephony. EDGE was developed to enable the transmission of large amounts of data at a high speed, 384 kilobits per second. EDGE uses the same TDMA (Time Division Multiple Access) frame structure, logic channel and 200 kHz carrier bandwidth as today's GSM networks, which allows existing cell plans to remain intact.

IMPLEMENTATION

The proposed system has been designed for image transfer using a UMTS type of 3G Wireless Network (Figure 2), which could provide a data transfer rate of at least 384 kbps, i.e. about 10 times the speed of a **Global System for Mobile Communications** (GSM) [Tong et al., 2003]. The security risk for this kind of connection to the hospital enterprise network is also ten times greater. One way to enhance the security is using One-time Two factors Authentication technologies for effective access control as shown in Figures 4, 5 and 6. The access right control and audit trail are based on the VPN technology. For management of terabytes of image data, a SQL-based database could be used. Users can search the database for any interested studies with the related images (Figures 7, 8, 9, 10, and 11). This design allows the clinician searching and retrieving the required studies without any help from the hospital side.

WEB technology, today, is the most robust client/server computing technology. WEB technology has been applied in many areas includ-

Figure 2. Schematic diagram of 3G medical image viewing system



ing desktop and mainframe computers. A WEB browser is also available in some mobile phones such as the PDA phone used in this project. In the proposed system, all CT and MR images were stored in a DICOM image archiving server. An Apache-based WEB server software [Kabir et al., 1999] was installed in an existing DICOM image server connected with a SQL server for image data management. The image data stored in the DICOM server would allow queries from the WEB browser through the 3G mobile PDA phone (Figure 3). The retrieved images were first converted from 16 bits into 8 bits and then JPEG-compressed before being sent to the remote client as a WEB page with configurable parameters such as image size, window level and width, rotation and flip using Common Gateway Interfacing (CGI) method [Tong et al., 2003]. CGI is one of the common techniques for the manipulation of data in WEB technology. The installed CGI software is an interface between the DICOM server and WEB client sides. It talks to the DICOM server using SQL and DICOM commands and communicates with the WEB clients using Hypertext Modelling Language (HTML). The clinical user can change the display parameters for image processing of the data dynamically in order to improve the viewability of the images.

RESULT

In the initial test of the system, transfer of 20 slices of CT images took about 45 seconds, which is much shorter than the travelling time required by the doctor driving from home to hospital. During a remote consultation, the clinician can also work independently to search for the target studies without any help from the hospital. Hence, unnecessary travelling of doctors to hospital can be minimized for non-urgent cases. For remote consultation using medical images, two major limitations have been identified: processing power of the portable image receivers and the speed of wireless network.

A typical 3G PDA phone was showed in Figure 3. In this phone, the display size is 2.9 inches with a resolution of 208 x 320. Comparing the resolutions of CT and MR images of resolutions of 256 x256 and 512 x 512 pixels, the PDA phone was sufficient or nearly sufficient to display those images. The one-time two-factor authentication token in which the displayed password was kept changing with an interval of one minute as shown in the Figure 4. The login page of the system was shown in Figure 5. The data entry was using a virtual keyboard, which was a basic function of the PDA phone as Figure 6. After login, user

Figure 3. A 3G PDA phone

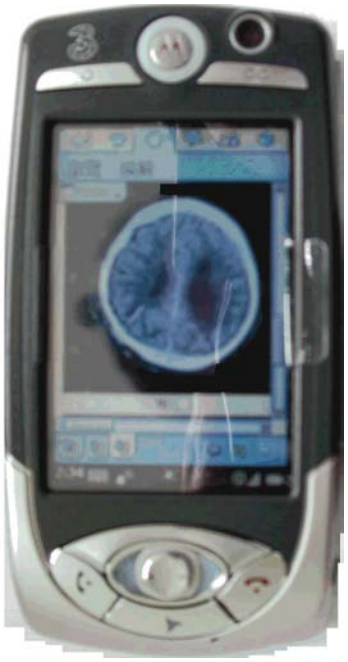


Figure 4. One-time two-factor authentication token

Figure 5. Log-in page of image-viewing system

Figure 6. Log-in of the image-viewing system using a virtual keyboard

Figure 7. Patient search in image-viewing system

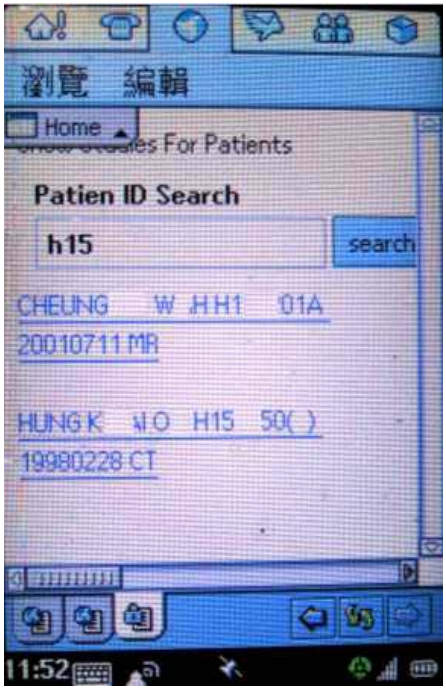


Figure 8. Series selection in image-viewing system

Figure 9. Multiple-image display

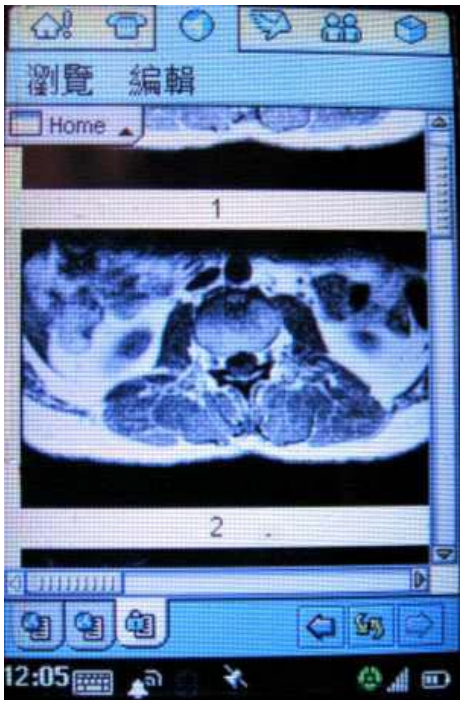


Figure 10. Display of CT image

Figure 11. Display of MR image

might require the search of related studies. He could use the data management system for the search of interested examination using the patient identity number or study date as shown in Figure 7. For the patients with multiple studies or series of images, the thumbnail image of the first image of each series was displayed as the Figure 8. After the display, user could select the interested study by clicking the thumbnail image. A few minutes later, the user could see multiple CT or MR images as shown in Figure 9, 10, and 11.

FUTURE TRENDS

Today, teleradiology is mainly limited by the speed of the wireless network. With the next generation of wireless network such as IEEE 802.11 with a speed up to 108 mbps, it is anticipated that teleradiology can be performed more efficiently.

4G is the next generation of wireless networks that will replace 3G networks sometimes in future. In another context, 4G is simply an initiative by academic R&D labs to move beyond the limitations and problems of 3G which is having trouble getting deployed and meeting its promised performance and throughput. In reality, as of first half of 2002, 4G is a conceptual framework for or a discussion point to address future needs of a universal high speed wireless network that will interface with wireline backbone network seamlessly. 4G also represents the hope and ideas of a group of researchers in Motorola, Qualcomm, Nokia, Ericsson, Sun, HP, NTT DoCoMo and other infrastructure vendors who must respond to the needs of Multimedia Messaging Service (MMS), multimedia and video applications if 3G never materializes in its full glory.

CONCLUSION

It can be seen that with the proposed scheme involving the integration of WEB, SQL, DICOM, 3G, and JPEG technologies, a practical solution has been developed to tackle the commonly noted teleradiology implementation problems using high speed wireless networking technologies. In addition, data integrity and accessibility have been enhanced significantly. In conclusion, the proposed 3G mobile medical image viewing system can provide an efficient tool for clinicians' remote consultations, thus allowing timely diagnosis for patients at critical conditions at anytime and anywhere.

REFERENCES

American College of Radiology (ACR), (2003), ACR Technical Standard for Teleradiology

Bidgood WD, Horii SC. (1992), Introduction to the ACRNEMA DICOM standard. Radiographics, 12, 34–355.

Collins D, Smith C, (2001), 3G Wireless Networks. McGraw-Hill Professional.

DuBois P., (2002) MySQL Cookbook. O'Reilly & Associates.

Kabir, M.J., (1999) Apache Server Administrator's Handbook. Hungry: Minds Inc.

Marcus E, Stern H, (2003) Blueprints for High Availability**,** 2nd edition, Wiley.

National Electrical Manufacturers Association (NEMA), (2004) The DICOM Standard.

Nelson M, and Gailly J.L., (1995) The Data Compression Book. 2nd ed. Hungry: Minds Inc.

Oguchi K, Murase S, Kaneko T, Takizawa M, Kadoya M. (2001), Preliminary experience of wireless teleradiology system using Personal Handyphone System, Nippon Igaku Hoshasen Gakkai Zasshi, 61(12), 686-7.

Reponen J, Ilkko E, Jyrkinen L, Tervonen O, Niinimaki J, Karhula V, Koivula A., (2000), Initial Experience With A Wireless Personal Digital Assistant As A Teleradiology Terminal For Reporting Emergency Computerized Tomography Scans, J Telemed Telecare, 6(1), 45-9.

Royal College of Radiologists (1999). Guide to information Technology in Radiology: Teleradiology and PACS. Board of Faculty of Clinical Radiology, RCR.

Tong C.K.S., Chan K.K. & Wong C.K., (2003) Common Gateway Interfacing And Dynamic JPEG Techniques For Remote Handheld Medical Image Viewing, Computer Assisted Radiology and Surgery, 815-820

Tong, C.K.S. & Wong, E. T.T. (2005) Picture Archiving and Communication System in Health Care. In M. Pagani(Ed.) Encyclopedia of Multimedia Technology and Networking, Idea Group Reference, 821-828.

KEY TERMS

3G Technology: 3G (or 3-G) is an abbreviation for third-generation technology. It is usually used in the context of mobile phones. The services associated with 3G provide the ability to transfer both voice data (a telephone call) and non-voice data (such as downloading information from the internet, exchanging email, and mutlimedia messaging).

Common Gateway Interface(CGI): a standard protocol for interfacing external application software with an information server, commonly a web server. This allows the server to pass requests from a client web browser to the external application. The web server can then return the output from the application to the web browser.

Digital Imaging and Communications in Medicine (DICOM): is the industry standard for transferral of radiologic images and other medical information between computers. Patterned after the Open System Interconnection of the International Standards Organization, DICOM enables digital communication between diagnostic and therapeutic equipment and systems from various manufacturers.

Enhanced Data rates for Global Evolution (EDGE): An enhancement to the GSM and TDMA digital cellular phone systems that provides data transmission up to 384 Kbps. Like cellphones in general, the EDGE service is more ubiquitous for mobile users than searching for Wi-Fi hotspots; however, data rates are much lower than Wi-Fi.

Global System for Mobile Communications (GSM) A digital cellular phone technology based on TDMA(Time Division Multiple Access), which

is a satellite and cellular phone technology that interleaves multiple digital signals onto a single high-speed channel. Operating in the 900MHz and 1.8GHz bands in Europe and the 1.9GHz PCS band in the U.S., GSM defines the entire cellular system, not just the air interface (TDMA, CDMA, etc.).

One-Time Two-factor Authentication (OTTFA) is any authentication protocol that requires two independent ways to establish identity and privileges. This contrasts with traditional password authentication, which requires only one factor (knowledge of a password) in order to gain access to a system.

Teleradiology: is a means of electronically transmitting radiographic patient images and consultative text from one location to another.

Universal Mobile Telecommunications System (UMTS): The European implementation of the 3G wireless phone system. UMTS, which is part of IMT-2000, provides service in the 2GHz band and offers global roaming and personalized features.

Wideband Code Division Multiple Access (WCDMA): also known as Universal Mobile Telecommunications System(UMTS) in Europe, is 3G standard for Global System for Mobile Communications(GSM) in Europe, Japan and the United States. It supports very high-speed multimedia services such as full-motion video, Internet access and video conferencing. It uses one 5 MHz channel for both voice and data, offering data speeds up to 2 Mbps.

This work was previously published in Web Mobile-Based Applications for Healthcare Management, edited by L. Al-Hakim, pp. 300-315, copyright 2007 by IRM Press (an imprint of IGI Global).

Section IV
Utilization and Application

This section introduces and discusses the ways in which information technology has been used to shape the realm of mobile computing and proposes new ways in which IT-related innovations can be implemented within organizations and in society as a whole. These particular selections highlight, among other topics, the implementation of mobile technology in healthcare settings, and the evolution of mobile commerce. Contributions included in this section provide excellent coverage of today's mobile environment and insight into how mobile computing impacts the fabric of our present-day global village.

Chapter 4.1
Dynamics of Mobile Service Adoption

Hannu Verkasalo
Helsinki University of Technology, Finland

ABSTRACT

This study utilized a newly developed handset-based mobile end-user research platform and obtained data from 548 Finnish smartphone users in 2006. In addition to descriptive adoption statistics, a path analysis model is developed that explains mobile service adoption contingent on a set of explanatory variables. The paper finds that user intentions have a strong impact on consequent adoption of the service. What is more, perceived hedonic benefits from the service are the strongest factor driving user intentions to use the service. The perceived technical capability to use the service and the role of the surrounding social network explain little why early-adopter users intend to use services. Interestingly multimedia services are strongly driven by newer more capable handsets and mobile Internet browsing benefits significantly from block or flat-rate (instead of usage-based) pricing plans for transmitted data. The paper develops several indices that measure time-varying characteristics of mobile services.

INTRODUCTION

Mobile services have evolved quite a lot from mere communication oriented services (circuit-switched voice, text messaging, voice mailbox) to today's multimedia, content retrieval, browsing and other advanced services. The mobile Internet (see Funk 2004) is emerging and the IP-based service delivery is likely to hit the mobile mass market domain very soon. Overlay networks existing already in the Internet (Clark et al. 2006) may have spill-over effects to the mobile industry. The mobile Internet scenario contrasts sharply with the dominant, vertically-oriented way of doing mobile business (see e.g. Karlson et al. 2003, Verkasalo 2007a and Vesa 2005). The emergence of the mobile Internet is driven by the wide-scale adoption of smartphones (i.e. converged devices) along with improvements in both cellular (GSM and 3G) and alternative (e.g. WiFi) radio networks. In terms of data services the same service evolution trends have been seen in the "wired" Internet earlier that can be seen in the mobile domain today. For example,

the movement from messaging data services to static content (Web) and further to multimedia streaming can already be seen in mobile service studies (Verkasalo 2007b).

Amidst the rapid evolution of the mobile industry many commercial service failures have taken place. It is difficult to pinpoint the reasons behind successes and failures. Typically not one but many issues affect the adoption of a particular mobile service. The reasons can be categorized into two main categories. First of all, a commercial/technical perspective includes issues that relate to marketing, positioning, developing, implementing, delivering and timing of the mobile service. These factors include e.g. demand forecasting, pricing, positioning of the service in the service provider's service portfolio, promotional activities, creation of end-user awareness, service quality management, and strategic push of the service in the value-chain (i.e. distribution management). These factors are called as t*echnological* or *business strategic* in Pedersen (2001). Second, the end-user perspective deals with end-user related factors driving or inhibiting service adoption. This perspective is called as *behavioral* in Pedersen (2001). Factors under this perspective include e.g. service usability, social pressure, network externalities, contextual environment, consumption choices, the user's motivation and technical capabilities. The first perspective deals more with the producer side of the market whereas the second perspective deals with the demand side of the market. Drivers and bottlenecks for service adoption might emerge in either domain.

Even though many potential factors explaining successes and failures of mobile services can be identified, it is often difficult to test hypotheses in practice. No suitable empirical research approaches have existed earlier to provide actual usage data to study the dynamics of mobile service adoption. Accurate data from end-users can be nowadays acquired with a handset-based mobile end-user research platform that was introduced in Verkasalo & Hämmäinen (2007). The new platform provides accurate usage statistics along with flexible tools to deploy questionnaire studies. The present paper attempts to provide descriptive results on mobile service adoption with data from Finland 2006. In addition, a path analysis model is built explaining the main drivers and bottlenecks of mobile service adoption based on empirical usage data and questionnaire studies.

EARLIER RESEARCH

Theoretical Models Explaining Technology Adoption

The adoption research can generally be divided into four categories:

- Diffusion research (market focus)
- Adoption approach (individual user focus)
- Gratification research (needs of users focus)
- Domestication research (consequence of adoption focus)

The diffusion research focuses on the market-level phenomena, and studies the diffusion of technology in the whole market. Adoption research, on the other hand, considers individual users as a focal research object. Gratification research contributes by analyzing the different kinds of benefits users seek from new technologies, and domestication research analyzes the role of new technologies in integrating to the every day life of people. Although this research paper mainly applies the statistical models introduced in the adoption research, elements from other research approaches are also applied in building the framework introduced in chapter 3.3. Therefore all these approaches are discussed now in detail.

First, Rogers (1962) introduced the idea of "***diffusion*** of innovations", and approached the adoption process in diffusion terms. In his research the adoption process follows a bell curve, suggesting that different kinds of people adopt new technologies at different pace. Only the most technology enthusiastic people adopt new products/services at first, the mass market being more cautious and thus adopting slower. The late-adopters are the most technology averse people, typically purchasing new technology only when it is inevitable. Rogers' research still serves as the general background for many kinds of adoption research. The diffusion research can be used in studying the emergence of new services. In Rogers (1995) it is argued that adoption is initiated by a new technology, after which the social setting and communication channels boost the diffusion. Rogers' theory has laid ground for many other research frameworks, for example Christensen's (1997) theory of "disruptive technologies" that take over dominant technologies by having a disruptive diffusion path. Time is the core factor in Rogers' idea of technology diffusion, as adoption (penetration of service) follows different patterns at different points of time after its introduction.

The second approach is the ***adoption perspective***, in which an individual user standpoint is taken. Each individual makes her own decisions in considering whether to try the service or not. A wide domain of research stems from adoption of information systems science, utilizing theoretical models developed a couple of decades ago. Earlier models attempting to explain adoption of technologies (particularly information system technologies and this is why the approach is sometimes called as IS adoption science) include the *theory of reasoned action, theory of planned behavior* and *technology acceptance model.*

A theory of reasoned action (Fishbein & Ajzen 1975) is based on the individual's attitude towards the action and subjective norm of the action (expected behavior of others in response to the individual's action). Together these determine the behavioral intention to use the technology. Ajzen later expanded the model; in a theory of planned behavior (Ajzen 1985; 1991 a third concept exists, namely the perceived behavioral control. This reflects the difficulty of performing the action. These models communicate that technology adoption depends both on the individual's own perceived benefit of performing the action and the social norm driven by people around the individual. All in all, these frameworks suggest that usage patterns not only depend on the individual's own capabilities and interests, but also on the sociological environment and norms in the culture.

Davis (1989) follows similar logic in his framework - a technology acceptance model (TAM). He distinguishes two concepts. First, the perceived usefulness reflects the expected benefits from using a certain technology. Second, the perceived ease of use reflects pretty much the same thing as the perceived behavioral control in the theory of planned behavior, i.e. how difficult it is to use the technology. In predicting information technology adoption the TAM model developed by Davis is the most used framework, and by 2000 more than 400 journal articles had cited the two original TAM articles (Venkatesh & Davis 2000). Almost all the information technology adoption articles that will be discussed later in this paper stem from the TAM model. Despite its popularity, further development of the model has taken place. For example, the original model as projected below suggested that perceived usefulness and perceived ease-of-use mediate all external factors (e.g. demographics), though this is not always the case (see e.g. Burton-Jones & Hubona 2005). The original model typically explains 40% of usage intentions and 30% of actual use (see e.g. Venkatesh & Davis 2000 and Meister & Compeau 2002). Applications of the framework might in the best case achieve better explanatory power.

Holistic adoption models (see e.g. Pedersen and Thorbjørnsen 2003) deployed earlier in the mobile context and stemming from the TAM model are close to the research approach applied

in developing the theoretical framework for this research paper. Holistic behavioral models typically utilize statistical methods such as structural equation modeling (SEM), and they derive from theoretical models illustrated above (particularly Davis' research). In Nysveen et al. (2005a) the adoption approach utilizing variations of the TAM model with SEM analysis is called as information systems research, as most studies utilizing structural equation models from the adoption perspective deal with the adoption of ICT services and systems. The research done by e.g. Pedersen and Thorbjørnsen (2003) is used as a basis for the model of this paper. The next section will discuss the theoretical approaches to study mobile services in detail. Because of e.g. increasing context-specific nature and various ubiquitous characteristics (Heinonen and Pura 2006; Rask and Dholakia 2001) mobile services should be considered carefully when applying earlier information systems adoption models.

The other approaches to study the adoption of technology include a uses and gratifications research approach (see e.g. Leung and Wei 2000; Höflich and Rössler 2001) and domestication perspective (Haddon 2001; Ling 2001; Skog 2002). The former approach (***gratification research***) deals with the gratifications that users look for when using mobile services. These gratifications can be either utilitarian (business value or direct utility) or hedonic (entertainment oriented) (Flanagin and Metzger 2001). The latter approach (***domestication research***) has close linkage to sociology, anthropology and ethnology. Sometimes domestication research does not focus on individual adoption only, but extends to the adoption from the cultural or societal point of view (see e.g. Ling and Yttri 2002). Domestication research attempts to tackle the consequences of service usage and the integration of the technology/services in the customer's every day life (Pedersen 2005). These two frameworks reflect softer scientific disciplines in studying mobile services than the diffusion or adoption perspectives. All approaches share same concepts and ideas, and they are thus not totally separate from each other.

Emergence of Mobile Services

Little research on the emergence of mobile services exists. This section briefly describes some of the main outcomes of earlier research, particularly in light of this paper. Many of the focal services studied in this paper have just been introduced to the market, and no research on them is available. However, this section attempts to emphasize generic observations in the adoption of mobile services that have relevance also with new emerging services.

The mobile Internet is defined in this paper to consist of new packet-switched mobile data services. The first mobile data services were hyped quite a lot in the public in the late 90s, but in practice the mobile Internet has not kicked off yet (Saarikoski 2006). Particularly the WAP technology and mobile email can be considered failures (Sigurdson 2001). Many reasons are suggested as possible bottlenecks, from pricing (see e.g. Gao et al. 2002) to general difficulties in terminal configuration (Verkasalo 2007b). These reasons were found as bottlenecks in mobile payment adoption, too (Mallat 2006b). MMS messaging experienced similar disappointing customer adoption rates at first to WAP, email and electronic payments are experiencing right now. In Japan NTT DoCoMo has achieved satisfactory demand for mobile Internet services, and Japan can be considered as the world's leading mobile market in many other dimensions, too (Saarikoski 2006; Minges 2005). The key difference to Western operators is that NTT DoCoMo has bundled service interfaces directly to the handset and actively pushed new value-added services. Consequently both the customer awareness has been increased and ease of adoption significantly improved.

The Internet service expansion to the mobile domain could serve as a prospective spark that could significantly push the extent of mobile data

service usage. The reasons for the potential impact of Internet services can be divided into two. First of all, the pricing/commercial models of Internet services are different from operator-based mobile services (in terms of pricing, for example, Internet services are free of charge, or sold in flat-rate bundles). Second, the spill-over effects and value already embedded in Internet services (e.g. instant messaging communities, the variety of WWW-based content services, webmail) serve as a strong force if suitable mobile access technologies (e.g. the network access and adequate terminals for the usability's sake) can be deployed with low price. Earlier research has pinpointed the need for alternative radio access methods to fixed Internet service access, the key alternative being mobile cellular connectivity (Kearney 2001). These two factors (commercial/pricing models and mobile extension of the Internet) are evident in the final conclusions of Aarnio et al. (2002), as they state on the future prospects of mobile services that "*…prices must come down to overcome the critical mass threshold and in most cases the services should be integrated to the Internet...*".

Niina Mallat (2006a) illustrates some interesting characteristics of mobile services in her dissertation. According to her studies, several successes and failures could be identified in the mobile payment service sector. Mallat concludes her dissertation with the note that usage situation and context are important in explaining the adoption of mobile electronic payments. The same can be generalized to the wider mobile domain. In studying mobile service adoption the special characteristics of mobile services – the freedom of context and location – should be internalized in the framework is possible. Mallat criticizes earlier mobile service adoption research for not taking contextual factors that well into account. Also Hejden et al. (2005) emphasized the importance of context in influencing the perceived value of services.

In tackling the drivers and bottlenecks of mobile services, Aarnio et al. (2002) studied 1 553 Finnish respondents and found five user segments from the sample that resembled classical adoption categories. Their paper studied both traditional Internet and new mobile services. The authors emphasize the important role of prices in mobile service adoption. Furthermore, they suggest that social norm should not be overlooked in mobile services. However, Nysveen et al. (2005a) suggest that different mobile services depend on explanatory factors differently. For example, some services are more dependent on social pressure/norms than others. For example in the research of Pedersen and Thorbjørnsen (2003) the social norm did not explain that much of the variance.

Pedersen and Thorbjørnsen (2003) developed a structural equation model incorporating motivational, attitudinal, social and resource-related influences on the intention to use mobile services. Their model explained about 62-75% of the variance in the dependent variable (intention to use). In all of their case studies (focusing on different services) they found that both extrinsic (utilitarian) and derived (i.e. expressiveness) motivations have an important role to play, whereas intrinsic (entertainment-oriented) motivations are less explanatory. Pedersen (2005) continues by suggesting that many of the factors suggested in earlier ICT adoption research work fine in the context of mobile services, but also many new dimensions should be taken into account (such as expectations and subjective norm). Pedersen (2005) argues that by extending the decomposed theory of planned behavior (Taylor & Todd 1995) with elements from domestication research the explanatory power of the models in mobile services can be increased. To criticize existing statistics models Nysveen et al. (2005b) argue that gender could have moderating effects in the adoption of mobile services, though the original TAM model suggests that all background variables should be fully mediated by the included explanatory factors. All in all, in statistically evaluating the determinants of usage, different services should

be treated individually and one should be careful with cross-service generalizations.

Some mobile services are more goal-directed (utilitarian) instead of experimental (hedonic) (Nysveen et al. 2005a). Pedersen and Thorbjørnsen (2003) acknowledge the same. Also Wong and Hiew (2005) argue for the emphasis on the expected benefits from mobile services. Hejden et al. (2005) comply by concluding that utilitarian and hedonic values have some correlation but it is particularly utilitarian value that has a significant impact on the intention to use the services whereas hedonic values relate to more ad-hoc use cases. The utilitarian importance might have, however, derived from the market context at the time of the study in Hejden et al (2005), as many new mobile services emerging today are actually quite entertainment instead of business-oriented. Most new mobile services are therefore targeted at consumers instead of business customers. Hejden et al. (2005) nevertheless found that perceived risks from the service negatively drive the service's utilitarian value, whereas hedonic value is not affected by the perceived risks (e.g. the service is justified to be more difficult to use if it relates to "killing time"). Some papers (e.g. Tseng et al. 2007) emphasize that in addition to direct value also network externalities matter in analyzing the benefits of electronic services. In mobile instant messaging, for example, the value embedded in the buddy network affects the utility from using mobile instant messaging.

Pedersen (2005) notes that adoption studies in the mobile domain might have differing objects of research. He distinguishes between users (Green et al. 2001; Bakalis et al. 1997), services (see Verkasalo & Hämmäinen 2007; Kim 2001; Pedersen et al. 2001) and terminals (see Chuang et al. 2001; Skog 2000). In this paper the approach is mainly from the user point of view, and the process through which end-users adopt services is the focal research objective.

EMPIRICAL ANALYSIS

Research Method and Dataset

A pioneering mobile end-user research platform was utilized in acquiring data for this research paper. The new mobile end-user research platform is based on a developed Symbian/S60 smartphone client that observes all kinds of usage actions taking place in the handset. Usage-levels stamps on any application, network or user interface level action is logged with accurate context-specific information (e.g. time). This data is sent to centralized servers every night for analysis purposes. Research is conducted in panels lasting typically 2-3 months, and a typical panel includes hundreds of interested customers which in the panel become panelists. All panelists participating in these study panels sign a contract and they are aware of the research process. Usage data is complemented with various WWW-based questionnaires through which to acquire data on issues that are not usage-related (e.g. motivations and attitudes). Though the accuracy and scope of acquired data is a clear contribution in the world of end-user research, the challenges include e.g. biased end-user domain (early-adopter users) and sample size (being still in the range of 400-1000 panelists). The research method has been used already in various papers, such as (Verkasalo 2006a, Verkasalo 2006b, Verkasalo & Hämmäinen 2006, Verkasalo 2007a, Verkasalo 2007b, Verkasalo 2007c and Verkasalo 2007d). For more information on the research method, see Verkasalo & Hämmäinen (2007).

In acquiring data for this research paper SMS recruitment messages were sent to 28 000 Finnish consumer subscribers who owned a Nokia S60 device. 1 071 (3.8%) customers visited the recruitment site and answered the beginning questionnaire. Out of the registered panelists 695 (65%) managed to generate at least three active weeks of smartphone usage data. The rest either did not manage to install the research client or then quit the panel study earlier than was supposed.

Some people might have used the smartphone with the research client installed as a secondary phone, and thus they were excluded, too. Out of the active panelists 548 (79%) answered the final questionnaire, which was more comprehensive than the beginning questionnaire. In addition to demographics, the beginning and final questionnaires covered various background questions related to user motivation, usage patterns and opinions on usability and performance of mobile services.

The details of the panel are introduced in Verkasalo (2007b). In brief, most panelists were young male consumer customers. The panel consisted of early-adopter customers, as in earlier handset-based service studies (see Verkasalo 2005). 37% of panelists considered themselves either experienced or very experienced smartphone users, whereas only 19% considered themselves as beginners. 56% considered themselves as normal in terms of smartphone usage patterns.

The questions asked from the panelists in background questionnaires are listed in Appendix A. The whole dataset used in this paper consists of a questionnaire that was filled in by the panelists, together with a comprehensive set of aggregated usage-level measurements that reflect objectively panelists' actual behavior during the panel. The background questionnaire was used in obtaining data on the factors that have been found important earlier in studying technology adoption (see particularly Pedersen's research). Because of respondents' limited time, only one question was asked for each of the identified factors. Earlier statistical adoption studies have had more comprehensive questionnaires, and therefore they have the possibility to use both the measurement and structural parts of structural equation modeling. In this paper only the path model is estimated. Data is collected for the services that were identified as being of great interest at that point of time (in 2006).

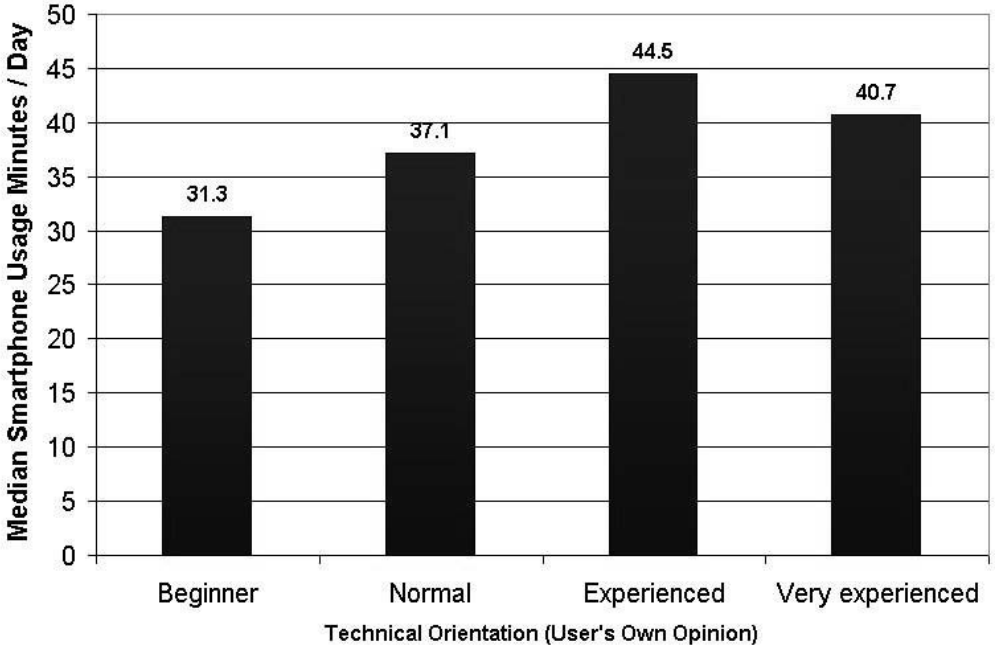
For each of these services a usage frequency variable was calculated measuring the share of days when a particular service was used out all active days spent in the panel. If a service is in constant use, this variable ranges from 50% (meaning usage every second day) to 100% (usage every day). If usage is less frequent, the usage frequency variable takes lower values. This variable well reflects the extent of use, and it is restricted to take values in the range of 0% to 100%. This variable is used as an indicator of usage in the estimated path model later in this paper.

Descriptive Statistics

Figure one depicts the user's own opinion about himself/herself in terms of technical experience with smartphones and estimated average minutes of smartphone service usage per day. More experienced users spend time with smartphone devices more than less experienced users by observation. People with stronger technical interests and capabilities should be more likely to explore new services on the one hand and more likely to learn using them and adopting the service into every day use, on the other hand. Therefore the observed results on total smartphone usage time are quite expected. Because no data on user experience was available for all of the panelists, the total smartphone usage minutes per day is used a proxy for user experience in path modeling later in this paper.

Next the focus is on particular services for which a comprehensive dataset was collected combining both questionnaires and actual usage data. Figure below projects the identified key smartphone services in two dimensions: the share of panelists who intended to use the service before the panel (*intention index*), and the share of panelists who actually used them (*usage index*). The projected diagonal line depicts the points in which the service's all users who have intentions to use the service actually used the service during the panel. In some cases it, however, might be that some users who did not intend to use the service actually used it anyways. The services that fall off the line have some bottlenecks in the

Figure 1. Correlation of technical orientation and extent of handset use



adoption process because not all of the interested panelists used them. Conversely, those services that are above the line have been adopted by more panelists than who actually intended to do so.

WEB/WAP browsing, MMS, offline multimedia and gaming have experienced usage approximately according to intentions. In other words, approximately those who intended to use the services also were able to do so. Therefore these services can be considered as sort of successes in this study. More interestingly, there are many services that lie relatively far away from the diagonal line. These include e.g. instant messaging, mobile email (both embedded email clients and webmail included) and streaming multimedia. People had intentions to use these services, but for a reason or another usage did not realize. Some bottlenecks for adoption remain for some services, whereas for some services bottlenecks are significantly lower. All services that received high usage indices (located in the upper right hand corner) have been available in the market for longer than those that received low usage indices. Clearly service maturity reflects in the results.

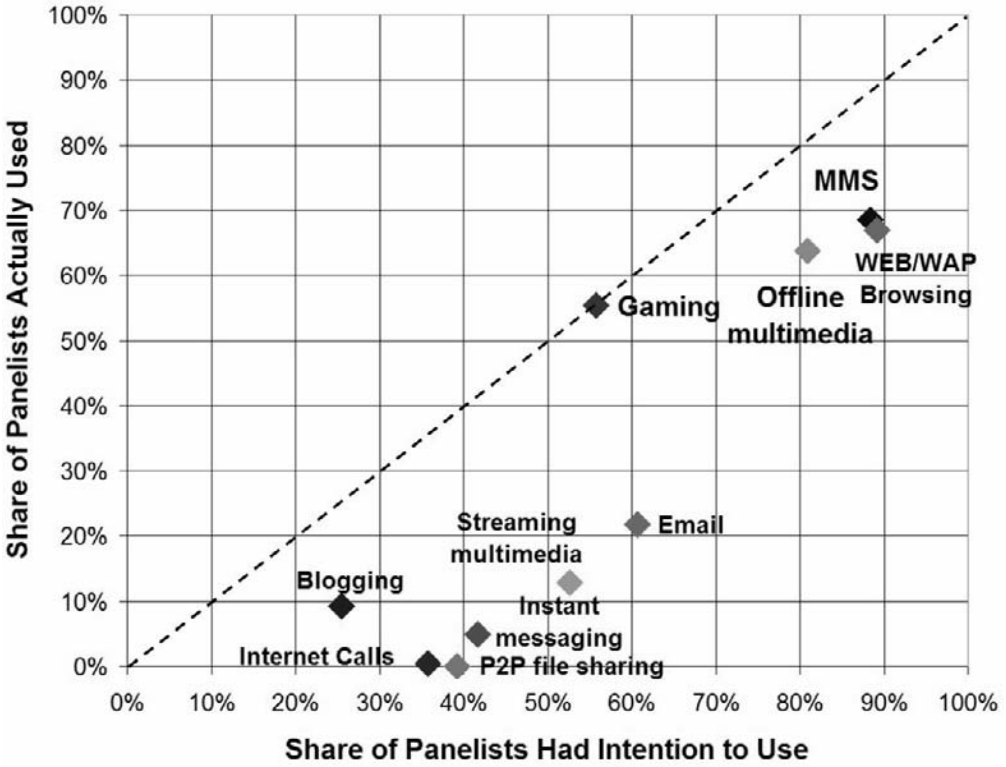
Figure two demonstrates how intentions correlate with service adoption. Those who stated they have strong interest towards the service had significantly higher adoption rates. Strong correlation between intentions and the consequent adoption process exists. As was found above, WEB/WAP browsing, MMS, gaming and offline multimedia playback catch a wide domain of those users who originally had intentions towards the particular service. On the other hand, only 60% of those who had a very strong interest to use mobile email and 33% of those who had a very strong interest to use various streaming multimedia services actually did use these services. There are many arguments to explain why certain services experience good adoption whereas some others do not. Some of these reasons include the technical difficulty to configure/use the service, immaturity of the service, unavailability of necessarily technical

help, problems in the marketing / communication functions and suboptimal pricing.

Figure three demonstrated that there exists a gap between user intention and actual usage. This is called here as the "adoption gap", as it reflects the problems in moving from intentions to actual usage. This paper develops a service *adoption index* measuring the extent of this gap. The index can be calculated by measuring the share of panelists who actually used the service out of those who intended to use it. For services that experience little bottlenecks in adoption, this should be close to 100%, whereas for services with significant bottlenecks this should be closer to 0%. Graphically all services located close to the diagonal line have high service adoption indices. The advantage of this index is that it can be used when comparing services with different customer awareness (e.g. different diffusion levels of original intention to use), as effectively the measure is normalized based to the number of people who intended to use the service (instead of all panelists).

In addition, there exists a gap between intended usage (reflecting current short term interest to explore the service) and expected future service diffusion (reflecting future expectations for the service, i.e. is the service going to succeed in the mobile market by extending the "wired" version of the service). A background question was asked with regards to users' expectations of different mobile services succeeding in the mobile market by replacing their use of the service in other platforms (e.g. desktop computers or MP3 players). This is called as an attitude towards the service variable in this research paper as it reflects positive expectations for the mobile version of the service. Rather than reflecting general attitude it reflects positive

Figure 2. User intentions and actual use of services



long-term perception of the service eventually being used by the panelist herself. *Attitude index* measures the share of panelists having a positive attitude towards the service. Further, *timing index* measures the share of panelists having positive intentions to use the service during the panel (in the short term) out of those panelists having a positive (long-term) attitude towards the service. This variable is very useful, as it normalizes the number of people intending to use the service by the number of people having long-term interest (i.e. attitude) towards the service. Therefore the timing index can be used in evaluating whether the market is ready to adopt a particular service right now, given that positive attitudes and therefore eventual latent demand for the service exist. The timing index is high if a relatively high share of interested panelists are likely to adopt a given service in the short-term, and low if service adoption is for a reason or another (e.g. unavailability of the service) not acute right now.

The table above summarizes the calculated indices for the identified services. By looking at the attitude index, it can be said that people have positive attitudes (future expectations for mobile services replacing their use of corresponding services with other devices) towards most services, the most so with regards to offline multimedia playback, MMS (replacing corresponding multimedia messaging clients in desktop computers, such as some IM clients) and mobile email. The second column (intention index), however, tells that most people have short-term interest to use only browsing, MMS, games and offline multimedia services. Indeed, the timing index is higher than 50% for these services. They are likely to generate usage right now. Some other services (having high attitude indices) might be successes further in the future according to the panelists, an example being e.g. mobile blogging which has a rather positive attitude index but low intention and consequently timing index. Also mobile VoIP (Internet calls) have promising prospects, as many people have positive long-term attitudes towards the service though no short-term interest (intention) to use the service exists. The timing index therefore reflects current propensity of taking service into use, given that there is a positive attitude towards the service. The fourth column projects the actual share of panelists having used the service during the panel, and finally adoption index in the fifth column states that only offline multimedia, browsing, MMS and gaming have been adopted without significant bottlenecks. All of these five indices can be measured over time. Longitudinal comparisons would provide visualizations on the development of services over their individual adoption curves.

The panelists themselves have opinions with regards to the key bottlenecks for new mobile services not to actually succeed in the market. All together 71 panelists were asked randomly (utilizing the pop-up questionnaire functionality in the research platform) why they think new mobile services do not survive in the market. Each respondent could pick only one answer, and the results are projected in figure four.

According to the panelists it is still pricing that has the most influence on actual usage. In other words, even though people have interests, too high a price level or suboptimal pricing structure might prevent them from using a particular service. The second most important thing the respondents identified as a bottleneck is technical implementation. Many services (according to the panelists particularly handset-embedded mobile email) are simply too difficult to use or configure, and this serves as a bottleneck. The research paper now turns to studying the adoption process in a more detailed manner by utilizing path modeling and comprehensive background data.

Path Analysis Model

Path analysis is used as a tool in verifying the theoretical hypotheses of this paper. Path analysis is an extension of multiple regression. Wright (1921; 1934; 1960) was first to utilize path analysis

in empirically studying direct and indirect effects of theoretical models. In essence, path analysis extends multiple regression by including a number of equations instead of only one (Schumacker & Lomax 2004).

Pedersen and Thorbjørnsen (2003) is used as a basis for the theoretical model developed here. Therefore factors perceived enjoyment, perceived usefulness, social push (called in other contexts as subjective norm) and behavioral control (i.e. perceived ease of use) are included to have either direct or indirect (mediated by attitude) effect on intention to use. The perceived expressiveness variable is not included in this model to keep it simpler. After all, it has not been included in all of the earlier TAM models. For the attitude variable a question was asked with regards to a particular service having already replaced or replacing potentially in the future the existing use of the service with e.g. computers or MP3 players. If people have a positive attitude towards the service, they answer positively to this question. Each factor is measured by a single question, listed in Appendix A. Due to limited data (and consequent unobserved uncertainty) the developed path model does not work that well by itself. However, the path models estimated for each of the services can still be compared against each other in analyzing the role of each factor in the context of a cross-service study. The original TAM model assumes that demographics (such as gender and age) and many other background variables are totally mediated out by the included independent factors in the extended TAM model. This simplifies the analysis.

Because of the variety of data available not only related to background variables but also on actual usage, the theoretical model is extended to include a factor reflecting actual usage (usage frequency of the service). Actual use is depicted to depend on intention to use, technical capability of the phone (the handset features serve either as a bottleneck or enabler), data pricing (lower marginal data prices should push services that generate data charging records), monetary pool available for smartphone usage (more monetary resources should have a positive impact on service usage that generate charging records) and user experience

Figure 4. Reasons for the failure of new mobile services

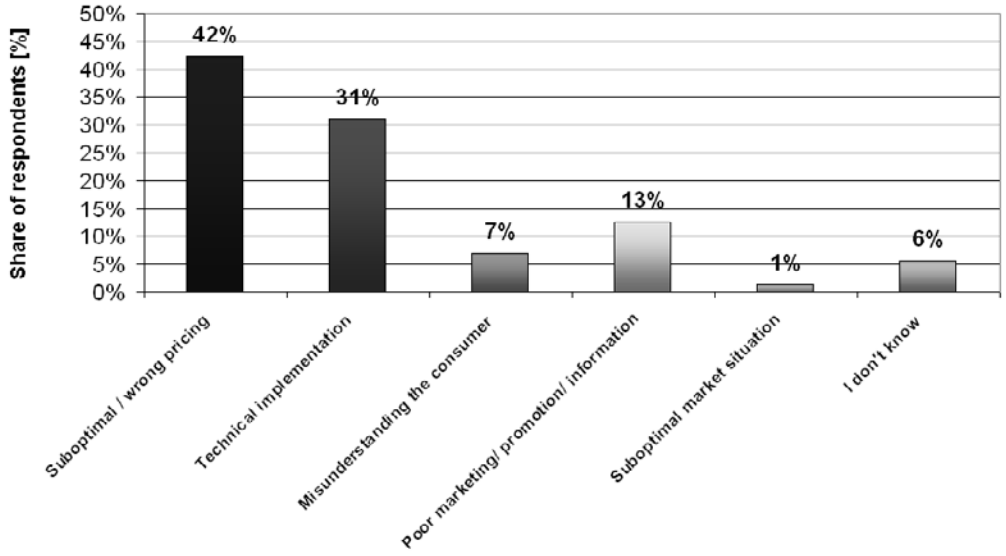
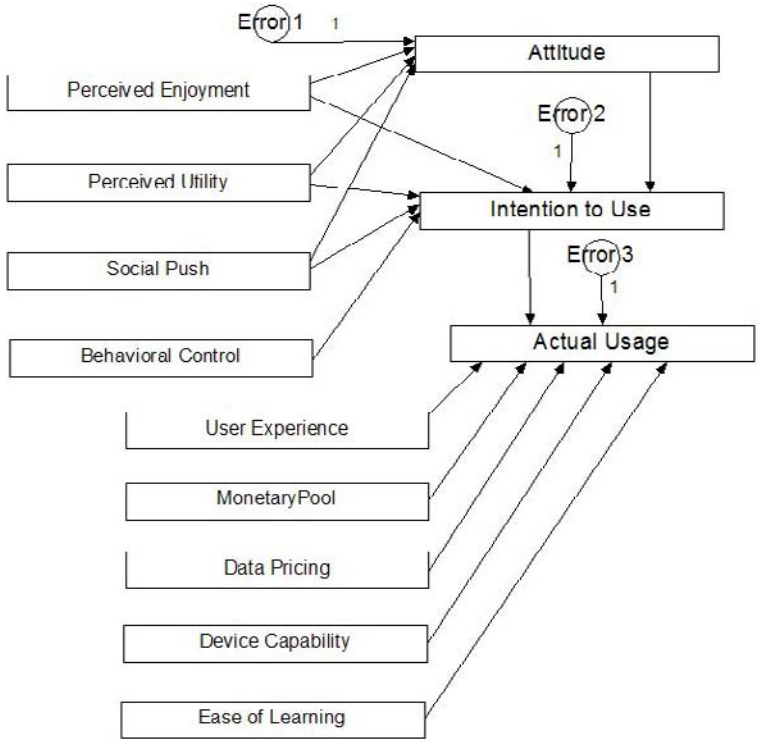


Figure 5. Developed theoretical path model of mobile service adoption



(more experienced users should be more likely to be capable of using the service). The technical capability of the phone is a boolean variable taking value 1 if the handset is based on Nokia S60 3rd edition platform (all new Nokia smartphones) and 0 otherwise (older Nokia S60 handset). Data pricing takes values 0 (no data plan, expensive usage-based charging), 1 (block-priced data plan, relatively cheaper to use data services) and 2 (flat-rate, unlimited usage with zero marginal cost). Monetary pool is proxied through ABPU (average billing per user) rates of subscribers, hypothesizing that high average bills of the past reflect the user's mobile service budget. User experience is here communicated through estimated average minutes of smartphone usage per day (that has strong correlation to user's experience according to a smaller sample of data).

One more variable is included to reflect a positive perceived easiness of learning and ease of getting help (i.e. facilitating conditions in actually using the service) for the service. This is projected to have a direct impact on actual use. The behavioral control variable introduced earlier better reflects the expected difficulty of use thus having implications in the initial intention to use, whereas the ease of learning component is hypothesized to be relevant if the initial interest towards the service exists and the problem is how to actually use the service given positive intentions.

The models are estimated with AMOS (add-on package for SEM in SPSS), using the maximum likelihood approach. Because there are missing data points (not all people answered the questionnaires), only panelists with full data are included.

All in all, 426 panelists are included in the path estimation.

Results of Path Analysis

The estimated path analysis diagrams for the chosen services are depicted in Appendix B. The figures include path coefficients (above the paths) and for dependent variables the share of variance explained by the model (above the box representing the variable). Some model fit indices are reported, too. First, a fit index CMIN/DF is included - the relative chi-square that represents how much the fit of data has been reduced because one or more paths have been dropped from the full model. Generally if this index is more than 3, too many paths have been dropped. The second included fit index is GFI (goodness of fit index). This tells how much of the variance in the sample variance-covariance matrix is accounted for by the model. Generally this should be above 0.9. Finally RMSEA (root mean square error of approximation) estimates lack of fit if compared to the full model. 0.08 or below is considered generally as adequate fit.

By looking at the results in Appendix B, the first observation is that the fit indices are quite poor. In other words, the developed model is perhaps not the best one to explain the involved complex relationships. GFI indices range from 0.82 (IM and email) to 0.86 (browsing). The second observation is that the model quite poorly explains the dependent variables under interest. The variance explained for the intention to use variable range from 0.05 (blogging) to 0.24 (gaming and MMS). The variance explained for the actual usage variable range from 0.01 (blogging) to 0.27 (gaming). Lots of uncertainty (e.g. non-internalized variables) exists that would explain intentions and actual usage, and the model therefore cannot depict all of the variance in the dependent variables. Almost all of the variables included in the framework model the demand-side, and therefore the missing (explanatory) variables are probably related to supply-side factors (e.g. service push, pricing and availability).

Without going into the details of each service (the estimated path models can be found from Appendix B), the table below summarizes the statistically significant path coefficients for each service.

It is surprising that perceived enjoyment has very strong impact on attitude for all services. People who think the enjoyment value is high (for any service) have high expectations that they are going to replace existing use of the service with the mobile version of the service (this is here called considered positive attitude). Perceived enjoyment positively (and statistically significantly) drives intentions directly, too. It can be generalized that the advanced mobile service market in Finland is building on expected hedonic benefits even in services which have been generally considered business-oriented (email).

Further, utilitarian benefits are important, too. The impact towards attitude is positive for all services expect for streaming multimedia, though the relationships are not as strong as with hedonic benefits. Interestingly the positive forward-looking attitude towards mobile instant messaging is strongly pushed by perceived utilitarian benefits. This confirms that instant messaging is a viable communication channel rather than a service meant for entertainment purposes solely. The direct impact of perceived utility on intentions is not that strong as the impact of perceived enjoyment (hedonic benefits), but the model nevertheless confirms that the intentions to use browsing, MMS, email (strong effect), blogging and IM are at least partly driven by utility-related expectations. Expectedly perceived utility does not drive offline multimedia playback, gaming or streaming multimedia usage intentions, because they are quite entertainment-oriented by nature.

Social push is all about the impact of supportive/driving social context with regards to usage attitude and/or intentions. Social push has positive impacts particularly on email and instant messag-

Table 2. Statistical significance of estimated paths

ing. This can be explained by the fact that these services are currently available, and the marginal impact of somebody in the close social neighborhood helping or recommending the service might be significant to drive attitude towards possible future use. The direct effect of the social push on intention to use is not very strong for any of the other services, communicating the fact that these early-adopter users are quite independent.

Behavioral control measures the user perception of herself being capable of using the service and to overcome technical challenges. Particularly for browsing and gaming this factor is a significant driver of intentions, whereas for most other services it has a positive though less significant effect. The statistical significance for the path originating from behavior control is generally not that strong. This might be due to the fact that Hejden et al. (2005) hypothesized that for hedonic services behavioral control does not have that strong an impact. For MMS there is no relation at all, as MMS is already well integrated into smartphones and thus there is little variance in end-user behavioral control.

The TAM model suggests that attitude towards the service should drive intentions. The attitude variable included in this research is geared towards current or future role of the mobile service possibly replacing existing use of the service with other devices (e.g. laptops or MP3 players). Therefore this variable is not reflecting general attitude in the purest sense. There is a statistically significant relation between attitude and intention only for mobile email. Positive mobile email attitude levels correlate strongly with positive intentions. People who think mobile email is or will be an important part of their smartphone usage in the future also have strong short-term intentions of using the service. For many other services the effect of attitude on intention is statistically insignificant. This can be explained by the fact that people generally perceive that these new advanced services are unavailable / technically poor, and therefore they lack short-term intentions of use though there is a positive general attitude. Another argument is relatively stronger effects of direct effects of background variables. In other words, rather than for many background factors having an indirect impact via positive attitude that the service would replace something (having a flavor that something else is sacrificed), for example perceived enjoyment or utility directly drive intentions and these relations are very strong.

When moving to actual use of services, it becomes clear that the variables hypothesized to have direct effect on realized usage are important. User experience (proxy through average usage minutes per day) has statistically significant relationships with actual usage of browsing, MMS, offline multimedia playback and gaming. These are services users have probably used already earlier with their smartphones, and therefore experience variable strongly drives usage. For some very new services such as email, IM or streaming the experience variable does not have a statistically significant effect. This might be because earlier experience relates to services that have been in the market already for some time (on which experience can accumulate), whereas experience does not help with regards to services that are just about to hit the market or are available later in the future.

Monetary pool has a strong effect on MMS usage, which has strong correlation with other mature service (such as voice and SMS) usage. In other words, higher consumption levels on mobile telephony do not drive any of the truly new service categories. First, some services are simply free and they do not depend on consumption levels or available monetary pool but on e.g. perceived enjoyment instead. Second, still most of the ABPU (average billing per user) rate consists of voice and SMS service charging, and therefore it does not reflect data service prices/consumption that well. All of the services that require network connectivity in this research are based on IP connectivity and are therefore independent of circuit-switched service pricing.

Data pricing (the lower marginal cost of using data services) has the most significant relationship expectedly with browsing usage. Browsing is currently the most important data (Internet) service. There is also weaker (but still statistically significant) effect on email usage. However, for other data services data pricing does not have an effect. This is partly due to little actual usage realized for other than browsing data service.

Interestingly device capability has the biggest influence on email and multimedia functionalities. The impact on multimedia is not a surprise, as if nothing else, new handsets include at least better multimedia functionalities (camera chips, memory capacity, MP3 player functionalities, better displays for movie/video playback) than older ones. The marginal email usage is coming from webmail oriented email usage. This is partly driven by better browsers included in newer handsets together with new high-resolution displays. This explains the path from device capability to email usage.

Ease of learning variable does not have an impact on actual usage at all. These early-adopter users, after all, all know that there is help available

for example in the Internet. On the other hand, they also tend to play around with the handset by themselves rather than think that they need somebody else for help, not to talk about manuals that are shipped with new handsets.

Finally, the most important variable in the model (intention to use) has statistically significant effects on all service usage expect for blogging. As concluded already earlier with descriptive statistics, people who have strong intentions to use the service tend to adopt the service (use it in practice) with a higher probability than those who do not have intentions. Blogging has not been used by that many, and therefore intentions cannot correlate with usage. Most blogging usage comes actually from Nokia's preinstalled Lifeblog application. Because of this it can be that those who explored blogging merely randomly launched Lifeblog, even though they did not have clear intentions. All in all, usage intentions are still the most important driver of advanced mobile service usage, therefore suggesting that the needs of end-users are the key issue in bringing new mobile services to the market instead of push-centric service introduction strategies.

A couple of reasons exist for the general poor fit of the estimated models. First, these are advanced services, and there are probably many random factors that lead into a usage event. The service must be in the market available for panelists and the technical performance/accessibility has to be in order. These are some of the variables the model cannot internalize, and this explains the poor share of variance explained for usage-level variables. Similarly intentions originate from many other factors (in advanced services) than those included in the model. For example, promotional events and availability of the service certainly do affect the short-term intention to use a service. These variables are not included in the model. In many cases these reasons explain the poor fit of the model. Indicative results suggest that the depicted theoretical model is not the optimal one. For example mobile blogging is a rather new service on which customers have very little or limited knowledge on. For this service most paths are insignificant and model fit is very poor. In addition, it is difficult to explain the variance of dependent variables if these variables have no or very little variance. Gaming is a well-known service and the market-level factors (e.g. operator competition, network performance, pricing, promotional events) have less impact. Therefore the theoretical adoption model receives much better results for gaming than for blogging, for example.

CONCLUSION

The paper utilized a newly developed handset-based mobile service research platform in deploying questionnaires and measuring actual usage of mobile services. This study differs from other handset-based research in its specific focus on the adoption process and associated measurements, utilizing also questionnaires. Descriptive statistics reveal that for most services not all mobile service demand (i.e. usage intentions) is fulfilled, i.e. not all panelists interested in services are actually using them. However, those strongly intending to use the service, ceteris paribus, expectedly have higher probabilities of adopting the service in practice.

The paper developed five indices that can be measured over time and therefore utilized in longitudinal analysis. These indices communicate the general attitude towards mobile services (attitude index), short-term intentions of usage (intention index), the current propensity of taking services into use (timing index), the extent of service usage (usage index) and the probability of using the service given intentions exist (adoption index). These indices communicate that panelists on average have positive attitudes towards mobile versions of the services included in the study, but only matured mobile services (introduction more than a year ago) experience positive short-term

intentions of usage and consequent adoption. Different services are clearly experiencing different phases of their life-cycle at this point of time. The developed indices can be utilized in projecting the phase of diffusion for each service emerging in the market.

The estimated path models reveal that perceived hedonic (enjoyment) benefit is the strongest driver of both attitude and intention towards the service. This is in contrast to Pedersen and Thorbjørnsen (2003) who found that utility-centric benefits drive intention to use mobile services. In the estimated models of this paper perceived utility drives intention to use only in the case of mobile email. In general, many of the newly developed mobile services should be considered as generating hedonic rather than business/utilitarian value to end-users. The social setting around panelists or expected technical difficulties do not explain intentions to use. This is most likely due to panelists in the dataset being early-adopters (quite independent and technically advanced users). User experience with smartphones explains actual use of services that have already been in the market for some time, but experience of the user does not explain the adoption of the most recent services. Pricing of data traffic has a strong effect only on browsing and email as they are the most visible mobile Internet services currently. Multimedia services, on the other hand, benefit from higher capability of handsets. The most important thing driving actual use of services, however, is the intention of end-users to use the service, and therefore user needs should be looked upon in the future instead of technology push strategies.

The estimated path models do not have very good fit. Instead of using standardized sets of questions and structural equation modeling in verifying the theoretical model, a more simplified path modeling exercise was chosen due to limitations in the panel study implementation. Improvement of the theoretical model remains as a future work task in addition to the acquisition of more comprehensive dataset supporting not only path but also measurement part of the SEM analysis. One of the key conclusions of this paper is, however, that the internalized variables of the estimated adoption model poorly explain the variance observed in intention and usage-level variables. Therefore most advanced mobile services have special characteristics and external sources of variance that were not observed in this analysis. These likely originate from the supply side of the market. Nevertheless, the demonstrated fitting of empirical data with path analysis was found to be an efficient method to compare the role of different variables in explaining the adoption of emerging mobile services.

Some mobile services are truly successful in the market, whereas some others are hyped quite a lot but few people actually adopt them. A need exists to acquire data on user opinions and service usage in order to understand the actual dynamics of mobile service adoption. This understanding is valuable in e.g. better commercializing mobile services and in improving mobile customer relationship management processes (for mobile CRM see Liljander et al. 2007). This paper showed that new handset-based data can help in better understanding the mobile service adoption process. The research differed from other handset-based research in focusing on the modeling of the adoption process, and combining efficiently questionnaire data with usage data. The paper introduced many new indices that reflect the different sides of the adoption process, and additionally demonstrated that different services can be compared against each other with the handset-based research approach. Future handset-based research should deploy wider questionnaires to make the statistical procedure of path modeling more reliable. In particular, the measurement part of the analysis should be deployed before estimating the structural part of the path model. In addition, a wider perspective towards the adoption process should be taken. For example, potential long-term usage should be better modeled against potential

short-term usage, and reasons why explorative usage sometimes does not lead to sustainable and repetitive usage should be looked upon.

REFERENCES

Aarnio, A. & Enkenberg, A. & Heikkilä, J. & Hirvola, S. (2002) Adoption and Use of Mobile Serivces - Empirical Evidence from a Finnish Survey. Proceeding of the 35th Hawaii International Conference on System Sciences 2002: 1454 - 1463.

Ajzen, I. (1985). From Intentions to Actions: A Theory of Planned Behavior. *In Action Control: From Cognition to Behavior*. Ed. J. Kuhl and J. Beckmann. New York: Springer

Ajzen, I. (1991). The theory of planned behavior. *Organization Behavior and Human Decision Processes*, 50: 179-211.

Bakalis, S. & Abeln, M. & Mante-Meijer, E. (1998) *Adoption and use of mobile telephony in Europe.* Cost 248 Report. Telia, Farsta, Sweden.

Burton-Jones, A. & Hubona, GS. (2005). Individual Difference and Usage Behavior: Revisiting a Technology Acceptance Model Assumption. *The DATA BASE for Advances in Information Systems - Spring 2005 (Vol. 36, No. 2): 75.*

Christensen, CM. (1997). *The Innovator's Dilemma*. Harvard Business School Press.

Chuang, MC. & Chang, CC. & Hsu, SH. (2001). Perceptual factors underlying user preferences toward product form of mobile phones, *International Journal of Industrial Ergonomics*, 27: 247-258.

Clark DD. & Lehr W. & Bauer SJ. & Faratin P. & Sami R. & Wroclawski J. (2006). Overlay Networks and Future of the Internet. *Journal of Communications and Strategies*, 3(63): 1-21.

Davis, FD. & Bagozzi, RP. & Warshaw, PR. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science. Vol 35: 982-1003.*

Davis, FD. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly , 13: 319-340.*

Fishbein M. & Ajzen, I. (1975). *Belief, attitude, intention and behavior: an introduction to theory and research.* Addison-Wesley, Reading MA.

Flanagin, AJ. & Metzger, MJ. (2001). Internet use in the contemporary media environment. *Human Communication Research* . 27: 153-181.

Funk, JL. (2004). *Mobile Disruption - the technologies and applications driving the mobile internet,* Wiley January 2004.

Gao, M. & Hyytinen, A. & Toivanen, O. (2006). Demand for and Pricing of Mobile Internet: Evidence from a Real-World Pricing Experiment. Helsinki Center of Economic Research Discussion Paper No. 122 (October 2006).

Green, N. & Harper, RHR. & Murtagh, G. & Cooper, G. (2001). Configuring the mobile user: sociological and industry views. *Personal and Ubiquitous Computing*, 5: 146-156.

Haddon, L. (2001). *Domestication and mobile telephony.* Presented at the conference "Machines that Become Us", Rutgers University, April 18-19, New Jersey.

Heijden van der, H. & Ogertschnig, M. & van der Gaast, L. (2005). Effects of Context Relevance and Perceived Risk on User Acceptance of Mobile Information Services. ECIS 2005.

Heinonen, K. & Pura, M. (2006). Classifying Mobile Services. Presented at Helsinki Mobility Roundtable, 1-2 June 2006, Helsinki.

Höflich JR. & Rössler, P. (2001). Mobile schriftliche Kommunikation oder: E-Mail für das Handy. *Medien & Kommunikationswissenschaft*, 49: 437-461.

Karlson, B. & Bri, A. & Link, J. & Lönnqvist, P. & Norling, C. (2003). *Wireless Foresight: Scenarios of the Mobile World in 2015*. John Wiley and Sons, 2003.

Kearney AT. (2001). *Satisfying the experienced*

on-line shopper: Global e-shopping survey, A.T. Kearney Ltd., Inc. London.

Kim, J. (2001). Analyzing mobile Internet users. Presented at the CHI 2001 Workshop on mobile communications.

Leung L. & Wei, R. (2000). More than just talk on the move: Uses and gratifications of the cellular phone. *J&MC Quarterly* , 77: 308-320.

Liljander, V. & Polsa, P. & Forsberg, K. (2007). Do Mobile CRM Services Appeal to Loyalty Program Customers? International Journal of E-Business Research, Vol. 3, Issue 2: 24.

Ling, R. & Yttri, B. (2002). Hyper-coordination via mobile phone in Norway. In *Perpetual contact*. Eds. J. E. Katz and M. Aakhus. New York: Cambridge University Press.

Ling, R. (2001). *"It is 'in.' It doesn't matter if you need it or not, just that you have it.":* Fashion and the domestication of the mobile telephone among teens in Norway. Oslo: Working Paper, Telenor R&D, Norway.

Mallat, N. (2006a). Consumer and merchant adoption of mobile payments. Doctoral dissertation. Helsinki School of Economics 2006.

Mallat, N. (2006b). Exploring Consumer Adoption of Mobile Payments. Presented at Helsinki Mobility Roundtable, 1-2 June 2006, Helsinki.

Meister, DB. & Compeau, DR. (2002). Infusion of Innovation Adoption: An Individual Perspective. *Proceedings of the ASAC, Winnipeg, Manitoba.*

Minges M. (2005). Is the Internet Mobile? Measurements from the Asia-Pacific region. *Telecommunications Policy 29 (2005): 113–125.*

Nokia Telecommunications. (1999). *The Demand for Mobile Value Added Services: A Market Study*, available at http://www.nokia.com/press/background/pdf/study-vas.pdf, 1999.

Nysveen, H. & Pedersen, PE. & Thorbjørnsen, H. (2005b). Explaining intention to use mobile chat services: Moderating effects of gender. *Journal of Consumer Marketing*, vol. 22, no. 5: 247-256.

Nysveen, H. & Pedersen, PE. & Thorbjørnsen, H. (2005a). Intentions to Use Mobile Services: Antecedents and Cross-Service Comparisons. J*ournal of the Academy of Marketing Science*, vol. 33, no. 3: 330-346

Pedersen, PE. & Thorbjørnsen, H. (2003). Adoption of Mobile Services. Model development and Cross-Service study. Norwegian School of Economics and Business Administration.

Pedersen, PE. (2001). An adoption framework for mobile commerce. *Proceedings of the 1st. IFIP Conference of E-Commerce*, Zürich, Switzerland, October 3-5, 2001.

Pedersen, PE. (2005). Adoption of mobile Internet services: An exploratory study of mobile commerce early adopters. *Journal of Organizational Computing and Electronic Commerce*, vol. 15, no 3: 203-221

Pedersen, PE., Methlie, LB. and Thorbjørnsen, H. (2001). *Understanding mobile commerce end-user adoption: a triangulation perspective and suggestions for an exploratory service evaluation framework* . Presented at HICSS-35, Hawaii, US, Jan 7-10, 2002

Rask, M. & Dholakia, N. (2001). Next to the customer's hart and wallet: Frameworks for exploring the emerging m-commerce arena. *AMA Winter Marketing Educator's Conference*, Vol. 12: 372-378.

Rogers, EM. (1962). *Diffusion of Innovation*. New York, The Free Press..

Rogers, EM. (1995). *Diffusion of innovations.* (4. edition). New York, The Free Press.

Saarikoski, V. (2006). *The Odyssey of the Mobile Inter*net. Doctoral Dissertation. University of Oulu, Finland.

Schumacker, RE. & Lomax RG. (2004). A Beginner's Guide to Structural Equation Modeling. 2nd Edition. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Sigurdson, J. (2001). WAP OFF - Origin, Failure

and Future. Prepared for the Japanese-European Technology Studies (JETS), October 9, 2001, University of Edinburgh.

Skog, B. (2000). "The mobile phone as symbolic capital in teenage cultures" (in Norwegian). In *Social consequences of mobile telephony* (in Norwegian). Eds. R. Ling and K. Thrane. Proceedings from a seminar on society, children, and mobile telephony, Telenor FoU - Notat nr. 38/2000.

Skog, B. (2002). Mobiles and the Norwegian teen: identity, gender and class." In *Perpetual contact*. Eds. J. E. Katz and M. Aakhus. New York: Cambridge University Press.

Taylor, S. and Todd, PA. (1995). Understanding information technology usage: a test of competing models. *Information Systems Research*, 6: 144-176.

Tseng, FC. Teng, CI. & Chiang, DM. (2007). Delivering Superior Customer Perceived Value in the Context of Network Effects. International Journal of E-Business Research, Vol. 3, Issue 1: 41.

Venkatesh, V. & Davis, FD. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science,* Vol. 46, No. 2: 186-204.

Verkasalo H. (2005). *Handset-Based Monitoring of Mobile Customer Behavior*. Master's Thesis Series. Networking Laboratory. Department of Electrical and Telecommunications Engineering. Helsinki University of Technology.

Verkasalo, H. & Hämmäinen, H. (2006). Handset-Based Monitoring of Mobile Subscribers. Presented at Mobility Roundtable, 1-2 June, 2006, Helsinki, Finland.

Verkasalo, H. & Hämmäinen, H. (2007). A Handset-Based Platform for Measuring Mobile Service Usage. *INFO: The Journal of Policy, Regulation and Strategy. Vol 9 No 1, 2007*.

Verkasalo, H. (2006a). Mobile Data Service Evolution - Empirical Implications from Europe and the USA, in 3rd International CICT Conference, November 30 - December 1, 2006, Denmark, 2006.

Verkasalo, H. (2006b). Empirical Observations on the Emergence of Mobile Multimedia Services in the U.S. and Europe, in The 5th International Conference on Mobile and Ubiquitous Multimedia. December 4-6, 2006, Stanford University, California, 2006.

Verkasalo, H. (2007a). *A Cross-Country Comparison of Mobile Service and Handset Usage*. Licentiate's thesis, Helsinki University of Technology, Networking Laboratory, Finland.

Verkasalo, H. (2007b). Empirical Insights on the Evolution of the Finnish Mobile Market. Presented at Conference on Telecommunication Techno-Economics (CTTE) 2007, 14-15 June 2007, Helsinki, Finland.

Verkasalo, H. (2007c). Empirical Findings on the Mobile Internet and E-Commerce. Presented at 20th Bled e Conference: eMergence: Merging and Emerging Technologies, Processes, and Institutions. Bled, Slovenia, June 4 - 6, 2007.

Verkasalo, H. (2007d). Contextual Usage-Level Analysis of Mobile Services. Accepted for publication at The 4th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MOBIQUITOUS 2007). August 6-10, 2007 - Philadelphia, PA, USA.

Vesa, J. (2005). *Mobile Services in the Networked Economy*. IRM Press, Hershey, PA, USA.

Wong, CC. & Hiew, PL. (2005). Correlations between factors affecting the diffusion of mobile entertainment in Malaysia. ICEC 2005: 615-621.

Wright, S. (1921). Correlation and causation. Journal of Agricultural Research, 20, 557-585.

Wright, S. (1934). The method of path coefficients. Annals of Mathematical Statistics, 5, 161-215.

Wright, S. (1960). Path coefficients and path regression: Alternative or complementary concepts? Biometrics, 16, 189-202.

APPENDIX A: QUESTIONS ASKED IN THE BACKGROUND QUESTIONNAIRE

Background factors are recoded into the following scale:

- ABPU... 1 (small) - 5 (big) (asked before the panel)
- Work... 0 (not) - 1 (yes) (asked before the panel)
- Age... 0 (young) - 1 (old) (asked before the panel)
- Gender... 0 (women) - 1 (men) (asked before the panel)
- Device_Capability... 0 (old device = lower capability) - 1 (new device = higher capability) (asked before the panel)
- Data_Pricing... 0 (usage-based), 1 (block-priced), 2(flat) (asked before the panel)
- Experience (of smartphone usage in the user's own opinion)... 1 = Beginner, 2 = Normal, 3 = Experienced, 4 = Very experienced (asked before the panel)
- Panel_Days... Number of days observed active in the panel (derived from usage data)
- Handset_Usage_Activity... Number of application activations during an average day (derived from usage data)
- Smartphone_Usage_Day... Minutes of smartphone usage on an average day (derived from usage data)

Questions are asked in relation to the following mobile services:

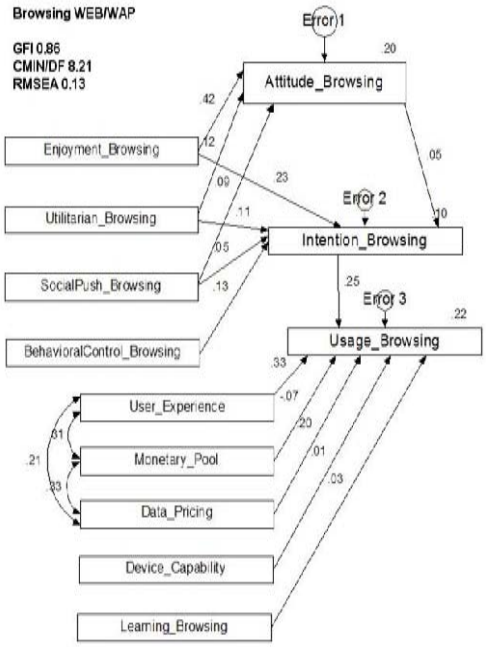
- WEB/WAP services, Internet browsing
- Email
- Instant messaging / chat
- MMS messaging
- Gaming
- Internet telephony
- P2P file sharing
- User-created content creation, blogging
- Streaming/Internet multimedia
- Offline multimedia

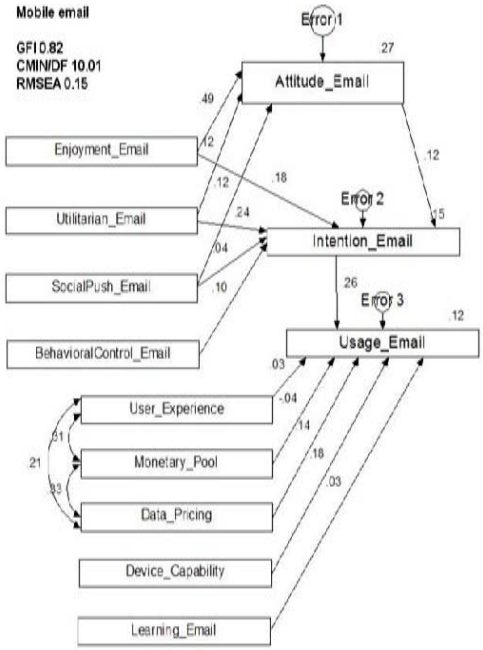
1. I intend to use the following services during the next couple of months… (intention to use) (asked before the panel)
2. Mobile versions of the following services are going to replace / have already replaced the use of those services with other devices (such as PCs, MP3 players, Radio)... (attitude) (asked after the panel)
3. The use of the following services generates enjoyment, pleasure and entertainment to me... (enjoyment value) (asked after the panel)
4. The use of the following services increases my work or study related productivity and performance... (utilitarian value) (asked after the panel)
5. I do not have / would not have significant technical difficulties in using the following services... (behavioral control) (asked after the panel)
6. It is easy for me to learn and develop my skills in using the following services... (ease of learning) (asked after the panel)
7. People around me (e.g. friends or family) have recommended the use or have helped me in using the following services... (social push) (asked after the panel)

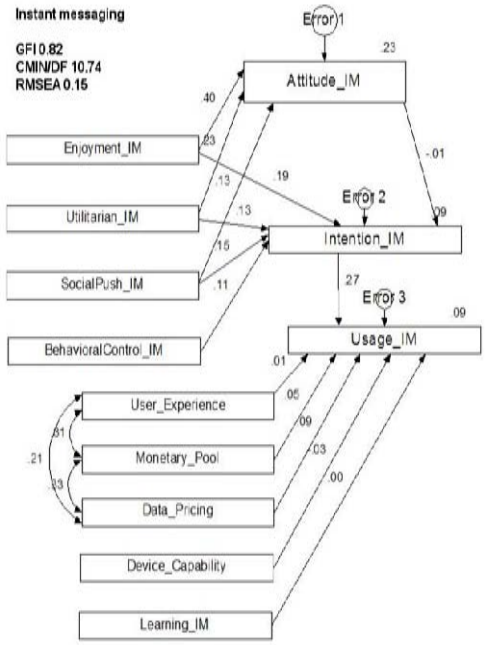
Likert scale answers:

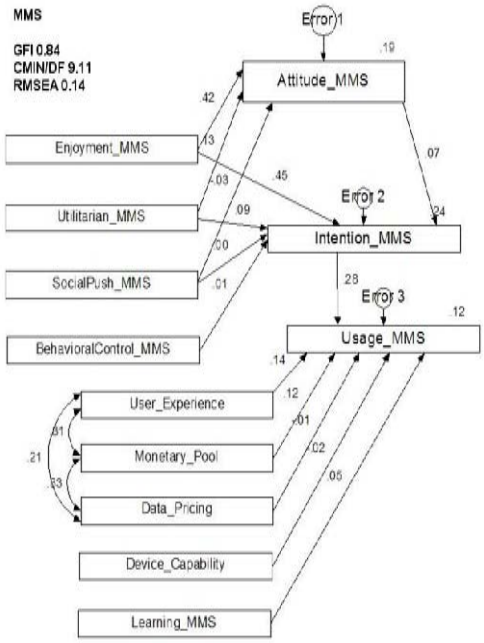
1 = Strongly disagree, 2 = Disagree, 3 = Slightly disagree, 4 = Neutral, 5 = Slightly agree, 6 = Agree, 7 = Strongly agree

APPENDIX B: RESULTS OF THE PATH ANALYSIS

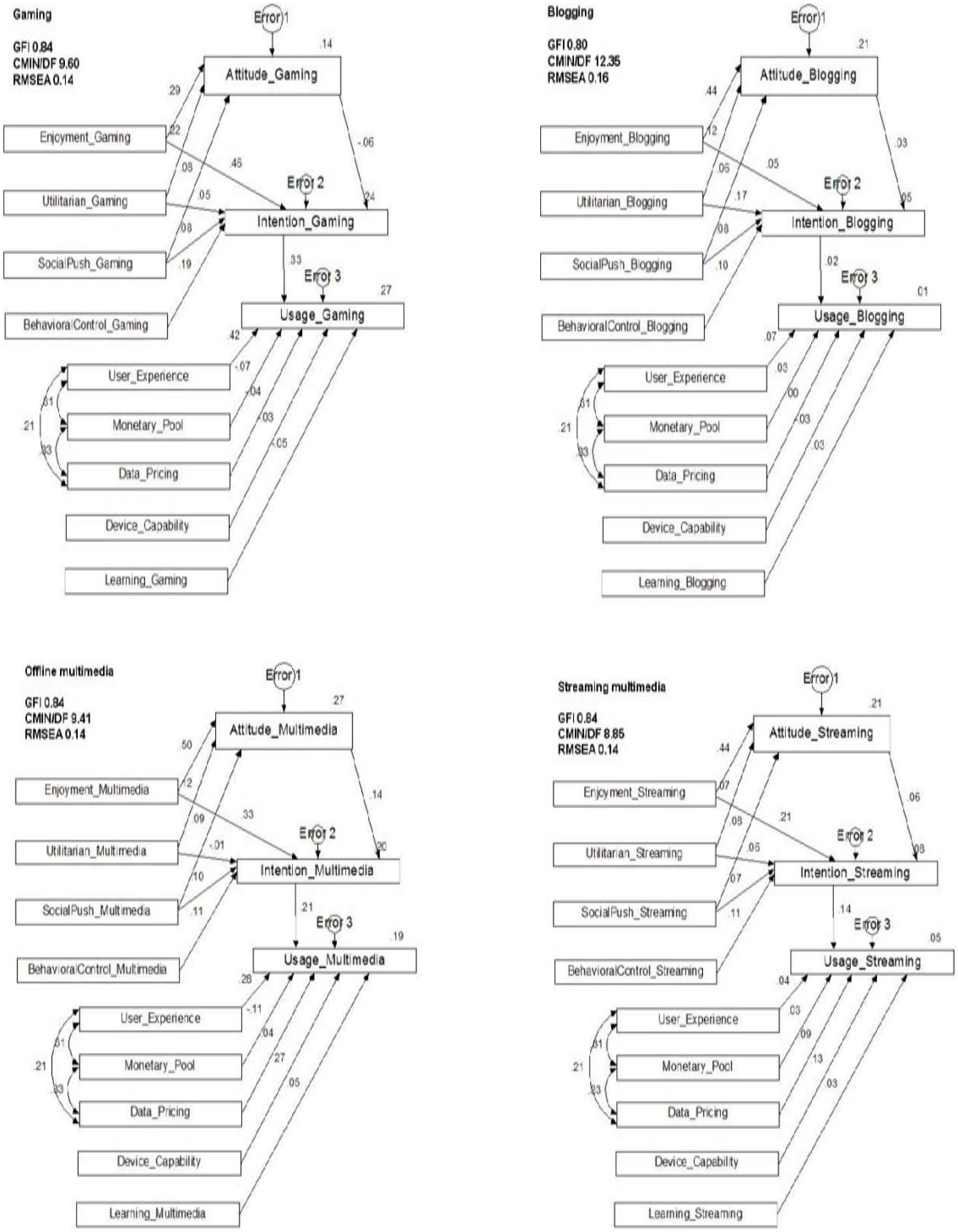








APPENDIX B: CONTINUED



This work was previously published in International Journal of E-Business Research, Vol. 4, Issue 3, edited by I. Lee, pp. 40-63, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.2
Exploring the Use of Mobile Data Services in Europe:
The Cases of Denmark and Greece

Ioanna D. Constantiou
Copenhagen Business School, Denmark

Maria Bina
Athens University of Economics and Business, Greece

ABSTRACT

Mobile data services seem a promising revenue source for the stakeholders that have heavily invested in mobile communications infrastructures. However, in the Western world those services have not reached the mass markets yet. This chapter focuses on two markets that are representative of the European socioeconomic environment, Greece and Denmark, with the aim to investigate and compare mobile data services use through the means of an online survey. We depict two user groups and observe their behaviors in both countries. The results indicate differences in locations and frequency of services use that can be attributed to specific socioeconomic characteristics. However, certain similarities exist in the experience derived from mobile data service use since users believe that services cannot cater to their specific needs. We conclude by underlining the current challenges faced by mobile service providers in order to increase mobile data services use and summarizing user groups' characteristics.

INTRODUCTION

Mobile communications markets have been in the spotlight for the last decade due to the impressive increase of users in most countries around the world. This dynamic environment led practitioners and academics to numerous predictions and speculations for the industry's growth potential. Many mobile operators, in pursuit of high returns on investment, upgraded their network infrastructures (e.g., third generation [3G]). They expected that this would stimulate high demand for advanced mobile services similar to those used

on the Internet, such as infotainment content and e-commerce transactions. However, these hopes have not been materialized in the Western world yet. It seems that mobile users are not interested or willing to massively adopt and extensively use the mobile data services (MDS) offered.

In the IS research, there has been considerable attempts to explore, understand, and analyze diffusion of innovation and technology acceptance through models and theories that offer constructs or factors affecting adoption and use based on individuals' expectancies and attitudes (Venkatesh, Morris, Davis, & Davis, 2003). This research domain has inspired many academics who have applied these models and theories in mobile communications markets in order to interpret or predict future trends and identify drivers and inhibitors for MDS adoption. In these research efforts, consumer behavior has been analyzed using conceptual frameworks inspired by the *technology acceptance model* (TAM) (Davis, 1986) or *diffusion of innovation theory* (DoI) (Rogers, 2003). Moreover, it has been pointed out that a cross-disciplinary integration of the four research directions on diffusion, adoption, uses, and gratification as well as domestication, was needed for understanding individuals' mechanisms for adopting MDS (Pedersen & Ling, 2003).

Following the same reasoning, Carlsson, Carlsson, Hyvønen, Puhakainen, and Walden (2006) applied a modified TAM model as an explanatory framework for mobile devices/services adoption. In addition, Lu, Yao, and Yu (2005) extended the TAM for wireless Internet adoption by incorporating concepts such as social influences and personal traits. Similarly, Wu and Wang (2005) enriched TAM with constructs regarding perceived risk, cost, and compatibility, while Yang (2005) added individual characteristics, such as innovativeness, past adoption behavior, and knowledge about technology, as external antecedents of TAM constructs. Furthermore Kim, Chuan Chan, and Gupta (2005) developed the *value-based adoption model* including benefits as well as sacrifices in the adoption process of mobile Internet, whereas Massey, Khatri, and Ramesh (2005) identified technology readiness and wireless Web site interface usability as key factors influencing the uptake of mobile commerce and services. This indicative, though non-exhaustive, listing of research efforts underpins how the domain of MDS adoption and diffusion has acquired an important position within the research agenda of IS researchers.

Overall, past research has mainly focused on investigating in detail the impact of several attributes on MDS adoption and use. However, in this emerging market landscape, it becomes pivotal to understand the types of existing users and their perceptions of MDS use. At this point, it is important to keep in mind that the MDS market is a voluntaristic setting, where the potential user chooses to adopt or not the service based on his/her individual perceptions about it. Thus, different behavioral characteristics and experiences of the users may affect usage patterns (Constantiou, Damsgaard, & Knutsen, 2007). Moreover, exploring different end-user types and behavioral patterns may provide a comprehensive market segmentation that can be used to improve matching of consumer needs to appropriate service offerings. In this line, there have been few research attempts to categorize mobile users based on characteristics such as their demographics and level of innovativeness (Constantiou, Damsgaard, & Knutsen, 2004).

In this chapter, we draw on earlier categorization research and focus on exploring users' perceptions and experiences from MDS. To this end, we adopt the categorization proposed by Constantiou, Damsgaard, and Knutsen (2005) and investigate MDS users in two European countries, namely Denmark and Greece. Our comparisons are based on the results from a global survey on MDS use that was conducted during November 2004. In particular, we apply the proposed categorization on two samples representing two countries with different information and communications technologies (ICT) market structures.

The chapter's contribution is twofold. First, it provides suggestive evidence on the applicability of the proposed categorization in different economic contexts. Second, it facilitates the uncovering of similarities and differences on MDS usage patterns between user categories in the two countries that can be interpreted based on national socioeconomic characteristics. Thus, we offer useful insights to both researchers in the mobile domain, by underlining the importance of socioeconomic context in the use of MDS, as well as the key players in mobile market arena, by informing their marketing campaigns and corporate strategies.

The remainder of this chapter explores mobile users' perceptions and experiences from MDS use. The next section offers a brief review of user categorization research in mobile markets and introduces the segmentation applied in this study. The following section describes the research design and depicts the MDS adopter segments and the hypotheses regarding inter-group similarities and differences. The empirical data is analyzed and discussed focusing on the group profiles, the MDS usage patterns, and the experiences in the next section. Finally, the chapter concludes with a summary of the resulting MDS adopter categories and highlights on the research and managerial insights offered by this research.

RESEARCH ON THE DEMAND SIDE OF MOBILE MARKETS

Research Efforts on User categorization in Mobile Markets

DoI explores adoption as a direct result of the five elements that characterize an innovation: (1) relative advantage, (2) compatibility, (3) complexity, (4) trialability, and (5) observability. These elements relate to the relative rate of adoption in a social system (Rogers, 1995). In turn, the rate of adoption refers to the relative speed with which members of a social system will adopt a new idea. In addition, innovativeness is a behavioral characteristic of individuals that relates to the time it takes to adopt an innovation relative to others in a social system.

Rogers has identified five categories of adopters: (1) innovators, (2) early adopters, (3) early majority, (4) late majority, and (5) laggards, all of whom co-exist in a social system. These groups are characterized by different socioeconomic profiles and are defined based on their degree of innovativeness (Rogers, 2003). The categorization of the adopters' population into different groups has a pivotal role within the context of diffusion of innovation research (Wolfe, 1994) since it is commonly believed that a "copy-behavior" or "imitation principle" guides the adoption process (Rogers, 1995): from innovators to early adopters and so forth. However, Rogers' categorization is not immaculate; it suffers from the "individual-blame bias" (2003) and it implies that innovators and early adopters benefit more from the adoption than the early and late majority, or the laggards. Furthermore, researchers have questioned the validity of Rogers' prescribed adoption pattern under the light of the current ICT environment; for example, Moore (1999) introduced "The Chasm," a critical stage between early adopters and the majority that needs to be crossed for an innovation to be massively adopted. Moore's chasm was introduced to explain why adoption for many innovations suddenly stops at the early adopters' category (e.g., wireless application protocol [WAP]).

Moreover, in the case of communications services that exhibit network effects (Katz & Shapiro, 1985, 1992), the value for users increases almost exponentially with the number of other users following Metcalfe's law (Shapiro & Varian, 1999). Additionally, value increases by indirect network effects generated from complementary offerings available over the network (Katz & Shapiro, 1992). Thus, whether innovators and early adopters are benefited more than all others,

as posited by Rogers, becomes quite doubtful. For example, the value of short message service (SMS) due to connectivity increased slowly with the first adopters, but then as the number of connected users grew it increased substantially. Furthermore, the wide use of SMS creates business opportunities for content providers to offer new services through SMS (e.g., voting for a contest), which in turn generates extra value to mobile users through indirect network effects. In sum, the value of mobile communications is not only attributable to the offering as such but also to the user's benefits from network effects. In the light of network effects, the role of innovators and early adopters according to Rogers is very important from an economic and business perspective. These groups may constitute the critical mass (Carter, 1998) enabling MDS to take off faster.

Although Rogers' categorization is not explicitly used in current MDS research, it is reflected in a number of recent studies. Gilbert and Kendall (2003) collected data from Singapore and Malaysia and outlined five adopter categories based on their intention to use WAP services, their specific service requirements as well as their demographics. In particular, "mobile professionals" required services useful in relation to work such as calendaring, e-mail and access to intranet/extranet; "sophisticates" focused on material style services; "socialities" were more interested in interpersonal contact; "technotoys" were driven by a need to know technological developments hands-on; and "lifestylers" focused on the always-mobile way of living. In addition, two segments unlikely to adopt mobile services were also identified; "misers" were the ones unwilling to pay, while "laggards" were the last to know and adopt new technologies. In recent research Gilbert and Han (2005) combined findings from adopter categorization of mobile gaming and entertainment services to dedicated, social, and casual gamers with cross-country comparisons of the five adopter categories described in Gilbert and Kendall (2003) to delineate a dynamic needs-based segmentation of the MDS adopter population. The proposed segmentation separates five adopter groups (mobile professionals, sophisticates, socialities, technotoys, and lifestylers) based on their degree of expectation for intrinsic or extrinsic rewards from using MDS and their individuality, or collectivity when learning and taking up a new technology.

In the same line, Aarnio, Enkenberg, Heikkilä, and Hirvola (2002) identified five adopter categories that had all adopted SMS text messaging and e-mail. In their research, advanced forms of SMS services were used by "innovative opinion leaders" representing 12% of the sample, "early adopters" (14%) and "late adopting students" (40%). Moreover, the "innovative opinion leaders" were the only category using WAP and data transmission services. Mort and Drennan (2005) combined evidence from a careful screening of related literature and empirical data on MDS usage patterns to identify five groupings of MDS adopters: "innovators" who tend to adopt technological innovations like MDS earlier than all others, "techno-confidents" who have high mobile self-efficacy and belief in their ability to use MDS effectively, "shopping lovers" who have strong emotional orientation towards shopping, "belonging seekers" who tend to use MDS as a vehicle for receiving others' approval, and "consulters" who tend to ask friends or significant others for advice for using communication services such as pictures, SMS, and multimedia message service (MMS). Okazaki (2006) offered an alternative approach based on a two-step procedure for identifying clusters of mobile Internet adopters. Both demographic and attitudinal variables describing individual mind-sets towards MDS innovation features were used as input to the analysis, which resulted to the uncovering of young people as the core segment exhibiting a strong positive inclination for taking up MDS.

Furthermore, a European study conducted in Finland, Germany, and Greece during 2001 (Vrechopoulos, Constantiou, Mylonopoulos,

& Sideris, 2002; Vrechopoulos, Constantiou, Sideris, Doukidis, & Mylonopoulos, 2003) proposed a basic two-group classification: "mobile users"—those using traditional (e.g., voice services) or "more sophisticated" data services (e.g., MMS)—and "mobile shoppers" those using at least one of the content (information and news) or transaction (physical goods or services such as travel, ticketing, banking, or entertainment purchasing) services. These two groups were found to differ on perceptions of the importance of key attributes affecting consumer behavior such as price, service quality, user interface, security, and personalization.

This research was followed up in Denmark during 2004 through a survey focusing on investigating technology and service use, innovativeness as well as technology service requirements of mobile users (Constantiou Damsgaard & Knutsen., 2004, 2006). The outcome was a broad segmentation between "basic" and "advanced users." "Basic users" were found to have matured in terms of their perceptions of the key attributes affecting MDS use and did not differ significantly from "advanced users." However, significant differences between the two user groups were traced in terms of innovativeness and technology-service requirements. Thus, it becomes clear that studying the adoption pattern of an innovation like MDS, developed and marketed within the complex and continuously evolving ICT landscape, involves improving our knowledge on the different categories of MDS users and potential users and the ways they interact with each other to allow the diffusion of MDS to take place.

The Applied Market Segmentation

To deepen their understanding of mobile users, Constantiou et al. (2007) refined their early broad segmentation and provided a more detailed categorization of mobile users. Accordingly, four mobile user categories were delineated based on the learning steps taken towards specific service use:

1. **Talkers:** Users of voice services only
2. **Writers:** Users of SMS in addition to voice services
3. **Photographers (PH):** Users of MMS services in addition to voice and SMS
4. **Surfers (SU):** Users of data services in addition to SMS, MMS, and voice services

Consequently, "talkers" have taken a primary learning step in terms of mobile communications use, whereas "writers," "photographers," and "surfers" have experienced one, two, or three additional changes in their behavior. Constantiou Damsgaard & Knutsen. (2005) concluded that the groups were different in terms of innovativeness, technology, and service use as well as technology service requirements. These results underline the differences observed in terms of learning steps taken by each group. In this chapter we focus on the two latter categories, namely, "photographers" and "surfers." These two categories consist of MDS users who are the main target group of mobile service providers in order to generate high revenues (Constantiou Damsgaard & Knutsen, 2007).

THE RESEARCH APPROACH

The Research Context

In this chapter we investigate and compare the behaviors of MDS users originating from two European countries, namely; Greece and Denmark. Demark is among the most advanced European countries in mobile communications (e.g., Denmark holds the first place in the Information Society Index). In particular, broadband penetration in Denmark was 25% (Organisation for Economic Co-operation and Development [OECD], 2005)

and Internet use reached 75% of the population during 2005 (Eurobarometer, 2006). Besides, the Danish mobile communications market is one of the most progressive in terms of liberalization (e.g., 13 mobile operators resulting in extensive price wars on contracts for SMS and voice services) and has high mobile penetration (95%) (i.e., SIM cards as percentage of the population).

Greece is less advanced in the ICT sector. It exhibits low broadband penetration of 1,4% (OECD, 2005) and low Internet use at home (25%) (Eurobarometer, 2006). Nevertheless, mobile communications penetration has reached saturation with 109% (i.e., SIM cards as percentage of the population) though the market is characterized as an oligopoly with only four mobile network operators (Kopf, 2005).

We believe that those two countries are useful examples of dynamic and mature mobile communications markets that have different structures. Additionally, they have differing ICT market growth rates, both of which are representative within the European context. Furthermore, they represent two traditional geographic poles: the Scandinavian and the Mediterranean territories. Thus, we can investigate emerging trends in MDS use that may be prominent and repeated in other European markets as well.

The Survey Instrument

Empirical data for the research presented in this article were collected within the 2004 Worldwide Mobile Data Services Survey (WMDS), a global Web-based survey designed to explore consumer behavior and the market environment for MDS around the world through cross-cultural and longitudinal trend analysis. WMDS is a rigorous academic study conducted annually since 2001. Ten countries (Australia, China, Denmark, Finland, Greece, Hong Kong, Japan, Korea, Taiwan, and USA) took part in WMDS during 2004. The survey theme was the effects of quality of life on MDS use. All countries are expected to use a jointly developed core set of questions on their surveys to facilitate cross-comparison of results, while each country is also free to include additional questions of local significance if required.

The Danish and Greek research teams adopted a common approach that led to the design of a survey instrument including 31 questions organized in different categories such as mobile communication usage patterns, mobile data services use, pricing considerations, assessment of quality of life, and demographics. Furthermore, both research teams followed similar processes for pilot testing (trials conducted using university staff) and revising the questionnaire before its official launch on the Internet. The survey ran during November 2004.

Our target groups were respondents that were MDS users and had prior experience with networked technologies and services such as the Internet and the use of online consumer services. In particular, we focus on Internet users since this group can be depicted as MDS early adopters (Constantiou Damsgaard & Knutsen, 2007). Moreover, according to Carter (1998) and Rogers (2003) early adopters are the target groups that should be addressed with specific strategies at the introductory phase of a new product or service in order to create the critical mass. Thus, we explore a sample including respondents with an early adopter profile to identify behavioral patterns that may significantly affect the diffusion process.

The samples are not representative of the total Danish or Greek populations since they include self-selected Internet users who are also mobile users. According to Hair, Bush, and Ortinau (2000), as well as Kinnear and Taylor (1996), self-selected sampling is suitable for exploratory research and when *ex ante* knowledge of the population characteristics is not sufficiently present. The sample is influenced by Internet and mobile penetration as well as the advertising effort put up by the two research teams that included a balanced mix of press announcements and Internet pages hosting links to the survey (e.g., information

portals and university Web sites). The resulting sample consists of 673 and 683 usable responses in Denmark and Greece respectively. Respondents who have used MDS at least one reached 78% of the Danish sample and 57% of the Greek sample. This suggests that MDS penetration has been faster in Denmark than in Greece. It remains to be seen whether this penetration is associated with sophisticated usage patterns in terms of frequency of use or it has yet to be transformed to an actual users' pool.

The Sampling Categorization

The available samples in both countries encompass all mobile users who have or have not used MDS. However, for the purposes of this research, we focus only on those mobile users who have at least once in their life used one or more of the commercially available MDS in both countries—in total, 42 services.

The statistical analysis performed in both data sets denotes low penetration scores for the majority of MDS categories. In particular, in a scale from 1 (never) to 7 (very often), the most popular MDS in Greece is taking photos and/or sending them for printing through MMS that scores a mean value of 3.44, while among Danish respondents the most popular service is receiving weather forecasts with a mean score of 2.26. Our research focuses on the top-scoring MDS in both data sets. In other words, our focus is on the services whose mean value is greater than 2.0 (Table 1). It appears that MDS market penetration is clearly very low in both countries. Moreover, Table 1 indicates that Greek and Danish consumer preferences for MDS exhibit similarities. Both top-ranking MDS are related to entertainment- and communication-oriented services, some of which are mobile extensions of Internet services (e.g., e-mail). This observation may be valuable for mobile marketing strategy design at a cross-national level. In addition, the strong preferences in both countries for logos, wallpapers, and ringtone-related services are yet another demonstration of the increasing importance that these services have acquired as means of expressing individual identity, inner values, and emotions (Hjorth, 2005).

Table 1. Top-ranking MDS in Greece and Denmark

Rank	Greece	Denmark
1	Taking photos and sending them to a third party for printing through MMS	Getting weather updates
2	Updating logos/wallpapers and/or ringtones	Updating logos/wallpapers and/or ringtones
3	Downloading logos/wallpapers and/or ringtones	Downloading logos/wallpapers and/or ringtones
4	Searching for infotainment content (movie updates, restaurants, concerts, etc.)	Exchanging e-mails through the mobile phone with colleagues
5	Getting weather updates	Searching for infotainment content (movie updates, restaurants, concerts, etc.)
6	Exchanging e-mails through the mobile phone with colleagues	Exchanging e-mails over my mobile phone with friends
7	Exchanging e-mails over my mobile phone with friends	Downloading updates for the software or firmware installed on my mobile phone
8	Downloading games for my mobile phone	Exchanging e-mails over my mobile phone with family members
9	Downloading and/or listening to music on the mobile phone	Taking photos and sending them to a third party for printing through MMS

Adhering to the categorization in "photographers" and "surfers" proposed by Constantiou et al. (2005) we formed an initial hyperset of MDS users by selecting all respondents who have used one or more of the proposed MDS at least once in their life (i.e., corresponding values of the MDS greater than 1). Photographers were found within the initial hyperset by selecting respondents who have used at least one in the MMS-related service, while their use of the remaining MDS is seldom (value < 3). Surfers are frequent users of MDS. They were found within the initial hyperset by selecting respondents whose frequency of use of MDS scores values greater than 3. Thus, we delineated four groups including "photographers" and "surfers" in both Denmark and Greece.

The Hypotheses

After the identification of the four groups, the next step in our research involved investigating their differences in terms of frequency of MDS use in various locations, purpose of MDS use, and experience from using MDS by formulating the corresponding hypotheses. According to the research efforts in the mobile domain, these behavioral characteristics of users are pivotal in understanding MDS market adoption and use dynamics.

Recent research has shown that the purpose of MDS use is different among categories of users and this may affect their adoption and usage patterns (Aarnio et al., 2002; Anckar & D'Incau, 2002; Gilbert & Kendall, 2003). A major dichotomy in terms of MDS use is between functional uses of mobile services, with the objective to increase work productivity, and efficient allocation of time during the day and entertainment, or hedonic use, among individuals who perceive the mobile device and the services enabled over it as an entertainment or lifestyle tool. In fact, Lee, Kim, Lee, and Kim (2002) have empirically shown that cultural differences can be at the heart of this differentiation in MDS purpose of use. We asked respondents to indicate on a 1-7 scale whether they use MDS more for personal- or business-related purposes. Thus, we investigated respondents' purpose of MDS use by setting the hypothesis:

H_0^I: *There are no significant differences between the four groups in their purpose of MDS use.*

Moreover, there is a large stream of research focusing on location-based services (Rao & Minakakis, 2003) that underlines the role of physical locations in the adoption of MDS. A broad classification of physical locations includes school/university, work, home, public places (e.g., streets and shops), and transit or transportation means (including waiting time). Recent research efforts on consumer choices of MDS by Blechar, Constantiou, and Damsgaard (2006b) underlined the importance of physical location on actual service use. In particular, the individual's decision to use a service may be affected by his/her physical location. For example, when the user has available Internet access (i.e., school, work), he/she may choose this infrastructure over mobile networks to access a service (e.g., e-mail, news). We asked respondents to mark on a 1-7 scale their frequency of MDS use in various locations, namely, workplace, school, in public (e.g., shopping, street), in transit (including waiting time), or at home. We explored the frequency of MDS use in different physical locations by hypothesizing that:

H_0^{II}: *There are no significant differences between the four groups on the frequency of MDS use in different locations*

Finally, experience from MDS use that relates to perceived content quality has been postulated as a key determinant of future use (Andreou et al., 2005; Cronin & Taylor, 1992, 1994). We asked respondents to reveal their experience from MDS use through four constructs. Two constructs are related to overall experience; "I am satisfied with my decision to use mobile data services,"

"Mobile data services provide excellent overall service." Another construct is based on earlier documentations of the relation between pleasure or joy from using a technology and its adoption (Davis, Bagozzi, & Warshaw, 1989); "I feel that my experience with using mobile data services has been enjoyable." The fourth construct captures user evaluation of MDS; "Compared to the money, effort and time I have to spend on mobile data services, their overall ability to satisfy my wants and needs is high" based on the empirical documentation that consumer's choice of a service is influenced by a cognitive process comparing the cost and benefits involved (Thaler, 1985). We explored users' experiences by hypothesizing that:

H_0^{III}: *There are no significant differences between the four groups on the experience from MDS use.*

The aforementioned hypotheses are expected to shed light on the observed low usage of MDS in both countries that has not reached the mass market and is not frequent.

ANALYSIS OF RESULTS

Group Profiles

Having delineated the four groups in the two countries, we start by depicting their demographic characteristics (Table 2).

Table 2 indicates that "surfers" are in both countries a considerably larger group than "photographers," which is in line with the purpose of this survey targeting users of MDS. Greek "photographers" are students with relatively low annual income whereas the respective Danish group includes mainly men working in the private sector with high annual income. Greek "surfers"

Table 2. The groups' demographics

		PH_{GR}	**PH_{DK}**	**SU_{GR}**	**SU_{DK}**
Group	size	62	50	185	103
Gender	male	60%	94%	59%	92%
	female	40%	6%	41%	8%
Age	<18	5%	0%	8%	0%
	18-24	32%	10%	22%	9%
	25-34	50%	70%	47%	52%
	35-49	8%	18%	21%	32%
	>50	5%	2%	2%	7%
Education	Primary, secondary, and no tertiary	34%	14%	31%	28%
	Tertiary	32%	66%	27%	46%
	Quaternary	34%	20%	42%	26%
Occupation	Private sector	45%	80%	62%	77%
	Students	50%	6%	28%	13%
	Public-semi public	5%	14%	10%	10%
AnnualHousehold Income	<20,000 €	40%	2%	37%	10%
	20,001 €-40,000 €	26%	6%	37%	6%

are young people with quaternary education working in the private sector that have medium annual income. Danish "surfers" are mainly men with tertiary education working in the private sector that have relatively high annual income. The Danish sample seems to be overrepresented by men. This trend is in line with previous research on this market indicating that "surfers" in Denmark are mainly men (Constantiou Damsgaard & Knutsen, 2005). Overall, the demographics structuring of the four groups is in accordance with previous research findings regarding the role of age and gender in the adoption of MDS (Anckar & D'Incau, 2002; Ling, 2004). In addition, the difference in annual income underlines the differences in the economic structure of the two countries.

Additional questions were also asked to explore the groups' characteristics in terms of MDS usage patterns. We asked "photographers" when they started using MMS. We observed different distributions between countries and groups. Upon performing chi-square tests we found significant differences between "photographers" experience with MMS use. In particular the highest percentage of Greek "photographers" (29%) started using MMS less than three months ago, whereas the majority of the Danish group (42%) started using them more than 24 months ago. We also asked them how much time they spent on MMS use on a monthly basis. The majority of Greeks (62%) and Danish (60%) use MMS for less than 9 minutes monthly. This is not a surprising result since "photographers" are only sending and receiving MMS which is not a time intensive activity.

Turning to "surfers" usage patterns, the Greek group is divided into two subgroups (30% each) with an MDS use period of either less than three or more than 24 months ago, whereas the majority of the Danish group (53%) have used MDS for more than 24 months. In terms of time spent on MDS monthly, the majority of Greeks (52%) and the highest percentage of Danes (27%) state less than 9 minutes. However, 25% of the Danish respondents state 10-29 minutes of monthly MDS use. Thus, in relative terms, Danish "surfers" use MDS more frequently. This result may relate to the fact that Danish respondents have longer experience with MDS.

Mobile Data Services Usage Patterns and Experiences

We then explored the purpose and the context of MDS use by performing a series of ANOVA tests and post hoc comparisons to identify significant contrast values between groups (Table 3).

According to Table 3, "photographers" purpose of use does not significantly differ. It seems that they mainly use MMS for personal reasons. This finding is in line with the main property of MMS which is to improve person-to-person communications (Yrjänäinen & Neuvo, 2002) and enhance the creation and maintenance of personal and group memory (Van House, Davis, Ames, Finn, & Viswanathan, 2005). However, "surfers" significantly differ from "photographers" since they use MDS in a more balanced mix of business and personal purposes. This result may relate to their profile (i.e., people working on the private sector) and the wide variety of MDS available covering both purposes. Thus, they appear to be more mature users who are capable of distinguishing among the various functions of mobile data services and achieve higher added-value from using them.

In terms of frequency and place of use, there are significant differences among all groups. A general observation is that the frequency of use for "photographers" is lower than "surfers." Besides, Danish "surfers" mainly use MDS while in transit or when waiting for transportation means and Greek "surfers" mainly at home. This can be partially explained by the relatively high Internet penetration at home in Denmark and low in Greece. It is empirically documented through longitudinal field studies that MDS are perceived by users as close substitutes of Internet services (Blechar, Constantiou, & Damsgaard,

Table 3. Group purpose and context of MDS use (significant contrast values identified with the Games Howell post-hoc comparison procedure)

	PH_{GR}	PH_{DK}	SU_{GR}	SU_{DK}	Asympt. F	P-value	Significant Contrast Values
Location (1: never, 7: very often)							
Workplace	2.56	2.93	3.64	4.06	8.76	0.00	PH_{GR} - SU_{GR} PH_{GR} - SU_{DK} PH_{DK} -SU_{DK}
School/University	1.52	2.48	2.25	3.00	5.60	0.00	PH_{GR} - SU_{GR} PH_{GR} - SU_{DK}
Public place	2.78	3.08	3.68	3.57	4.43	0.00	PH_{GR} - SU_{GR} PH_{GR} - SU_{DK}
In transit	2.72	3.81	4.00	5.08	21.14	0.00	PH_{GR} - PH_{DK} PH_{GR} - SU_{GR} PH_{GR} - SU_{DK} PH_{DK} - SU_{DK} SU_{GR} - SU_{DK}
Home	3.49	3.21	4.59	3.63	12.33	0.00	PH_{GR} - SU_{GR} PH_{DK} - SU_{GR} SU_{GR} - SU_{DK}
Purpose (1: only business, 7: only personal)							
	5.19	5.00	4.47	4.16	6.36	0.00	PH_{GR} - SU_{GR} PH_{GR} - SU_{DK} PH_{DK} - SU_{DK}

2006a; 2006b). A similar trend is also observed for "photographers."

Moreover, in places like school, university, or work, where Internet access is available, we observe low frequency of use. These observations indicate that the user may compare the two access means and the Internet may be preferred over mobile networks. This choice may relate to the perceived value derived from the reference price and situation of the mobile user (Blechar et al., 2006a). For example, when accessing news the user may find mobile network access more expensive compared to the Internet that is perceived as free of charge at the university or at work. This comparison may induce the user not to use the former means.

Furthermore, we asked respondents to comment on their experience from using MDS.

Observing Table 4 shows that "photographers" experience with MDS does not significantly differ between the two countries and is rather neutral. This observation may be explained by the fact that they only use MMS and, thus, their experience is relatively narrow in scope and not intense. The average experience and satisfaction observed may relate to the nature of MMS, which has a very short duration and simplicity in content provided. In addition, we found average scores for the last construct comparing the user's perceived costs incurred from MMS use to the benefit derived. This result implies that the "photographers" do not perceive MMS as a high value-adding service which may in turn affect the decision to adopt other MDS service as well.

"Surfers" experiences between the two countries significantly differ. It appears that Greeks have more positive experience than Danes. This may relate to content per se that is offered in Greece by a wider variety of independent service providers (i.e., an operator with high market share provides access to i-mode platform that allows mobile users accessing independent content pro-

viders) than in Denmark (i.e., the operators offer their own content services and accessing third party providers in not encouraged in technical and economic terms). Besides, Greeks may value MDS use more since in specific locations with high observed frequency of use such as home there are no other data communications means (e.g., Internet). Moreover, the significantly lower evaluations of Danish surfers may also explain the relatively lower use of MDS that we observe in this sample.

However, both groups indicate neutral experience when asked to compare their satisfaction in terms of wants and needs with money, time, and effort related to MDS. This finding suggests that respondents are not generally satisfied from the use of MDS when comparing the costs incurred and the benefits derived. This result highlights the challenges faced by mobile service providers both in terms of content and variety of services that can cover users' needs and wants as well as improving the technical characteristics, (e.g., interface, network availability) to decrease the cost incurred in terms of money, effort, and time spent.

The low satisfaction may in turn affect their willingness to pay for MDS. We asked respondents to calculate how much of their total monthly expenditure was related to MDS use. In this question there are no significant differences between the two countries' "photographers" since the majority of Greeks (82%) and Danish (66%) stated less than 15 Euros. This in turn relates to the fact that MMS is offered on a relatively low flat price per use in both countries. Turning to "surfers," there are also no significant differences between the two countries on monthly expenditure for MDS, since the majority of Greeks (50%) and Danish (39%) state less than 15 Euros.

In sum, although our research suggests that there are differences in several aspects of MDS use in the two countries under study, it appears that both markets are rather immature. Several steps need to be taken for MDS use to reach a critical mass.

CONCLUSION

This chapter aimed at offering an empirical argumentation regarding the particularities of the MDS adoption trajectory in two European countries, namely Denmark and Greece. To this

Table 4. Group mean values and the results of ANOVA tests

Experience from MDS use (1: Not at all , 7: Quite a lot)	**PH_{GR}**	**PH_{DK}**	**SU_{GR}**	**SU_{DK}**	**Asympt. F**	**P-value**	**Significant Contrast Values**
I am satisfied with my decision to use mobile data services	4.32	3.54	5.15	4.10	18.81	0.00	PH_{GR} - SU_{GR} PH_{DK} - SU_{GR} SU_{GR} - SU_{DK}
I feel that my experience with using mobile data services has been enjoyable	4.45	3.56	4.82	3.88	12.42	0.00	PH_{DK} - SU_{GR} SU_{GR} - SU_{DK}
Mobile data services provide excellent overall service	3.94	3.78	4.43	4.57	4.27	0.01	PH_{DK} - SU_{DK} SU_{GR} - SU_{DK}
Compared to the money, effort, and time I have to spend on mobile data services, their overall ability to satisfy my wants and needs is high	3.56	3.34	4.07	3.65	3.63	0.13	

end, the analysis presented in the previous sections was structured around the application of a segmentation specific marker to distinguish two broad genres of MDS adopters in the two countries under scrutiny, "photographers" and "surfers." Although these groups are meant to reflect advanced MDS users, groups' mean frequencies of specific service use are relatively low indicating that the adoption process is still at a nascent stage in both countries. Furthermore, our research investigated key attributes and parameters in MDS usage patterns through the development and empirical testing of three hypotheses set up to explore intra- and inter-country group differences relating to purpose and location of, as well as experience from using MDS. Table 5 highlights the primary findings from our categorization research.

It appears that while the socioeconomic differences among the two countries can partially explain the emerging usage patterns, there is a common trend between advanced mobile service users. They are not satisfied with service offerings because they cannot generate the expected benefits. This also implies lower willingness to pay for MDS. It seems that in current market settings there are limited opportunities to generate high revenues from MDS as expected. There is no critical mass of MDS users to generate benefits from network effects. These findings set new challenges for mobile service providers. They may have to alter their development and marketing efforts to increase exposure and personalization of service offerings to meet their customers' needs.

This chapter identified country-related differences between advanced mobile service users that mainly affect the context of MDS use. In particular, different infrastructural development indicates that in Greece MDS may substitute Internet services at home, whereas in Denmark MDS are used when there is no Internet access available (e.g., in transportation means). It is important to underline the possible substitution effect between Internet services and MDS that is reflected by the low usage rates at locations where there is Internet access (e.g., work).

Furthermore, the significant differences between the two groups at a country level relate to the fact that "surfers" have taken an extra learning step and are actual MDS users whereas "photographers" are technologically advanced users that have not adopted MDS yet. This may relate to negative experiences. Thus, mobile ser-

Table 5. Group characteristics

	PH_{GR}	PH_{DK}	SU_{GR}	SU_{DK}
Demographic profile	**students low income**	**men working in private sector high income**	**individuals working in private sector medium income**	**men working in private sector high income**
Experience with MDS	**<3 months**	**>24 months**	**> 3 months**	**>24 months**
Expenditure on MDS	**<15 Euros**	**<15 Euros**	**<15 Euros**	**<15 Euros**
Monthly time spent on MDS	**<9 minutes**	**< 9 minutes**	**<9 minutes**	**>9 minutes**
Purpose of MDS use	**personal**	**personal**	**personal and business**	**personal and business**
Place and frequency of MDS use	**Home, sometimes**	**Transit, sometimes**	**Home, often**	**Transit, often**

vice providers should develop different marketing strategies pushing "photographers" to adopt MDS while maintaining and increasing "surfers" current use.

MDS represents a new category of information and communication services and applications that combine well-known and established Internet features (e.g., instant access to information, communication, or entertainment facilities) with the particularities of the mobile device as the access medium. The understanding MDS adoption occupies an important position within the IS research stream and researchers have attempted to explore the challenges associated with MDS consumer acceptance using a wide portfolio of theories and methods. Research findings are then applied to explain why markets are still struggling to increase MDS adoption rates above certain thresholds that will enable the stakeholders (enterprises and consumers) to capitalize on MDS value elements. Results of this chapter contribute to this area by offering some initial insights on why the MDS market suffers from low use rates in European countries.

Furthermore, our work highlights the need for incorporating cross-cultural surveys and comparisons of MDS usage patterns since it confirms early observations regarding the non-universalism of the MDS adoption process. MDS geography is not only characterized by the Europe-Asia dichotomy, since important differences can be traced within countries belonging to the same periphery, such as Denmark and Greece. Thus, future research needs to take into account not only individual traits and behaviors but also the wider socioeconomic or cultural context within which MDS use takes place.

REFERENCES

Aarnio, A., Enkenberg, A., Heikkilä, J., & Hirvola, S. (2002). *Adoption and use of mobile services. Empirical evidence from a Finnish survey.* Paper presented at the 35th Annual Hawaii International Conference on System Sciences (HICSS-35'02), Big Island, HI.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. *Journal of Information Technology Theory and Application, 4*(1), 43-65.

Andreou, A. S., Leonidou, C., Pitisillides, A., Samaras, G., Schizas, C. N., & Mavromoustakos, S. M. (2005). Key issues for the design and development of mobile commerce and applications. *International Journal of Mobile Communications, 3*(3), 303-323.

Blechar, J., Constantiou, I., & Damsgaard, J. (2006a). Understanding behavioural patterns of advanced mobile service users. *Electronic Government: An International Journal, 3*(1), 93-104.

Blechar, J., Constantiou, I. D., & Damsgaard, J. (2006b). Exploring the influence of reference situations and reference pricing on mobile service user behaviour. *European Journal of Information Systems, 15*(3), 285-291.

Carlsson, C., Carlsson, J., Hyvønen, K., Puhakainen, J., & Walden, P. (2006). *Adoption of mobile devices/services—Searching for answers with the UTAUT.* Paper presented at the 39th Annual Hawaii International Conference on System Sciences (HICSS-39'06), Big Island, HI.

Carter, J. (1998). Why settle for early adopters? *Admap, 33*(3), 41-44.

Constantiou, I. D., Damsgaard, J., & Knutsen, L. (2004). *Strategic planning for mobile services adoption and diffusion: Empirical evidence from the Danish market.* Paper presented at the Mobile Information Systems (MOBIS), Oslo, Norway.

Constantiou, I. D., Damsgaard, J., & Knutsen, L. (2005). *Beware of Dane-geld: Even if paid, m-*

service adoption can be slow. Paper presented at the European Conference on Information Systems (ECIS), Regensburg, Germany.

Constantiou, I., Damsgaard, J., & Knutsen, L. (2006). Exploring perceptions and use of mobile services: User differences in an advancing market. *International Journal of Mobile Communications, 4*(3), 231-247.

Constantiou, I. D., Damsgaard, J., & Knutsen, L. (2007). The four evolution steps to advanced mobile services' adoption. *Communications of the ACM, 50*(6), 51-55.

Cronin, J., & Taylor, S. (1992). Measuring service quality: A reexamination and extension. *Journal of Marketing, 56*(3), 55-68.

Cronin, J., & Taylor, S. (1994). SERVREF vs. SERVQUAL: Reconciling performance-based and perceptions-minu-expectations measurement of service quality. *Journal of Marketing, 58*(1), 125-131.

Davis, F. (1986). *Technology acceptance model for empirically testing new end-user information systems: Theory and results.* Boston.

Davis, F. D., Bagozzi, R., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Eurobarometer. (2006). *Safer Internet.*

Gilbert, A. L., & Han, H. (2005). Understanding mobile data services adoption: Demography, attitudes or needs? *Technological Forecasting & Social Change, 72,* 327-337.

Gilbert, A. L., & Kendall, J. D. (2003). *A marketing model for mobile wireless services.* Paper presented at the 36th Hawaii International Conference on System Sciences (HICSS'03), Big Island, HI.

Hair, J. F. J., Bush, R. P., & Ortinau, D. J. (2000). *Marketing research: A practical approach for the new millennium*: McGraw-Hill International Editions.

Hjorth, L. (2005). Postal presence: A case study of mobile customisation and gender in Melbourne. In P. B. Glotz, S. & Locke, C (Eds.), *Thumb culture: The meaning of mobile phones for society.* Bielefeld, Germany: Transcript-Verlag.

Katz, M. L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *American Economic Review, 75*(3), 424-440.

Katz, M. L., & Shapiro, C. (1992). Product introduction with network externalities. *Journal of Industrial Economics, 40*(1), 55-83.

Kim, H.-W., Chuan Chan, H., & Gupta, S. (2005). Value-based adoption of mobile Internet: An empirical investigation. *Decision Support Systems, In Press.*

Kinnear, T. C., & Taylor, J. R. (1996). *Marketing research: An applied approach* (5th ed.). McGraw-Hill.

Kopf, W. (2005). *The European mobile industry—A case for consolidation?* T-Mobile International AG & Co. KG.

Lee, Y., Kim, J., Lee, I., & Kim, H. (2002). A cross-cultural study on the value structure of mobile Internet usage: Comparison between Korea and Japan. *Journal of Electronic Commerce Research, 3*(4), 227-239.

Ling, R. (2004). *The mobile connection. The cell phone's impact on society* (3rd ed.). Morgan Kaufmann.

Lu, J., Yao, J.-E., & Yu, C.-S. (2005). Personal innovativeness, social influences and adoption of wireless Internet services via mobile technology. *Journal of Strategic Information Systems, 14*(3), 245-268.

Massey, A. P., Khatri, V., & Ramesh, V. (2005). *From the Web to the wireless Web: Technology readiness and usability.* Paper presented at the

38th Annual Hawaii International Conference on System Sciences (HICSS-38'05), Big Island, HI.

Moore, G. A. (1999). *Crossing the chasm* (2nd ed.). Oxford: Capstone.

Mort, G. S., & Drennan, J. (2005). Marketing m-services: Establishing a usage benefit typology related to mobile user characteristics. *Database Marketing & Customer Strategy Management, 12*(4), 327-341.

OECD. (2005, December). *Telecommunications and Internet policy: OECD broadband statistics.*

Okazaki, S. (2006). What do we know about mobile Internet adopters? A cluster analysis. *Information & Management, 43*(2), 127-141.

Pedersen, P. E., & Ling, R. (2003). *Modifying adoption research for mobile Internet service adoption: Cross-disciplinary interactions.* Paper presented at the 36th Hawaii International Conference on System Sciences (HICSS'03), Big Island, HI.

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York: Free Press.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.

Shapiro, C., & Varian, H. (1999). *Information rules: A strategic guide to the network economy.*

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science, 4,* 199-214.

Van House, N., Davis, M., Ames, M., Finn, M., & Viswanathan, V. (2005, April 2-7). *The uses of personal networked digital imaging: An empirical study of cameraphone photos and sharing.* Paper presented at the CHI 2005, Portland, OR.

Venkatesh, V., Morris, M., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Towards a unified view. *MIS Quarterly, 27*(3), 425-478.

Vrechopoulos, A. P., Constantiou, I. D., Mylonopoulos, N., & Sideris, I. (2002). *Critical success factors for accelerating mobile commerce diffusion in Europe.* Paper presented at the 15th Bled E-commerce Conference, e-Reality: Constructing the e-Economy, Bled, Slovenia.

Vrechopoulos, A. P., Constantiou, I. D., Sideris, I., Doukidis, G. I., & Mylonopoulos, N. (2003). The critical role of consumer behavior research in mobile commerce. *International Journal of Mobile Communications, 1*(3), 329-340.

Wolfe, R. A. (1994). Organizational innovation: Review, critique and suggested research directions. *Journal of Management Studies, 31*(3), 405-432.

Wu, J.-H., & Wang, S.-C. (2005). What drives mobile commerce? An empirical evaluation of the revised technology acceptance model. *Information & Management, 42,* 719-729.

Yang, K. C. C. (2005). Exploring factors affecting the adoption of mobile commerce in Singapore. *Telematics and Informatics, 22,* 257-277.

Yrjänäinen, J., & Neuvo, Y. (2002). Wireless meets multimedia. *Wireless Communications and Mobile Computing, 2*(6), 553-562.

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, and J. Day, pp. 134-149, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.3
The Mobile Phone Telecommunications Service Sector in China

Michelle W. L. Fong
Victoria University, Australia

EXECUTIVE SUMMARY

Technology leapfrogging by a late adopter of technologies means skipping intermediate technologies and adopting the latest technologies. In this way, this late adopter would be exposed to unprecedented opportunities offered by the new technologies. This case study focuses on China's attempt at leapfrogging to mobile phone telecommunications technology. It provides a description of the underlying forces involved in shaping and influencing this leapfrogging attempt. Students or readers are encouraged to analyse this case from their contextual perspective—may it be from the standpoint of a competing country, foreign investor, competing marketing corporation, policy maker, or consumer. Instructors of teaching cases may perhaps consider assigning different roles to students in discussing this case within a group.

ORGANIZATION BACKGROUND

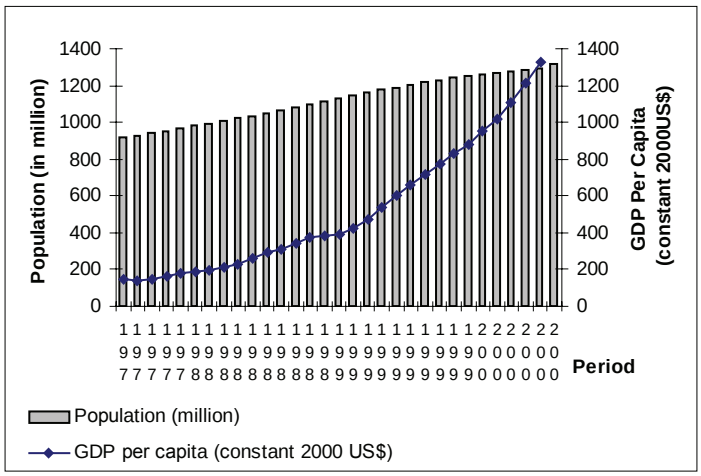
The Chinese Economy

China is a rising economic power in the world. With its massive geographical landscape, it is the fourth largest country in the world and the world's most populated country with its 1.3 billion citizens. Figure 1 shows the growing population and increasing GDP (gross domestic product) per capita (in constant U.S. dollars using 2000 as the base year).

Information and Communication Technology Spending and Telecommunications

The Chinese government recognizes that the adoption of information technology for an interconnected economy will sustain and add impetus to its development (Ministry of Information

Figure 1. China: population, in million, and GDP per capita, in US$ (Source: The World Bank Group, 2006)



Industry, 2005b; "Striving for a Nation Stronger in Information Industry," 2006). As shown in Figure 2, spending on ICT within the Chinese economy between 2000 and 2004 was on average about 4% of the GDP.

Although Figure 2 shows that China experienced a slight decline in telecommunication revenue after 2002, this decline is attributed to the fall in telecommunication prices rather than the fall in demand for telecommunication services. Average telecommunication revenue between 2000 and 2004 was about 3% of the GDP (*The World Bank Group,* 2006). China's telecommunication revenue (3.2% of GDP) in 2004 was higher than that in the East Asia and Pacific region (2.6% of GDP) and the world (3.0% of GDP).

ICT Adoption

Prior to the emergence of fibre optic cable for fixed-line communications, many places in China were not connected via copper cable. Even today when this technology is commercially available, China has not been able to establish an interconnected economy through fibre optic cable due to its massive geographical landscape and resource constraint. Fixed-line telephones and faxes are comparatively widely available in major cities and provinces, but not in the rural inland areas where there has been the additional problem of underdeveloped supporting infrastructure such as electricity supply (Peng, 2003). Mobile technology represents an infrastructure alternative to fixed-line communications for this country. Its embarkation onto this communication platform constitutes an act of technology leapfrogging as the Chinese essentially skipped over wire-based communications technology to a wireless network.

Figure 3 shows the rate of adoption of mobile phones, telephone mainlines, and computer Internet in China. Each of these ICTs has varying capabilities or potential in enabling e-commerce. The trend lines in Figure 3 show that mobile phones have been experiencing a rapid rate of adoption as compared to telephone mainlines and computer Internet. Mobile phone technology may be China's technology springboard for e-commerce because it is capable of providing a quicker and less costly solution for overcoming the slow development or inadequacy of the current fixed-line infrastructure. Instead of spending

Figure 2. Telecommunications Revenue and Expenditure on Information and Communication Technology (% GDP) in China (Source: The World Bank Group, 2006)

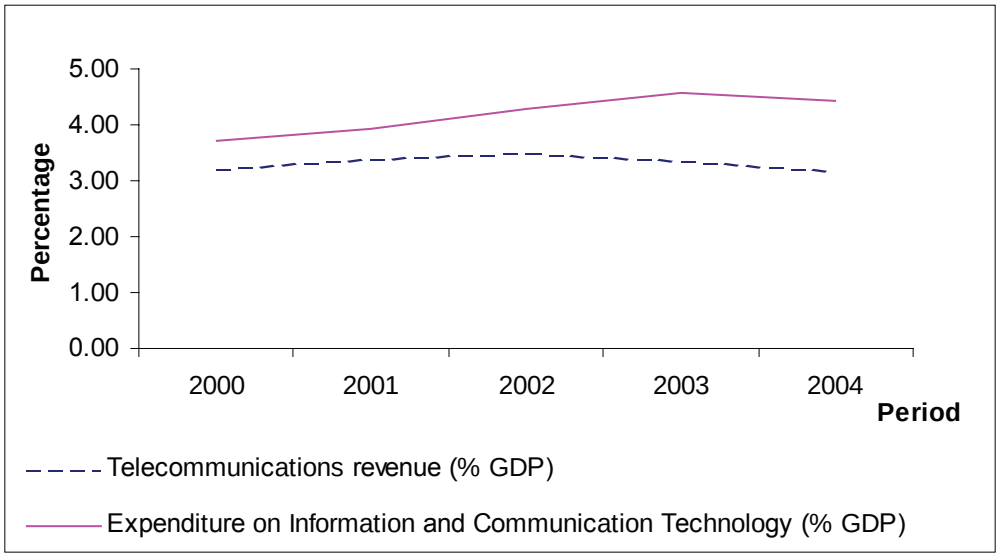
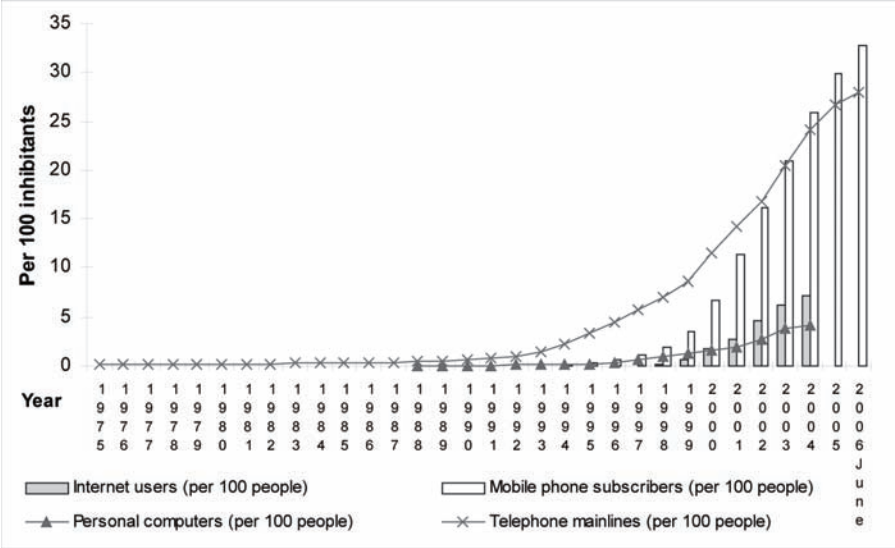


Figure 3. Number of Internet users, mobile phone subscribers, personal computers and telephone mainlines (per 100 inhibitants) (Source: The World Bank Group, 2006)



vast amounts of resources and time to establish fixed-line networks to facilitate telecommunications, China can substitute this infrastructure with easily deployable cellular towers, which are relatively cheap and easy to develop.

China now has the world's largest mobile phone user population, many of whom do not have a fixed-line telephone. Although the mobile phone was only introduced into this country in 1987, it has been experiencing a relatively rapid rate of adoption. By October 2003, the number of mobile phone subscribers exceeded the number of

fixed-line telephone subscribers. By June 2006, there were 33 mobile phones per 100 inhabitants as compared to 28 telephone mainlines per 100 inhabitants ("Nationwide First," 2006). If the trend in mobile phone adoption continues to increase into the future and exceed the adoption rates of all other ICTs, this technology may become the common base for e-commerce. In fact, the Chinese government and telecom market players are pouring resources and support into 3G (third generation) technologies for a robust e-commerce infrastructure.

SETTING THE STAGE

Market Players

Although general economic reforms commenced in 1978 within the Chinese economy, the telecommunications sector remained heavily protected and monopolized by the then regulator the Ministry of Posts and Telecommunications (MPT) of China until the mid-1990s (Hao, 2005). Its commercial arm, China Telecom, operated as a monopoly in four distinctive market segments: fixed-line telecommunications, mobile telecommunications, paging services, and satellite telecommunications. In 1993, MPT allowed non-MPT state-owned enterprises to compete against China Telecom in this lucrative sector. However, the initial services offered by these non-MPT state-owned enterprises were restricted in scope. China United Telecommunications Corporation (China Unicom) was established in 1994 with approval from the State Council to provide mobile telecommunications services. A series of restructures and reforms, most of which were state mandated, began to take place in the telecommunications sector from 1998. The first significant reform was the establishment of the Ministry of Information Industry (MII) through the amalgamation of the MPT and the Ministry of Electronic Industry (MEI), which became the principal regulator in this sector. This reform also resulted in the transferring of the different business segments within China Telecom to other non-MPT state-owned enterprises in the telecommunications sector. The mobile business segment was transferred to China Mobile, the paging business segment to China Unicom, and satellite business segment to China Satellite. China Telecom was downsized to focus on the fixed-line telecommunications business segment. The market reform has also resulted in the emergence of China Netcom Communications Group and the entry of China Tietong, a small player, into this telecommunications sector. Table 1 shows the market shares of these main players.

Despite this breaking down of the monopoly structure in the Chinese telecommunications service market, competitors were either associated with or created from the legacy of the former Chinese monopolist. Each of these competitors has the government as its major shareholder despite being public listed. The Chinese telecommunications service sector has remained highly restrictive to foreign private investment.

China Mobile and China Unicom have dominated China's mobile telecommunications service market, while China Telecom and China Netcom are the dominant players in fixed-line telecommunications services. In terms of revenue contribution, the mobile telecommunications segment has been a significant source of revenue. Figure 4 shows the composition of telecommunications revenues in the first quarter of 2006, and specifically that mobile telecommunications is the highest revenue earner in the sector.

Consumer Usage

Chinese consumers' initial experience with a mobile telecommunications device was with the pager in 1984, and the mobile phone was introduced in 1987. Pager subscription experienced a rapid growth during the 1990s but began to decline in 2000 (as shown in Figure 5) when SMS

Table 1. Share of China's telecommunications service market among major players in 2005 (Source: Lui, 2006)

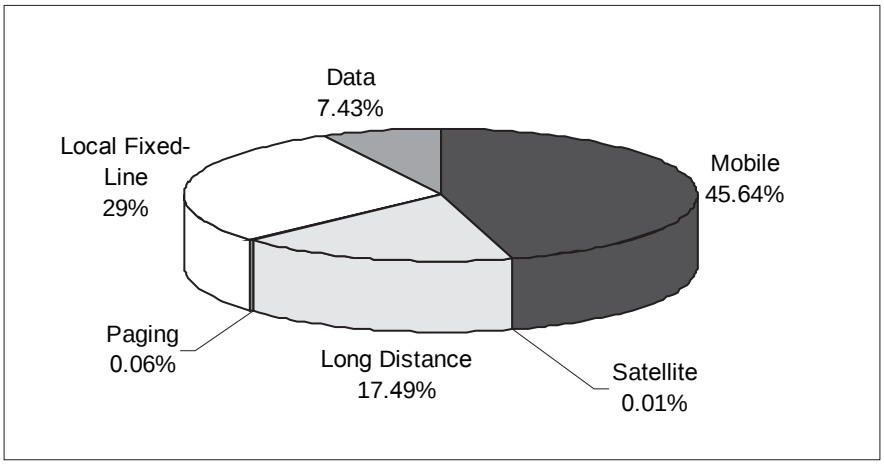
Company	Revenue (in billion U.S. $)	Market share (%)	Main services provided by each company
China Mobile	29.70	40.2%	GSM (Global System Mobile)
China Telecom	20.70	28.0%	Fixed lines, PHS, Internet
China Unicom	10.90	14.8%	GSM, CDMA (synchronous code division multiple access), and others
China Netcom	10.90	14.8%	Fixed lines, PHS, Internet
China Tietong	1.73	2.0%	Fixed lines and others

(short messaging service) in mobile networks and cheaper mobile communications alternatives (for example, a Personal Handyphone System [PHS] such as the Xiaolingtong[1] communications platform) propelled the rapid adoption of mobile phones in China.

SMS has been a major source of growth in the mobile phone telecommunications segment. The number of SMS messages sent in China in 2003 was 137 billion, and this number grew to 218 billion in 2004. The volume of SMS messages grew further to 304.7 billion in 2005. In the first 6 months of 2006, the number of SMS messages sent in this country had already reached 202.9 billion and registered an increase of 45.8% over the same period of the previous year (Nationwide First, 2006). At 0.10 yuan per SMS message, this translates into revenues of billions of yuan for mobile service operators. Based on a survey conducted in April 2004 of 1,063 residents between 18 and 60 years old in six cities (Beijing, Shanghai, Guangzhou, Taiyuan, Chengdu, and Changsha), it was found that 68.1% of mobile phone subscribers have used value-added data services. Of this group, 95.2% of them actually used them in the form of SMS messaging (Fan & Wang, 2005). The SMS feature was utilized for chatting and playing games (57.1% of respondents), circulating entertainment information such as jokes and humour (44.6%), downloading information such as news, including financial reports (25.3%), downloading ringtones (19.8%), and posting quizzes and riddles to each other (15.5%). Chinese seldom use their mobile phones to make voice calls because such calls are relatively expensive. It was found that less than 30% of mobile phones were actually used to make calls (Li, 2003). These subscribers prefer to communicate via SMS messages. Voice calls are only made when the communication is urgent, long, and complex.

Mobile phones were considered a luxury item when they first emerged in the Chinese economy. The fall in prices of phones with basic functions in recent times and the introduction of a cheaper communications system but with restricted mobility, the PHS (such as Xiaolingtong), have enabled the less wealthy to afford and experience this mode of telecommunication. However, this technology is expected to phase out in 5 years time (*MII Zhang Xing Sheng*, 2006). Mobile phones with sophisticated functions and global roaming ability are still not within the reach of the less wealthy, especially inhabitants in the rural areas. Such phones are seen as both status symbols and fashion statements in China (Castells, Ardevol, Qiu, & Sey, 2004; Katz & Sugiyama, 2005). Chinese mobile phone purchasers are influenced by the latest and flashiest models in their purchase decision-making process in contrast to their Western counterparts who tend to base their purchase decisions on the principle of value for money (Li, 2003). Mobile phone replacement cycles in China were claimed

Figure 4. Composition of telecommunications revenues in first quarter of 2006 (Source: MII: Research,



to be 6 to 12 months faster than in Europe and North America (Salkever, 2004).

Cost and Price Issues

Although prices of mobile phones have fallen over the years, owning them is still regarded as costly to a majority of the Chinese, whose average annual disposable income in 2004 in the urban areas was 9,421.6 yuan ($1,177.7) and in the rural areas 2,936.4 yuan ($367.0). In 2003, the cost of using mobile phone telecommunications was 6.2 times that of using fixed-line telecommunications (Xu & Tao, 2003). In a recent survey, it was found that 71% of users spent between 5 to 10 yuan per month on mobile phone communication. Another survey found that close to one quarter of potential new mobile phone subscribers prefer to spend less than 1,000 yuan on a new mobile phone. In September 2005, 35% of mobile phones on the Chinese market were selling for prices between 1,000 ($125) and 1,500 yuan ($188), while 36% were selling below 1,000 yuan ("Overview of China's Mobile Phone Market," 2005). PHS phones were selling at an average price of 750 yuan.

Prior to October 2005, the Chinese government set tariffs for telecommunications services. In that pricing regime, communication charges were imposed not only for making calls but also for receiving them. Table 2 shows the tariffs for PHS telecommunications, local fixed-line telecommunications, and regular mobile phone telecommunications in China before 2005.

In order to enable China Unicom to compete effectively with China Mobile, the dominant player in the market, the government allowed China Unicom to set tariffs at 10% below China Mobile. However, there have been cases where operators ignored the government's pricing rules and competed on prices lower than the stipulated prices in the local markets. In other instances, fees for incoming calls were waived and purchases of handsets were subsidized to attract new subscribers (*Mobile Communications*, 2004).

As for Xiaolingtong, its communications fees were 45% to 50% lower than regular mobile phone services with global roaming ability. Its monthly connection charges were 25 yuan ($3.13) and communication charges were 0.20 yuan ($0.025) per minute. In addition, subscribers to this service were able to use discounted IP (Internet protocol) service for domestic long-distance calls, which is cheaper than prepaid IP phone cards for fixed-line communications. Subscribers to this service were

Figure 5. Number of pager and mobile phone subscribers (Source: China Statistical Yearbook, 2005)

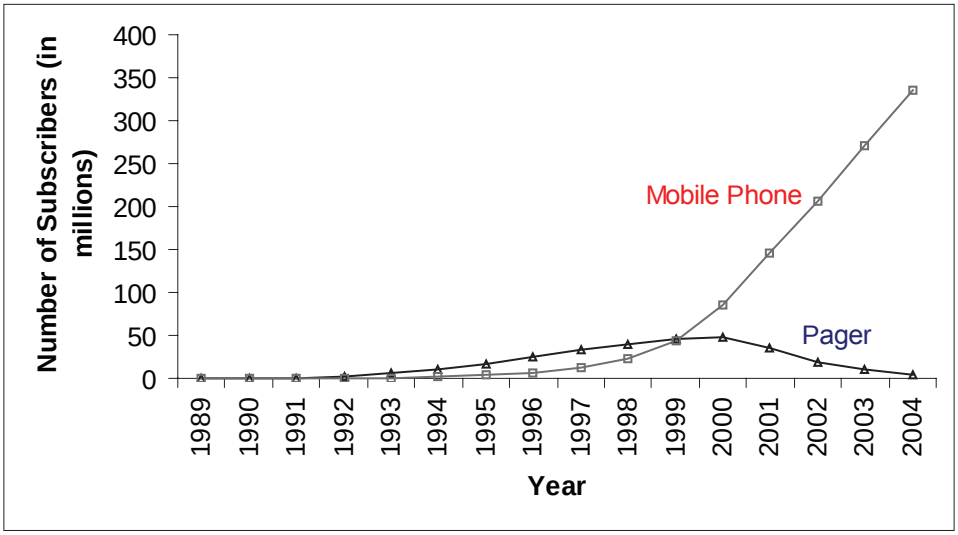


Table 2. Telecommunications tariffs before 2005 (Source: K. Chen, 2004; Wang, 2002)

	Monthly rental (yuan per month)	Communication charges (yuan per minute)
PHS: Xiaolingtong	25 yuan ($3.13)	0.20 yuan ($0.025)
Fixed-line local call: China Telecom China Netcom	 20.25 yuan ($2.53) 21.60 yuan ($2.70)	0.20 yuan ($0.025) per first 3 minutes and 0.10 yuan for each additional minute
Mobile phone: China Mobile China Unicom	 50 yuan ($6.25) 45 yuan ($5.63)	 0.36 yuan ($0.045) 0.40 yuan ($0.050)

largely medium- and low-income consumers, who make up half of China's population.

Price Competition

In October 2005, the Chinese government announced that it would set ceiling prices rather than fixed prices for the mobile phone telecommunications sector (Ministry of Information Industry, 2005a). Local mobile phone telecommunication service providers were allowed to set prices, at or under the ceiling prices, based on market forces prevailing in their local markets. The Chinese government's decision to allow operators to set their own prices has brought about a new dimension in competition in the local market. Mobile phone telecommunications operators were able to offer varying types of service packages tailored to different consumers' needs. In addition, pricing competition has resulted in lower mobile phone communication fees for subscribers. For example, mobile phone subscribers in Guangdong were offered 0.20 yuan (previously 0.36 yuan) per minute for making a call. Their counterparts in Shanghai, on the other hand, were offered 0.05 yuan (previously 0.10 yuan) per SMS message

and 0.10 yuan per minute for making a call. This offer of 0.10 yuan per minute for making a call in Shanghai could even be comparable to the cost of making a fixed-line call. It was reported that mobile phone telecommunications fees between January and February 2006 decreased by 5.89% on average compared to 2005 (*Looking at the Pattern of Fees*, 2006). This development has created a significant impact on the revenue earned by mobile phone telecommunication operators, who were already experiencing declining average revenue per user (ARPU) over the years. For instance, China Mobile's ARPU decreased from 0.603 ($0.075) per minute in 2001 to 0.41 yuan ($0.051) per minute in 2003, and continued to decline to 0.27 yuan ($0.034) per minute under the new pricing regime in 2005. This is despite the continuing increases in the volume of mobile phone telecommunications business over these years (L. M. Chen, 2006; Zhang, 2006). China Unicom experienced a similar situation with its ARPU, which declined gradually from 0.535 yuan ($0.067) in 2001 to 0.255 yuan ($0.032) per minute in 2005 (Zhang). Long-distance calls made through mobile phone networks at a fee of 1.30 yuan ($0.163) per minute have fallen in some areas under the new pricing regime. In some areas, fees for making long-distance calls were as low as 0.10 yuan ($0.013) per minute. This reduction in long-distance call charges has brought about an increase in the volume of long-distance calls made through mobile phone telecommunications networks. The ratio of the volume of long-distance calls made through fixed-line and mobile phone networks was 1.6: 1 in 2001, and is about 1:1 in 2006 (Zhang).

Despite the lowering of these fees and charges, a 2006 survey in China revealed that 66% of respondents found the mobile phone telecommunications fees structure unreasonable and inconsistent. In-depth interviews revealed that these respondents tend to be confused by the various packages being offered by the operators in the market, as well as the varying fee structures in different localities (*Discussion of China's Pressing Mobile Fees Problem*, 2006). In the same survey, 64% of the respondents felt that there is room for a further decrease in mobile phone telecommunications fees. In addition, 75% of the respondents expected operators to abolish fees charged for receiving calls on their mobile phone in the future, because this practice has already been implemented in some regions.

Complaints and Government Intervention

The level of complaints from mobile phone users regarding unreasonable and inconsistent fees and charges was particularly high in Beijing after the implementation of the new pricing regime. One survey revealed that mobile phone fees in Beijing were 7 times higher than in Tianjin and Chongqing, and that this city had the highest mobile phone fees (*MII Confirmed China Mobile to Lower Phone Call Charges in May*, 2006). This considerable difference has also resulted in Beijing's mobile phone subscribers preferring SMS communication to voice communication (*Statistics Revealed Substantial Difference in Handphones' Penetration Rates*, 2006). In May 2006, the MII intervened and held talks with the offices of China Mobile and China Unicom in Beijing to consider lowering fees and charges for their services (*Analysis of China's Current Mobile Fees Situation*, 2006). The talks resulted in a positive outcome for Beijing's mobile phone subscribers. The two mobile telecommunications giants began to lower fees and charges for their services. Subscribers in Beijing were then able to enjoy mobile phone communication rates as low as 0.02 yuan ($0.0025) per minute for just receiving a call and 0.24 yuan ($0.03) per minute for making or receiving a call (*Beijing Mobile Phone Operators to Slash Charges*, 2006).

On the whole, the change in pricing regulation has triggered more intensified pricing competition within the mobile phone telecommunica-

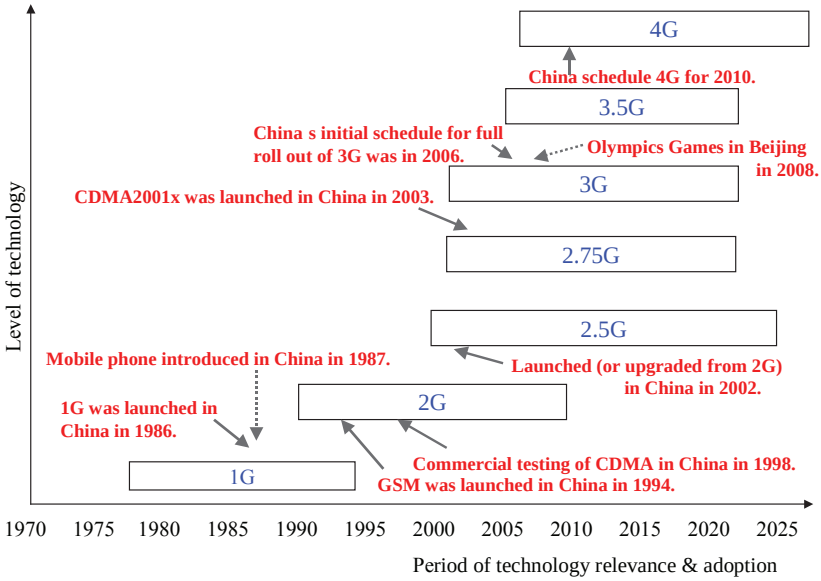
tions market, to the advantage of consumers. Competitors' responses to rivals' new marketing strategies, in defense of their market shares, are being undertaken at a quicker rate than before. The positive outcome from competitive pricing in this mobile phone telecommunications segment had prompted the MII to relax pricing regulation in the fixed-line telecommunications segment in the desire to replicate this outcome in the latter segment.

Technology

Figure 6 shows the adoption of different levels of mobile phone technologies in China. The length of the bar for each technology indicates the approximate period of relevance of the technology. For example, the bar for 1G (first generation) technology, which is analogue technology (such as the Total Access Communications System), was technologically relevant from the late 1970s and throughout the 1980s. During this period, countries around the world adopted 1G technology. It was in use in a limited number of cities in China between 1987 and 1995, before the country launched digital GSM in 1994.

China Mobile operates a GSM network and China Unicom operates both GSM and CDMA networks. GSM was introduced in 1994 and now has 80% of the market coverage. Because of the variety of mobile wireless standards, only a small proportion of mobile phone subscribers used unified wireless network services. Currently, a number of countries are transiting or preparing to transit from 2G (second generation, such as CDMA, TDMA, or GSM) to 3G technologies (such as CDMA2000, UMTS, or TDSCMA). China is one of these countries, and it has a 3G implementation deadline that must be met before the Beijing Olympic Games in 2008. 3G technologies enable the simultaneous transfer of data, sound, text, pictures, audio, and video, and support high-end value-added services at high speed, such as high-speed Internet access, entertainment, videoconferencing, mobile shopping, and information updates. It has faster speed and greater capacity than 2G technologies, and has the potential to support a wireless electronic platform

Figure 6. Adoption of mobile phone technologies in China



for realizing e-commerce. However, evolving 3G networks require more expensive network equipment investments than previous generations as they are far more complex and sensitive to poor configuration (Forge, 2004). 3G technologies are not capable of efficiently integrating preceding intermediate technologies such as 2G and 2.5G. Some experts have suggested that China should abandon its attempt at 3G infrastructure and proceed with 4G (fourth generation) technology, a network unifier capable of integrating earlier technology such as 2G and 2.5G. On the other hand, 4G presents technical challenges that are even more daunting than 3G.

In China, 3G services were expected to roll out to all mobile phone users in the fourth quarter of 2006. It is also envisaged that Xiaolingtong will be phased out in 5 years time after the implementation of 3G technologies (MII Zhang Xing Sheng, 2006). It was claimed that the Chinese government's preference for homegrown TD-SCDMA (time division synchronous code division multiple access) to be the national 3G technology standard for the mobile phone telecommunication industry has resulted in interference with the launch of foreign-developed standards, such as WCDMA and CDMA2000 (Burns, 2006). However, technology obstacles have hampered the deployment of TD-SCDMA. Test results carried out by the Ministry of Information Industry in 2004 revealed that networks using this homegrown TD-SCDMA standard were unstable and unreliable (*Mobile Communications*, 2004). It was found that far too few mobile phone handsets were compatible with this technology, and that they were not as good as handsets produced for the other two international standards (WCDMA and CDMA2000). For example, it was claimed that the chips of TD-SCDMA mobile phones have problems with supporting 3G value-added applications (*Domestic Problems Grow for China's 3G*, 2005). In addition, there were interoperability problems between terminals produced by a number of manufacturers (*The Sprint in the Trial of TD Network*, 2006). The shortage of talent caused by brain drain and labour turnover has aggravated the delay in the establishment of a nationwide 3G network (*Zhongxing TD's R&D Frustration*, 2006). The TD-SCDMA standard was supposed to be ready by mid-2005 and fully rolled out by 2006, but this target is no longer expected to be achievable. On January 20, 2006, the MII announced TD-SCDMA as the chosen national technology standard for telecommunication, but was silent on the date for granting 3G's mobile phone licenses (*TD-SCDMA, 3G in China*, 2006).

The government is keen to roll out 3G technologies for applications by 2008 for the Beijing Olympic Games. However, the trials of TD-SCDMA in Qingdao, Baoding, and Xiamen in 2006 did not produce satisfactory results. The unsatisfactory performance of this 3G infrastructure may continue to delay the award of mobile phone licenses by the Chinese government. In a closed meeting on August 8, 2006, the government indicated that it is no longer prepared to adjust its testing and implementation schedule for mobile phone manufacturers, who need more time to get their products up to application standards, unless the setback produces very serious consequences (*MII Zhang Xing Sheng*, 2006).

Censorship

The Chinese government has viewed the wireless SMS mobile platform as a two-edged sword. It can potentially work to the advantage of the country, but can also work against the government. The SMS platform has helped the Chinese government to disseminate announcements of impending natural calamity and epidemic, as well as propaganda. For example, authorities in the Fujian province sent 18 million SMS messages about weather information during five typhoons in 2006 (*Warning, Storm Ahead…TNX*, 2006). In addition, the government has used text messages

to reassure the public about bird flu outbreaks and to discredit the banned Falun Gong movement.

China exerts strict censorship on politically sensitive or inappropriate content transmitted over mobile phone networks. The ruling party has viewed such material as capable of undermining its political power, as well as creating social discourse. SMS has been used to bypass government censorship and to access specific information that is normally not available in the public media. Like Internet spamming, SMS messages transmitted from one mobile phone to another may lead to mass hysteria and urban legends[2] whereby well-intentioned users unknowingly pass untrue or damaging pieces of information onto colleagues and friends. In addition, the Chinese government is concerned that the SMS platform might be used for illegal activities. Over the years, the increasing popularity of SMS for communications has been accompanied by an increase in the number of illegal activities facilitated by it. There have been cases where unscrupulous third-party providers transmit deceptive SMS messages deliberately designed to entice responses from unsuspecting mobile phone users, and to commit these users to paying for goods or services that they do not need or would not have ordered. These third-party providers are not necessarily the scammers themselves but may be contracted for their services in sending deceptive junk SMS. One of the common cases of SMS deception is where a fraudster sends short SMS greetings, seemingly from a friend or acquaintance, to mobile phone users. If an unsuspecting mobile phone user responds to the message under the mistaken belief that he or she is replying to a known person, it may be found out later that the fraudster has committed the user into paying for goods or services that he or she was not aware of and had no intention of buying. SMS has also been used in the prostitution business, which is illegal in China.

CASE DESCRIPTION

The declining cost to performance and increasing user friendliness of technologies are providing opportunities for China, a late adopter of technologies, to leapfrog technology generations and arrive at state-of-the-art networks. Adopting state-of-the-art technology would avail this once technologically backward country of the unprecedented opportunities offered by the new technology. Developed nations, on the other hand, have found it difficult to exploit leapfrogging opportunities or adjust to the leapfrogging process because they are entangled in old systems. The unprecedented technological opportunities offered by the new technology are not easily or readily accessible to such countries without themselves incurring costs associated with displacing the last generation's technology infrastructure. For example, the United States is so saddled with extensive investments sunk into fibre optic cable that investment in mobile Internet infrastructure (particularly in 3G) is lagging (Dholakia, Dholakia, Lehrer, & Kshetri, 2002).

The evolutionary process in traditional technological advancement is a time-consuming process, whereas technology leapfrogging allows latecomers to skip intermediate technologies and bypass undesirable constraints connected with these technologies. Technological leapfrogging in the telecommunications and computing infrastructure is technically feasible in terms of physical facilities in the developing countries. However, leapfrogging to the latest technologies would be a remote possibility if a country does not have the technological and institutional capabilities to operate or harness these technologies. There are predictable barriers when introducing an advanced technology into a developing country. On the technical front, it is important not only to apply the advanced technology efficiently, but also to maintain and update it. All these require talents and skills, which could be a pertinent

problem in a developing country. The new technology may also require radical adjustments in the lifestyle, behaviour, and mindset of people in the developing country, and it is important that a careful planning approach be adopted to roll out such technology.

CURRENT CHALLENGES AND PROBLEMS FACING THE COUNTRY

Achieving Critical Mass

As competition becomes more and more intense, the Chinese mobile phone telecommunications operators are increasingly being challenged in fulfilling customer expectations, retaining existing customers, and recruiting new ones. These operators are keen to achieve critical subscriber mass for sustaining and growing their existing and new business units, as well as creating competitive barriers to entry, before the local market is fully opened to foreign competition. Reaching critical mass depends heavily on customer adoption and retention of their products. Without a reasonably stable customer base, an operator will find its new technology being underutilized and resources being constantly expended to retain existing customers and identify new customers.

The Chinese mobile phone telecommunications operators are likely to struggle to maintain critical mass in a highly competitive market because Chinese consumers display little brand loyalty. A recent survey (refer to Table 3) conducted on consumers' attitude toward China Mobile's services (Quanqiutong, Benditong, and Shenzhouxing) revealed that although respondents gave high ratings of satisfaction for the product/brand of this dominant market player, their ratings on product/brand loyalty were lower than the former (L. M. Chen, 2006). The ratings on product/brand loyalty do not suggest strong brand loyalty among Chinese consumers, and this was substantiated by findings in the same survey that approximately one third of the users had previously used another mobile phone network operator. This signals to incumbent operators that customers are likely to switch to a service operator who can provide a more attractive offer. This suggests that mobile phone telecommunications operators are likely to encounter greater difficulty in achieving critical mass if they have not established a strong foothold in the local market by the time it is fully opened to foreign competition.

Local businesses bemoan the increasing level of difficulty in competing within the changing landscape of the telecommunications market. The local operators are making an effort to shift from a production-oriented mentality (that was commonly associated with the past centrally planned regime) to a market-oriented mentality as competition intensifies (Hao, 2005).

In the telecommunications market, Chinese consumer behaviour and expectations have evolved. Subscribers are gradually becoming confident and aware of their rights, and they express their dissatisfaction by switching to another operator and/or lodging complaints with the authorities. For example, complaints against China Mobile in the first quarter of 2006 grew by 189% and in the second quarter of the

Table 3. Survey on users of mobile China's services (Source: L. M. Chen, 2006)

	Quanqiutong (Global roaming)	Beditong (Local roaming)	Shenzhouxing (Nationwide roaming)
Product/brand satisfaction	82.49%	81.13%	77.07%
Product/brand loyalty	66.64%	67.26%	54.07%

same year by 220%. Complaints against China Unicom in the second quarter of 2006 grew by 44% (Lang, 2006). Billing disputes constitute a significant portion of consumers' complaints. 70% of complaints against China Mobile and 50% of complaints against China Unicom were related to billing disputes. Attempts at adopting market-oriented practices by these local operators were also dampened by the inexperience of consumers and errant behaviour of contractors. For instance, China Unicom has pointed out that some of the complaints arose because of the inexperience of users. For example, some users were not aware that offers accepted by them have expiry dates and associated consequences (Lang). In addition, some of the complaints about unfair business practices were actually due to the inappropriate behaviour of agents or third-party service providers rather than the mobile phone telecommunications operator itself.

Technology Readiness of Participants

In technology leapfrogging, participants' attitude toward and acceptance of new technology can have an impact on the successful harnessing of unprecedented e-commerce potentials from this technology. Mobile phone technology has the potential to create a wireless electronic platform for realizing e-commerce such as paying for goods and services through mobile phones (m-payment). However, most Chinese are not used to paying for transactions through electronic means such as the Internet and mobile phone network because of security concerns and their traditional habit of making payments by cash. It would be a challenge to change Chinese consumers' perceptions about using their mobile phones to make payments.

The usage of m-payment systems through mobile phones has been basically confined to small-value purchases or simple transactions such as purchasing information (for example, weather reports, stock information, and transportation schedules), buying lottery tickets and admission tickets, paying utilities bills, topping up prepaid mobile phone accounts, and checking bank balances. Payments are often made from the payer's preestablished escrow (debit) account rather than based on credit facilities because the latter is relatively underdeveloped. Payment

Table 4. Price basket for different ICTs as a percentage of GDP per capita (Source: The World Bank Group, 2006)

Countries	Price basket for mobile as % of GDP per capita in 2003	Price basket for residential fixed line as % of GDP per capita in 2004 (other year in bracket)	Price basket for Internet as % of GDP per capita in 2004
China	0.80%	0.24%	17.33%
Australia	0.07%	0.09%	2.58%
Germany	0.05%	0.07% (in 2002)	2.60%
Japan	0.06%	0.07%	1.98%
United Kingdom	0.08%	0.08%	2.88%
United States	0.04%	0.06%	1.55%
India	1.55%	0.50%	6.84%
East Asia & Pacific	1.61%	0.40% (in 2002)	17.17%
World	0.44%	0.17% (in 2003)	4.31%

transactions take place under the instructions of the payer via SMS communications rather than on direct mobile credit or banking platforms. For example, ICBC (Industrial and Commercial Bank of China) allows its customers to send instructions in formatted short messages to its special service number, where their enquiries, transfers, remittances, donations, consumption, and payments are processed and the bank confirms or informs customers of the result via SMS (ICBC, 2006). The processing of these transactions is done manually. On the whole, China's major financial institutions are cautious about m-payment due to security and interoperability issues with the present networks. Mobile payment is one of the high-end value-added service segments that mobile phone operators are keen to cultivate for e-commerce. The major drivers of mobile payment have been the mobile phone telecommunications operators and mobile payment service companies, rather than the banks themselves. Besides interoperability issues associated with banking networks, another technology obstacle is that many mobile phones are currently not equipped to handle such transactions (Hendrickson, 2006). In addition, mobile phones with m-payment capability are likely to be more expensive.

Cost and Price

Owning and using a mobile phone for communication could be an expensive affair in China. Table 4 shows the expenditure (represented by a price basket) in using different ICTs as a proportion of the income (represented by percentage of GDP per capita) of a subscriber exhibiting the same usage pattern in different countries. The expenditure as a proportion of income is also interpreted as the cost of using the respective ICT in these countries. On the basis of the same usage behaviour in different countries, the cost of using a mobile phone for communication in China in 2003 (0.80% of an individual's income) would be more expensive than in developed countries such as Australia (0.07%), Germany (0.05%), Japan (0.06%), the United Kingdom (0.08%), and the United States (0.04%). Although this cost was lower than the average cost of using mobile phone telecommunication in the East Asia and Pacific region (1.61%), it was higher than the world's average cost (0.44%). Table 4 shows that the cost of using residential fixed lines in China in 2004, based on the same usage behaviour in other countries, was higher than the average cost in the developed countries. The costs of using the Internet in China in 2004 were also higher than in other countries.

China's mobile communication cost is higher than the world average and this can have a significant impact on its competitiveness in global e-commerce trade.

Regulation

Much effort needs to be devoted to establishing a clear regulatory policy, which is critical to the success of a secured and transparent mobile phone usage environment for e-commerce activities. Negative experience with or perceptions of mobile phone network security can seriously limit consumer acceptance of the use of e-commerce. However, to create an adequate legislative and regulatory framework for the protection of the environment is a difficult process in any country. This is particularly true for developing countries like China where institutional structure and infrastructure are still at a nascent stage of development (Zhu, 2005).

The Chinese government has struggled to develop strategies to control and combat illegal activities occurring on this wireless platform. In 2004, the Chinese government issued new rules for controlling transmission of SMS content. In 2006, further new rules were issued that require mobile phone users to use their real names when registering to set up prepaid or postpaid mobile phone accounts. Other countries such as Japan, South Korea, and Singapore have already

implemented this practice in their mobile phone telecommunications industries. A survey undertaken prior to the effective date of this new regulation in 2005 revealed that only 7.2% of the respondents in China supported this new regulation, but other respondents preferred to protect their privacy or believed that there are alternative solutions to combat illegal activities (Real-Name Mobile Phone Subscription Questioned, 2005). The MII acknowledged that this new measure of using one's real identity for subscription may inconvenience mobile phone service operators and providers, but deemed it necessary to combat crimes and inappropriate content being carried out and transmitted via this wireless platform. In fact, this requirement of using real identification to set up accounts is not new to the Chinese. Since 2000, the government has required the use of a real name and identification when an individual opens a banking account. The rationale behind this decision was similar to the mobile phone situation: It was necessary to combat financial crimes and fraudulent practices.

Digital Divide

Technology leapfrogging can open up development opportunities for a developing country. However, if technology leapfrogging is not properly managed, it can generate imbalance in regional development and create or aggravate any digital divide within an economy.

In China, the rise of rural-urban inequality in income constitutes a grave challenge to its economic and social development. In 2004, statistics show that the average annual disposable income of urban residents was 3.2 times that of rural residents (*China Statistical Yearbook*, 2005). In addition, about 60% of China's population lives in the rural regions and 10% of this rural population lives below the poverty line of $105 per annum. This population has been isolated from the urban economy and mostly engaged in semisubsistence farming, with relatively little cash income available. The narrowing of the economic gap between these two socioeconomic groups requires improved communications for the commercialization of rural food markets for the rural farmers and increased interchange between rural and urban populations.

However, teledensity coverage in the rural areas is significantly lower than the urban areas (as shown in Table 5). The fixed-line telephone penetration rate in rural areas was about 3 times below urban areas in April 2006.

Mobile phone technology may seem to be the technology springboard for rural areas. However, the ratio of the mobile phone penetration rate between urban (50%) and rural areas (7%) has been about 7:1 (*Prospect of Mobile Telecommunications Market in 2006*, 2006). Due to network coverage and high fees, the mobile phone penetration rate in the rural areas, particularly in the poor midwestern region of China, has been significantly lagging behind the urban areas.

Table 6 shows the expenditure (represented by a price basket) in using different ICTs as a proportion of income (represented by percentage of GDP per capita) between an urban and rural subscriber exhibiting the same usage pattern in China. On the basis of the same usage behaviour, the cost of telecommunications constitutes a significant portion of the rural per capita disposable income, particularly for mobile phone usage (37.12%).

Although a mobile phone provides a quicker and less costly solution for overcoming the slow development or inadequacy of the current fixed-

Table 5. Number of fixed-line telephone subscribers per 1,000 persons in April 2006 (Source: Striving for a Nation Stronger in Information Industry, 2006)

Areas	Fixed-line telephone subscribers per 1,000 persons
Urban	440
Rural	153

Table 6. Comparison of price basket for different ICTs as a percentage of GDP per capita between urban and rural areas (Source: The World Bank Group, 2006)

China	Price basket for mobile as % of per capita disposable income in 2003	Price basket for Internet as % of per capita disposable income in 2004	Price basket for residential fixed line as % of per capita disposable income in 2004
Urban	11.49%	3.77%	3.63%
Rural	37.12%	12.10%	11.64%

line infrastructure, the cost burden of mobile phone telecommunications is shifted to the users by way of high fees. To encourage mobile phone adoption in the rural areas, the high telecommunications cost requires some form of subsidization as operators will take a considerable period of time to build a critical mass in those regions to achieve a break-even point or economies of scale on their investment. The expansion of telecommunication coverage, whether fixed line or wireless, into the rural areas is not so much for the sake of increasing business revenue but more on the grounds of social responsibility in closing the economic gap between the haves and have-nots. It is expected that the purchasing power and demographic background of rural inhabitants would not generate high demand for high-end value-added services or services enabled by 3G technologies for a considerable period of time. As a result, government support and intervention would help the rural communities in leapfrogging to mobile phone technology in order for them to be integrated into the mainstream of economic activities.

Market Opening

Although the Chinese government has undertaken a series of market reforms, it continues to exert considerable influence on the mobile phone telecommunications industry. This is despite its undertaking to the World Trade Organisation (WTO) in 2001 to schedule direct foreign participation in value-added and basic services, and to establish an independent and transparent regulatory authority and a procompetitive regulatory regime over a 6-year time frame. China's opening of its telecommunications sector has been considered slow and bureaucratic.

This protective attitude would deprive the industry of opportunities for technology spillovers and human development from foreign direct investment (FDI). Studies have supported that FDI provides important opportunities for knowledge transfer to domestic firms and helps improve local productivity (Blomström & Kokko, 1998; Organisation for Economic Co-operation and Development [OECD], 2001; Wei & Liu, 2006; Torlak, 2004), and these attributes can be beneficial to developing local capabilities in advanced technologies.

Despite the enhanced competition, as a result of the breakup of China's telecommunications monopoly structure and changes in pricing regimes, the optimal benefits from unrestricted market competition in the basic and value-added mobile phone segment are yet to be realized. A highly competitive market will likely provide advanced means of communication (high-end value-added services) at a cost-effective price, or at a price that is no longer an adoption barrier in itself to the cost-conscious Chinese users. In this way, new technology would not be underutilized given its potential applications.

However, local mobile phone telecommunications operators have been worried about their

decreasing ARPUs. In addition, their worries have been compounded by delays in the roll out of 3G technologies, which are capable of supporting high-end value-added services as well as bringing down the prices of these services. These operators want to explore the potential of this technology for new initiatives and to generate more revenue from existing subscribers. They are keen to establish a strong foothold in the market with this technology, and the liberal opening to foreign competition would stifle their chance of realizing such objectives. In addition, the experienced and seasoned foreign competitors are likely to drive radical changes to this market system and siphon some of their current market share.

REFERENCES

Analysis of China's current mobile fees situation. (2006, July 26). Retrieved August 11, 2006, from http://www.cttl.cn/cjgl/cjgc/t20060726_405447.htm

Beijing mobile phone operators to slash charges. (2006, May 9). Retrieved August 8, 2006, from http://www.chinaview.cn

Blomström, M., & Kokko, A. (1998). Multinational corporations and spillovers. *Journal of Economic Surveys, 12*, 247-277.

Burns, S. (2006, February 24). *Vnunet.com analysis: China's 3G conundrum.* Retrieved August 30, 2006, from http://www.vnunet.com/vnunet/news/2150838/government-delays-china-3g

Castells, M., Ardevol, M. F., Qiu, J. L., & Sey, A. (2004). *The mobile communication society: A cross-cultural analysis of available evidence on the social uses of wireless communication technology.* CA: University of Southern California, Annenberg Research Network on International Communication.

Chen, K. (2004, April 26). *Xiaolingtong's existence and development.* Retrieved August 14, 2006, from http://www.cttl.cn/tegd/shchgch/t20060709_394514.htm

Chen, L. M. (2006, July 25). *Comments and analysis of China mobile's market segmentation strategy.* Retrieved August 11, 2006, from http://www.cttl.cn/cjgl/zlgl/t20060725_405112_4.htm

China Statistical Yearbook. (2005). Beijing, China: China Statistics Press.

Dholakia, N., Dholakia, R. R., Lehrer, M., & Kshetri, N. (2002). *Patterns, opportunities, and challenges in the emerging global m-commerce landscape.* Retrieved August 4, 2006, from http://ritim.cba.uri.edu/wp2002/pdf_format/M-Commerce-Global-Landscape-Chapter-v07.pdf

Discussion of China's pressing mobile fees problem. (2006, July 26). Retrieved August 11, 2006, from http://www.cttl.cn/cjgl/cjgc/t20060726_405465.htm

Domestic problems grow for China's 3G. (2005, June 22). Retrieved August 15, 2006, from http://www.3gnewsroom.com/3g_news/jun_05/news_5984.shtml

Fan, Y. Z., & Wang, Z. F. (2005, January 12). *Develop tactics for mobile SMS and fixed network SMS to face the competition in 2005.* Retrieved August 14, 2006, from http://www.cttl.cn/tegd/shchgch/t20060709_394618.htm

Forge, S. (2004). Is fourth generation mobile nirvana or…nothing? *Info, 6*(1), 12-23.

Hao, W. (2005). Meeting challenges under the new environment. *China Communications*, 13-17.

Hendrickson, D. (2006, August 1). *Mobile TV's curtain call.* Retrieved August 22, 2006, from http://www.cityweekend.com.cn/en/beijing/cib/2006_08/mobile-tv2019s-curtain-call

Industrial and Commercial Bank of China. (2006).

Mobile banking (short message). Retrieved August 18, 2006, from http://www.icbc.com.cn/e_center/gerenjingpin/grl-5e.html

Katz, J. E., & Sugiyama, S. (2005). Mobile phones as fashion statements: The co-creation of mobile communication's public meaning. In R. Ling & P. Pedersen (Eds.), *Mobile communications: Renegotiation of the social sphere* (pp. 63-81). Surrey, United Kingdom: Springer.

Kellerman, T. (2002). *Mobile risk management: E-finance in the wireless environment* (Financial sector discussion paper). The World Bank.

Li, W. (2003). China's burgeoning mobile phone industry. *China Today*. Retrieved August 11, 2006, from http://www.chinatoday.com.cn/English/e2003/e20039/9p12.htm

Looking at the pattern of fees in voice-based business from the recent adjustment in fees. (2006, July 26). Retrieved August 11, 2006, from http://www.cttl.cn/cjgl/cjgc/t20060726_405456.htm

MII confirmed China Mobile to lower phone call charges in May. (2006, May 10). Retrieved July 3, 2006, from http://www.cn-c114.net/market_html/mii2006519142317-1.html

MII Zhang Xing Sheng: Value TD important test period, otherwise the possibility of lost opportunities. (2006, August 10). Retrieved August 14, 2006, from http://www.cttl.cn/hydt/txyw/t20060810_413327.htm

Ministry of Information Industry. (2005a, December 22). *State Development and Reform Commission, Ministry of Information Industry on the revision of the Telecommunications Tariff management notice [2005] No. 408*. Retrieved February 2, 2007, from http://www.mii.gov.cn/art/2005/12/22/art_541_3180.html

Ministry of Information Industry. (2005b). Telecom development targets and focus in China in 2005. *China Communications*, 22-27.

Mobile communications. (2004, December 16). Retrieved July 3, 2006, from http://www.cn-c114.net/market_html/re0420041216114452-1.html

Mobile phone payment has entered into financial channel. (2006, July 3). Retrieved July 3, 2006, from http://www.c114.net/technic/ZZHtml_20067/T2006739592115847-1.shtml

Nationwide first six months: Mobile phone subscribers reached 0.426 billion, penetration rate is 3.27 units and 200-plus billion in SMS. (2006). Retrieved August 11, 2006, from http://www.cttl.cn/hydt/ywzb/t20060801_407719.htm

Organisation for Economic Co-operation and Development. (2001). *Foreign direct investment for development: Maximizing benefits, minimizing costs* (OECD Policy Brief). Paris: Author.

Overview of China's mobile phone market. (2005, November 29). *CRIENGLISH.com*. Retrieved June 9, 2006, from http://en.chinabroadcast.cn/855/2005/11/29/262@33515.htm

Peng, S. (2003). Universal telecommunications services in China: Trade liberalization, subsidy, and technology in the making of information equality in the broadband era. *Asian-Pacific Law & Policy Journal*, *4*(1), 21-49.

Prospect of mobile telecommunication market in 2006. (2006, July 13). Retrieved August 11, 2006, from http://www.cttl.cn/tegd/zhjft/t20060713_397167.htm

Real-name mobile phone subscription questioned. (2005, December 21). *Xinhua News Agency*. Retrieved June 14, 2006, from http://in.china-embassy.org/eng/zgbd/t227713.htm

Salkever, A. (2004, July 22). Apple's slow boat to China. *BusinessWeek online*. Retrieved August 11, 2006, from http://www.businessweek.com/technol-

ogy/content/jul2004/tc20040722_8277_tc056.htm

The sprint in the trial of TD network. (2006, August 9). Retrieved August 14, 2006, from http://www.cttl.cn/txis/jsdt/t20060714_400802.htm

Statistics revealed substantial difference in handphones' penetration rates, and Beijing and Shanghai unwilling to make calls. (2006, August 8). Retrieved August 14, 2006, from http://www.cttl.cn/cjgl/xwpd/t20060808_410473.htm

Striving for a nation stronger in information industry: A report from the 2006 National Information Industry Working Conference. (2006). *China Communications,* 27-32.

TD-SCDMA, 3G in China. (2006, February 16). Retrieved July 3, 2006, from http://www.cn-c114.net/markethighintro.asp.asp?id=30

Torlak, E. (2004). *Foreign direct investment, technology transfer and productivity growth in transition countries: Empirical evidence from panel data.* Retrieved December 9, 2006, from http://www.cege.wiso.uni-goettingen.de/Dokumente/Diskussion/Torlak_26.pdf

Wang, X. G. (2002). Xiaolingtong: Rise from windy storm. *Mobile Communication, 8.* Retrieved June 14, 2006, from http://www.mc21st.com/magazine/2002-08/02.pdf

Warning, storm ahead…TNX. (2006, July 28). Retrieved August 9, 2006, from http://cnn.com

Wei, Y., & Liu, X. (2006). Productivity spillovers from R&D, exports and FDI in China's manufacturing sector. *Journal of International Business Studies, 37*(4), 544-557.

The World Bank Group. (2006). Retrieved December 8, 2006, from http://devdata.worldbank.org/dataonline/

Xu, A. Z., & Tao, W. H. (2003). *China's mobile telecommunications competition situation.* Retrieved June 12, 2006, from http://www.mc21.st.com/magazine/2003-05/21.pdf

Zhang, M. Z. (2006, August 10). *Direction of telecommunications reform: "Favouring" unfair competition?* Retrieved August 11, 2006, from http://www.cttl.cn/cjgl/cjgc/t20060810_412948.htm

Zhongxing TD's R&D frustration: Skilled talents were lured away by high salaries. (2006, August 9). Retrieved August 25, 2006, from http://www.cttl.cn/cttlcds/scgc/sckx/t20060808_410782.htm

Zhu, G. (2005). ICI initiatives in China. *China Communications*, 4-12.

ENDNOTES

1. Xiaolingtong (meaning *little smart* in Chinese) is a wireless extension of the fixed-line system with no roaming capability. The operators of this telecommunications system are China Telecom and China Netcom. Its phone's range is limited to the local geographic region, which is usually a single metropolitan area.
2. An urban legend is a story, which may at one time have been true, that has grown from constant retelling into a mythical yarn (http://www.netdictionary.com).
3. SIM Tool Kit technology can be used to provide encryption security through the SMS channel, but it is a transport layer security mechanism and does not provide end-to-end confidentiality (Kellerman, 2002).

This work was previously published in International Journal of Cases on Electronic Commerce, Vol. 3, Issue 4, edited by M. Khosrow-Pour, pp. 19-38, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.4
United States of America:
Renewed Race for Mobile Services

Mats Samuelsson
Mobio Networks, USA

Nikhilesh Dholakia
University of Rhode Island, USA

Sanjeev Sardana
Mobio Networks, USA

ABSTRACT

Somewhat behind in the mobile telephony adoption than leading European and Asian markets, the U.S. market caught up in the 2000s. While simple types of mobile data services — such as messaging and downloads — had made some headway, the preexisting popularity of PC-based Internet made the U.S. users somewhat resistant to m-commerce offerings that did not match the richness of PC e-commerce. By the mid-2000s, however, network and technological capabilities were in place to usher in rich, new m-commerce offerings in U.S. markets. By taking advantage of new technologies, the U.S. mobile industry had the opportunity to become an innovator in m-commerce offerings.

INTRODUCTION: MOBILE COMMERCE AND DATA SERVICES IN THE UNITED STATES

After a late start, by mid 2000s the United States' mobile phone market quickly caught up with the rest of the world and started undergoing the same consolidation experienced in all other major markets of the world. From a field of six major service operators — Verizon, Cingular, Sprint, AT&T Wireless, T-Mobile and Nextel — at the beginning of 2005, there were four major carriers left at the end of 2005: Cingular acquired AT&T Wireless, and Nextel acquired Sprint. At the same time, the U.S. market reached the 70-80 percent mobile phone penetration level common in all industrial countries around the globe.

The year 2004 represented a breakthrough in terms of wide introduction and deployment

Box 1. Mobile Phones as M-Wallets in Boston

The Super 88 Market, Infusion Tea Spa, Paris Café, and Marty's Liquors– these stores in Boston's Allston neighborhood accept the MobileLime payment system developed by Vayusa, a startup company from the Boston area. Super 88 customers save up to 5 percent when they buy with MobileLime.

To purchase items, MobileLime users dial a toll-free number. The customer's account is linked to either a credit card or to a prepaid account. A PIN number is required to authorize purchases. The cash register at the store also connects with MobileLime and receives the sale approval information. An e-mail sent to the customer verifies the transaction.

Bob Wesley, who joined Vayusa as its CEO after many years with financial services firms such as American Express, realizes that not everyone is likely to jump on the m-wallet bandwagon. There would, however, be a segment that would find it convenient not to carry credit cards and just use the phone as a payment device.

of data services based on GPRS and CDMA by U.S. operators. Initial plans for data volume-based pricing (dollars per MB) were quickly abandoned in favor of flat data rates and unit service charges for messaging. In late 2005, these charges settled around $5/month for mobile phone Web browsing and $20-$30/month for unlimited mobile data access for handheld PDAs and laptop computers.

Initial hope for mobile data services was high, and AT&T Wireless introduced a U.S. version of the Japanese i-Mode called m-Mode. Except for e-mail and some WAP browsing, however, the m-Mode service uptake was disappointing and m-Mode was practically shelved when Cingular acquired AT&T Wireless. On-demand services did not fare much better. The most successful examples of such services were downloadable ringtones that generated about $217 million in 2004, followed by mobile games at $72 million. These accounted for about 10 percent of the non-access data revenues for mobile carriers. Messaging — SMS and MMS — accounted for 65 percent of about $3 billion in data service revenues circa 2005 that were not associated with data access fees. As MP3-capable handsets began entering the market during 2005-06, downloadable MP3 tunes held the promise of being the next large download service, although there were no clear mobile business models developed to effectively compete with Apple Computer's successful iPod/iTunes.

The future for mobile services continues to be driven by the critical areas of handset technology, network technology, business models and service innovation. Current mobile data services in the U.S. have taken advantage of some of the developments in these critical areas (see Box 1 "Mobile Phones as M-wallets in Boston"), but there is a potential to do a lot more. In Japan, South Korea and some European markets, the mobile data and m-commerce envelope has been pushed a bit further than in the U.S.

Because of the very large user base and the "ecosystem" of rapid innovation in the U.S., however, there exist real prospects for spawning many

new m-commerce applications and services in the American setting. This chapter reviews the current mobile data and m-commerce business models in the U.S., examines some international business models briefly and finally offers suggestions for fresh and innovative approaches towards 3G and higher-generation m-commerce for the U.S. and global markets.

SERVICE OVERVIEW: CURRENT STATUS

By 2005, the U.S. market for mobile communications had characteristics similar to other major global markets. A range of fairly popular services was available, and some services had been tried without much success.

Messaging

SMS and MMS messaging continue to gain ground in the U.S. market, particularly messaging from camera phones. Interoperability between all U.S. mobile operators is now in place so the market is expected to grow substantially.

Web Browsing

As in the rest of the world, WAP has failed to achieve large usage in the U.S. market. Low speed and cumbersome user interfaces have kept WAP from developing any substantial usage.

Downloads: Ringtones and Games

Downloading ringtones has turned out to be a huge success, with yearly service revenues growing from zero in year 2000 to $300 million in 2004, part of the surprisingly large $4 billion worldwide market for ringtones. As with many other data services, such as SMS, ringtones became an unexpected success for the mobile industry. Some optimistic voices even claimed that ringtones had become a music category on its own and were a potential future savior of the music industry as well as an advertising medium (Scott, 2005).[1]

Downloading games was slower to catch on in U.S. compared to markets such as Japan and South Korea, mostly because of fierce competition with handheld and console game-players such as Game Boy and PSP from Sony. In the U.S., downloadable mobile games remained at a disappointing market level of under $100 million in 2005.

Tried and Failed: Payments

Till 2005, mobile payments — micropayments using the mobile phone — had failed to take off in the U.S. market despite numerous attempts. While the notion of turning the mobile handset into a mobile wallet is a powerful idea, this idea has pitted mobile operators and banks against each other in the marketplace. These two industries have not been able to agree on any common approach and have parted ways over issues like security, information stored in handsets and ways of sharing profits. Failed attempts have included proposals for dual-chip handsets with encrypted customer information. One of the business challenges is also the mobile phone bill — operators are loath to load up the bill with third-party related purchase charges for customers used to fixed-rate 500-minute plans. This quite different than, for example, Japan, where mobile phone customers are quite used to seeing all manner of third-party service charges on their monthly mobile communications bills.

Future

With the positive experience of downloads, the industry was looking for the next natural service and revenue opportunities. Eyes were cast at the booming MP3 download market (iPod, Napster, etc.) as well as video clip success stories from South Korea. Despite strong initial successes,

the long-term prospects of some of these services were less than clear. The iPod already had a strong hold on the market and the video clip as a stand-alone service was yet to be proven by compelling content that could be viewed on the small mobile phone screen.

CURRENT BUSINESS MODELS

Over and above the revenue from voice calls, mobile operators have gradually developed other sources of data communications-based revenues.

Transactions and Transport: E-Mail, SMS and MMS

Driven by the huge international success of SMS, data services provided as transactions continue to be a way to charge premium prices for bandwidth. Mobile operators need revenues from services in addition to mere voice telephony to justify the very substantial investments in upgrading the mobile network for data services. The unfortunate reality has been that apart from modest success with MMS in the U.S., most volume usage of data has been in the form of flat rate plans with unlimited usage — many targeted at business users. Given competition from Wi-Fi offerings (that often make Internet and data services available at low or no costs at locations such as coffee shops, hotels and airports), the most successful mobile data applications have been e-mail driven, initially offered over separate data networks but migrating toward GPRS and EDGE. In the foreseeable future, our view is that transaction-based business models will continue to have their relevance in the mobile space, but will be under constant pressure from flat-rate transport "needs."

Portals and Downloads: Yahoo and Google

WAP represented the first attempt to mimic the Internet business model in the mobile space. Unfortunately it turned out that traditional Internet content had become optimized for the PC with its good graphics, a mouse and, most importantly, a full keyboard. WAP-enabled mobile Internet content, by comparison, was of poor quality and was unappealing — especially to American users who had been used to rich desktop and laptop Internet content for years.

After initial failures of WAP portals, mobile portals again started receiving attention in the mid 2000s as both Google and Yahoo announced efforts in this area. It is clear that success of these would depend on either simplicity of user interface (no text input required), or very specific transaction-oriented usage. The Internet itself needs to become mobile-friendly — content modified to fit the mobile handset (screen size and numerical input only) and enrich the experiences of users on the move — before widespread Internet usage over mobile phones would take off. This means that mobile portals would have to be dramatically improved before they could become a viable service revenue source (see the first chapter in this volume, also Rask & Dholakia, 2004).

Mobile Operator Service Aggregators: m-Mode, t-Zones, Media Net

At one stage of the mobile technology evolution, service aggregation was viewed as the next savior of mobile data services, moving such services beyond the message-oriented revenue models of SMS and its multimedia MMS sibling. Most service aggregation approaches were modeled after Japan's NTT DoCoMo i-Mode approach (see the Japan chapter in this volume, also Lennon & Dholakia, 2004) and this Japanese operator made a big investment in AT&T Wireless in the late 1990s, a key part of which was to bring the

i-Mode approach to the U.S. The USA i-Mode variant, dubbed m-Mode, never took off in the U.S. and the AT&T-Wireless's implementation was in a state of limbo until the acquisition of AT&T Wireless by Cingular. By the mid 2000s, there were no signs of m-Mode activity in the U.S. market. Instead the state of mobile operator service aggregation in 2005 can be represented by excerpts from a Website of a leading U.S. mobile operator (see Figure 1).

As can be seen from this offering, service aggregation has become a collection of disjoint services with one or two providing the great majority of content aggregation revenues. The other services have failed to get traction beyond small segments of the user market. Furthermore, the service aggregation model is being challenged in the U.S. market by an ever-increasing number of Web-based, third-party providers of ringtones and games that completely bypass the operator portal. Thus, this business model continues to be under pressure as suppliers operating outside the loop are challenging existing revenue successes.

Business Model Issues: Operator vs. Third Party Revenues, Transport vs. Content

One of the most challenging issues for mobile data services in the U.S. is associated with business models, in particular revenue sharing aspects of such services and the notion of charging for transport or content.

With the early lead of Internet-based business models, U.S. mobile operators have had to adapt to the preexisting Internet models rather than create new data services models as in the rest of the world. An example of this is data usage pricing. Early ISPs tried this approach but it quickly fell by the wayside in favor of flat rate monthly access fees. In the U.S., early attempts to only offer fixed quantities of data download service plans for mobile data services quickly gave way to flat rate plans. This was not the case in Japan, for example, where NTT DoCoMo's i-Mode succeeded in the late 1990s by charging according to the size of the data packets downloaded — the Japanese users, not addicted to flat-rate Internet like the U.S. users are, were quite willing to pay

Figure 1. Example of Mobile Service Aggregation in USA, circa 2005

MEdia Net brings Web sites, e-mail, messaging, downloads, and more to the mobile phone.

- Mail & Messaging: Check Yahoo! Mail, MSN Hotmail, and chat with friends with Yahoo! Messenger and Upoc.
- Sports: Get the latest scores from CBS SportsLine and ESPN.
- Ringtones, Games & Graphics: Personalize the phone by downloading favorites.
- News & Finance: Stay informed with round-the-clock headlines from CNN.
- Entertainment: Get local movie times and reviews, dining recommendations, and more.
- Weather & Travel: Check forecast from The Weather Channel, get flight times, and traffic reports.

And much more! Start browsing today and discover what MEdia Net has to offer. And be sure to visit What's Hot! where you'll find the latest and greatest MEdia Net content.

Source: https://www.cingular.com/media/media_net and authors' research

such charges (see Bradley & Sandoval, 2002; Lennon, Dholakia, & Dholakia, 2004). While usage pricing still exists in the U.S. mobile markets, it will most likely disappear as mobile data usage increases.

Similarly, attempts by ISPs to share transaction revenues with service or content providers have failed for several reasons. The Internet model easily lent itself to direct interaction between user and Website, and any intermediary introduced unnecessary complexity without any obvious value. American users have therefore come to regard the ISPs — the fixed-line data service providers — as just communication providers and not as value-adding services. In the mobile markets, these same expectations have spilled over and attempts by mobile operators (WSPs) to act as value-added service providers have generated no enthusiasm among mobile users.

The early lead of the Internet (and other existing business models) will continue to be a challenge for new data services. A good example of this is the next expected set of killer data services: MP3 downloads. MP3 playback capability is already available in high-end handsets and is expected to migrate downward following the path set by the color screen and camera phones. The problem for mobile operators is that iPod has already laid claim to a substantial part of this market with a carefully crafted terminal, distribution and content relationship model. At $0.99 per song, there is little room outside this model for additional revenue sharing with mobile operators. Further complicating the matter is the issue of digital rights, solved in the iPod's "closed and proprietary" model but unresolved in the open world of mobile phones.

Ringtones show that it is possible to create a major service revenue success when the mobile technology and service firms are able to participate in the creation of the business models and be critical components of the early service deployment models. Even with this business model, however, the "proprietary ownership of content" days are numbered as more and more ways become available for users to bypass the operator and download directly to their phones or via the PC.

The mobile data services business thus faces the challenge of operating at the crossroads of the fixed-rate Internet business models and usage-based (and mainly non-U.S.) models created from within the mobile business. This will continue to challenge suppliers as they try to create successful mobile-specific business models.

EVOLVING CONSUMER NEEDS

Voice Calls

Voice communications continue to be the core of the mobile business and will continue to be the main reason why people carry and use mobile phones. Most voice services remain stuck in the same domain as traditional fixed line services but as VoIP enters the mobile domain (in the form of IMS) it is clear that the same easy-to-use call features, configurations and services that are available in the fixed-line world will become available in the mobile domain. It is easy to envision the same types of profiles that are available for ringtones — conferencing, forwarding, messaging and automatic dialing — all in one integrated profile invoked at the push of a button.

Handset as Gadget

This is probably the most underestimated part of the mobile market — the constant ability of the handset to add new functions and capabilities. With 700 million handsets shipped globally in 2004, the mobile phone is becoming one of the largest consumer electronic categories around. It is second only to the television set in terms of availability, and exceeds even the TV set because the mobile phone is a personal rather than a family device. In the developed world, the majority of these sales are in the form of replacement phones.

It is therefore critical for the suppliers to keep enhancing their phones with features, capabilities and designs that appeal to existing customers and provide reasons for switching. As with all advanced consumer product markets, there is constant fragmentation of tastes and finer re-segmentation of markets; and the numbers of phones, types and suppliers keep increasing. The "gadget competitive envy factor" forces a combination of features and designs and other trends. Future design "innovations" would no doubt continue to replace recent successes, such as the craze for flip-the-lid type "clamshell" phone designs.

SMS/IM/E-Mail

It is difficult to envision further evolution of SMS, IM and e-mail needs with the exception of ways to restrict junk messages and improvement of user interfaces. Keyboards are standard on high-end handheld devices and this feature will no doubt migrate downwards to popularly priced handheld and palmtop device models.

MMS

"Fun with Photos" can be expected to increase in popularity as cameras get better and ways of sharing pictures and experiences improve. Good video and true multimedia will provide the next natural evolution once bandwidth restrictions are removed (in 3G and later 4G networks) to allow for transfers of large data files at blazing speeds.

Data Services in General

Data services remain an area of untapped growth potential for U.S. mobile operators. With steadily increased usage of the various services described so far, data revenue is becoming a larger share of the U.S. mobile operators' average revenue per user (ARPU). Hindering wider adoption of mobile data and m-commerce in the American mobile markets are a couple of challenging factors:

- Overly complex user interfaces (number of clicks to take a picture with a mobile camera phone and send it to a family member or friend is often 10-15 instead of 2-3); and
- Business models have to evolve to meet existing customer needs (ease of use and convenience) as well as new ones (impulse use, content download, payment, location-specific services, etc.).

Handset Evolution

Phone, Display, Camera, Media-Player and User Interface

In looking at the most sophisticated contemporary phones, it is difficult to see any dramatic evolution in feature sets and capabilities. What can be expected is some major innovation in user interface, away from today's complexity and more oriented toward specific usages at any given time. The desktop and laptop PC Web browser has evolved into a very versatile interface for supporting nearly any type of Website in an effective manner. The same cannot be said for the mobile phone. The mobile interface needs to evolve to become as versatile and convenient as the desktop Web browser.

Processing Power, Memory, Platform and Software

Several underlying supply-side technological forces are driving the evolution of mobile data services and m-commerce. These forces include an increase in processing power (100-400MB clocking) and memory (64-256MB and beyond), improving the performance of existing platforms (Symbian, MS-Mobile and BREW), and allowing for new software (such as games) to perform at acceptable levels (comparable to desktop performance levels).

Network Evolution

1G, 2G, 3G and IMS, have evolved as networks from 1G to 2G and 3G; we have experienced increases in data speeds from 9.6 Kbps to 56Kbps and 200Kbps and beyond. These are the advertised rates and it will take some time for network buildouts and improvements before the 3G data speeds are commonly available to all users of 3G phones (Samuelsson & Dholakia, 2004). The introduction of IMS promises to dramatically improve these speeds into the 1Mbps domain and fully integrate voice as an IP service. From the perspective of the mid 2000s, it appears that IMS will suffer the "3G syndrome" — considerable hype and talk for a number of years before it is suddenly implemented on a widespread basis, as it becomes clear that it offers substantial service revenue opportunities to mobile operators. It is clear that this is not the case in 2005.

OPERATORS' DILEMMA

With a high level of service penetration in the U.S. market, mobile operators are facing a stagnating market — not so much in terms of market size increase but certainly in terms of growth of ARPU. The initial reaction to this has been a consolidation of the market with the number of U.S. mobile operators reduced to four, the largest with 60 million subscribers and the smallest with 20 million. At the current time, these operators are facing a key problem. There is little differentiation amongst service providers, and this is leading to price-based competition and no operator loyalty. Adding to this problem are the following ongoing issues and changes:

- Need for continued high network investments: Building and maintaining wireless networks, while less expensive than building wired and fiber-optic networks, is nonetheless a high cost proposition. This is especially the case in geographically vast and dispersed national markets such as the United States. Large user bases sustained over long periods are needed to justify such investment;
- Impending and costly migration: IMS;
- Challenges posed by new radio spectrum availability, coverage area issues;
- Intense competition, often leading to price wars and declining ARPUs (multiple competitors in every market — four in the U.S., more and often bigger mobile operators globally);
- Marketing (customer acquisition and retention) costs remain at high levels and are unabated. Coupled with stagnant or declining ARPUs, these high marketing costs create major drags on the finances of mobile operators;
- Handset subsidies (U.S. consumers are hooked on "free" handsets as part of service plans);
- High channel costs (reaching new or existing customers for new services is difficult and expensive);
- Voice revenues eroding (marginal rates are under 5 cents minute in the U.S.);
- Data and other service revenues are relatively small (in the mid 2000s, these varied between 2-6 percent in the U.S.);
- Business models changing in unfavorable directions: Consumer preference for, and industry migration toward, flat-rate pricing models that can be sustained only with huge user bases; very limited emergence of usage-based, segmented, value-adding and specialized service models that could boost average revenue per user (ARPU);
- Stagnant messaging services models (SMS and MMS not growing fast enough); and
- Uncertain transaction ownership and revenue sharing models (not clear who owns specialized content/services and how to share their revenues and transaction costs. Mobile operators, banks, content providers

and special third-party applications providers often do not get along very well).

POTENTIAL SOLUTION: OFFERED SERVICES

Mobile communications and m-commerce markets in the U.S. have to move from this state of non-differentiation, intense competition and stagnant/declining ARPUs into a stage where there is clearer differentiation through unique product and service attributes.

Only those differentiation points that offer meaningful and appealing differences to individual and business users have a chance of supporting successful new business models. In technological and competitive terms, it is imperative for the United States to have such new and growing business models that can support the ongoing business and technology evolution of mobile networks and services. What are some of the potential differentiators available to mobile operators? The following lists the main ways that mobile offerings need to create differentiation:

- **Networks:** Difficult for users to feel the difference;
- **Handsets:** Scale economies of manufacturing these devices work against significant differentiation;
- **Customer Service:** Some differences, but all operators are deemed unsatisfactory by most users; and
- **User Experience (Offered Services):** Difference in voice services tied mainly to network quality attributes; data and other applications can be meaningfully differentiated (e.g., i-Mode in Japan).

Out of these, "Offered Services" is the only category that offers meaningful differentiation opportunities and it is data services in particular that have the potential to create new services that could jolt individual and business users to snap out of their inertia and take notice.

To date, successful "Offered Data Services" differentiation has come in two basic forms: i-Mode or customized enterprise applications. These have very different product attributes, operator participation and varying results.

The i-Mode Way

The success of NTT DoCoMo's i-Mode service in Japan was based on a number of key attributes. While these attributes worked well in Japan, it should be noted that it has proved difficult to transport this business model to other settings. The factors responsible for i-Mode success in Japan include the following:

- Essentially an m-commerce platform;
- A closed system, run by NTT DoCoMo;
- Micropayments aggregated, billed and collected with subscriber's telephone bill by NTT DoCoMo. In turn, NTT DoCoMo retained a substantial chunk of these micropayments as a fee for doing all this work;
- Offers subscription and consumption-based (usage-based) pricing models;
- Generally limited to information services that can be delivered to the specialized m-commerce platforms;
- Handset-based, simple-to-download services were the main offerings available on i-Mode (ringtones, stock quotes, horoscopes, etc.);
- Qpass and many other players have similarly succeeded with limited simple-to-download transactions like ringtones, fortune cookies, etc.; and
- Third parties providing such services via NTT DoCoMo's i-Mode platform were basically acting as application servers for mobile applications (Web download/browse model).

The Enterprise Applications Way

These are business-oriented mobile data services. The key attributes of successful mobile enterprise applications include the following:

- **Stovepipe Solutions:** Single application services developed for a specific purpose such as sales-call scheduling;
- **Hosted behind the firewall, weakly linked to the network:** Stovepipe solutions tend to depend on transport services only, connecting a handset application with a corporate server behind the corporate firewall. The mobile operator's network adds little or no value beyond providing connectivity;
- **Custom solutions with little leveraging across applications, hence expensive to build and operate:** Each service tends to be custom and different services are not integrated (for example e-mail, order entry and repair tracking are three different applications — they remain distinct in the mobile environment);
- **High ROI hurdle:** The low adoption rate of enterprise services creates an impossible-to-cross return on investment (ROI) hurdle.
- **Most success in field service applications (for large fleets):** UPS, FedEx or repair fleets tend to be the main users of mobile enterprise applications;
- **Mobile e-mail most desired but usually limited to subset of employees:** Corporate e-mail security issues and limitations prevent the use of full-fledged mobile e-mail applications for enterprises;
- **Little operator participation:** User experience is "carrier agnostic." As can be expected from a service that only uses data transport, one mobile operator can be substituted for another without any loss of functionality or features; and
- Does not create loyalty or branding benefits for the mobile operator since the operator provides just plain vanilla transport of data.

Implications for USA Mobile Operators

As can be seen, both of these approaches have had limited success, if any, for operators in the U.S. It is clear that enterprise applications have provided operators with substantial contracts for data services. Unfortunately, these have occurred without any branding or other operator advantage.

NEED FOR A FRESH APPROACH

Given the overall vitality of the mobile communications industry with its strong operators and strong supplier infrastructure, as well as its strategic position smack in the middle of the mobile devices, entertainment and Internet industries, there are plenty of opportunities ahead. But in order to address these opportunities, there is a need for a "fresh approach," an approach that capitalizes on new and evolving network/Internet technologies and provides mobile operators with sustainable advantages. A critical part of this is a reinvention of the user interface and a service delivery platform approach. Such reinvention should aim to reduce the time and effort to develop new services and integrate these with new network and handset capabilities as well as the emerging area of enterprise Web services. Such integration would enable mobile service providers to reach almost every conceivable content and business transaction. Some key attributes of this new approach are:

- Operator "ownership" of user experience is critical (similar to i-Mode);
- Platforms need to be closely tied to newer network capabilities like presence and location in addition to the critical control

channels and protocols allowing for full usage of all network services;
- Platforms need to provide a richer handset experience by taking advantage of the newer smartphones that are becoming pervasive. These handsets can render user-friendly user interfaces with images and usage directions;
- Platforms need to integrate fully with Web services, the rapidly evolving standard for access and delivery of enterprise applications and content;
- Reusable rich capability sets, residing in the network-based applications servers, would reduce the third party/enterprise cost of applications development and thus economically enable the creation of additional services. The challenge is to have service platforms that can serve a wide variety of applications;
- The new applications and services need to be parts of existing and evolving mobile network cores (2.5/3G and IMS). If services are part of the evolving infrastructure, they can begin introducing new features and services as soon as the supporting capabilities are introduced in the network; and
- Integration with existing and evolving Web-based business/enterprise computing (Web/XML).

In order to accomplish this, a new and different platform approach is needed, a platform that can provide a unique, easy-to-use user experience for a large number of different data services and allow operators to create and package horizontal services working with business partners providing content, information and other capabilities.

Scenarios of New Services

The following service scenarios provide a flavor of what can be expected in the United States mobile markets with greater integration of network capabilities, mobile device features, content provider offerings and third-party technologies:

Bob is driving along the highway and sees a billboard advertisement that his favorite country artist is coming into town six weeks from now. He presses the music button on his mobile phone and punches 7322 (the first name of the star). A large menu pops up: "Tickets, Music, Both or All." He selects all by hitting the "0" button. A $37.45 price tag flashes on the screen and Bob hits "OK" to accept this charge. One minute later the MP3 version of the singer's latest hit starts playing on the mobile phone music player. An electronic ticket to the concert, with premium seating, has been sent to Bob's mobile phone and a T-shirt announcing the tour and city is in the mail. As Bob walks into the arena six weeks later he receives a personal greeting from the artist and as he leaves the concert, the final number, just played at the concert for the first time ever, is already available on his mobile phone.

Suzie wants to meet up with some friends after work. She presses the dinner button and selects her first friends list. Two minutes later the screen shows that Rick, Steve, Pamela and Carol have accepted her invitation. She is at the same time told that the preference of three of these four is for Italian food and a list of nearby Italian restaurants is displayed, one with an offer of free desserts for a party of five. With a click of a button a table is reserved for 6:30 and a message, with directions, is generated to all the invitees. On the way to the restaurant, knowing her friends' preferences, Suzie selects the initial glasses of wine — three Frascatis and two Chiantis. When she walks in to the restaurant, she is greeted by name and shown to a table where her four friends are waiting with their respective wines in front of them. And by the way, at the end of the dinner Suzie presses pay and the bill shows up on each person's cell phone split in 5 parts according to what was ordered! A simple acknowledgement by

each person settles the bill and as a compliment, a five-dollars off discount coupon to the nearby dance club flashes on each screen. Tomorrow, however, is a working day so each person drives home listening to the same music that played in the background in the restaurant, downloaded to his or her mobile phone.

Towards a Fresh Marketing/Business Model

By the mid 2000s, technological developments as well as market sophistication levels in the U.S. had advanced enough to where services portrayed in the scenarios for the consumer markets above — as well as comparable enterprise applications — could be launched successfully. To do so, the mobile telecom sector needs to think carefully and innovatively about the CLIP functionalities: how to take the ubiquitously available communication (C) devices, seamlessly and securely integrate location (L) and payment (P) functions by taking full advantage of device *and* network capabilities and then deliver compelling information (I) content that users would gladly pay premium prices for.

SUMMARY

Mobile data services are at a crossroads between past successful offerings, such as messaging, e-mail and downloadable services on the one hand, and newer, richer capabilities offered by evolving network and handset technologies supported by new types of service platforms. This approach, married to evolving Internet technologies, has the potential to revolutionize the future of data services, not just in the United States, but globally.

QUESTIONS FOR DISCUSSION

1. Why did mobile payment systems not have much success in the U.S. mobile markets until the mid 2000s? Offer some recommendations to create robust, secure and appealing mobile payment options for the U.S. market.
2. The chapter argues that the Internet itself needs to become "mobile friendly" for m-commerce applications to take off in a big way. Offer a set of recommendations to make the Internet more "mobile friendly."
3. How is the "fresh new approach" to mobile commerce discussed at the end of this chapter different from existing business models as well as international business models such as Japan's i-Mode?

REFERENCES

Bradley, S.P., & Sandoval, M. (2002, Spring). Case Study: NTT DoCoMo — The future of the wireless Internet? *Journal of Interactive Marketing, 16,* 80-96.

Jette, J. (2005). Ka-Ching! Mobile commerce gets closer, *HBS Working Knowledge*. Retrieved from http://hbswk.hbs.edu/pubitem.jhtml?id=4631&t=special_reports_convergence2005

Lennon, M. M., & Dholakia, N. (2004, May). *Pay-for-play: The Japanese way to m-commerce success.* Paper presented at the 2004 IRMA: Information Resources Management Association International Conference, New Orleans.

Pickering, M. (2003). Vayusa turns cell phones into store discount cards. *INDIA New England Online*. Retrieved November 1, 2005, from http://www.indianewengland.com/media/paper549/

news/2003/11/01/Business/Vayusa.Turns.Cell.Phones.Into.Store.Discount.Cards-537154.shtml

Rask, M., Dholakia, N., & Dholakia, R. R. (2004). Configuring m-commerce portals for business success. In Nan Si Shi (Ed.), *Wireless communications and mobile commerce* (pp. 76-94). Hershey, PA: Idea Group Publishing.

Samuelsson, M., & Dholakia, N. (2004). Assessing the market potential of network-enabled 3G m-business services. In Nan Si Shi (Ed.), *Wireless communications and mobile commerce* (pp. 23-48). Hershey, PA: Idea Group Publishing.

Scott, A. O. (2005, Aug 7). Post-popism. *New York Times Magazine*, Section 6, 11-X.

ENDNOTES

[1] See also http://www.msnbc.msn.com/id/8348206/

This work was previously published in M-Commerce: Global Experiences and Perspectives, edited by N. Dholakia, M. Rask, and R. Dholakia, pp. 240-258, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.5
M-Learning with Mobile Phones

Simon So
Hong Kong Institute of Education, Hong Kong

INTRODUCTION

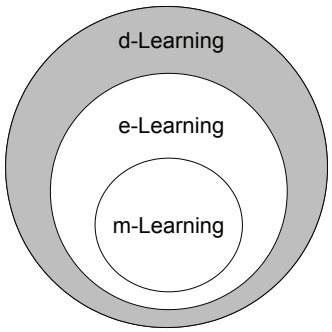
The Internet is a major driver of e-learning advancement and there was an estimate of over 1000 million Internet users in 2004. The ownership of mobile devices is even more astonishing. ITU (2006) reported that 77% of the population in developed countries are mobile subscribers. The emergence of mobile, wireless and satellite technologies is impacting our daily life and our learning. New Internet technologies are being used to support small-screen mobile and wireless devices. In a field marked by such rapid evolution, we cannot assume that the Web as we know it today will remain the primary conduit for Internet-based learning (Bowles, 2004, p.12). Mobile and wireless technologies will play a pivotal role in learning. This new field is commonly known as mobile learning (m-learning).

In this article, the context of m-learning in relation to e-learning and d-learning is presented. Because of the great importance in Web-based technologies to bridge over mobile and wireless technologies, the infrastructure to support m-learning through browser-based technologies is described. This concept represents my own view on the future direction of m-learning. An m-learning experiment, which implemented the concept, is then presented.

BACKGROUND

Many researchers and educators view that m-learning is the descendant of e-learning and originates from d-learning (Wikipedia M-Learning, 2006; Georgiev, Georgieva, & Smrikarov, 2004). The m-learning space is subsumed in the e-learning space and, in turn, in the d-learning space, as shown in Figure 1. This may be true chronologically. D-learning has more than hundred years of evolution starting from the printed media of correspondence (signified by carefully designed and produced materials by specialists to support the absence of instructors and independent study [Charles Wedemeyer] and the industrialization of teaching [Otto Peters]), to mass and broadcast media (marked by the open-

Figure 1. M-learning space as part of e-learning and d-learning spaces



ing of British Open University in 1961 [Daniel, 2001]), and to the telecommunication technologies supporting asynchronous and synchronous learning through teleconferencing, computer mediated communication and online interactive environments for students to create and re-create knowledge individually or collaboratively. In d-learning, the teacher and students are separated quasi-permanently by time, location, or both (Keegan, 2002; ASTD Glossary, 2006). With the advent of computer and communication technologies, e-learning covers a wide set of applications and processes, such as Web-based learning, computer-based learning, virtual classrooms, and digital collaboration (ASTD Glossary, 2006). The delivery of content is through a media-rich and hyperlinked environment utilizing internetworking services. M-learning can be considered as learning taking place where the learner is not at a fixed, predetermined location, or where the dominant technologies are handheld devices such as mobile phones, PDAs and palmtops, or tablet PCs. It can be spontaneous, personal, informal, contextual, portable, ubiquitous and pervasive (Kukulska-Hulme & Traxler, 2005, p. 2).

In my view, new concepts in teaching and learning can be generated from m-learning. For example, mobile phones can be used as voting devices for outdoor learning activities or in classrooms without computer supports, as interactive devices in museums, positioning or data logging devices at field trips or in many pedagogical situations. The justification of m-learning being descendent of e-learning and d-learning is rather thin, and Figure 2 is better represented. Furthermore, not everything can be delivered through m-learning. The small form factor, one-finger operation in some cases—slow computational and communication speed, short battery life and limited multimedia capabilities in contrast with computers do not really suit applications requiring heavy reading, high over-the-air communication and a lot of typing or texting.

In summary, m-learning is restricted and expedited by its nature. Different teaching and learning applications require different approaches, whether it is in d-learning, e-learning or m-learning. We must keep in mind their salient characteristics in different teaching and learning contexts, as shown in Table 1.

M-LEARNING INFRASTRUCTURE

In order to support m-learning, mobile devices such as PDAs, mobile phones and tablet PCs, together with servers such as Web servers, streaming servers and database servers on top of applications such as specific adaptation of LMS must

Figure 2. Overlapping and differential spaces of m-learning, e-learning and d-learning

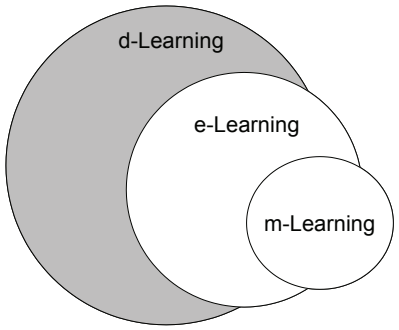
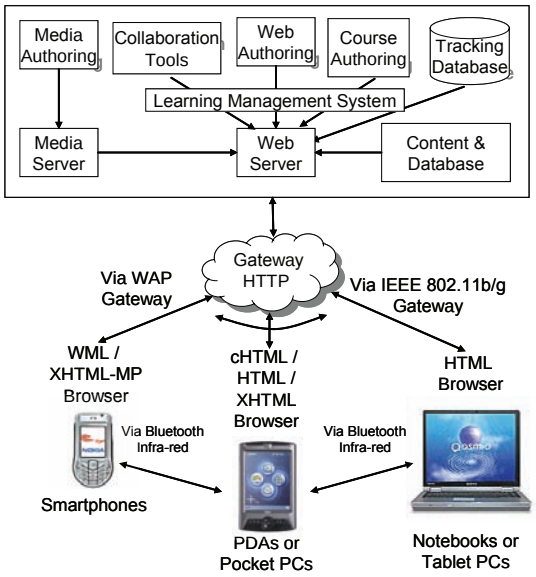


Table 1. Different teaching and learning contexts

	Salient characteristics
d-learning	• Separation of teachers and learners • Learning normally occurs in a different place from teaching • Formal educational influence and organization
e-learning	• Multimedia-rich • Hypermedia • Independent • Collaborative
m-learning	• Mobile • Portable • Ubiquitous • Pervasive

Figure 3. Browser-based support for m-learning



be employed (Horton & Horton, 2003; Chen & Kinshuk, 2005). Despite the rapid development in mobile technologies, Figure 3 provides a typical browser-based architecture to support m-learning. It represents a full-scale implementation of any learning system formally. Processing and logic are controlled from the server-side and the mobile devices act as interfaces (Hodges, Bories, & Mandel, 2004, p. 2).

It is also possible that the learning applications are run locally on mobile devices with or without accessing network resources. Applications can be built using Java, such as mobile information device profile (MIDP), C++ on Symbian or native OSs, and Adobe Flash for mobile devices. Feature-rich applications can be implemented to take advantage or avoid limitations of the hardware.

Many researchers believe that, in order to support m-learning, a mobile learning management system (mLMS) is necessary. The logical derivation of mLMS is through the extension of conventional LMS (Trifonova & Ronchetti, 2003; Trifonova, Knapp, Ronchetti, & Gamper, 2004). Direct presentation of materials from computers to mobile devices is likely not legible, aesthetically pleasant, or technically not feasible. Adaptation according to the hardware and device profiles is required. This view is also supported by Goh & Kinshuk (2004). CSS, XSLT and XSL transformation in XML technologies are used to support WML, XHTML and HTML through server pages (Shotsberger & Vetter, 2002). Open standards, including e-learning standards such as SCORM (Fallon & Brown 2003), are the keys for the success of any mLMS.

M-LEARNING WITH MOBILE PHONES

To illustrate the concept discussed above, an m-learning experiment using phone simulators with one of my classes in a computer lab was conducted, as shown in Figure 4. The purpose of the experiment is to find out how my students react to the concept of m-learning. Three activities were developed to address different

Figure 4. An m-learning experiment using phone simulators

Figure 5. The system architecture for the m-learning experiment

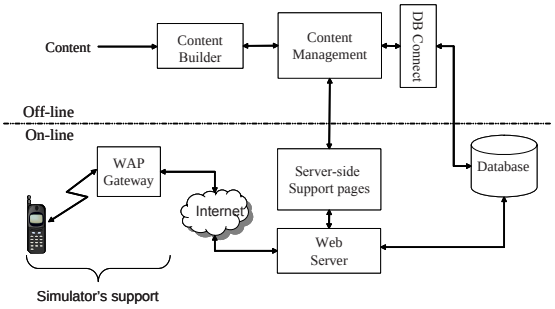


Figure 6. Voting activity

Figure 7. The corresponding voting result

applications of mobile phones for teaching and learning. Simulators developed to execute in real mobile phones are used for this study (Openwave, 2006). There are three reasons for this. Firstly, the chosen software has been implemented in a number of real phone models. It behaves like a real phone. Secondly, some students may not have mobile phones with advanced features to support WAP 2.0 (Wapforum, 2006) and XTHML-MP, or connect to the mobile service providers with the features turned on. Some students may still have text-based mobile phones! Thirdly, as long as students operate the simulator (e.g., one-finger operation) as the experiment intended, I have a much better controlled environment to answer my research questions.

To support this experiment, a WAP gateway connected to a Web server is needed. Figure 5 outlines a practical and partial implementation of the architecture described in the previous section. Apache, PHP and MySQL are chosen as the Web server, server-side programming and database support respectively.

Among the three applications developed for this experiment, the first application is a voting system. Students can cast their votes on their simulators and teachers can interactively check the voting results as illustrated in Figure 6 and Figure 7. Students can use the quick access keys ("1" to "X") on the keypad to cast their votes. This acts as if the voter has a simple voting machine at hand. Teachers can retrieve the voting results from the database onto their handsets as well.

The second application is an interactive game called "15/16" which is a popular game on Hong Kong's television. Instead of two players per game,

it was modified that the whole class can participate in each game. Students make their selections and the teacher (or any student) suggests the explanation. Students can change their mind depending on whether they believe the teacher/students or not. Figure 8 illustrates two questions. Teachers can show or refresh the selections at anytime. Figure 9 shows the students' selections for Question #1 in Figure 8.

The third application is a system to administrate tests. Students attempt the questions stored in the database. The overall score can be sent to the students at the end of the test, as shown in Figure 10 and Figure 11. The scores are kept in the database as well.

The applications described above are currently being rewritten in English and implemented for Nokia handsets. Figure 12 provides some of the snapshots.

CONCLUSION

M-learning has attracted a lot of research interest recently. It is a fashionable term in education. We can expect a lot of research work in this area will emerge for years to come. It is an exciting

Figure 8. Two questions for the game

Figure 9. Students' selections

Figure 10. A test on mobile phones

(login) (subject selection) (questions)

Figure 11. The score

(student's score)

Figure 12. Nokia's handset implementation corresponding to Figures 8 and 9

field. It also poses a lot of challenges to educators, instructional designs, software engineers and network specialists.

The main concept of m-learning has been highlighted in the article. The browser-based approach to m-learning is presented. It is illustrated by the experiment conducted with my students. This article serves as an example for those researchers to pursue further studies in this direction.

REFERENCES

ASTD Glossary. (2006). *ASTD's source for e-learning.* Retrieved on June 30, 2006, from http://www.learningcircuits.org/glossary.html

Bowles, M. S. (2004). Relearning to e-learn: Strategies for electronic learning and knowledge. Melbourne: Melbourne University Press.

Chen, J., & Kinshuk. (2005). Mobile technologies in educational services. *AACE Journal of Educational Multimedia and Hypermedia*, *14*(1), 89-107

Daniel, J. (2001). The UK Open University: Managing success and leading change in a mega-university. In C. Latchem & D. Hanna (Eds.), *Leadership for 21st Century: Global Perspectives from Educational Innovators.* London: Kogan Page.

Fallon C., & Brown S. (2003). *E-learning standards: A guide to purchasing, developing, and deploying standards-conformant e-learning.* FL.: St. Lucie Press

Georgiev, T., Georgieva, E., & Smrikarov, A. (2004). M-learning: A new stage of e-learning. In *Proceedings of International Conference on Computer Systems and Technologies, CompSys-Tech'2004.* Retrieved on June 30, 2006, from http://ecet.ecs.ru.acad.bg/cst04/Docs/sIB/428.pdf

Goh, T., & Kinshuk. (2004). Getting ready for mobile learning. In *Proceedings of the 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA2004)* Lugano, Switzerland (pp.56-63).

Hodges, A., Bories, J., & Mandel, R. (2004). Designing applications for 3G mobile devices. In R. Longoria (Ed.), *Designing Software for the Mobile Context: A Practitioner's Guide*. London: Springer.

Horton, W., & Horton, K. (2003). *E-learning tools and technologies: A consumer's guide for trainers, teachers, educators, and instructional designers*. New York: Wiley.

ITU. (2006). Executive summary. In *World Telecommunication/ICT Development Report 2006: Measuring ICT for Social and Economic Development.* Retrieved on June 30, 2006, from http://www.itu.int/ITU-D/ict/publications/wtdr_06/material/WTDR2006_Sum_e.pdf

Keegan, D. (2002*). The future of learning: From eLearning to mLearning.* Retrieved on June 30,

2006, from http://learning.ericsson.net/mlearning2/project_one/book.html

Kukulska-Hulme, A., & Traxler, J. (2005). *Mobile learning: A handbook for educators and trainers.* London: Routledge.

Openwave. (2006). *V7 simulator.* Retrieved on April 15, 2006, from http://www.openwave.com

Shotsberger, P., & Vetter, R. (2002). The handheld Web: How mobile wireless technologies will change Web-based instruction and training. In Allison Rossett (Ed.), *The ASTD E-Learning Handbook: Best Practices, Strategies, and Case Studies for Emerging Field*. New York: McGraw-Hill

Trifonova, A., & Ronchetti, M. (2003). A general architecture for m-learning. In *Proceedings of the Second International Conference on Multimedia and ICTs in Education*. Badajoz, Spain.

Trifonova, A., Knapp, J., Ronchetti, M., & Gamper, J. (2004). Mobile ELDIT: Transition from an e-Learning to an m-Learning. In *Proceedings of the 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA2004)*, Lugano, Switzerland (pp.188-193).

Wapforum. (2006). *WAP 2.0 Standard.* Retrieved on June 30, 2006, from http://www.wapforum.org

Wikipedia M-learning. (2006). *M-learning.* Retrieved on June 30, 2006, from http://en.wikipedia.org/wiki/M-learning

KEY TERMS

Distance Learning (D-Learning): The teacher and students are separated quasi-permanently by time, location, or both. The content can be delivered synchronously or asynchronously.

Electronic Learning (E-Learning): Processes of learning through Web-based learning, online learning, computer-based learning and/or virtual classrooms. The delivery of content is through media-rich and hyperlinked environment utilizing internetworking services.

Learning Management System (LMS): LMS allows the tracking of learner's needs and achievement over periods of time.

Mobile Learning (M-Learning): Learning takes place where the learner is not at a fixed, predetermined location, or where the dominant technologies are handheld devices such as mobile phones, PDAs and palmtops, or tablet PCs.

Wireless Application Protocol 2.0 (WAP 2.0): This is a new version of WAP, which facilitates a cut-down version of XHTML and makes it work in mobile devices.

XHTML Mobile Profile (XHTML-MP): XHTML-MP is a markup language for mobile phones and the like.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 419-423, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.6
Using Mobile Communication Technology in Student Mentoring

Jonna Häkkilä
University of Oulu, Finland

Jenine Beekhuyzen
Griffith University, Australia

INTRODUCTION

Information technology (IT), computer science, and other related disciplines have become significant both in society and within the field of education. Resulting from the last decades' considerable developments towards a global information society, the demand for a qualified IT workforce has increased. The integration of information technology into the different sectors of every day life is increasing the need for large numbers of IT professionals. Additionally, the need for nearly all workers to have general computing skills suggests possibilities for an individual to face inequality or suffer from displacement in modern society if they lack these skills, further contributing to the digital divide. Thus, the importance of IT education has a greater importance than ever for the whole of society.

Despite the advances and mass adoption of new technologies, IT and computing education continually suffers from low participant numbers, and high dropout and transfer rates. This problem has been somewhat addressed by introducing mentoring programs (von Hellens, Nielsen, Doyle, & Greenhill, 1999) where a student is given a support person, a mentor, who has a similar education background but has graduated and is employed in industry. Although the majority of these programs have been considered successful, it is important to note that it is difficult to easily measure success in this context.

In this article, we introduce a novel approach to mentoring which was adopted as part of an ongoing, traditional-type mentoring program in a large Australian university. The approach involved introducing modern communications technology, specifically mobile phones having

an integrated camera and the capability to make use of multimedia messaging services (MMS). As mobile phones have become an integrated part of our everyday life (with high adoption rates) and are an especially common media of communication among young people, it was expected that the use of the phones could be easily employed to the mentoring program (phones were provided for the participants). Short message service (SMS), for example text messaging, has become a frequently used communication channel (Grinter & Eldridge 2003). In addition to text, photo sharing has also quickly taken off with MMS capable mobile phones becoming more widespread. The ability to exchange photos increases the feeling of presence (Counts & Fellheimer, 2004), and the possibility to send multimedia messages with mobile phones has created a new form of interactive storytelling (Kurvinen, 2003). Cole and Stanton (2003) found the pictorial information exchange as a potential tool for children's collaboration during their activities in story telling, adventure gaming and for field trip tasks.

Encouraged by these experiences, we introduced mobile mentoring as part of a traditional mentoring program, and present the experiences. It is hoped that these experiences can affirm the legitimacy of phone mentoring as a credible approach to mentoring. The positive and negative experiences presented in this article can help to shape the development of future phone mentoring programs.

BACKGROUND

Current education programs relating to information technology continue to suffer from low applicant numbers in relation to the available enrollment positions. In the USA alone, the number of computer science graduates dropped from a high of 50,000 in 1986 to 36,000 in 1994, reported by the Office of Technology Policy in 1998 (von Hellens et al., 1999). Many general IT degrees also have high dropout rates, particularly in the transition from the first to second year of undergraduate studies. Student statistics also show that university IT degree programs are not attracting the high achieving students, some possible reasons include the low entrance level scores needed to enter the program, the attraction to high-entrance level degree programs such as medicine, law, and psychology and the confusion and uncertainty relating to what a career in IT will entail (ASTEC, 1995).

Misconceptions associated with understanding IT as a field specialized for those with masculine attributes exist and are reinforced by the teachings at secondary school level (Beekhuyzen & Clayton, 2004; Greenhill, von Hellens, Nielsen, & Pringle, 1997), thus often having a negative effect on students, particularly on females. Consistent results have been obtained in studies concerning high school physics, which faces similar difficulties and biased ideas as IT (Häkkilä, Kärkäs, Aksela, Sunnari, & Kylli, 1998). A remarkable number of university students choose their area of study without any preliminary experience in the particular field. With information technology, the students also often have unclear or distorted perceptions of what to expect later in their studies or after graduation, including what kind of employment their area of study can offer (Nielsen, von Hellens, Pringle, & Greenhill, 1999).

Within the IT context, university student mentoring has been introduced to offer students insight into the industry and to employment possibilities enabling them to have them a closer look at the everyday life of working in the field. The aim is to dispel some of the misconceptions associated with what IT work is all about. When entering into this mentoring program, the student is matched with a personal mentor who has a similar educational background and is currently employed in the IT industry. Conventionally, mentoring is carried out with face-to-face meetings, e-mail and telephone conversations between mentor and mentee. In line with many published

studies, early results from our studies suggest that mentoring can provide valuable information on career possibilities, thus increasing the motivation of study and working in the area. It also clarifies and enhances student perceptions concerning the realities of the field. Note: all participation in the program is of a voluntary basis, and no financial benefits are obtained.

When commencing the traditional part of the mentoring program, mentors and mentees participate in an initial short training session. In this session, the mentoring partners are introduced, and the role and expectations of mentors and mentees is discussed. Mentor and mentee generally meet thereafter on a regular basis during one semester period (usually 13-15 weeks) which is arranged as suits best for both parties. Face-to-face communication is also usually complimented by e-mail conversations. A mid-program event is organized by the Alumni Association, usually with a presentation by an industry representative on a pertinent topic such as networking (in terms of meeting people, making contacts, etc.—a skill particularly useful within the IT industry). A final session is held to close the program and gather together all program participants to discuss their experiences.

ENHANCING COMMUNICATION WITH MOBILE TECHNOLOGY

In addition to the traditional mentoring methods being employed by the mentoring program in the university, we have introduced the use of mobile communication technology into the mentoring program. The primary aim in introducing the novel approach was to augment communication during the mentoring process. There was no aim to replace the conventional communication mediums but to add value with features offered by the mobile communication device. A pilot study was conducted in 2003. Due to positive feedback, the approach has continued to be integrated in the traditional program in 2004.

The equipment used in the experiment consists of two Nokia 7650 Mobile Phones, of which one was given to the mentee and one to the mentor for the duration of the program. The mentor was advised to communicate with the student about all which (s)he felt was a relevant part of their work and leisure, and especially to use picture messaging as an illustrative supplement. Using this type of technology to communicate brings about many issues relating to human-computer interaction. For example, the size of the screen, the structure of the information being viewed/sent (Chae & Kim, 2003), and the increasing complexity of functionality can lead to ineffective use of the mobile device. However, benefits include the ability to access information from anywhere without the need to physically sit at a computer workstation (Chae & Kim, 2003).

The student mentee was given a certain monetary amount (AUD $15—Australian dollars) of pre-paid credit on the mobile phone, which they were allowed to use during the study. For example, the price for sending an SMS message and MMS message were 0.20 AUD (20 cents) and 0.75 AUD (75 cents), respectively. The phone mentoring period lasted for one week for one mentor-mentee pair. In the beginning of the experiment period, the functions of the phone were explored together to ensure seamless communication. At the completion of the one-week period, the participants gave their feedback about the experience via a questionnaire.

The media used in communications between the mentor and mentee were short messages (SMS) and multimedia messages (MMS), the latter to be more common. Conversations consisted mainly of one message or a message and a reply, where the reply included feedback or comment to the previous messages. The typical number of sent messages was two per day from mentor to mentee, and one from mentee to the mentor, although more

messages were exchanged if a message gave rise to a longer, more detailed conversation. The time for messaging was found to be varied from morning hours to past midnight and also sometimes during the weekend, as shown in Figures 1 and 2.

The majority of messages sent contained a short description of the work task the mentor was currently involved in, accompanied by a picture. In addition to the actual task, the messages often described the atmosphere at that particular moment and also included short opinions (see Figure 1). Some of the messages were not primarily related to the work tasks, but described more the mentor's personal interests—free time, hobbies, and personal preferences.

The initiative for conversations containing professional information was taken by the mentor, and was not motivated by, for example, a question from a mentee. However, mentees took initiative in reporting about their duties related to studying. For instance, assignments and projects they were working on. Conversations relating to free time were initiated equally by both parties. Examples of messages relating to free time are illustrated in Figure 2.

FEEDBACK

The results obtained from both sets of participating parties were positive and encouraging. The most positive aspects reported by the mentees were on increasing the frequency of the communications and thus developing a closer relationship and gaining a deeper insight for the mentor's work. Comments collected from two students at the end of the mobile mentoring period are presented in the following:

Figure 1. Two multimedia messages describing the work tasks of a mentor.

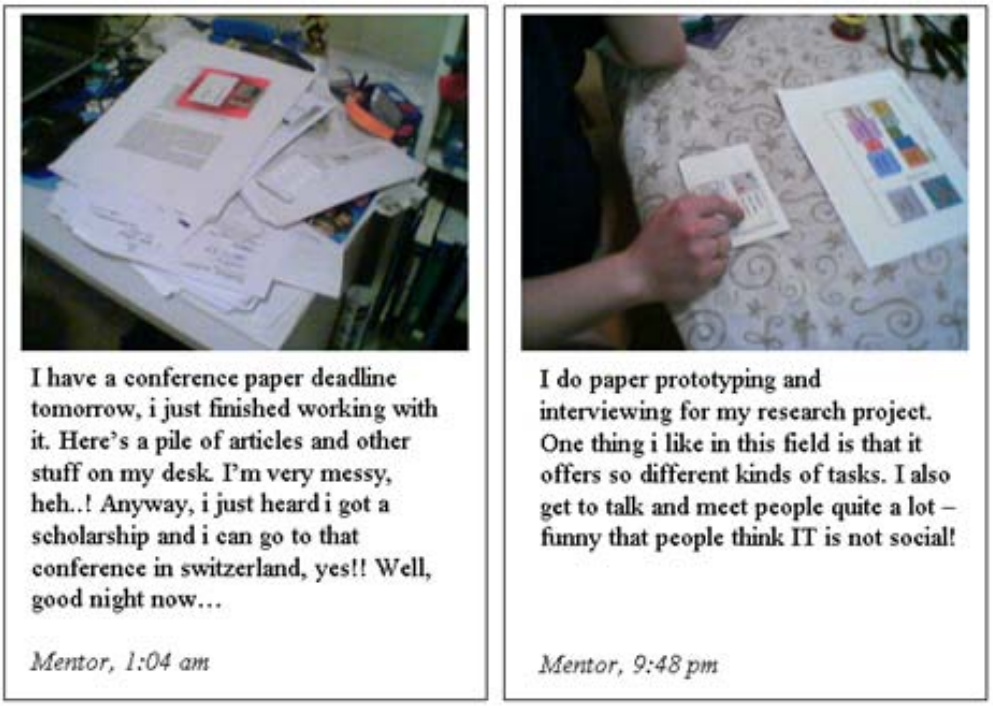
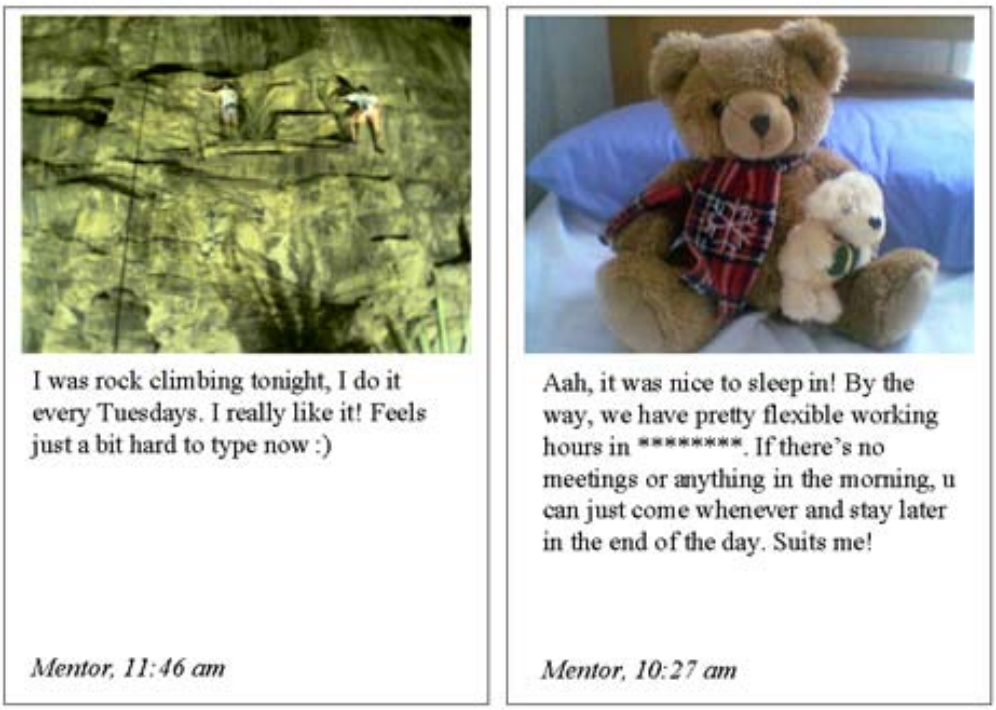


Figure 2. Two free time orientated multimedia messages from a mentor (Company name of the employer replaced with asterisk)



- **Mentee #1:** "I believe that mobile communication is quick and easy. It gives you an opportunity to learn more about your mentor and what they do and vice versa. It is especially good when both parties are unable to meet on a regular basis due to time constraints, commitments, etc."
- **Mentee #2:** "The best thing about the phone mentoring was that I was able to see how another person, in the field I want to work in, interacts with their life as well as being able to share aspects of my life with my mentor. It helped break the ice, enabling my mentor and myself to get to know each other."

Positive feedback obtained from mentors particularly concerned the flexibility in regard to the place and time of the communication, ease of use, and the extra personal touch it gave to the conversations. The amount of credit was reported to be sufficient for a one-week experimental period. Overall, the integration of mobile technology was suggested to offer a valuable tool for the mentoring program.

Other reported positive aspects of this novel approach include:

- Easy way to communicate
- Minimal effort required
- You can do it at any time and you don't miss the person
- (The phone is) very popular with people, so it is an advantage to use it with mentoring,
- Quicker, more efficient, because people have their phone on them more than they check their emails
- Teaches responsibility, taking care of the phone

- More informative, a picture can say a thousands words.

As a weakness, mentees referred to the short length of the experimental period, and were suggesting it to be elongated from one to, for example, two weeks. This was argued by explaining that it would offer a longer period of time to get used to both using the technology (the MMS phone) and the mode of communication. A wish for a longer lasting experimental period was also mentioned by mentors, as one week was not considered to be a long enough time to cover the different aspects related to the diverse work in IT. The suggestion of a longer experimental period is also supported when examining the messages, as the communication become relaxed towards the end of the week. This feedback from participants in 2003 was integrated into the program run in 2004, as longer experimental periods (1 week/10 working days) are used.

The technical barriers noted were battery time/length and lack of network coverage in some areas, which were described to limit the communication in some instances. These, in addition to small screen size, low bandwidths, limited storage and cumbersome input facilities are common barriers that have been presented in the literature (Chae & Kim, 2003; Tarasewich, Nickerson, & Warke, 2001). However, the participants' perception was that the technical barriers did not have significant impact to the overall experiment. A criticism from mentors was that if the use of mobile communications would be the only medium of interaction in mentoring, the conversations between mentor and a mentee would remain too light and no deep knowledge or "big picture" would be obtained from the short communications. For instance, the following comment was obtained from a mentor when asked the weaknesses of mobile mentoring:

- **Mentor:** "The nature of conversations differs a great deal in comparison to ones had in face-to-face meetings and e-mail exchange, where the discussion is held for longer. However, I highly recommend this system as an additional part of communication, as it offers a possibility for more intense and frequent interaction and highlights the aspects which otherwise hardly were considered, e.g., the work environment, task descriptions and the time schedule of the day."

FUTURE TRENDS

When the mobile mentoring program first began in early 2003, the number of multimedia messaging capable phones was minimal. It is expected that when they become more common, mobile mentoring can be adopted on a larger scale, and it may come a natural part of the interaction process. However, as a starting point, it was important to lend out the MMS capable phones and get people actively involved in the process.

In 2004, yet another novel approach to mentoring was added to the ongoing Alumni Association mentoring program in the form of international mentoring. In addition to a local mentor to communicate with (either traditionally and/or phone), volunteering student mentees were given contact persons working abroad as mentors. The mentoring contacts were obtained through the university department's connections to the international IT industry. The communication employs mainly e-mail, but also additional mobile messaging techniques. Due to difficulties in connectivity between mobile phone operators, the MMS between two phones was found inoperative, thus picture messages were exchanged by sending a MMS from a phone to e-mail. Initial results of this additional experiment will be reported.

CONCLUSION

This article introduces how mobile communication technology has been embedded into a university student mentoring program which was held among first-year information technology students within an Australian university. The study was implemented by giving mentor-mentee pairs mobile phones with MMS functionality. Participants communicated with each other over a one-week period. Participants were advised to incorporate visual information into the communication by a form of multimedia messaging.

Although effective in many situations, mentoring can be rather unproductive and thus unsuccessful for many reasons. One common reason for failure is a lack of structure. Many communications between mentor and mentee are adhoc and generally unplanned which can and often does result in long periods between communications. Lack of structure can also distort perceptions of outcomes and results, with no clear aim being achieved.

The results show that integrating mobile communication into the mentoring process has provided added value to the traditional program. Participants suggest that it enhances the mentoring experience and that it can be regarded as a valuable tool in communications between the mentor and mentee. Positive aspects of the program were identified as increased frequency and flexibility in communication, which are highly valued because of the time constraints of both mentor and mentee. Mentors emphasised also the easy access and speed of use, as sending a message with mobile phone was regarded as more easy and flexible than e-mail, which took more time and was limited to the work situations and a computer. Both parties reported on the development of a deeper personal relationship and relaxed communication between mentor and mentee over time. Including visual information to the communication in a form of MMS, new aspects of both mentor's work and her/his lifestyle were highlighted.

Generally, mobile communication technology was found to offer a valuable tool for a mentoring program as a supporting tool of communication, even though some weaknesses were identified. The results have encouraged the authors to continue the integration of mobile communication into mentoring to enhance the information exchanged between mentor and mentee. Continuing and future research in this area includes the continuation of the study using both MMS and conventional styles, with improvements according to feedback received from this pilot phase concerning issues such as the time period devoted to the experiment. The aim is to increase the amount of participants from a relatively small sample to larger numbers of mentee-mentor pairs and to attempt to better measure the benefits of the program.

REFERENCES

ASTEC. (1995, September). *The science and engineering base for development of Australia's information technology and communication sector.* A discussion paper by Australian Science and Technology Council (ASTEC).

Beekhuyzen, J., & Clayton, K. (2004, July 28-30). ICT career perceptions: Not so geeky?!?. In the *Proceedings of the 1st International Conference on Research on Women in IT*, Kuala Lumpur.

Chae, M., & Kim, J. (2003, December). What's so different about the mobile Internet? *Communication of the ACM, 46*(12), 240-247.

Cole, H., & Stanton, D. (2003). Designing mobile technologies to support co-present collaboration. *Personal and Ubiquitous Computing, 7,* 365-371.

Counts, S., & Fellheimer, E. (2004). Supporting social presence through lightweight photo sharing on and off the desktop. In the *Proceedings of the Conference on Computers and Human Interaction (CHI)* (pp. 599-606).

Greenhill, A., von Hellens, L., Nielsen, S., & Pringle, R. (1997, May). Australian women in IT education: Multiple meanings and multiculturalism. In the *Proceedings of the 6th IFIP Conference on Women, Work and Computerization*, Bonn, Germany (pp. 387-397).

Grinter, R. E., & Eldridge, M. (2003). Wan2tlk? Everyday text messaging. In the *Proceedings of the Conference on Computers and Human Interaction (CHI)* (pp. 441-448).

Häkkilä, J., Kärkäs, M., Aksela, H., Sunnari, V., & Kylli, T. (1998). *Tytot, pojat ja fysiikka. Lukiolaisten käsityksiä fysiikasta oppiaineena.* University of Oulu, Finland. Abstract in English: Girls, Boys and Physics.

Kurvinen, E. (2003). Only when Miss Universe snatches me: Teasing in MMS messaging. In the *Proceedings of the Conference on Designing Pleasurable Products and Interfaces* (pp. 98-102). ACM Press.

Nielsen, S. H., von Hellens, L., Pringle, R., & Greenhill, A. (1999). Students' perceptions of information technology careers. Conceptualising the influence of cultural and gender factors for IT education. *GATES (Greater Access to Technology Education and Science) Journal, 5*(1), 30-38.

Tarasewich, P., Nickerson, R. C., & Warke, M. (2001, August). Wireless/mobile e-commerce: Technologies, applications, and issues. In the *Proceedings of the Seventh Americas Conference on Information Systems (AMCIS)*, Boston.

von Hellens, L., Nielsen, S., Doyle, R., & Greenhill, A. (1999, December). Bridging the IT skills gap. A strategy to improve the recruitment and success of IT students. In the *Proceedings of the 10th Australasian Conference on Information Systems,* Wellington, New Zealand (pp. 1129-1143).

KEY TERMS

Mentee: Participant of the mentoring program; student or equivalent; being "advised."

Mentor: Participant of mentoring program; the "advisor."

Mentoring Program: A process where a mentee is given a personal guide, a mentor, who has professional or otherwise advanced experience and can advise the mentee on the specifics about the particular field of study and work in the industry.

Mobile Communication Technology: A medium to communicate via mobile devices.

Mobile Mentoring: Mentoring which uses mobile communication technology as an integrated part of the communication between mentor and mentee.

Multimedia Messaging Service (MMS): A form of mobile communication, where each message can contain picture, audio, video, and text material with certain data size limitations. A multimedia message is typically sent from one camera phone to another.

Short Message Service (SMS): A form of mobile communication, where mobile phone user is able to send and receive text messages typically limited to 160 characters. A short message is typically sent from one mobile phone to another.

This work was previously published in Encyclopedia of Human Computer Interaction, edited by C. Ghaoui, pp. 680-685, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.7
A Mobile Portal for Academe

Hans Lehmann
Victoria University of Wellington, New Zealand

Stefan Berger
Detecon International GmbH, Germany

Ulrich Remus
University of Erlangen-Nuremberg, Germany

INTRODUCTION

Today, many working environments and industries are considered as knowledge-intensive, that is, consulting, software, pharmaceutics, financial services, and so forth, and the share of knowledge work has risen continuously during the last decades (Wolff, 2005). Knowledge management (KM) has been introduced to overcome some of the problems knowledge workers are faced when handling knowledge, that is, the problems of storing, organizing, and distributing large amounts of knowledge and its corresponding problem of information overload and so forth (Maier, 2004).

At the same time, more and more people leave (or have to leave) their fixed working environment in order to conduct their work at changing locations or while they are on the move. Mobile business tries to address these issues by providing (mobile) information and communication technologies (ICTs) to support mobile business processes (e.g., Adam, Chikova, & Hofer, 2005; Barnes, 2003; Lehmann, Jurgen Kuhn, & Lehner, 2004,). However, compared to desktop PCs, typical mobile ICT, like mobile devices such as PDAs and mobile phones, have some disadvantages, that is, limited memory and CPU, small displays and limited input capabilities, low bandwidth, and connection stability (Hansmann, Merk, Niklous, & Stober, 2001).

So far, most of the off-the-shelf knowledge management systems provide just simple access from mobile devices. As KMS are generally handling a huge amount of information (e.g., documents in various formats, multimedia content, etc.), the management of the restrictions described becomes even more crucial (Berger, 2004).

Based on requirements for mobile applications in KM, an example for the implementation of a mobile knowledge portal at a German university

is described. The presented solution offers various services for university staff (information access, colleague finder, campus navigator, collaboration support). With the help of this system, it is possible to provide users with KM services while being on the move. With its services, it creates awareness among remote working colleagues and hence, improves knowledge sharing within an organization.

MOBILE KNOWLEDGE MANAGEMENT

A mobile working environment differs in many ways from desk work and presents the business traveller with a unique set of difficulties (Perry, O'Hara, Sellen, Brown, & Harper, 2001). Throughout the last years, several studies have shown that mobile knowledge workers are confronted with problems that complicate the fulfilment of their job.

Mobile workers working separated from their colleagues often have no access to the resources they would have in their offices. Instead, business travellers, for example, have to rely on faxes and messenger services to receive materials from their offices (Schulte, 1999). In case of time-critical data, this way of communication with the home base is insufficient. In a survey about knowledge exchange within a design consulting team, Bellotti and Bly (1996) state that it is difficult for a mobile team to generally stay in touch. This is described as "lack of awareness." It means that a common background of common knowledge and shared understanding of current and past activities is missing. This constrains the exchange of knowledge in teams with mobile workers. In addition, mobile workers have to deal with different work settings, noise levels, and they have to coordinate their traveling. These "logistics of motion" lower their ability to deal with knowledge-intensive tasks (Sherry & Salvador, 2001) while on the move. The danger of an information overflow increases.

Mobile knowledge management is an approach to overcome these problems (e.g., Berger, 2004; Grimm, Tazari, & Balfanz, 2002,). Rather than adding to the discussion of what actually is managed by KM-knowledge workers, knowledge, or just information embedded into context–in this chapter, mobile KM is seen as KM focusing on the usage of mobile ICT in order to (Berger, 2004, p. 64):

- provide *mobile access* to knowledge management systems (KMS) and other information resources;
- generate *awareness* between mobile and stationary workers by linking them to each other; and
- realize *mobile KM services* that support knowledge workers in dealing with their tasks.

THE CASE OF A MOBILE PORTAL AT A GERMAN UNIVERSITY

In recent years, the German universities, which are financed to a large extent by public authorities (federal states and federal government), have been severely affected by public saving measures. As a result, lean, efficient administrative procedures are more important than ever. KM can help to achieve these objectives. One example is to provide easy access to expert directories, where staff members with certain skills, expertise, and responsibilities can be located (e.g., "Person X is responsible for third-party-funding") in order to support communication and collaboration.

However, there are several reasons why the access to information of this type is limited at the University of Regensburg. First, there is the hierarchical, but decentralized organizational structure. All together about 1,000 staff members are working in 12 different schools and about 15 research institutes at the university, serving for

about 16,000 students. As most of the organization units are highly independent, they have their own administrations, and the exchange of knowledge with the central administration is reduced to a minimum. Likewise there is hardly an exchange of knowledge between different schools and departments. As a result knowledge, which would be useful throughout the whole university, is limited to some staff members ("unlinked knowledge," Figure 1).

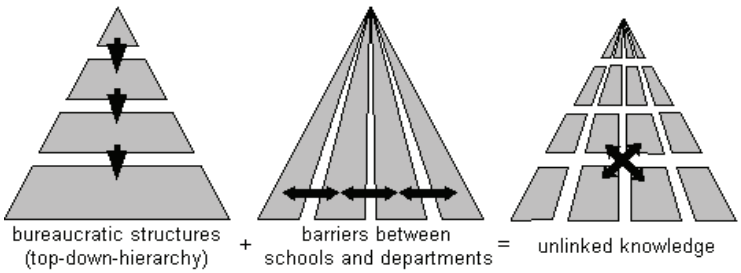
A second problem is that many scientific staff members work on the basis of (short-term) time contracts. This leads to an increasing annual labour turnover, comparable to the situation that consulting companies are facing. Important knowledge about past projects, courses, and scientific results is lost very easily. Due to this fact a high proportion of (new) staff members are relatively inexperienced to cope with administration processes that can be described as highly bureaucratic and cumbersome.

To solve some of these problems—the lack of communication between departments and the need to provide specific knowledge (i.e., administrative knowledge) for staff members—the University of Regensburg decided to build up a knowledge portal called U-Know (ubiquitous knowledge). U-Know is meant to be a single point of access for all relevant information according to the knowledge needs described.

The portal should support staff members by managing documented as well as tacit knowledge. A knowledge audit was conducted in order to get a better picture of knowledge demand and supply. This was mainly done with the help of questionnaires and workshops, where staff members were asked to assess what kind of (out of office) information is considered as useful. In order to support the exchange of tacit knowledge (which is hard to codify, due to the fact that this knowledge lies solely in the employees' heads, often embedded in work practices and processes), the considered KM solution should also enable communication and cooperation between staff members.

However, when conducting the knowledge audit, it became obvious that a large amount of knowledge is needed when knowledge workers are on the move, that is, working in a mobile work environment. Staff members are frequently commuting between offices, meeting rooms, laboratories, home offices, they visit conferences, and sometimes they are doing field studies (e.g., biologists or geographers). Hence the picture of one single resource-rich office has to be extended towards different working locations, where a large number of knowledge-intensive tasks are carried out as well. Consequently, the considered solution should meet these "ubiquitous" knowledge needs of current mobile work practices at a university, and should try to enhance the knowledge portal by mobile knowledge services in order to (see chapter "Mobile portals for knowledge management" in the same book):

Figure 1. Unlinked knowledge because of independent organization structures (Berger, 2004)



- support the social networking of knowledge workers and to create awareness (e.g., mobile access to employee yellow pages, skill directories, directories of communities, via e-mail, SMS, or chat);
- enable mobile access on various knowledge sources via different devices (e.g., knowledge about university organisation and processes, internal studies, proposals, and lessons learned);
- support location-oriented information delivery;
- support heterogeneous technologies and standards, for example, different devices, protocols, and networks;
- to provide proactive and adaptive information delivery (using mobile devices focusing on push services, profiling, personalization, and contextualization); and
- to use speech technology in order to simplify mobile access of knowledge portals.

In order to meet these requirements, U-Know offers KM services to support information, communication, collaboration, and search (Figure 2).

- **Information Services:** The first category comprises all services that are responsible to manage simple information in the knowledge base. By invoking these services, staff members obtain the information they need to perform their daily tasks, for example, news, notifications about changes in rooms or phone numbers. Very important are "*yellow pages*" (Figure 3), where all staff members are listed. This list can be browsed by names, departments, fields of research, and responsibilities, respectively.

Frequently asked questions (FAQs) try to give answers to questions that are typically asked by new staff members. The Campus Navigator helps locating places and finding your way around the campus. Each room at the university carries a

Figure 2. Features of U-Know (Berger, 2004)

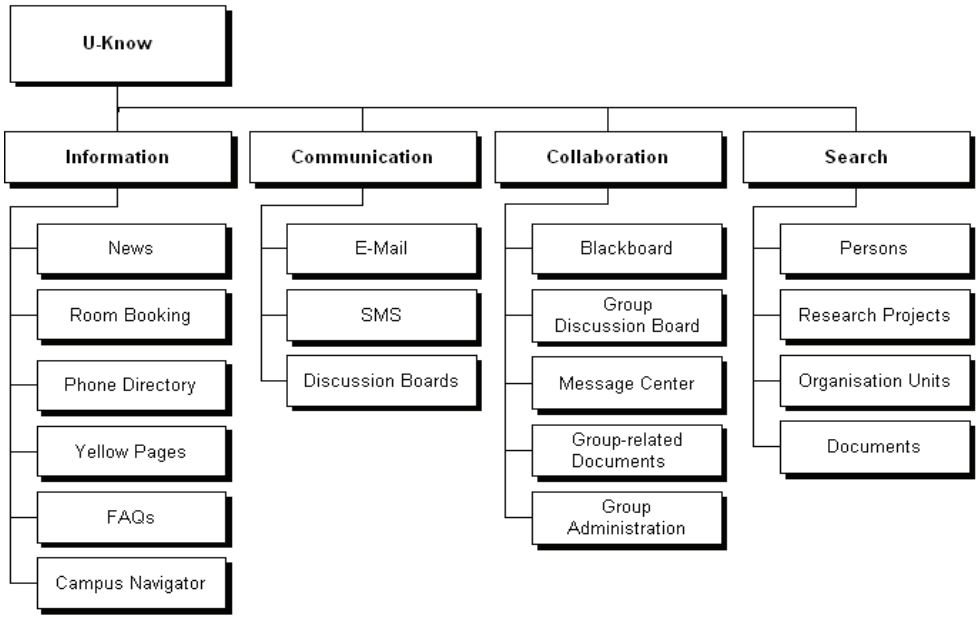
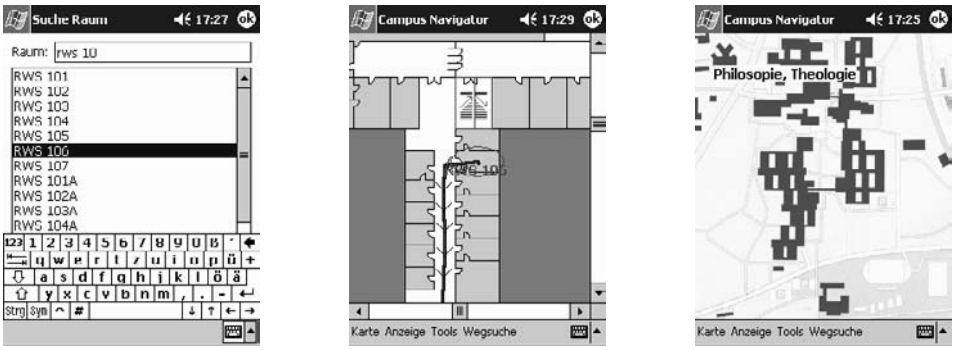


Figure 3. U-Know "Yellow Pages" (Berger, 2004)

Figure 4. U-Know campus navigator (Berger, 2004)



doorplate with a unique identifier. After entering a starting point in the form of the identifier and a destination in the form of the name of a person, of an office (e.g., "office for third-party-fundings," "academic exchange service") or just another room number, the shortest way to the destination is calculated and shown on maps of different sizes (Figure 4).

- **Communication Services:** Communication-oriented features like e-mail, short message service (SMS), and discussion boards are intended to support the exchange of tacit knowledge between staff members.
- **Collaboration Services:** To foster collaboration, for example, in temporary project groups, staff members can initiate workgroups by inviting colleagues via SMS or e-mail to join a virtual team space. After forming a workgroup, the participants can use their team space for (electronic) group discussions and sharing documents. The blackboard displays all recent events, including new group members, new files, discussion entries, and administrative actions that are taken.
- **Search Services:** In the search section, queries can be limited to persons, research projects, organization units, or documents.

To support different networks, there are several ways to access the portal. University staff can use the campus-wide WiFi-network with WiFi-capable devices, such as laptops. Users can also

deploy a mobile phone and access the portal via a GSM-network and the wireless application protocol (WAP). Hence it is possible to use the portal even when users are outside the university, at a conference for instance. The phone directory or the yellow pages can be accessed via voice, as the entry of longer words may be cumbersome in many situations. An integrated speech-recognition-system "translates" the user's spoken words into database requests and the results back into speech, respectively.

Different application scenarios are possible: Staff can use the system within the campus, for example, to get up-to-date information about the library, such as opening times or finding the appropriate book shelves. An SMS push service is implemented to inform staff and students about books that have to be picked up and returned. The integration of this kind of information service with personal information management of contacts, tasks, and dates by using PDAs or similar mobile devices will bring KM closer to the personal sphere. On the other side, staff and students may use the system outside the campus on their way to or back from university by participating in discussion boards and joining virtual team spaces. Here, they can retrieve news about lectures and seminars, discuss course related topics, and communicate with their peers while on the move.

CONCLUSION AND OUTLOOK

All in all, the implemented solution provides mobile access to a broad range of different knowledge sources in a mobile work environment. University staff can use the KM services provided by U-Know in order to access information, to find colleagues, to navigate the campus, to collaborate, and so forth. With its services like Yellow Pages, messaging features, and so forth, it creates awareness among remote working colleagues and hence, improves knowledge sharing within an organization. These KM services mainly support the human-oriented KM approach. In fact, typical knowledge services were adapted with regard to the characteristics of mobile devices, that is, small display, bandwidth, and so forth.

However, an adaptation of these services according to the user's location did not take place yet, whereas a customization of services according to the location of the user would enable a mobile knowledge portal to supply mobile knowledge workers with appropriate knowledge in a much more targeted way. At the same time, an information overload can be avoided, since only information relevant to the actual context and location is filtered and made available. Think of a researcher who is guided to books in a library according to his own references, but also according to his actual location. Location-orientation is the next consequent step in pushing mobile KM portals towards more comprehensive mobile KM solutions.

What are the experiences so far? The main users of U-Know are those who already own a mobile device, especially a PDA, in order to organize their appointments and contacts (personal information management). In contrast to staff members without this experience, this group perceives the additional KM-related services as an extension of the capabilities of their devices.

The WiFi-access within the university campus soon became the most popular way of accessing the system, mainly because of the free access for university members and the higher bandwidth (and therefore faster connections) of WiFi in comparison to a GSM-based access via WAP. However, decreasing connection fees and higher bandwidths of 3G-Networks (UMTS) would encourage staff to use the system from outside the university.

What are the next steps for improvement? Still, a more proactive information delivery using push services, as well as more adaptive information delivery using mobile devices focusing on profiling, personalization, and contextualization,

is desirable. The initial prototype introducing speech technology should be refined and improve the ease of use of the portal by providing more advanced services, for example, to read out e-mails and information subscriptions and use speech-to-text technologies.

REFERENCES

Abecker, A., van Elst, L., & Maus, H. (2002). *Exploiting user and process context for knowledge management systems*. Paper presented at the 8th International Conference on User Modeling, Sonthofen, Germany.

Adam, O., Chikova, P., & Hofer, A. (2005). *Managing inter-organizational business processes using an architecture for m-business scenarios.* Paper presented at the International Conference on Mobile Business (ICMB'05), Sydney, Australia.

Amberg, M., Remus, U., & Wehrmann, J. (2003). *Nutzung von Kontextinformationen zur evolutionären Weiterentwicklung mobiler Dienste*. Paper presented at the 33rd Annual Conference Informatics 2003 Workshop, Mobile User—Mobile Knowledge—Mobile Internet, Frankfurt a.M., Germany.

Barnes, S. J. (2003). The mobile commerce value chain in consumer markets. In S. J. Barnes (Ed.), *mBusiness: The strategic implications of wireless communications* (pp. 13-37). Oxford: Elesevier/ Butterworth-Heinemann.

Belotti, V., & Bly, S. (1996). Walking away from the desktop computer: Distributed collaboration and mobility in a product design team. *CSCW'96* (pp. 209-218). Boston: ACM Press.

Berger, S. (2004). *Mobiles Wissensmanagement. Wissensmanagement unter Berücksichtigung des Aspekts Mobilität.* Berlin: dissertation.de.

Grimm, M., Tazari, M.-R., & Balfanz, D. (2002). Towards a framework for mobile knowledge management. *Fourth International Conference on Practical Aspects of Knowledge Management 2002 (PAKM 2002)* (pp. 326-338). Vienna, Austria.

Hansmann, U., Merk, L., Niklous, M. S., & Stober, T. (2001). *Pervasive computing handbook*. Berlin: Springer.

Lehmann, H., Jurgen Kuhn, J., & Lehner, F. (2004). *The future of mobile technology: Findings from a European Delphi study.* Paper presented at the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)—Track 3.

Lueg, C., & Lichtenstein, S. (2003, November 26-28). *Location-oriented knowledge management.* A workshop at the 14th Australasian Conference on Information Systems (ACIS 2003), Perth, WA, Australia.

Maier R. (2004). *Knowledge management systems, Information and communication technologies for knowledge management.* Berlin: Springer.

Maier, R., & Remus, U. (2003). Implementing process-oriented knowledge management strategies. *Journal of Knowledge Management, 7*(4), 62-74.

Open Text Corporation. (2003). *Livelink wireless: Ubiquitous access to Livelink information and services.* White Paper. Waterloo, ON Canada.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001) Dealing with mobility: Understanding access anytime, anywhere. *ACM Transactions on Human-Computer Interaction, 8*(4), 323-347.

Rao. B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Schulte, B. A. (1999). *Organisation mobiler Arbeit. Der Einfluss von IuK-Technologien*. Wiesbaden: DUV.

Sherry, J., & Salvador, T. (2001). Running and grimacing: The struggle for balance in mobile work. In *Wireless world: Social and interactional aspects of the mobile age* (pp. 108-120). New York: Springer.

Wolff, E. N. (2005). The growth of information workers. *Communications of the ACM, 48*(10), 37-42.

KEY TERMS

Enterprise Portal: An enterprise portal is an application system that provides secure, customizable, personalizable, integrated access to a variety of different and dynamic content, applications, and services. They provide basic functionality with regard to the management, structuring, and visualization of content, collaboration, and administration.

Knowledge Management System (KMS): Knowledge management systems (KMS) provide a single point of access to many different information and knowledge sources on the desktop together with a bundle of KM services.

Mobile KM Service: The core of the KMS architecture consists of a set of knowledge services in order to support discovery, publication, collaboration, and learning. Personalization services are important to provide a more effective access to the large amounts of content, that is, to filter knowledge according to the knowledge needs in a specific situation and offer this content by a single point of entry (portal). In particular, personalization services, together with mobile access services, become crucial for the use of KMS in mobile environments.

Mobile Knowledge Management: Mobile knowledge management is a KM approach focusing on the usage of mobile ICT in order to provide mobile access to knowledge management systems and other information resources, generate awareness between mobile and stationary workers by linking them to each other, and realize mobile KM services that support knowledge workers in dealing with their tasks.

Mobile Portal: A mobile portal is an enterprise portal focusing on the mobile access of applications, content, and services as well as the consideration of the location while on the move. Mobile access is about accessing stationary KMS whereas location-orientation explicitly considers the location of the mobile worker.

Mobile Portlet: Mobile portlets are portlets enabling the mobile access of mobile workers. Special portlets can be implemented to support location-orientation in mobile portals.

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 577-582, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global)

Chapter 4.8
Accessing Learning Content in a Mobile System:
Does Mobile Mean Always Connected?

Anna Trifonova
University of Trento, Italy

ABSTRACT

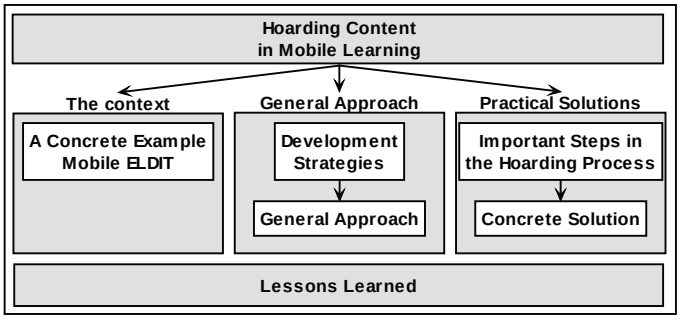
This chapter has the aim to point out an important functionality of a *ubiquitous* mobile system, and more specifically, its application in the learning domain. This functionality is the possibility to access the learning material from mobile devices, like PDAs (personal digital assistants) during their off-line periods and the technique to approach it, called hoarding. The chapter starts with the overview of a concrete mobile learning system—Mobile ELDIT, so as to give a clear idea of when and how this problem appears and why it is important to pay attention to it. Later, a description of the development approaches for both general and concrete solutions are discussed, followed by more detailed description of the important hoarding steps.

INTRODUCTION

The use of mobile devices for educational purposes was explored for the first time quite a long time ago, but the term mobile learning can be more and more often found in the literature of recent years. This is due to the fast advances of the mobile devices industry. On the market a large variety of devices with already reasonably powerful characteristics is available. The prices also allow almost everyone to be in possession of such a toy. Of course, this leads to the growing desire to use mobile devices more widely in our everyday activities.

At the same time, in the learning domain the research on the use of those mobile devices and technologies for educational purposes is also growing. Learning happens at every time and in every place of our life and the concept of ubiquitous computing overlaps very well the ways we would like to support the learning processes.

Figure 1. Chapter content



As mobile becomes so important, we should consider what makes a mobile learning system different from what we are used to having in an e-learning system, and how we should adapt to the coming changes. One of these differences is the possibility to become disconnected, and in order to allow the user to continue using the system without disturbance, a technique called hoarding might be used. Here we will define what hoarding is, when it will be needed, and how to integrate it into a mobile learning system.

While mobile learning is mainly discussed within universities and research organizations, there are also commercial m-learning products that appear on the market. They include downloadable m-learning modules, online access to learning material especially designed for mobile devices, supportive tools, and complex frameworks for mobile content creation and management. Some examples are given in Table 1 at the end of the chapter.

BACKGROUND

The field of mobile learning is growing with every passing day. New ideas, approaches, and solutions are continuously appearing, involving different mobile devices, different target groups, and having different pedagogical or technology-testing goals. A review of the literature (Trifonova & Ronchetti, 2003b) shows that there are as many common points researched as there are differences.

Mobile devices, including cell phones, PDAs, and even notebooks, are used for different purposes in different m-learning projects. In certain cases, content is accessed online through the local area network or by using the Internet. In other cases, the devices are used for communication between students and teachers or for cooperation with other students for completing common tasks. Voice or SMS might be used for receiving important educational information, images might be interchanged for sharing experiences, or common spaces might be used for collaborative work. Some of the important research directions are the following:

- The adoption to context, in particular providing location-aware learning
- The pedagogical side of m-learning – new approaches to teaching and studying
- Integration to e-learning and reuse of learning materials
- Usability issues, like facilitation of the input and output
- Provisioning of supportive to learning services
- and so forth

Nevertheless, two things are often overlooked—the support of the user during off-line

periods and the possibility to access the study content even if the device is not connected at the moment. The problem appears when the learning material is of large size, especially compared to the available memory of the device, and cannot be fully loaded locally. Generally, the issue is dealt with in two ways: in some cases, the researchers/developers rely that the devices' limitations, like small memory and intermittent connection, will soon disappear. The other alternative often used is that the full content is packed in small predefined modules that fit into the local memory of the device.

Our approach aims to provide a more flexible and adaptable solution. We would like to allow dynamic selection of the "portion" of learning material needed by the current user that should be loaded in the device memory. In order to discuss this problem, we should start from the architecture approaches for building a mobile learning system and the functionalities it includes.

MOBILE ELDIT: A REAL MOBILE LEARNING SYSTEM

The project that will be described here is called Mobile ELDIT. It aims at development of a mobile version of an online language learning system, so that the content of the e-learning platform is available to the mobile users in a ubiquitous way. Here, some details about the system will be given, as this system is the example for the architectural approaches and the working solutions described throughout the following sections. The system should be seen as a proof-of-concept, rather than a model to follow.

Mobile ELDIT (or m-ELDIT) is a mobile version of an existing innovative e-learning platform (Gamper & Knapp, 2003). From the users' point of view, ELDIT consists of two types of data—a searchable dictionary of words, both in German and Italian, and a set of texts, also in German and Italian, divided thematically into groups. The texts are especially designed for the preparation for the exam in bilingualism that is required in the South Tyrol region in Italy as a precondition for employment in the public sector. The content is prepared in such a way that will allow the user to optimally acquire needed lexical set, work on texts that had appeared on previous exams, and practice with the questions which might be asked for every single text. Though this system is especially suitable for people preparing for the bilingual exams, it can be used also by anyone interested in learning and practicing the Italian and German languages. The user might use the system as a normal electronic dictionary and search for unknown words or might browse the texts for more systematic studying. Currently, in Mobile ELDIT, only one of these two parts is included—the part comprising texts and associated to them words, which allow the users to switch, or at least to shift part of, their methodical study from a PC to a PDA device and to do it "on the go."

Here is a possible scenario of usage of Mobile ELDIT:

Scenario 1: A girl form Germany found nice work in the South Tyrol parts of Italy. In order to keep this work she needs in short time to pass a special exam and acquire a certificate of bilingualism. She starts to prepare herself for the exam by using the especially designed online e-learning system from her desktop PC. For her Christmas holidays she plans to go by train back home to Germany. On the last work day, she synchronizes her PDA with her desktop PC. The next day she gets on the train and, as her travel will be few hours, takes out her PDA device to continue her systematic study. She starts reading an Italian text, which comes next in her study plan, and clicks on *unknown word* to see its meaning. As she is preparing for the exam, she needs to have deep knowledge on every word, different forms, and senses, so she also reads a few examples and other cases of usage of the same word. She also takes a look at the list of words that derive from

the chosen one, and later continues reading the text. Continuously, she takes notes on the words she is learning, so that she can re-read them again later. At home she connects the PDA to her PC and some synchronization happens in the background while the device battery charges. On her way back, she continues reading other texts that are needed for her preparation.

This scenario has the goal to give a concrete example for a way to present learning content to the user in a ubiquitous way. It also has to show a few important points that we will discuss further.

1. How is the learning content presented to a mobile device and to a desktop PC?
2. What differs in these two cases?
3. If the content is Web-based, do we always need Internet connection to access it?

In the next section, a description will be given of the architecture that sits behind Mobile ELDIT and the concrete technological solutions to every module.

GENERAL MOBILE LEARNING ARCHITECTURE AND ITS APPLICATION IN PRACTICE

In general, e-learning systems have a wide spectrum of functionalities and responsibilities (Aggarwal, 2000). Maybe the distribution of didactic material stays in first place, but other important functionalities include, but are not limited to, the management of the learning resources, the support of different user roles and thus the identification of users and authorization of their access to the system, and often also the personalization of the learning experience based on the knowledge about the user collected during the system usage or through questionnaires and assessments, the support of collaboration between the participants, and so forth.

Let's imagine another simple scenario where a mobile device is used:

Scenario 2: A user at the university requests an interaction with the mobile learning system from her PDA. The system shows to the user the services which it can provide, and the user selects to request more information about a seminar. The system provides to the user the data about the subject, speaker, and location of the seminar, and asks the user if he is interested. When the user responds positively, the system also creates a reminder, which is triggered depending on what time the user needs to get to the seminar room. Later, the systems give the user directions for how to get to the seminar room. Though in the seminar room no Internet connection is available during the seminar, the user is able to watch the slideshow of the presentation on the PDA display. The student takes notes and later is able to see them from the desktop PC in the library from a standard Web browser.

Confronting the scenario of the first section and the one above with common e-learning functionalities (for details see Trifonova & Ronchetti, 2003a), we reveal three main differences, namely the context, device hardware, and software characteristics and connectivity.

- **Context:** In the mobile scenario, it is important to obtain the context information that might be dynamically changing and that might influence different behaviour of the user to which the system should react accordingly. By contextual information we mean, for example, the spatial data (i.e., the location in the scenario given previously), other environmental information which might be obtained with special sensors, like surrounding noise level or changes in the available light, availability of resources, like parameters of the infrastructure or the battery condition of the device, and so forth.

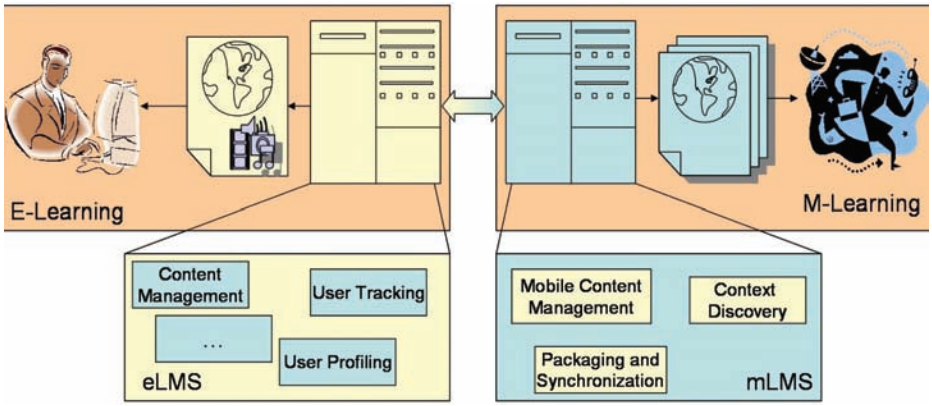
- **Device hardware and software characteristics:** The most obvious difference between e-learning and m-learning is the usage of "mobile devices for learning." There is often a discussion about exactly what devices might be considered "mobile devices for learning." Should we stress the size of the device, or underline the fact that the device is mobile (not fixed)? Should laptops be considered "mobile devices for learning?" To be more clear, we define a "mobile device for learning" as "*any device that is small, autonomous and unobtrusive enough to be carried in everyday life with the user and that will be used to support the educational processes, teaching or studying*" (Trifonova & Ronchetti, 2003b). Typically compared with a desktop PC, these devices are much smaller; for example, PDAs with a 16 bit colour 240x320 pixels screen. And the screens are not the only hardware difference. Some devices have touch screens instead of keyboard. Other, like mobile phones, do have keyboard, but with a different layout, and in this case the input speed is much more limited. It is visible that the input sometimes could be difficult, but in all cases it is different from the PC. Another very important difference is in the available software, both in sense of existing programs and difference in versions. Even if a mobile version of an application exists, most often it is very limited. For example, most e-learning systems strongly depend on frames to present their content, but the current Internet Explorer version of the Pocket PC has no support for frames.
- **Connectivity:** The connectivity is one of the main prerequirements for any e-learning system. Recently, the high bandwidth requirements are quite strong. Nowadays, there are lots of technological ways to access Internet from a mobile device: WAP, GPRS, UMTS, WiFi, Bluetooth. Though the options grow fast, still very often the bandwidth is comparatively smaller, and the user often gets disconnected, either because the infrastructure is not provided or because the expenses are still high and the user prefers to connect when a cheap connection is again available. We can distinguish disconnected periods that are intentional or not.

Depending on the concrete application that is being created, it is possible that only some of these differences will appear between the desktop and a mobile version of a system. For example, if the mobile devices that will be used are laptops, then the software and hardware characteristics are practically the same as in e-learning. Still, the connectivity is not guaranteed, and the environment of the user might be changing periodically.

To support the described differences, we proposed a general mobile learning architecture (Trifonova & Ronchetti, 2003a) that, if needed, sits on the top of an e-learning system, thus reusing some of the e-learning system functionalities, like, for example, the repository with the study materials, or in other cases the authorization of the user and assigning of proper access rights.

The so-called mobile learning management system (mLMS) consists of three modules that map the three main differences we talked about. The modules are called: (1) "Context Discovery" – the module responsible for finding out the context data, (2) "Mobile Content Management and Presentation Adaptation" – the module where the presented data should be especially adapted to fit the devices limitations (hardware or software), and finally, (3) "Packaging and Synchronization" – the module which should prepare the system for the periods of disconnection, such that the user can study even in those circumstances. The modules should communicate between themselves for optimal performance. The architecture is presented in Figure 2. On the left, you can see the e-learning system and the user with a desktop computer, which receives Web pages, possibly with multi-

Figure 2. Mobile learning: General view



media content. On the server (see the lower part of the figure) different modules exist (not limited to what is shown on the picture) and interact between themselves. On the right, the mobile user connects to the m-learning system, and receives specially designed pages, the content of which the mobile server might request form the e-learning server. The functionalities of the mLMS modules will be described better with a deeper look at *Scenario 2*, previously described.

Scenario 2 explanation: Considering the previously introduced scenario, here is how it will be supported by the proposed architecture. First, the user request is captured and, in order to proceed, the system needs to know who the user is and what device is being used. This is done automatically by the "Context Discovery" module, which (based on the first request or additional interaction) already holds the information about the user id and the capabilities and limitations of the device (both software and hardware). Based on this data, the system can check the user role (student, teacher, guest, etc.) and access rights in the eLMS to decide what services can be offered at this moment, and propose a list to the user. After the next interaction with the user, the m-learning system requests information about the seminar from the eLMS and triggers the "mobile content management and presentation adaptation" module. Knowing the capabilities of the device (from the "context Discovery" module), the data is redesigned and returned to the user. Afterward, the user requests a reminder be set up for her. The system needs additional context information, namely the user location, in order to calculate the needed time to get to the seminar room. Once again, the "context discovery" module is triggered to track the user current position. Meanwhile, as the system "knows" that the network is not accessible in the seminar room, it activates the "packaging and synchronization" module. The eLMS might contain a large amount of materials concerning the seminar—the presentation itself, including explanations from the lecturer, related links, additional papers and examples, and so forth. As the system already knows the limitations of the device, the "packaging" module selects (with certain confidence) what part will be more useful and important during the seminar (for example, only the presentation). In order to fit the device memory, the system also "asks" the "presentation adaptation" module to resize the images used. Afterward, the presentation is seamlessly uploaded to the user's PDA and is accessible when needed. During the presentation, user's notes are saved

locally on the device, but on next connection to the Internet, synchronization is done and the notes are uploaded to the server in device independent format. The system also saves the interesting and important parts of the presented material together with the notes in the student's personal folder on the server, so that they are accessible later from the desktop PC in the library.

Here is how the modules proposed in this general architecture and described in the above map work to meet the needs of mobile ELDIT, and the solutions that were used in its technical development.

Context Discovery

For the adaptation needs of Mobile ELDIT, the only context information that has to be discovered is the device hardware and software limitation. Knowing the screen size, the browser type and the device's browser support for scripts and frames will allow the "Adaptation" module to create the proper "look" for the Mobile ELDIT pages. As a first step, we chose the easiest way to discover the context—through the device browser's HTTP request that is captured on the server side.

As shown in Figure 3, the HTTP request contains what we needed, that is, what the device is (Windows CE device), what the screen is (240×320), the colour resolution (colour 16), what the browser is (Mozilla/2.0), and so forth. In other mobile learning systems, or in a more complicated version of Mobile ELDIT, it is possible to use other context discovery methods. There are quite a lot of technological solutions nowadays (for example, the device independence initiative *www.w3.org/2001/di/*). One can imagine also other scenarios where adaptation can be used to show to the user context-dependant (e.g., location-dependant) language learning material, like, for example, the system proposed by Jung (2004). Thus, other methods of context discovery will be also needed, but it is out of the scope of our current work.

Figure 3. HTTP request from a mobile device (iPAQ Pocket PC)

```
GET http://science.unitn.it/mELDIT/text.056 HTTP/1.1
Accept: application/vnd.wap.xhtml+xml, application/xhtml+xml;
   profile="http://www.wapforum.org/xhtml",  text/vnd.wap.wml,  image/vnd.wap.wbmp,
        */*
UA-OS: Windows CE (POCKET PC) - Version 3.0
UA-color: color16
UA-pixels: 240x320
UA-CPU: ARM SA1110
UA-Voice: FALSE
UA-Language: JavaScript
Accept-Encoding: gzip, deflate
User-Agent: Mozilla/2.0 (compatible; MSIE 3.02; Windows CE; PPC; 240x320)
Host: science.unitn.it
Proxy-Connection: Keep-Alive
```

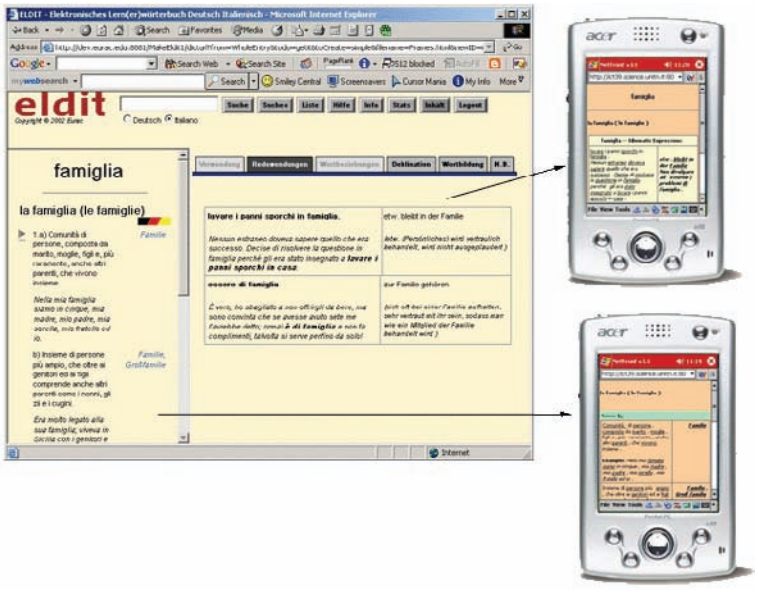
Content Adaptation

In order to keep the Mobile ELDIT users' experiences as close to the experiences with the online ELDIT system, we chose to use a browser as an interface to the learning material. Most of the browsers on the mobile devices nowadays still do not support frames, and support only limited versions of script languages. This leads to the need of specific adaptation of the content. The adaptation is also needed because, commonly, Web pages are designed for screen size at least 800x600, and in most cases are hard to read and navigate from devices with a smaller screen. ELDIT is not an exception. Different adaptation techniques can be used to attack this problem (see Butler, 2001). The adaptation can be server-side, or can be done in a proxy between the server and the client, or can be done on the client side. All of these solutions have their pros and cons.

The data of the ELDIT system consists of XML files, both for the texts and for the word entries. For displaying the data to a desktop PC, on the server side on every user request (on the fly), HTML pages are produced containing frames and Java scripts. The data is mainly text, but the entries are highly interlinked. For the Mobile ELDIT, we have decided to use server-side adaptation, namely XSLT transformations of the XML data on a Cocoon server (for more details, see *http://cocoon.apache.org*). Our decision was pushed by two facts—first, our data was already in XML format, which allowed us to easily create the adaptation rules using XSLT; second, the adaptation on the server side is a much better solution in the mobile context, as the adaptation process consumes quite a lot of computational power and will not fit well on a mobile device, as the devices are limited in CPU speed, operational memory, and battery.

Figure 4 shows a screenshot of a word entry from the ELDIT system (Figure 4a), displayed in a desktop PC browser. One can see that it is made out of three frames that contain the main

Figure 4. M-ELDIT content adaptation. 4a (on the left): Browser view of ELDIT word entry with three frames; 4b (right top): m-ELDIT additional information (idiomatic expressions) for a word entry; 4c (right bottom) m-ELDIT basic word entry screen



information about the selected word in the left-hand frame, and additional information in the right-hand frame. The frame above is dedicated to the searching functionality of the system. In the mobile version, we do not support searching. This decision is because the row data of ELDIT is a much larger amount than the available memory of standard mobile devices. Thus, we are not able to provide the entire dictionary on the device anytime (see next section about packaging and synchronization). It is also impossible to predict what word a user might be searching for. Thus, we have "converted" the screen on the left (4a) into a series of interconnected screens on the mobile device (4b and 4c). When a user wants to see a word entry, first, the main screen is displayed and the user might select to view more detailed information by clicking the links that were added during the XLST transformations on the server.

Packaging and Synchronization

The last functionality in our architecture is called "Packaging and synchronization" and it will be discussed in detail separately in the following section.

HOARDING: WHAT IS IT AND WHY DO WE NEED IT?

In the literature, one can see that quite a lot of years pass, but the supposition that "very soon" every device will always have connection has still not come true. In the fall of 2000, Clark Quinn says:

The vision of mobile computing is that of portable *(even wearable) computation: rich interactivity,* total connectivity, *and powerful processing. A small device that is* always networked, *allowing easy input through pens and/or speech or even a keyboard when necessary (though it may be something completely different like a chord keyboard), and the ability to see high resolution images and hear quality sound.*

In the above citation, we have italicized few words in order to focus your attention on the key expectations about the future of mobile computing and, as a result, also about m-learning. Though this sooner or later will happen, the current situation is not like we would like it to be. The devices had really become mobile in the sense of light and small for an impressively short period of time and, though there are quite a lot of technological ways to connect to the Internet, through WAP, GPRS, Wi-Fi, and so forth, still users have long periods of disconnection. These periods might be intentional or not—because of the lack of proper infrastructure or because the connection has high costs. Nevertheless, it is obvious that the vision of ubiquitous learning comprises the importance to give the user the possibility to access the learning materials that he/she wants to study even when the connection is not presented. Moreover, a good situation would be developed in such a way that the user does not bother and does not even understand whether connection exists or not.

To solve the problem described above, there comes to play a technique called hoarding (Trifonova & Ronchetti, 2005). Hoarding in practice is a procedure for automatic selection and caching of the data that the user will need during his off-line periods. Generally, hoarding is needed whenever the full data set of a certain application is bigger than the device's available memory; that is, it is not possible to have all the data on the device all the time. In such a case, it is necessary to select only the most relevant information and to consider how much memory is available. In our mobile learning context the data is the learning material that the user intends to study during the next off-line session. We should emphasis that, in the context of ubiquitous scenario, our interest is in the automation of the process. In other words, we do not want the learner to explicitly say what he/she wants to study, but on the contrary, the

system should be able to predict this. The fact that the process is automatic is good for two main reasons. First, we will free the user from tedious operations, and second, often we can not even trust the user in his/her own judgment for his/her knowledge and future needs.

The hoarding process should consist of few steps that we can formalize as follows:

1. **Predict the "starting point":** The algorithm should start by finding the entry point of the current user for the next learning session.
2. **Create a candidate set:** All related documents (objects) should be found and a "candidate for caching" set of objects should be created.
3. **Predict the most probable session path:** The algorithm should discover the most probable sequence of LO the user will be following.
4. **Prune the set:** The candidate set should be pruned; that is, the objects that will not be needed by the user should be excluded from the candidate set, thus making it smaller. This should be done based on user behaviour observations and domain knowledge.
5. **Find the priority to all objects still in the hoarding set:** When the candidate set is pruned, using all the knowledge available about the user and the current learning domain, every object left in the hoarding set should be assigned a priority value. The priority depends on how important the object is for the next user session, and should be higher if we suppose that there is a higher probability that an object will be used sooner.
6. **Sort the objects, based on their priority:** The hoarding algorithm produces an ordered list of objects.
7. **Cache, starting from the beginning of the list (thus putting in the device cache those objects with bigger priority):** and continue with the ones with smaller weights, until available memory is filled in.

As one can see, the hoarding process and its predictions should be based on the system knowledge about the learner style, preferences, and previous experience. Very useful knowledge can be extracted from the students' previous interactions with the system. These interactions are usually written and saved in log files that can be analyzed. For Mobile ELDIT, such log files are gathered by a local proxy on the mobile device that captures the browsers requests, and is also responsible for the system's cache. Some preprocessing of the log files is needed for doing analysis. The preprocessing is commonly one of the most time and computationally consuming processes, though in our context, this process will most likely be performed on the server, and not when the user is interacting with the online system, and thus it will not be disturbing for the user.

In the context of hoarding, we recognize two groups of characteristics that should be "known" to the system about the user. We schematically call the first "user behaviour," which will be kept in "usage patterns" profiles. The second is "user knowledge," which will be kept in an individual user's profile. The two groups of characteristics will be used differently by the hoarding algorithm. The user behaviour can be described in terms of browsing styles (e.g., consecutive, random, interest driven, etc.), preferred type of educational media (e.g., prefers video to combination of text and pictures), and so forth. Based on the user behaviour, we can group the learners and analyze the similarities and differences between the groups and between the members of the same group. This should help us predict what will be needed; that is, this data will be used to fill in the hoarding set. On the other hand, the user knowledge profile should consist of everything that the system knows about what the user already

knows. An example is the system awareness of the user's competence in a certain subject (i.e., beginner, intermediate, advanced) or a list of all the topics already covered by the user previously. In contrast of the user behaviour, the profile of the user knowledge will be used for pruning the entries from the "candidates for hoarding" set; that is, for excluding objects in order to decrease the size of the hoard.

For the Mobile ELDIT application, we tried a few different strategies for hoarding (Trifonova & Ronchetti, 2005), which include the generation of the candidate set, pruning, and prioritizing of the learning objects. First, we discovered that in our scenario the users have often very consecutive browsing behaviour. In other words, when shown a list of texts to be reviewed, the learners almost always read the texts in the order they are listed. This seems to be logical behaviour for students accessing other types of learning materials, where to understand the information presented later in the material there is a prerequisite that they have a good understanding of the information that preceeded it. This consecutive behaviour is a concrete usage pattern that helps to decide the inclusion of material into the candidate hoarding set. Some of the parameters that should be discovered for every user are the depth of the browsing, the number and types of learning objects requested, the time usually spent, and so forth. As far as the pruning is concerned, in the Mobile ELDIT we wanted to test the possibility for fully automatic hoarding, and also the automatic discovery of "user knowledge." A special feature of m-ELDIT is that the learning content is divided into small chunks (texts and connected words) and some of them are repeatedly shown to the user. Because of this fact, we used as a pruning rule the following logic: if the user had the option to review a chunk of the material, but decided not to do it, there is a big probability that the learner knows this chunk and will not need it in the future, and thus it will be pruned next time. Even this simple rule made the hoarding set decrease in size quite fast. However, its simplicity has a negative side also. In certain cases the deduction made with this rule that the user "knows" a certain word is not correct, but the word gets excluded from the hoarding set in the next iteration. This leads to an increased miss rate; that is, the number of unsatisfied user requests. A more sophisticated rule might take into consideration also the time the users spent for reviewing certain content chunks, or the number of times the same chunk was requested.

Possible further improvements in the hoarding might be done by grouping the users by similarity in their behaviour or knowledge. Predictions on what actions will be taken by a user, and thus decisions on what to include or exclude from the hoard might be taken based on the behaviour or the knowledge of another user that previously had shown very big similarity with the current one. Note that the best similarity measure will differ from one application to another—one case might be the type of reviewed learning material, in another its quantity, in a third the time spent on every portion.

As the learning material and the users of every specific mobile learning system will differ, all these processes, and the decisions taken, will be based often on a big number of parameters that should be defined based on analyzes of the tracking data collected of the specific system. For example, the size of the hoarding set should be a function of the available memory on the mobile device and the size of the learning content chunks. The behaviour of the user might differ based on the learning tasks, and thus things like grouping of the users might be done based on specific classification strategies. Also, the discovering of the user knowledge might be done in various ways—our strategy was automatic discovering, but also other methods are possible, like questionnaires or tests.

Details on the results of the experimentations with hoarding strategies and parameters can be found in Trifonova (2006).

Table 1. Commercial m-learning examples

Downloadable content modules
http://hotlavasoftware.com Hot Lava Software provides offline course modules for Palm and PocketPC. A number of modules can be downloaded for free, while others are commercial. Some examples are Cisco® mobile learning (example: CCNA® Prep 4-PACK: Networking, TCP/IP, Ethernet), Kids Mobile Learning; (example: 1st Grade Language Arts: Standardized Practice Test), Microsoft mobile learning (example: MCSE Prep Data Networking), COMPTIA A+ mobile learning, English as a Second Language, Business and Sales Skills Mobile Learning, and so forth. *http://www.italyguides.it* Italy Guides provides small audio tourist guides that can be downloaded and listened to with iPod. Information is available for some of the most interesting Italian cities – Rome, Florence, and Venice. *http://www.ipreppress.com* - Merriam-Webster Inc., in collaboration with iPREPpress, offers its learning material on iPod. The commercial 2006 Edition Pocket Dictionary is already available on the site. The 2006 Pocket Thesaurus and the Pocket Atlas are expected soon. Free modules are available also in subjects like the Declaration of Independence (1776), the Constitution of the United States, (1787); the Social Security Act (1935), etc.
Online accessible content
http://en.wapedia.org/ The popular lately Wikipedia provides a cell phone accessible version of its materials. *http://www.alc.co.jp/eow/pocket/* Japan is one of the leading places offering m-learning. Examples are many of the services offered by DoCoMo to i-mode enabled mobile phones. Pocket Eijiro is an English language learning site provided by ALC, where also small multi-choice quizzes can be used to test users' knowledge.

Content creation and management platforms
http://www.axmor.com AXMOR provides the tools to deliver mobile learning content to PocketPC and Palm devices. The platform consists of .Net-based Web site and Pocket PC part. Content maintenance, user management and reporting of different activities is done online. On the other hand, locally on the mobile device the purchased content is managed and played. The modules are in Macromedia Flash. *http://hotlavasoftware.com* Hot Lava Software, mentioned above, also offers mobile content authoring, publishing, delivery, and tracking for many mobile devices types, including PDAs and cell phones. *http://www.symexuk.com/* Symex offers to teachers a mobile system to facilitate teaching activities, like planning teaching, collecting students data, and managing their results. The system is available for PDAs and can be synchronized by connecting the device to the PC or via the Wi-Fi network.
Other tools
www.macromedia.com Macromedia provides a Lite version of Flash for creating multimedia movies on mobile devices, including cell phones. *http://www.pocketmobility.com* Different free and commercial tools for mobile education and learning are provided by Pocket Mobility Inc., like Quizzler Maker, to create easily quizzes for any platform. The bundle allows the teacher to collect the average scores in the class, to see which questions were missed the most, and so forth. *http://classinhand.wfu.edu/* DataInHand, created at Wake Forest University, allows from a PDA with wireless connection to control presentations, to receive feedback from classroom answers, and see the distribution of the answers.

CONCLUSION

Mobile learning seems to be an integral part for the future of learning. And it is a step further on the road from e-learning to ubiquitous learning. Ubiquitous access to learning material supposes that, regardless of whether Internet connection is available or not, the user will have access to the learning material needed at the moment from the device used at the moment. Throughout this chapter it was shown that the technology needed for realizing a ubiquitous learning system, the pieces of the puzzle, are already available. Possibilities are wide and waiting to be explored. Often, applications would be developed in such a way as to utilize the newly appearing fast Internet connections from the mobile device. But when such an option is not possible or is not sufficient, the needed piece of the puzzle is called hoarding.

Hoarding should be based on deep understanding of user behaviour, both specifics of the concrete user and the common patterns in the behaviour of all users of a concrete system. We have shown a possible approach used in a real mobile learning system, which showed us the viability of hoarding in practice. Nevertheless, specific parameters should be extracted in every specific case, as the user behaviour might differ drastically based on the proposed study material and connections between chunks.

ADDITIONAL SOURCES

Here we give ideas of the existing commercial m-learning tools and systems. We do not intend to give a complete list of available products, but rather examples of available possibilities and starting points for future work. The entries are grouped by provided functionality and are supported with URL and short description.

REFERENCES

Aggarwal, A. (2000). *Web-based learning and teaching technologies: Opportunities and challenges.* Hershey, PA: Idea Group.

Butler, M.H. (2001). *Current technologies for device independence* (HP Labs Tech. Rep. HPL-2001-83).

Gamper J., & Knapp, J. (2003). A data model and its implementation for a Web-based language learning system. In *Proceedings of the Twelfth International World Wide Web Conference (WWW2003).*

Jung, L. (2004). Context-aware support for computer-supported ubiquitous learning. In *Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'04).*

Quinn, C. (2001). *Mobile, wireless, in-your-pocket learning. LiNE Zine: Learning in the new economy.* Retrieved October 16, 2006, from http://www.linezine.com/2.1/features/cqmmwiyp.htm

Trifonova, A., & Ronchetti, M. (2003a, August 30-September 1). A general architecture to support mobility in learning. In *Proceedings of the 4th IEEE International Conference on Advanced Learning Technologies (ICALT 2004 - Crafting Learning within Context),* Joensuu, Finland.

Trifonova, A., & Ronchetti, M. (2003b, November 7-11). Where is mobile learning going? In *Proceedings of The World Conference on E-learning in Corporate, Government, Healthcare, & Higher Education (E-Learn 2003),* Phoenix, Arizona.

Trifonova, A., & Ronchetti, M. (2005, June 27-July 2). Hoarding content in an m-learning system. In *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-Media 2005)* (pp. 4786-4794). Montreal, Canada.

Trifonova, A. (2006, March 21). *Towards hoarding content in m-learning context.* PhD thesis, University of Trento, Italy.

This work was previously published in Ubiquitous and Pervasive Knowledge and Learning Management: Semantics, Social Networking and New Media to Their Full Potential, edited by M. Lytras and A. Naeve, pp. 198-215, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.9
Using Learning Objects for Rapid Deployment to Mobile Learning Devices for the U.S. Coast Guard

Pamela T. Northrup
University of West Florida, USA

William T. Harrison, Jr.
University of West Florida & U.S. Navy, USA

ABSTRACT

This chapter introduces the use of a learning objects content development tool, the eLearning Objects Navigator, (eLON™) as a strategy for creating, classifying, and retrieving reusable learning objects and reusable information objects. The use of eLON™ provides a context for rapid deployment of these SCORM-conformant packages to mobile learning devices as well as to learning management systems for a beta test with the U.S. Coast Guard Institute. Presented in this chapter is the underlying theoretical framework for the development of eLON™ as well as the specific design decisions made regarding the deployment of PDA mobile learning devices to military personnel. Furthermore, initial results from the beta test yield positive results as well as a series of lessons learned.

INTRODUCTION

The field of distance education continues to grow as emerging technologies present new opportunities to distribute learning anytime, anywhere. As more students choose distance learning to achieve college and career goals, universities are now faced with challenges to distribute learning using a variety of strategies to accommodate student

needs. Military students represent a large segment of many institutions in the United States through the Department of Defense off-duty voluntary education programs. Each year, approximately 300,000 service personnel enroll in voluntary education with universities making it one of the largest continuing education operations in the world (Department of Defense, 2003). As a result, universities with a strong military presence must be flexible to accommodate deployments, temporary duty, lack of Internet access, intermittent Internet access, and more. The level of flexibility required by military personnel pursuing educational degrees presents unique challenges to those who design distance learning instructional materials on university campuses. From creating blended learning opportunities to duplicative design across numerous delivery approaches, the time spent developing fully online programs and courses can easily exceed man-hours available on university campuses. Currently, the majority of higher education programs offered to the military are self-contained, nonflexible existing programs that may not meet the needs of individual service personnel that may be deployed, underway, or unable to access the Internet for an extended period of time. In an attempt to meet the need, the University of West Florida has partnered with the U.S. Coast Guard Institute and two community colleges, Florida Community College at Jacksonville and Coastline Community College to develop and beta test college level courses on a personal digital assistant (PDA). Given that few models currently exist for this mode of course development, the University of West Florida chose to develop all content using a learning objects content development tool, eLON™ for purposes of consistency and reuse across multiple delivery platforms.

Within the partnership, the community colleges agreed to offer a selection of general education courses, while UWF agreed to offer graduate level courses. For the beta test, UWF selected to offer a 12 semester hour graduate certificate in human performance technology.

THE BETA TEST WITH THE U.S. COAST GUARD

The U.S. Coast Guard Institute provided several specifications in the partnership to beta test the PDA as a viable mobile learning solution for Coast Guard personnel. The participants in the UWF study included those interested in the program area offered on PDA, including a graduate certificate in human performance technology (HPT). UWF students were recruited from several Coast Guard sectors including Key West, Islamorada, Miami, and the Yorktown Training Center. Students participating had to enroll at UWF to receive their tuition assistance or VA benefits. Students were then afforded access to all UWF student services. Since UWF is a SOCCoast Afloat institution, the programs offered had already been moved through the program approval process with other partnering institutions.

There were several restrictions placed on the selection and use of a mobile device that may be used on a Coast Guard cutter. All devices were required to have both Bluetooth and wireless disabled prior to use on a cutter for purposes of shipboard security. This presented some unique difficulties as most mobile devices offered these features with few companies or software applications in place at the time to disable both Bluetooth and wireless.

There were several design requirements that would ultimately affect the design models selected for development. We were asked to provide a bookmarking feature to enable users to pick up where they left off as many service personnel shipboard may have limited times available for study and may need to stop working at a moment's notice. With regard to access, we were required to design all materials for stand-alone use on

the PDA in the event that Internet access was not available. For student assessment purposes, we were required to work with the educational services officer (ESO) to proctor and certify all exams. The ESO also served as the primary point of contact for the institution on behalf of the Coast Guard. The ESO received all shipments of PDAs, print materials, assessments, and directions for the return of materials. As an advocate for the Coast Guard personnel, the ESO in many cases also worked to assist students in receiving their tuition assistance or in filing any additional paperwork required.

ISSUES FACED

The Academic Technology Center is tasked with designing, developing, and implementing all distance learning endeavors at UWF and has been very successful in designing fully online courses, blended courses, and interactive distance learning classrooms using two-way interactive video. However, until this beta test, the opportunity to design instruction for mobile devices did not exist. Issues related to the overarching design needed to be made to compensate for the lack of Internet access and subsequent lack of interaction between students and the instructor. As well, the PDA operating system had limited software applications available to deliver best-fit instructional strategies. For example, the Pocket PC did not include a PowerPoint viewer, thus requiring conversion of the PowerPoint presentations to a compatable format for the PDA. Finally, issues of faculty and student use of this new model of learning took some time. Faculty learning how to design content as a subject matter expert and to include everything needed for individualized instruction required a great deal of work, even for faculty members with expertise in developing online learning. As well, the role of the faculty member and student shifted in this more individualized environment. The issue of providing feedback to students in a timely fashion became a major hurdle for faculty. For students, issues included how to seek advice, guidance and direction from the instructor when Internet access was limited or not available.

An additional issue faced included a change in the development process typically used in the Academic Technology Center. The reality of work effort required to develop courses on a PDA and provide flexible, duplicate courses available for those able to access on the Web through the learning management system required us to rethink our whole design process. With this change, two major issues emerged. First, the development of content for multiple delivery containers required a strategy to develop content systematically so that it could be exported and reused in a variety of ways. This required significant thought as design for a mobile device such as the PDA requirements may be different from the requirements used for designing Web-based instruction. This issue enabled us to think more broadly about learning objects, issues of granularity, consistency, and technical specifications for specific assets, such as sizing PowerPoint presentations or creating Flash applications. Considerations for using graphics, the amount of text and overall screen geography immediately became in an issue.

In line with the new considerations for design, not only did faculty and instructional designers collaborate on overall design, it was essential for faculty members to finish their efforts early enough to allow technical personnel the time needed to export the developed content from our learning objects content development tool, eLON™ to the secure digital (SD) chip, which at the time, could only be made one at a time.

FOUNDATIONAL MODELS

Without a body of research or best practice on developing mobile learning, there was extensive research and development work to be done before

proceeding. The first step in this endeavor was to create a theoretical framework to serve as a guide for all design decisions made in the project. As a theoretical framework we chose Gagne's (1985) *events of instruction* as the major frame for *pre-instruction*, *instruction*, and *post-instruction* as a sound model representing the external events necessary to align to the internal processes of learning. These events should satisfy or provide the necessary conditions for learning and serve as the basis for designing instruction and selecting appropriate media (Gagne, Briggs, & Wager, 1992).

The flexibility of the model to align to a variety of learning outcomes and instructional tactics and strategies met the needs of this design effort. An additional benefit of using the events of instruction included its alignment to the Cisco model of creating reusable learning objects and reusable information objects. The Cisco model was constructed based on the foundational work of Merrill's (1994) component display theory and others. The linkage of Gagne's events of instruction and Cisco's model gave way to the inclusion of specific learning outcome specification templates that are embedded within eLON™. These eLON™ specifications include one template for each major learning outcome specified by Bloom's taxonomy enabling designers to select the most appropriate reusable learning object specification template.

Interaction and Feedback

With intermittent to no Internet access available, the instructional design and content development process required a complete reconsideration of the role of interaction in a distance environment. According to Moore (1989), typical online interaction includes: (1) peer-to-peer opportunities such as dialog in threaded discussions, small group assignments, and the theme of working with a partner; (2) student-to-instructor interaction including opportunities for assignment clarification, chat room events, threaded discussion with faculty participation to scaffold the process, and ongoing feedback on correctness of assignments; and (3) student-to-content interaction where the student interacts with the instructional materials by reading, participating in online simulations, and searching the Web for specific information. Since all possible interaction scenarios are not available in the mobile learning environment, design decisions were made to compensate for limited student-to-student interaction and student-to-instructor interaction, while designing strong processes for student-to-content interaction. Since we know that interaction is a key component in successful online learning (Northrup, 2002a), it is critical to make good decisions at this point to ensure student success in the mobile learning environment.

The feedback literature has long prescribed feedback models for purposes of learning and assessment. Kulhavey and Wager (1993) suggest that feedback on incorrect responses assists in further understanding specific concepts, which is the method designed into instruction. Mory's (1992) review of the feedback literature suggests that feedback has long been advocated as an important part of the learning process that historically has been used for purposes of reinforcement in behaviorist learning environments, but now being used more for error correction.

WHAT WE KNOW ABOUT DISTANCE LEARNING

Distance learning has been around for quite some time, initially entering the mainstream of universities as correspondence courses. The correspondence model, an example of individual learning, was used nearly exclusively for the first 120 years of distance education in the United States (Moore & Kearsley, 1996). According to a recent Sloan Report (2005), over 2.35 million students in the U.S. participated in online education in 2004, which is up from 1.98 million in

2003. Most online program and course offerings use the Internet as the type of access for students through a learning management system such as Desire2Learn, WebCT, or Blackboard. What has been found from years of research on distance learning are some guiding principles that became design requirements in the PDA course development model with the U.S. Coast Guard. There are many benchmarks that align to quality and distance learning (Kane, 2004) including the need for quality curriculum and student support. However, one of the most significant attributes for success in online learning has been the student's ability to interact with other students, in both a social and instructional environment. A recent study indicated that 90% of students cited interaction with the instructor and other students as one of the major reasons for staying in the course (Northrup, 2002a).

Interaction serves many purposes. First and foremost, it engages the learner with other students and the instructor and enables the instructor to provide the necessary scaffolding to achieve successful learning outcomes. When interaction is limited or not available, course retention and overall student satisfaction may be negatively impacted. Interaction also serves to assist in developing the students' social network. Distance students are isolated from others just by the nature of participating in a course remotely. In a campus-based course, students naturally form social groups by talking with one another after class, sitting together, or forming after class study groups. Each of these approaches help cement student success in college. Online, these social groups may not naturally emerge and it is essential to design experiences for getting to know one another into an online course. Student connection to other students remains a need, whether online or face-to-face.

Research also indicates that technical issues present a series of challenges to the student and the student's desire to stay engaged in school. Providing technical support such as a help desk, a student support center, an 800 number, or some type of FAQ repository will increase student success.

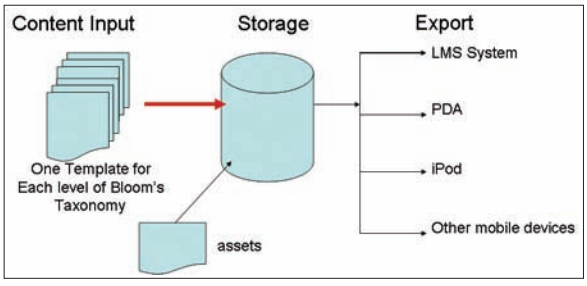
Another item that holds true across teaching and learning endeavors includes early success. Students with early, incremental success will build confidence and to continue with their schooling.

Finally, designing instruction that includes significant student engagement will increase opportunities for deeper processing of knowledge and ultimately remembering and hopefully applying what is learned.

LEARNING OBJECTS

Learning objects have been a major topic of discussion for the past several years with heated debates focusing on data standardization, interop-

Figure 1. eLON™ learning objects export architecture



erability, metadata, and SCORM. Most absent in the discussion are those responsible for designing instruction using learning objects. As a result, less attention has been given to standardization and optimization of instructional elements to be included, used, and reused. In the effort to frame the role of learning objects in the development of mobile learning instruction, we adopted and modified Cisco's model (Barritt & Alderman, 2004) for developing reusable learning objects and reusable information objects.

UWF developed and is implementing a rapid learning objects content development tool called eLearning Objects Navigator (eLON™). eLON™ enables UWF instructional designers in the Academic Technology Center and content experts to create instructional content and learning objects within eLON™ and export it to the PDA Secure Digital (SD) chip as well as running a secondary export to the university's learning management system, Desire2Learn for a second section of the course to be offered fully online. A unique benefit is that students needing ongoing flexibility can take part of the course on the PDA and the remainder online as dictated by military duty assignments.

Learning Objects Framework

To accommodate a pedagogically sound framework and given the Coast Guard requirement to provide bookmarking and location identification, we also included an extensive menu system, classified into menu items by learning object title and type. We adopted a modified version of the Cisco model (Barritt & Alderman, 2004) to classify objects as reusable learning objects (RLOs) and reusable information objects (RIOs). The classification scheme enabled us to classify RLOs as complete sessions, with five to seven RIOs making up a complete RLO (see Figure 2).

The classification scheme allowed us also to organize the menus around complete sessions making up an introduction, five to seven topics (RIOs), and a summary. This classification scheme was used to generate the menu system and subsequent bookmarking to the RIO level.

Additionally, we incorporated six templates, one for each learning outcome type specified by Bloom's (1956) Taxonomy as a way to establish explicit categories for each learning objects type. Each template represents a level of learning to accommodate the type of instruction that may be designed for specific learning outcomes. For example, for knowledge-level outcomes, selections may include tutorials, while higher order outcomes may include case studies. The templates

Figure 2. Reusable learning objects

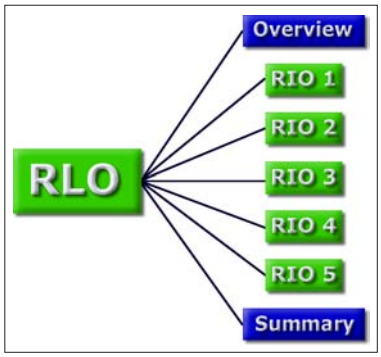
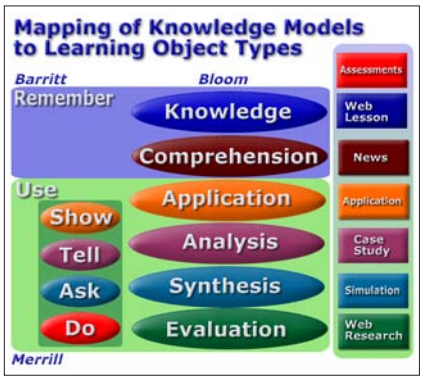


Figure 3. Mapping of knowledge models to learning object types



provide an appropriate instructional strategy as well as maintaining consistency across reusable learning objects (see Figure 3).

Issues

Developing learning objects for a project using multiple instructional designers and faculty members presented challenges as many items needed resolution quickly. Major issues included decisions on metadata, granularity, context, and reuse.

- **Metadata:** Metadata was one of the first issues requiring resolution. *Metadata* is generally defined as "data about data." In some disciplines, this is translated to describe information about a set of data in a particular representation. However, in content management and information architecture, metadata generally means "information about objects," that is, information about a document, an image, a RLO, a RIO, an asset, or other object type. Dublin Core, as one well recognized standard for metadata was intended to support information retrieval or discovery of resources. However, it is now commonly agreed that the metadata are as useful for the management of content as they are for the discovery of them after publication, and so metadata in practice tend to be used for both purposes.

  Defining the level of metadata to provide at the beginning of the development process proved challenging. Faculty resisted completing all metadata fields, and when they did complete, there were major inconsistencies in keyword identification and object descriptions. When course content was similar, keywords were so similar that it was difficult to differentiate from the titles, thus creating an ineffective strategy for resource discovery or management. In our lessons learned, in the next effort, we will assign one person to develop the keywords from an index of course-related topics.
- **Granularity:** Although granularity is routinely an issue when developing learning objects, in this effort, with a commonly defined template and agreement to use the guiding principles that specify limitations on text, large graphics, and PowerPoint, there were not as many issues. It was our decision to create RLOs with accompanying RIOs and couple them as intact reusable learning objects, translated in the PDA course to a complete session. In later situations, if RIOs are reused, the architecture of the RIO as defined by some as a sharable courseware

object (SCO) will allow consumers to use more granular RIOs, or even select to use the assets embedded within the RIOs themselves. Although the RLO with accompanying RIOs were intact for this effort, we have intentionally left the option available to drill down to the asset level in the eLON™ tool as well as providing the ability to create subsequent versions while retaining the original version of the asset or RIO. At any point, the RLOs could be reconfigured with other existing RIOs as well.

- **Context:** The context issue is resolved when this decision is made as introductions and summaries can be tied to the specific topics at hand. The issue with context was more tied to reuse on multiple delivery platforms. For example, descriptive information on what to do in the online version of the courses was a bit different from descriptive information on the PDA version of the course.
- **Reuse:** In many articles on learning objects, the focus is on initial development using highly defined specifications and metadata. What is not addressed as well in the literature is the capacity of learning objects to be reused. Initially eLON™ was designed with four major goals:
  - To promote c*onsistency* across a single course and among a group of courses in a program
  - To promote *efficiency* in the process of designing and developing Web-based courses
  - To promote *reuse* and *content sharing* across courses and between faculty members if chosen by content owners
  - To improve *overall course quality* by providing a series of pedagogically sound *templates* aligned to specified learning outcomes that *store* instructional *content*

For this project, an additional goal was included to reuse content across learning platforms. This has in effect been the major outcome of our investigation with the Coast Guard that we can successfully design high quality content once, thus increasing design efficiency and rapid deployment of content to multiple delivery containers. In working with the military, there are numerous available learning technologies; therefore our goal is to export learning objects across these major platforms using robust metadata tied to specifications within stylesheets and hooked through the manifests available in the export function of eLON™. Our design decision to export the RIOs together with the RLOs as a complete SCORM package will encourage reuse across major platforms but reduces the granularity of the object. However, the objects can be separated into discrete reusable information objects and assets, thus increasing the level of granularity and ultimately the capacity to reuse in other learning environments.

GUIDING PRINCIPLES FOR PDA DEVELOPMENT

From this brief knowledge base, a set of guiding principles have been developed in an attempt to address the specifications identified by the U.S. Coast Guard Institute. Considerations included the issues faced initial research and development of mobile learning, the constructed theoretical framework along with what is known about distance learning and translate to practice in a mobile learning environment using the following guiding principles for PDA development:

1. All course materials designed will use an RLO template available in eLON™ for purposes of consistency.
2. All course materials will be cataloged for retrieval using common identifiers that will export to the menu.
3. All cataloged RIOs will be hooked to a larger

RLO for purposes of export and to maintain state with the menu specifications allowing for bookmarking.

4. Each session will include a brief video to personalize the instruction and to motivate students.
5. Each session will select multiple media approaches to deliver instructional content, limited text and capitalizing on audio narrated PowerPoint short lectures, video, and Flash simulations and animations.
6. PowerPoint presentations will be focused on individual topics or RIOs, will minimize text and bullets, and will be no longer than 10 minutes in length.
7. PowerPoint presentations will attempt to tell stories to align bullets to real experiences.
8. Each session will have minimal use of detailed graphics due to the size of the screen display. Detailed graphics should be placed in print materials.
9. Each session will use self-check questioning to accommodate for limited interaction.
10. Each self-check question should provide corrective feedback enabling the learner to continue learning while participating in the self-check portion of the session.
11. Multiple choice assessments will be administered and proctored by a designated education services officer (ESO).
12. All content for a single course will be placed on a SD chip, duplicated and sent to students via surface mail.
13. All course materials will have an accompanying print-based guide and textbook to support the course.
14. All course materials will be packaged and sent to the ESO contact for distribution to students on base.

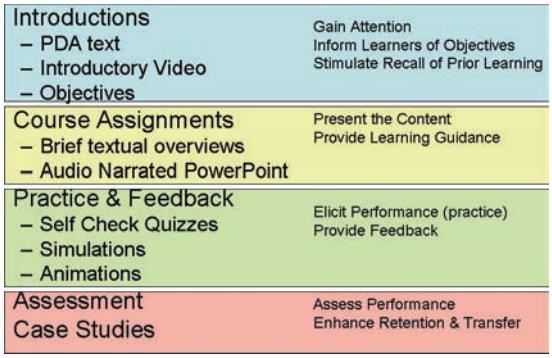
THE PDA LEARNING ENVIRONMENT

The PDA learning environment incorporates the foundations of learning and considerations for learning objects design while maintaining a simple interface and set of materials. Included in the materials sent to the student is a bag containing instructions for using the PDA and taking the course, the SD card containing the course materials, a textbook and a print packet including all PowerPoint lecture handouts, project assignment descriptions, and any worksheets required for class practice. As well, all additional readings are included given the fact that some students will be deployed, underway, or on temporary duty at some other location and do not have Internet access.

As noted in Figure 4, each session includes an introduction that is intended to provide a motivating anchor for the session that includes text, video, and written session objectives. The actual course assignments include course content presented with minimal text and supported fully by audio narrated PowerPoint presentations. Course assignments may have several topics all tied to the major session topic and is fully documented through metadata when the content is developed.

Course practice and feedback is achieved within each session by offering a series of self-check questions to allow students to determine the level of understanding achieved at that point. If students succeed on the self-check questions, they are ready to move forward to the next session. If not, corrective feedback is intended to guide students to the correct response, or students can repeat the session. This section was incorporated due to the lack of student-to-student and student-to-instructor interaction. This effort, at a minimum, allowed students to judge whether or not they were on track and gain immediate feedback on their responses to self-check items. Additional strategies used within the course practice and feedback section included short Flash-based simulations, a study guide, and encouragement to partner with another student in collaboration on course-based issues.

Figure 4. Overarching framework for PDA session design



Assessing retention and transfer was encouraged through the use of standard multiple choice assessments, papers, projects, and field work. Standard closed-book assessments were paper-based and proctored by the U.S. Coast Guard's ESOs and returned to UWF via surface mail. Papers, projects, and field work were returned to us either electronically (through e-mail or the learning management system drop box) or delivered via surface mail. To encourage transfer, one major case study was embedded in each course. Case studies are intended to engage students fully in a real-world case and apply newly learned information immediately.

Components of the PDA Course

Each PDA delivered course is developed in 12 complete sessions. Each session is equal to one reusable learning object. Each RLO contains an introduction, five to seven RIOs, and a summary. Evidenced in the following examples are the components of a session RLO. Students enter the course through a main menu that provides them with basic information about the instructor, the course syllabus, the course assignment sheet, about the class, and the course content materials.

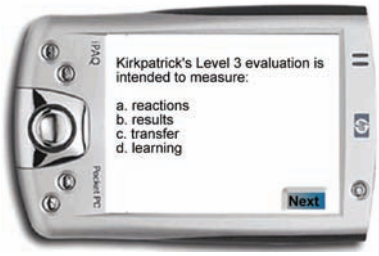
Each section is intended to align to items a student would need to be successful in class and to encourage and personalize the content as much as possible. Another significant variable while maintaining a strict representation of the events of instruction, Keller's (1987) ARCS model was woven throughout each session by presenting a motivating introduction, embedding video, promoting opportunities for successful small steps, and provided relevant case studies to apply newly learned knowledge.

- **Session introduction:** The session introductions include a brief video by the instructor talking about the upcoming session. The video is typically two to three minutes and provides the hook for each session. Students have noted that this short clip provides a connection to the instructor and to provide them with a sense of connectedness to the course. So, the brief videos as well serve as a motivational segment and a layer of interaction, albeit a small one. Also in the session introduction is the goals for the session and a paragraph or so describing upcoming events. Within the RIO, the introduction is tied to the content and the practice as a complete reusable information object package.

- **Content presentation:** The content RIOs include several "topics" that are further classified into small content objects to align to the Cisco model. The presentation of content presented some unique challenges as we tried to balance the media to limit the amount of text scrolling on the PDA, to take advantage of the outstanding video and audio available on the PDA, and to ensure multiple modes of media are presented to the learner. Within this section, audio narrated PowerPoint presentations were used consistently as the course lecture, while limited text extended the presentations providing further descriptions as needed (see Figure 5).
- **Immediate feedback:** As a major strategy for interaction, students are presented with self-check questions and feedback at the end of each instructional session. The self-check questions are tied to each RIO although they receive the items at the end of a session. If students do not perform well on the RIO, they can go back into the individual sections to review and retake the self-check. The self-check is not graded, but is used as an embedded cognitive strategy for students to self-assess and self-regulate their learning (see Figure 6).
- **Case studies:** During the second half of each 12 week course, students are presented with a case study of a real-world situation in an attempt to apply newly learned information to encourage retention and transfer. Case studies provide relevance to the course by engaging students in real-world practice, discussing real-world issues that are typically ill-structured and not always completely defined with one accurate, correct response. Cases cut across a range of fields including the military, financial management, corporate training, and the information technology industry. Case studies done online or face to face can be easily scaffolded by the instructor. In a mobile learning environment, this was a challenging task. Much scaffolding

Figure 5. Content presentation example

Figure 6. Immediate feedback example



had to be pre-prescribed rather than guiding students to "ah-ha" moments. Much data are provided, worksheets, surveys, schematics, notes, and whatever else is required to filter through the information to discuss and ultimately make a decision. For students with Internet access at some point in the course, clarification and communication with others could be provided in that manner. For the full PDA courses, students had to make some of their own assumptions and describe their positions in their final analysis and case study reports.

INITIAL FINDINGS

The beta test of U.S. Coast Guard students is complete with 20 of 20 students finishing their certificates in human performance technology. The HPT is a graduate level certificate and most of the completers were midcareer military and civilian professionals or retired military. The experience and age range of students participating in the beta test allowed us to look at the midcareer student, in contract with students served by the community colleges. As students mature, their experiences with the technology would be different than students who are undergraduate, 18 to 21 year olds.

As an incentive, each student was allowed to keep the PDA for participation in the effort. Overall, students participating in the beta test were satisfied with their experience and achieved equivalent grades to students participating in UWF's fully online HPT program. Reported will be student satisfaction comments as measured through a self-report survey and a focus group interview, test score comparisons, and their issues and suggestions for change.

Satisfaction

Overall, 100% of students completing the survey indicated satisfaction with the PDA course, with 83% indicating they would take another course in the program. Ironically, not only did 100% of the students participating in the beta test complete the program, more students joined as a result. Once students became comfortable with the technology, the fears of learning on a PDA virtually disappeared. 83% of the students reported that the courses met their goals and that it was relevant to them.

With regard to media usage, 100% of the students indicated satisfaction with the audio-narrated PowerPoint presentations and the self-check questions. Both were perceived to be very practical and usable for their course experience. The introductory course video that we incorporated for motivation was reported by 91% of students to align to their course needs. The bookmarking feature was successfully used by 83% of the students.

Eighty-four percent of students reported the overall course structure to work for them incorporating the menus, the cookie crumb trail as a notation of location information such as "where you have been" and "current location." Seventy-five percent of students reported that the directions were clear and understandable. The 12 week timeframe was reported to be just the right amount of time for 92% of the students. In focus group sessions, discussion focused on the amount of work required for the course. In later courses, more rigorous course requirements existed with less specific instructions requiring students to think more deeply about course direction. Students indicated that possibly later courses may need to be lengthened to provide time for reflection and completion of course materials.

Interaction

Recalling that interaction is a key component in successful online learning (Northrup, 2002b), there were three types of interaction described by the literature including student-to-instructor, peer-to-peer, and student-to-content. The PDA environment lent itself well to the student-to-content interaction, but was less able to design a robust system for peer-to-peer and student-to-instructor interaction. However, students who had Internet access at any time during the course were highly encouraged to log on and participate with the group. As well, students were encouraged to check in with the instructor for any clarification needed on assignments and projects.

In an attempt to find out how students interacted with the instructor, a series of questions were asked about the frequency of both peer-to-peer and student-to-instructor interaction through e-mail, online through the learning management system, or through phone calls. An additional question was posed asking how many students requested feedback from the instructor. Students reported interaction with the instructor as follows: 91% of students had the opportunity to interact via e-mail, 78% through the LMS, and 27% of students had the opportunity to interact via phone call.

An additional line of questioning attempted to determine how students interacted with peers throughout the course. Ninety-two percent of students had the opportunity to interact with peers throughout the course. The beta test provided strong cohorts where several students from single locations would participate together. In some cases, the social interaction evidenced through the course included students riding in to work together while one drove the other would be playing the lectures on the PDA, and then talking about the assignments together. In other cases, students collaborated in the office and online via the threaded discussions when possible. At no time during this test was Internet participation a requirement, but it was encouraged when available to encourage camaraderie and the community of learners that form in cohort environments. One hundred percent of students reported that they were able to discuss course content with peers at some level with 42% of students indicating that they interacted with peers through e-mail while 58% of students interacted through the learning management system's threaded discussions.

Open-Ended Comments

Course comments related specifically to the level of access and flexibility provided to empower students to continue learning where ever they may be. In some cases, the ability to take the course materials to little league fields and on the road for traveling encouraged students to continue learning and afforded them the opportunities they would not have had if having to sit in a classroom at a designated time and place. Overwhelmingly positive statements about convenience, access and flexibility were provided.

Suggestions for Change

Students did make several suggestions for change; however, most of them were directly related to the technology itself. Some of the issues included the fact that the screen was too small and the battery life was not as long as hoped. Of course, newer technologies will likely assist with both screen size and battery life. One of the suggestions was to include an additional battery or a car charger cable. Another suggestion was to include a keyboard attachment so that responses could be generated directly on the PDA. In the courses offered by UWF, we purposefully did not have a great deal of writing requirements, however, had a keyboard been included, we likely would have included some additional assignments and is under consideration for future courses.

SUMMARY

In conclusion, the U.S. Coast Guard beta test did suggest that mobile learning devices such as PDAs could successfully be used for voluntary education. Given the lack of research in the field of mobile learning design, this chapter presented a framework for consideration using learning objects defined as reusable learning objects and reusable information objects to align to pedagogically sound instructional practice. Embedded as well were instructional strategies most appropriate to achieve educational learning outcomes. The use of learning objects accomplished several tasks including design efficiency and consistency across designers. The most significant aspect of using learning objects for the design of mobile learning instruction is the ability to export the same instruction to other mobile devices and to learning management systems for online learning.

REFERENCES

Barritt, C., & Alderman, F. L. (2004). *Creating a reusable learning objects strategy: Leveraging information and learning in a knowledge economy*. San Francisco: Pfeiffer.

Bloom, B. S. (1956). *Taxonomy of educational objectives* (Handbook I: The cognitive domain). New York: David McKay.

Department of Defense. (2003). *DoD voluntary education*. Retrieved on February 10, 2007, from http://www.voled.doded.mil/voled_web/VolEdProgramScope.htm

Gagne, R. (1985). *The conditions of learning* (4th ed.). New York: Holt, Rinehart & Winston.

Gagne, R., Briggs, L., & Wager, W. (1992). *Principles of instructional design* (4th ed.). Fort Worth, TX: HBJ College Publishers.

Kane, K. (2004). *Maryland online. Quality matters.* Sponsored in part by the Fund for the Improvement of Postsecondary Education (FIPSE), U.S. Department of Education. Retrieved on February 10, 2007, from http://www.qualitymatters.org/index.html

Keller, J. M. (1987). Development and use of the ARCS model of motivational design. *Journal of Instructional Development, 10*(3), 2-10.

Kulhavey, R. W., & Wager, W. (1993). Feedback in programmed instruction: Historical context and implications for practice. In J. V. Dempsey & G. C. Sales (Eds.), *Interactive instruction and feedback*. Englewood Cliffs, NJ: Educational Technology Publications.

Merrill, M. D. (1994). *Instructional design theory.* Englewood Cliffs, NJ: Educational Technology Publications.

Moore, M. G. (1989). Three types of interaction. *The American Journal of Distance Education, 3*(2), 1-6.

Moore, M. G., & Kearsley, G. (1996). *Distance education: A systems view*. Belmont, CA: Wadsworth Publishing.

Mory, E. H. (1992). The use of informational feedback in instruction: Implications for future research. *Educational Technology Research and Development, 40*(3), 5-20.

Northrup, P. T. (2002a). An initial investigation of online learners' preferences for interaction. *Quarterly Review of Distance Education, 3*(2), 219-226.

Northrup, P. T. (2002b). A framework for designing interactivity into Web-based instruction. In A. Rossett (Ed.), *ASTD's e-learning handbook: Best practices, strategies and case studies for an emerging field*. Upper Saddle River, NJ: McGraw-Hill.

Sloan Consortium. (2005). *Growing by degrees: Online education in the United States.* Retrieved February 10, 2007, from http://www.sloan-c.org/publications/survey/pdf/growing_by_degrees.pdf

This work was previously published in Learning Objects for Instruction: Design and Evaluation, edited by P. Northrup, pp. 140-158, copyright 2007 by Information Science Publishing (an imprint of IGI Global).

Chapter 4.10
Using Mobile Phones and PDAs in Ad Hoc Audience Response Systems

Matt Jones
University of Waikato, New Zealand

Gary Marsden
University of Cape Town, South Africa

Dominic Gruijters
University of Cape Town, South Africa

ABSTRACT

This chapter investigates how to create ad hoc audience response systems using nonspecialist devices. The chapter revolves around two case studies: one involving the use of mobile phones, and the other based on PDAs. Both case studies are carried out in tertiary education institutions, showing how these devices can be used to facilitate audience participation using devices that students might, themselves, bring to lectures. Both are evaluated from the perspective of the student and the educator, using a mixture of observational and interview-based techniques.

INTRODUCTION

Anyone who has given a talk or lecture to a large audience will be well acquainted with the uncomfortable silences, embarrassed glances, and nervous shuffling that greet requests for audience participation. This anecdotal evidence is supported by survey findings presented by Draper and Brown (2004), indicating that if a lecture class is asked for a verbal response, 0% to 3.7% of students are likely to respond: even for the less exposing, "hands-up" response style, the participation rate might also be a low 0.5%-7.8%.

Not all audiences are so shy, though. In the late 1990s, the television game show, "Who Wants to

Be a Millionaire?" attracted large, viewing numbers throughout the world. As part of the game format, the contestant could "ask the audience," getting each member to answer the multichoice question using a handset.

Draper and Brown have taken similar handsets out of the TV studio and into the classroom. In Draper and Brown (2004), and an earlier paper (Draper, Cargill, 2002), they present pedagogic motivations for their work, which we share, and will not elaborate on here, beyond noting the value of interactivity and engagement between the learners (students) and the learning-leader (lecturer).

In a long-term, extensive study, summarized in Draper and Brown (2004), the personal response system they used for multiple-choice questions (MCQs) was seen as being of benefit: for example, 60% of 138 first-year computer students rated the system "extremely" or "very" useful; and, similar responses were seen in other disciplines as varied as medicine and philosophy. Handsets are also likely to increase the participation levels: when asked whether they would work out an answer if asked to vote using the system, between 32%-40% agreed.

Of course, specialized handsets have many advantages such as providing simple, direct ways for students to respond (they just press a button): however, there are some drawbacks, including large costs involved in providing handsets ubiquitously, for every student and every lecture; organizational-overheads (e.g., handing out and collecting handsets); and, the impoverished range of responses possible (a single selection for MCQ use).

Inspired by Draper and Brown's experiences, we sought to address these sorts of drawbacks by using a technology that most students now carry with them to every lecture — the mobile telephone. We were interested in whether the pervasiveness and easy familiarity students have with this technology would allow it to serve as a replacement for the purpose-built handsets. Furthermore, we wanted to explore the possibilities beyond MCQs such as students sending free-text questions or, perhaps suggestions and comments to the lecturer. Although other researchers have considered the use of mobile phones in a university setting, for example (Cheverst et al., 2003), we believe this to be a novel application.

Mobile phones are becoming increasingly sophisticated, with a number of current models, sometimes termed "smartphones," providing the sorts of functionality, such as web browsing and document editing, and wireless connectivity, like Wi-Fi and Bluetooth, as well as conventional mobile telecom networking, seen on the handheld personal digital assistants (PDAs). In light of these technological advances, we developed MISPE — the mobile information sharing in the presentation environment, to explore future interaction possibilities for audiences.

The use of personal technologies, like advanced mobile phones and PDAs, has the potential to help all students play a more active role in their education experiences. For people in developing countries though, for example those in South Africa or India, the mobile is a "bridging technology" that can span the digital divide (Marsden, 2003). In these contexts, access to traditional information technology is limited: meanwhile, in South Africa, for instance, over 40% of the population owns a cell phone (rising to 70% for Europe). Staggeringly, over one billion people worldwide own a GSM handset!

In this chapter, we present our experiences in terms of two case studies: the first involves the use of mobile phones to enable the audience to give real-time feedback and responses; the second considers the role of an ad hoc network consisting of the audience's personal technologies, and the lecturer's computer, using MISPE. We discuss both technology issues such as infrastructure requirements and limitations, as well as others relating to the users' experience.

CASE STUDY: TEXT MESSAGING

While the specialized handset studies provided us with a very useful set of functional and non-functional possibilities, we decided to also run some sessions bringing together a group of eight experts in both human-computer interaction and education (all of which were also lecturers), to brainstorm requirements. In the process, we developed scenarios such as this one:

Dr. Monday begins her lecture on advanced linguistic analysis to 300 first-year students. "Before we go any further, are there any questions about last week's topic? Send me a text now from your mobile phone to 444." After a minute, Dr. Monday checks the computer display and sees there are 25 questions, listed in the order they arrived: she can reorder the list alphabetically and by size of message as well. She selects one of the questions to answer.

Later in the lecture, Dr. Monday wants to test the students' understanding of "focus." "Here's a quick quiz," she says. "If you think focus is related to the subject, text 1 to 444; if you think it is related to the topic, text 2; and if you think it is related to the verb, text 3 to 444." Moments later, Dr. Monday can display a bar chart showing the students what the most popular choice was. "Most of you are wrong," she says, wryly, "The correct answer is 2 — the topic."

Several times in the lecture, Monday asks the students to text their current "happiness level": "send a text message to 444 now to show how well you understand the lecture so far," she says, "enter H followed by a number from 0 to 9, where 0 is the worst." She can view the changing level of "happiness" over time as a line graph.

After the lecture, Monday returns to her office, and can access all the questions sent by students: she can also review the bar charts for each multiple-choice question, and see the "worm" trace plotted over time. All this information helps her review the lecture content, and plan for next week's session.

Such discussions clarified some of the additional forms of interactivity mobiles might provide over specialised handsets:

- allowing multiple responses to an MCQ, for example, "choose 2 of the 5 features listed below";
- parameterised responses, for example, "text your answer (1-5) and how confident you are in your answer (0-100%)";
- open-ended "conversations" between the lecturer and audience; and
- as an active, lecture-experience feedback device.

Pilot-Study System

Before building a full-scale system tailored specifically to the lecture-context, we decided to acquire a third-party, commercial text-polling system to first explore the issues and feasibility of our ideas. The software chosen was the SMS PollCenter by Code Segment. (For information and a demonstration, see http//www.codesegment.com/) The system runs on a PC (we ran it on a laptop in the field studies), and also requires a mobile phone to be connected to the computer via a serial cable, so that sent text messages can be gathered. MCQ results can be displayed in a range of forms such as bar chart and a pie chart. The "SMS Chat" facility displays incoming texts in a scrolling whiteboard format. Software such as this has been used commercially to provide audience response facilities in a range of situations, including television programmes and conferences. Figure 1 illustrates the system in use.

Left-hand image shows a mobile phone being used to send user's response to the lecturer-posed MCQ: background shows lecturer and live

Figure 1. Pilot system use

results chart summarizing audience's overall response. Right-hand image shows free-form question: "How do I write a function in C++?" being entered on mobile phone: when sent, it is displayed in the system's "SMS Chat" window (which can be displayed to just the lecturer or the entire audience).

FIELD STUDIES

Initial Experience Gathering

In the first deployment of the system, we studied its use over six, 1-hour sessions spread over 2 months. Our aim was to gather impressions in a range of contexts, so we chose situations with different characteristics, and used the system in a variety of ways (Jones & Marsden, 2004). Three courses were involved:

- *A*: first-year programming class run in New Zealand (NZ);
- *B*: first-year programming class run in South Africa (SA); and
- *C*: a fourth-year human-computer interaction class in South Africa.

For courses *B* and *C*, we carried several trials, each separated by around a week. During each session, researchers set up and operated the system for the lecturer: they also observed the class interaction, and were involved in interviewing students at its end. In classes *A* and *C*, the authors were the lecturers — we wanted to experience the system from the front, as it were: two other lecturers were involved in presenting class *B*. Figure 2 shows the system in use in the first-year programming class in Cape Town.

In each session (e.g. session 2), there was one or more uses of the system (e.g., 2.1, 2.2). Questions were either factual (based on lecture content), or personal (eliciting subjective opinion). Text messages sent were either single selections relating to an MCQ, or free text (chat style). Messages/poll results were either fully visible (results shown during polling and dynamically updated), partially visible (final results shown at end of polling), or hidden (only the lecturer saw the messages).

A summary of each session, and use of the system within them, is shown in Table 1, along with data on the number of text messages received during each use. While this table gives some raw indications of interactivity, it is worth highlighting some of the specific behaviours and effects we noticed. First, 19% of all logged responses to MCQ style questions were in a form that was not recognized by our answer matching filters: for example, in Session 2.1, the students were asked to enter a single integer, but one sent "Turn 72 degqees" (sic). Second, on average, 10% of respondents sent more than one message in response to a question (either resending their initial response, or changing their vote). Third, in SA, 6% of all messages were spam (e.g., "Let the universe decide SMS "oracle" to 34009"); no

Figure 2. The pilot system in action at the University of Cape Town. Lecturer is discussing results of an MCQ poll (shown on the RHS display; the poll question is displayed on the LHS screen).

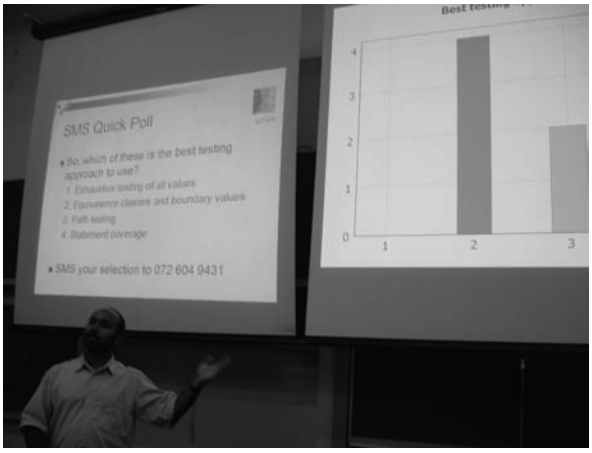


Table 1. Summary of sessions and system use

Session/system use	Course	Question type	Response elicited	Visibility	# people in class	#unique respondents (% of total)
1	A	factual	MCQ	Full	155	35 (23%)
2.1	B	factual	MCQ	Full	180	32 (18%)
2.2	B	personal	chat	Full	180	16 (9%)
3.1	B	personal	MCQ	Partial	150	17 (11%)
3.2	B	factual	MCQ	Partial	150	10 (7%)
4.1	C	personal	MCQ	Full	40	15 (38%)
4.2	C	personal	chat	Full	40	3 (1%)
5.1	C	factual	MCQ	Full	40	6 (15%)
5.2	C	personal	chat	Hidden	40	3 (1%)
6.1	C	personal	MCQ	Full	33	10 (30%)
Mean					101	15 (15%)

spam was received in NZ. Fourth, in most of the MCQ cases, as the lecturer discussed the results of the poll chart, additional messages would arrive —sometimes this was a mobile telephone network effect (5%-10% of messages were delayed), but there was also evidence of a "playfulness" as students attempted to "disrupt" the lecturer by altering the results.

At the end of each session, we asked for volunteers to remain behind and give feedback on the system. Overall, we spoke to around 50 people in this way. Views were consistent, in that students liked the idea of the approach (it gave them more of a role in the lecture, changed the pace of the session, etc.); strongly preferred the MCQ style of interaction over the chat scheme (as texting a freeform question could take too long, and the display of comments to the whole class could be distracting); but, they had concerns over the cost of sending messages (over and over again we were told "if sending a message was at a reduced rate, or free, I'd use it a lot more").

We also discussed the experience with the class *B* lecturers. They were less enthusiastic and more cautious about the scheme than the students. Their main concerns were the potential negative impacts of the technology on the "natural" flow of the lecture, and the need for more flexibility in the software to respond dynamically.

Longitudinal Study

Following this probing set of sessions, we carried out a more focused, longer trial during another course, using the most successful of the methods: multiple-choice questions where results are visible to all.

Over the first 5 weeks of a first-year programming course, one of us used the system to support lecture materials. The course included two lectures every week, and during one of the lectures, students were presented with an MCQ that they could answer using their mobile phone. On average, the number of people attending the lecture was 112 (which represented around 50% of those enrolled in the course). Table 2 presents the data for each weekly session.

During the sessions, we further observed the impact of the system on the lecture experience. Even though the numbers of people responding to the MCQ were low, there was a noticeable effect on the entire audience. The approach helped to strengthen the rapport between the lecturer and the students.

To gain a further perspective on the relative usefulness of the new system, we also deployed two other methods of gathering audience feedback during weeks two to five of the course. Each week, in the lecture that did not include the mobile MCQ system, we distributed post-it notes to each student, and asked them to write an anonymous, short comment, question, or suggestion to the lecturer. These were then collected up after the lecture when the students had left the room.

In addition, we used a Web-based polling system accessible to all the students in the course via their own or university computer. The same MCQ that had been asked in the class lecture via the mobile system was presented. Students could answer the question, and then the current results (showing the frequency of people selecting each choice) were then displayed. We recorded the number of unique respondents one week after each question was posed.

Both of these more conventional methods achieved higher participation rates than our new approach: on average, both achieved around 27% of the total number of possible respondents, where the total number in the post-it note case was the number of attendees in the lecture, and in the Web poll context, the total number of enrolled students.

Discussion

The results suggest that using the handsets to SMS responses to MCQs could improve the level

Table 2. Week-by-week response rates to MCQ used in first-year programming class

Week #	people in class #	unique respondents	response rate
1	110	16 1	5%
2	105	5	5%
3	110	19 1	7%
4	110	12 1	0%
5	126	8	6%
Mean 1	**12 1**	**2**	**11%**

of participation: in the initial study we saw a response rate of 7%-38% (much higher than that predicted by Draper and Brown for "hands-up"). The system was most successful when the results were always on display to the students (from the start to the end of the poll): we discovered that students liked watching their messaging change the display dynamically.

Even when the messaging rate was low, the technique appeared to have a positive impact on the lecture experience: the sessions became more participative, with the lecturer engaging the students in a discussion of the poll results, for instance.

While a novelty effect might well have been in play, in the initial study the response rate seen in 6.1 (30%) compares favorably with that of the earlier session for that class (4.1 (38%)), even though the second session took place approximately 1 month after the earlier one. In the second study, as the weeks went by, there were fluctuations in response rate, but we did not detect a steadily decreasing pattern of enthusiasm. Given Draper and Brown's experience, we predict the enthusiasm for the approach will grow, particularly if charging issues can be resolved (e.g., by providing free texting for students).

The "chat" form of interaction was disappointingly received in the initial study (and we did not proceed with it in the second study). However, we intend to explore this form further with a tailored system, as its potential was undermined by the constraints of the pilot system (e.g., lack of filtering or censoring facilities for the lecturer). Another area for potential was discovered in the form of interesting emergent "community" behaviour when the chat screen was visible to all students: as well as communicating with the lecturer, students posed questions to *each other*, and received replies from within the audience. While there is much exciting work on mobile communities for noncollocated people, this experience suggests there is some useful work to be done on supporting *immobile* mobile communities, such as crowds in football stadia.

Unlike when using specialized handsets in closed networks, designers of mobile phone-based response systems will have to accommodate "invalid" input, both from the users and spammers. In setting up software to process student MCQ responses, for instance, the aim should be to accommodate the variety of answer messages likely to be sent (e.g., "1," "one," "the first choice").

While the more conventional feedback methods used in the second study led to greater participation, they did not, however, foster higher in-class interaction.

CASE STUDY: PDAS

One of our motivations for using mobile phones as the basis of an audience response system, rather than the purpose-built handsets seen elsewhere, was to consider the richer forms of interaction they might facilitate. The previous case study, for instance, illustrated some potential roles of text messaging.

The study also showed, though, that entering responses, particularly free text questions or selections, can be problematic due to the impoverished text-entry facilities of conventional handsets. PDAs and advanced phones provide more sophisticated input and output facilities to the user — stylus-based handwriting recognition and larger, higher resolution displays, for example.

To consider the potential for these emergent, more advanced sorts of personal technology, we built the mobile information sharing in the presentation environment that connects students and their PDAs to the lecturer, with their laptop machine, in a lecture setting. In the sections that follow, we focus on describing the usage scenario and evaluation of the approach: a detailed discussion of the architecture and implementation can be found in Fry, Gruijters et al. (2004).

Usage Scenario

Students arrive at a lecture with their PDA or smartphone — these devices are the "clients" of the system. The lecturer turns on his or her laptop, and uses it to run the "server" application. Students then connect to this local, lecture server, wirelessly, the system supporting both Wi-Fi, and the shorter-range Bluetooth protocol.

The class begins, and as the lecturer presents the materials to the students in the form of a slide presentation, their devices receive the slides. Once the slide has arrived on a student's PDA, they may annotate it with pertinent information in much the same manner as they would usually annotate hard-copies of slides. At any point, the student can also write a question that is directly displayed on the lecturer's laptop.

As the lecture proceeds, the lecturer presents questions — eliciting free-form text answers — and MCQ polls for the audience to respond to, watching the results appear in his/her slideshow in real time (see Figure 3 for an example).

Once the class is over, the lecturer can save all information created during the class (student answers, student questions, and voting results) enabling them to reflect on the impact of the presentation on its audience.

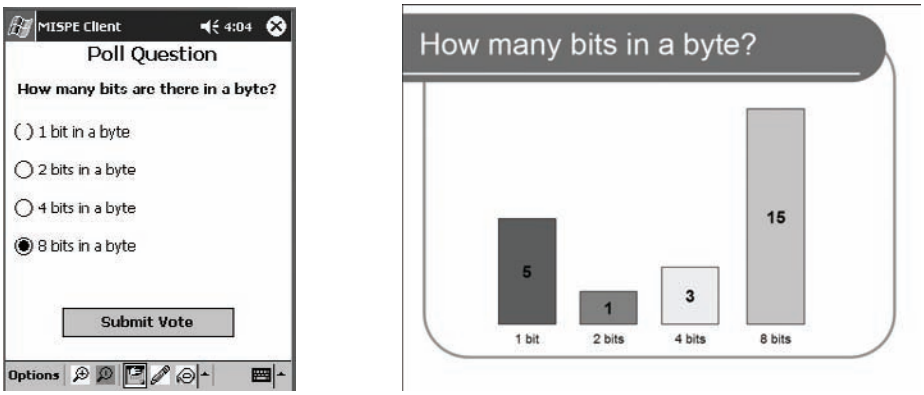
The image on the left shows the display on the student PDA. This screen is generated automatically, without the lecturer having to explicitly place controls on the screen. Instead, the lecturer works through a dialog box configuring the type of question they require. The software then renders the question in the most appropriate way for the handheld device. The image on the right shows the results received from the PDAs. The lecturer's software is a shell wrapped around PowerPoint™, with audience response slides being created automatically.

Related Work

The Pebbles project has done a substantial amount of work in the areas of collaborative environments involving handheld computers and other devices (Myers, 2001). Among the Pebbles applications is the Slideshow Commander, which allows educators to run a PowerPoint™ presentation from his/her laptop while controlling the presentation from a PDA connected via a wireless link to the laptop. The PDA provides thumbnails of the slides available, and allows the educator to move freely between slides using their PDA.

A system that allows multiple people to share an interactive whiteboard using PDAs is described

Figure 3. MISPE client (PDA) and server (laptop) example views



in Myers, Stiel et al. (1998). This system allows multiple contributors to take turns to use a pen or drawing device on a virtual whiteboard, using PDAs to wirelessly access this single whiteboard display. PebblesDraw, as the system is named, allows users to take turns drawing on their PDAs, and having their contribution appear on a single, central whiteboard or display.

While these systems are useful in an educational context, they supplement a very small proportion of the educator's overall workflow. MISPE differs from the systems developed by Pebbles by designing and evaluating a system that addresses more of an educator's workflow than just presenting information.

In terms of authoring presentation material, Hexel, Johnson, (2004) describe a tool that allows presenters to create materials that can be customised for specific members of an audience. This tool provides a server that delivers customised presentation data, based on the specifics of the audience members, with relation to language, culture, or disabilities. The tool provides no authoring capabilities in terms of questions or voting. In addition, the system provides only one-way communication from lecturer to audience member.

The work by Hexel et al. (2004), and that reported in Webster and Ho (1997), provides evidence that the sorts of features seen in MISPE may enhance the educational experience. This work suggests that learners experience higher engagement in multimedia presentations which (1) are more challenging, (2) provide more opportunities for feedback, (3) allow more presentation control, and (4) vary in multimedia features.

MISPE uses ad hoc networking. Ad hoc networks are networks that are able to exist without a specific network infrastructure, and without a fixed network topology (Buszco et al., 2001; Doshi, Bhandare, 2002). They also do not require a central, authoritative body. This makes them suited to a highly mobile environment where network nodes may come and go as they please.

Evaluation

An initial, small-scale, user-centred evaluation of MISPE has been carried out to assess its usefulness and usability in real lecture settings.

Method

Two lecturers volunteered to use the system during a lecture. The class size for both lectures was small: five in the first lecture and six in the second. The mean age of the student participants was 18, and all reported having moderate levels of computer experience (the course they were participating in was computer-related).

Researchers attended both sessions to carry out naturalistic observations: that is, to record impressions of how the system impacted on both the lecturer and students during the lecture itself. The observers recorded any unusual uses of the systems, as well as common trends of normal use. They were asked to collect data on the features most used, comments made, and overall user reactions to the system. They were also asked to note how the students' interaction with the technology affected their participation in the lecture: were they able to effectively make annotations, and ask and answer questions, while still listening to the lecturer, or did they fall behind, losing the flow of the content?

Observations of the educators specifically noted the normal teaching style and actions of the educator, as well as when and how the educators interacted with the system.

At the end of both sessions, the lecturer and participants were questioned about the usefulness and usability of the system.

IMPRESSIONS AND FEEDBACK

Lecturers' Perspective

Overall, the response of the two lecturers was enthusiastic: they were keen to use the technology again. In terms of the audience response features provided, they rated the facilities that allowed them to pose questions, and to carry out in-class votes, as the most useful.

In terms of student submitted comments and questions during the lecture, however, the usefulness of the system was hampered by the limited way the system accommodated the dynamic, "performance" nature of lectures. The technology, then, was seen to disrupt the natural flow of the lecture: lecturers would stop interacting with the class for periods of between 10-15 seconds, as they focused on their computer in an attempt to select a comment to discuss, or a student's answer to display to the rest of the class.

The immobile nature of laptop also caused problems: the lecturers often moved around the lecture room during the class. While they stood next to the computer when they wished to control the slideshow, they would often move away from the system to talk about the new slide displayed. This made it difficult for them to view any audience feedback — spontaneous comments or questions — that occurred as they were speaking.

Students' Perspective

The students felt the system led to them being more engaged during the lecture. Being able to ask and answer questions, and to receive feedback using their own device, made them feel more personally involved. Most students used the system to submit many comments and questions during the lecture: they were uninhibited as, in contrast to conventional verbal question asking, they felt their comments would be less disruptive.

There were, though, two usage trends that may lead to a negative impact on the lecture experience. First, students "played" a lot with the system when the lecturer talked for a long while about a slide's content. When they became bored, that is, the system became a distraction. Second, despite the more flexible, easier to use text input methods seen on the PDA, compared to those of the mobile phones in the first case study, the authoring of slide annotations and questions still took too long, causing students to fall behind, and to lose the context of what the lecturer was discussing.

Discussion

As in the texting case study, the use of MISPE provides some evidence that personal technologies can enhance the audience's participation. From the lecturer's point-of-view, though, we need to design a better way for them to control and interact with the system when they are in full-flow of a lecture performance. Specifically, the system should accommodate the lecturer's mobility: as in the SlideShow commander (Myers, 2001), it would seem important that the lecturer has a handheld device that they can use to orchestrate the slideshow and audience participation. Simple interactions with this control device, for example, one-handed button presses, need to provide them with direct access to, say, incoming student comments.

The frustration observed by students as they wrote or answered free-text questions could be overcome by providing a much more lightweight way of authoring. In the present version of the system, submissions have to be entered as textual strings, either by the student tapping on the letters of an onscreen keyboard, or by using the handwriting recognition facilities of the device. A faster approach would be to allow users to use sketching facilities: a scrawled note or comment can be created with the image being sent the lecturer without any preprocessing.

CONCLUSIONS

Personal technologies — mobile phones, both conventional ones, and the increasingly sophisticated smartphones, along with wireless-capable handheld computers — offer the potential for increasing audience participation in lecture settings.

Unlike the special-purpose handsets used in other trials, these devices offer practical benefits (such as lower cost set-ups, flexibility in deployment, and richer forms of audience response) as well as less tangible impacts on the audience experience arising from the personal nature of the device itself. People relate to these technologies they own and carry everywhere in engaging ways.

In the text-messaging studies, we saw higher participation rates than might be experienced using traditional verbal or hand-show methods. While the response rate was not overwhelming, with, on average, 15% of the audience directly taking part, the impact on the overall lecture experience was significant. In our studies, students had to pay the cost of every message sent, and they indicated that if the service was free, they would more readily respond using it.

In the PDA trial, where all student messages were free, we saw a much higher level of question answering and comment giving. In the next several years, most mobile phones will be equipped with Wi-Fi and Bluetooth capabilities, so that messaging sending costs can be eliminated.

Lectures are often dynamic, lively performances with their own rhythms and flow. Lecturers and audiences will want to use a range of participation methods, both technologically-based and traditional: one moment, the lecturer will ask for a show of hands; the next for questions to be sent to a shared, digital whiteboard.

There is a need, then, for any audience response system to fit within this ecology of resources. Too often, technology fails to accommodate the context, fails to fit in: instead of playing its part in the ecology, it devours, in the process destroying the user experience. In our case studies, we saw some examples of the impact of suboptimal contextual design, but also suggested ways of improving later prototypes.

ACKNOWLEDGMENTS

Thanks to Hussein Suleman and Donald Cook, who set aside time in their lectures. Dave Nichols and Mike Mayo helped with the NZ observations, and the Waikato HCI group worked on scenarios.

REFERENCES

Buszko, D., Lee, W., & Helal, A. (2001). Decentralized ad hoc groupware API and framework for mobile collaboration. *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Groupwork* (pp. 5-14). ACM Press.

Cheverst, K., Dix, A., Fitton, D., & Rouncefield, M. (2003). Exploring the utility of remote messaging and situated office displays. *Proceedings of Mobile HCI 2003* (pp. 336-341). Springer.

Doshi, S., Bhandare, S., & Brown, T. X. (2002). An on-demand minimum energy routing protocol for a wireless ad hoc network. *Mobile Computing and Communications Review, 6*(3), 50-66.

Draper, S. W., & Brown, M.I. (2004). Increasing interactivity in lectures using an electronic voting system. *Journal of Computer Assisted Learning, 20*, 81-94.

Draper, S. W., Cargill, J., & Cutts, Q. (2002). Electronically enhanced classroom interaction. *Australian Journal of Educational Technology, 18*(1), 13-23.

Fry, B., Gruijters, D., & Reid, S. (2004). *MISPE - Mobile Information Sharing in the Presentation Environment.* Technical report CS04-22-00. Cape

Town: University of Cape Town, Department of Computer Science.

Hexel, R., Johnson, C., Kummerfeld, B., & Quigley, A. (2004). PowerPoint™ to the people: Suiting the word to the audience. *Proceedings of the Fifth Conference on Australasian User Interface* (pp. 40-56). Dunedin, NZ: ACM Press.

Jones, M., & Marsden, G. (2004). Please Turn ON your mobile phone: First impressions of text-messaging in lectures. *Proceedings of the 6th International Symposium on Mobile Human-Computer Interaction (Mobile HCI '04)* (pp. 436-440). Glasgow, UK: Springer.

Marsden, G. (2003). Using HCI to leverage communications technology. *Interactions, 10*(2), 48-55.

Myers, B. (2001). Using hand-held devices and PCs together. *Communications of the ACM, 44*(11), 34-41.

Myers, B. , Stiel, H., & Gargiulo, R. (1998). Collaboration using multiple PDAs connected to a PC. *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW)* (pp. 285-294). Seattle: ACM Press.

Webster, J., & Ho, H. (1997). Audience engagement in multimedia presentations. *ACM SIGMIS Database, 28*(2), 63-77.

This work was previously published in Audience Response Systems in Higher Education: Applications and Cases, edited by D. Banks, pp. 359-372, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global)

Chapter 4.11
Perception of Mobile Technology Provision in Health Service

Astrid M. Oddershede
University of Santiago of Chile, Chile

Rolando A. Carrasco
University of Newcastle-upon-Tyne, UK

ABSTRACT

In this chapter the user interface perception and resources for mobile technology (MT) support in health care service activities is investigated. Most procedures oriented to provide better operation and quality of health service depend on the existing information and communication technology (ICT) system. However, the implementation of new technology competes with funding available for health institutions resources, hence introducing them is complex. The technical difficulties encountered in using ICT are: an inadequate physical infrastructure, quality of service (QoS) issues, and insufficient access by the user to the hardware/software communication infrastructure. A case study by multi-criteria approach was investigated involving three categories of hospitals in Chile and empirical data was collected comprising diverse health sector representatives. The main contribution is the proposed research decision-making model using the analytic hierarchy process (AHP) to evaluate and compare information and communications systems as fixed, wireless, or computer-assisted provisions for health-related activities and to identify the high priority dimensions in a health care service.

INTRODUCTION

A communications and IS provides an essential role for health-related activities. Most of the actions oriented to improve the operation and the quality of health service depend largely on the level of the information available and a commu-

nication system. The demand for health services increases each year by diverse demographic factors; (growth, immigration, ageing, etc.), cultural (greater information and expectations), technological (new therapeutic and diagnostic procedures), business (aggressive marketing of new procedures), professional (induced demand, preventive medicine), and organization (information deficiencies and management). The provision and management of services in the health sector implies processing enormous quantities of economic, welfare, clinical, and administrative data impossible to carry out by manual procedures (Oddershede, Carrasco, & Ontiveros, 2006). Nowadays, health centers are committed to putting into practice actions to facilitate clinical care activities, to satisfy professionals' aspirations, and citizen necessities.

The health communication system has changed from sending simple messages point by point, for example laboratory results, to the creation of virtual electronic records (Del Llano Señarís, 2003). The use of great data bases to collect health, social, and economic data, developed at a cost that is a fraction of the previous costs, means that files concerning the health of millions of people can be useful to predict future health requirements in a given population and to consequently assign and prioritize resources.

A demanding task of current innovation in health care processes is to improve the time to treatment in view of the fact that appropriate medical intervention immediately following an emergency or urgent situation significantly increases the chance of recovery for the patient.

The health sector has only recently had access to advanced ICT and there is confidence that modern ICT can progressively improve their performance, although a weak telecommunications infrastructure has shown in the past to be difficult to implement any plans, and therefore offer a good quality service to fulfil user expectations. (Suh, Suh, & Baek, 1994)

An emerging concept for health care provision is mobile health (m-health), which includes mobile computing, communications, and multimedia technologies in order to provide better access (Chan, 2000). This new evolutionary research area will provide new patterns for health care (Istepanian & Lacal, 2003). This will make available resources for both the health care professionals and patients with an efficient, secure, ubiquitous, and robust infrastructure coupled with tools for the assessment and management of patient health status and the support of preventive programs. (Istepanian, Jovanov, & Zhang, 2004).

MT provides an easy information flow that has yet to be exploited to its full extent. Applications of mobile ICT and IS in health care can be recognized as both emerging and enabling technologies (Ammenwerth, Gräber, Herrmann, Bürkle, & König, 2003), which have been applied in several countries for either emergent care or general health care. For example, the variety of wireless technologies such as mobile computing, wireless networks and global positioning systems (GPS) have been applied to ambulance care in Sweden (Geier, 2003) and emergent trauma care in the Netherlands (Jan ten Duis, & Van der Werken, 2003). Relative information about the patient and the ambulance location can be transmitted to the hospital in real time. Then, the hospital can be well prepared for the arrival of the ambulance at any time. The challenge is to provide the appropriate treatment to the patient at the right time at the right hospital (Jan ten Duis, & Van der Werken, 2003). A system with secure mobile health care services has been tested in Finland, including health consulting, electronic prescription, and so forth. Authorized individuals can easily access the system via mobile devices such as mobile phones (Jelekäinen, 2004).

The recent expansion of mobile communications and computing technologies to support highly specialized health-related requirements has generated a substantial interest in understanding the factors related to accepting a suitable ICT and m-health system.

A number of researchers have studied user acceptance of MT and services including mobile Internet, text messaging, contact services, mobile payment, mobile gaming, and mobile parking services based on IS adoption models (e.g., Pedersen, 2002; Pedersen & Nysveen, 2003; Pedersen, Nysveen, & Thorbjørnsen, 2003). They found that usefulness and ease of use are very important factors to determine user acceptance of MT. Khalifa and Cheng (2002) found that exposure of an individual to mobile commerce (m-commerce) positively influences the individual's intention to adopt m-commerce.

However, at the present time, there are still some questions to be answered: How do contemporary health professionals accomplish their mission in highly mobile work settings? Does their distinct mode of mobility function characterize their work practices in relation to ICTs and MT? How will mobile communications change health care in the future? Will e-prescribing be adopted and accepted? How to integrate different mobile health care devices? What would be the areas of growth in the health care IT market?

Using as a reference frame other Latin American countries, the hospitals lack an appropriate methodology to measure the satisfaction of the patients. The used procedures are limited to fulfill the regulating minimum requirements. Nevertheless, the health subject has been a reason for controversy between the politicians, doctors, and other citizens of the countries (Alleyne, 1998). The mass media publish/broadcast innumerable articles that present the different opinions, impatience, and preoccupations of the population in relation to the health system (Colomer, 2002; Prados de Reyes & Peña Yánez, 2002). Doctors and health workers agree with these articles and suffer greater discontent, since they consider that the resources that are assigned to them are not sufficient to provide a good service. (Sosa, 2004).

The existing literature does not provide studies that reveal the derivations of MT value in health-related activities in terms of user satisfaction and user perception.

Nowadays, health systems are seen universally pressured to improve service as part of their activity. Therefore, the greatest challenge that health institutions face is to offer good service. However, measuring quality in health care systems involves competing goals to satisfy citizen necessities and to proportionate appropriate resources for health professionals aspirations (Oddershede &Carrasco, 2006). In order to provide a base for health service it is necessary that health organizations recognize clearly their objectives; also, the activities and tasks that move toward the achievement of the stated objectives; the necessary resources and viability projects; the priorities, calendars, and responsibilities; and the mechanisms of control and evaluation of the fulfillment of the plan of its suitability. (Are we going well?)

Customer satisfaction is an important consideration of service quality in health care organizations. From a management perspective, customer/user (patient, doctors, clinical researchers, etc.) satisfaction with their health system is important for several reasons: satisfied users are more likely to keep a consistent relationship with a specific health service (Oddershede, Carrasco, & Ontiveros, 2006). Through detecting sources of users' dissatisfaction, an organization can deal with system weak points. User satisfaction dimension adds important information on system performance (Strasser & Davis, 1991).

Accordingly, developing a model that manages to capture the perception of satisfaction from the perspective of health-related representatives (patients, physicians, professionals, clinical researchers, etc.) towards the ICT system and MT services offered by the hospitals constitutes an important contribution (Weinstein, Toy, Sandbergea, et al., 2001).

Studies made by the Pan-American Organization of Health and 34 companies (Organización Panamericana de la salud, 2001) related to infor-

mation on health in Latin America and the Caribbean show that the expectations in ITC, MT, and e-health is too high in many institutions.

In any health care system with limited resources, priorities for investment must be set on the basis of clear evidence of benefit to patients and good value for the money invested. The implementation of new technology competes for funding available to health services and simply introducing all of them is too expensive. This has encouraged the development of a case study that is being carried out in Chile to investigate user perception concerning ICT and MT in the health sector. The study is concerned with the quality of IT services and MT, as a means to evaluate service performance to assure customer satisfaction. The results will give information about user acceptance and will serve as a tool for medical decision makers in generating course of actions when facing critical and complex decisions, such as to create or change activities in budget allocations or for distributing resources according to user requirements.

The aim of this chapter is to provide a decision-making model to examine, compare, and evaluate user perception with regards to ICT and MT provision in health-related activities. The evaluation instrument will help health care users to identify the benefits of an adequate ICT network and to point out weaknesses in service. The information obtained will provide a starting point to analyze those ICT system parameters considered as more relevant for the QoS of ICT in a health care service**.** In addition, this pilot study will help the decision makers to decide on the courses of action required for resource distribution relative to user requirements.

This chapter describes the system in study and the methodology employed to reflect high priority dimensions concerning MT contributions to health-related activities. Empirical data was collected from three types of public and private hospitals in Chile that were categorized according to their ICT/MT system provision. The case study results will answer three important questions: Who are the main ICT/MT agents/users in health-related activities? What are the main activities they are involved in and now require ICT/MT support? Which is the more important ICT/MT system support for each activity?

A multi-criteria approach is applied for formulating a model by the use of the AHP (Alexander, Biggers, Forman, & Schleicher, 1990; Oddershede, Soto, & Carrasco, 2001; Saaty, 1990). The method developed by Saaty (1990) is used to state criteria and rank user preferences. The second section describes the problem faced by the user in the case study. The third section introduces notions of the AHP approach to solve the problem. The fourth section presents a simplified hierarchical decision model based upon human expert's knowledge and experience. The results given in the fifth section generate information that is not currently available. In the final section, the conclusions are provided.

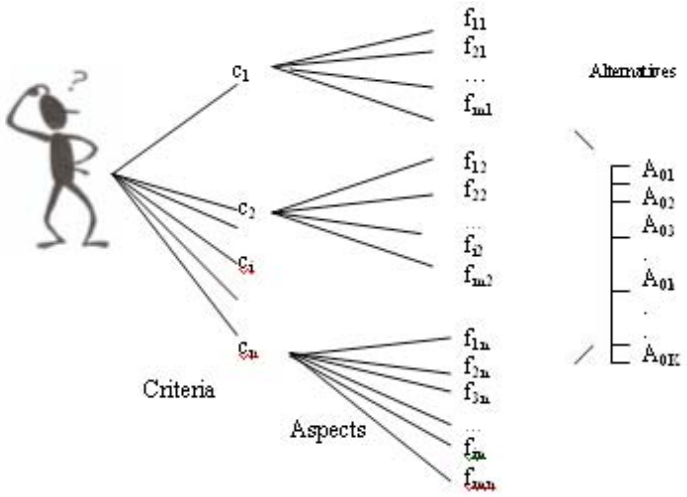
SYSTEM DESCRIPTION

This section outlines a framework for the development of a decision model regarding the MT and ICT system support for health-related activities.

Any health care requirement and service involves a wide range of participants, activities, resources, technologies, and others linked together within a complex environment.

The system under study is shown in Figure 1. Consider a user/consumer that has a health service requirement and there are many modus operandi that would, partially or totally, satisfy this demand. The user may desire many different features to improve the health care system. Moreover, the institutions will offer a variety of features showing differences in type, quality level, performance, cost, and so forth. These features rarely match with user requirements and depend on the user perspective. Many doubts need to be clarified such as, which features would be offered? Which

Figure 1. The system model



feature would be more important for satisfying a particular service demanded? Which features are the users willing to ignore? How much are they willing to sacrifice? Hence, the goals surrounding a health care system are complex and conflicting since they are different in each subsystem. As a result, the user is faced with a decision-making process that is obliged to select a course of action within an uncertain environment.

Figure 1 illustrates the problems to be studied The set of criteria stated by the users to attain a specific goal, is denoted as $\vec{C} = \{c_1, c_2, ..., c_j, c_n\}$ where c_j denotes the user perspective through criteria ***j*** and ***j***=1, 2…n (i.e., cost, quality of service, etc.) .The system in study could be analyzed through ***n*** different criteria.

Taking into consideration that the actions surrounding health care requirements will rely to a greater extent on the ICT system provision, we can classify diverse users and an assortment of activities to be performed. Each activity, type of user, and necessity will have different requirements for ICT and MT provision.

For the implemented system in Figure 1 the criteria will be considered from the different MT and ICT system user perspectives. Also, for each criterion there are many features or aspects that could contribute to the achievement of the main goal stated according to the criteria considered. The set of elements related to each criteria is $\{(f_{11}, f_{21}, ..., f_{ij}, ..., f_{mn})\}$, for this case $\boldsymbol{f_{ij}}$ represents the element ***i*** that contributes to the achievement of the goal stated according to the criteria ***j*** and ***i*** = 1,2……..m.

The set of all the alternatives or course of action that will lead in some extent to the achievement to the goal expressed by the users is denoted by $\{(A_{01}, ..., A_{0K})\}$.

In Figure 1, if we consider physician perspective criteria, he/she will desire to have access to the best MT and ICT system to perform a particular assignment. Then the aspects could represent the activities related to meet the assignment. The alternatives would be the optional ICT and MT system that provides support to achieve a specific requirement. If we consider a patient's perspective their needs would be different, so we will have other criteria to take into consideration. An analogous assessment comprises the other system users.

In view of the fact that modern ICT and MT can support the operation and quality of health service, the participants want the system to offer attributes that make their commitment more efficient and

satisfactory. Furthermore, the health center ICT system and MT system could rely on attributes that do not often agree with the user requirements (promptness, quality level, performance, cost, and others) for a specific assignment. At this point, the questions are related to the appropriate ICT and MT attributes for the execution of each activity. Do ICT and MT QoS have an effect on each activity? In what activity does the ICT support have more impact? Is MT important for health requirement and at what cost?

In general, from the ICT and MT QoS-user perspective, they expect fast, reliable, and easy access to online resources, applications, and Internet (databases, e-mails, voice, file transfer, browser, etc.). The QoS expectations will depend upon the area of use.

Having in mind that is not possible to satisfy the entire requirements simultaneously for all the participants, the omission of some characteristics in preference to others will occur (Oddershede, Carrasco, and Soto, 2005). Modeling under the existence of multiple conflicting objectives and subjective judgements becomes more complex (Birch & Gafni, 2003; Clemens, 1998). The multiple-objective decision method AHP is appropriate to help state criteria and rank user's preferences.

ANALYTIC HIERARCHY PROCESS APPROACH FOR MT SYSTEM PROVISION FOR HEALTH CARE

This section firstly reviews the AHP approach. Following the case study developed in the National Health Service in Chile the AHP process is applied to the collected data. It is initially structured to facilitate the empirical evaluation for obtaining results.

The Analytical Hierarchy Process Approach

The AHP engages decision makers in breaking down a decision into smaller parts, proceeding from the goal to criteria to sub-criteria down to the alternative courses of action. Decision makers then make simple pair-wise comparison judgements throughout the hierarchy to arrive at overall priorities for the alternatives. The decision problem may involve social, political, technical, and economic factors. This approach (Clemens, 1998; Saaty, 2001) provides the structure and the mathematics for helping decision makers make rational decisions. A rational decision is one which best achieves the multitude of objectives of the decision maker(s), (Claxton, Sculpher, & Drummond, 2002; Saaty, 1998). The three basic principles of AHP are:

1. **Hierarchy representation and decomposition:** A hierarchy is a representation of a complex problem in a multilevel structure whose first level is the goal followed successively by levels of factors, criteria, and sub criteria, and so on down to a bottom level of alternatives. Figure 2 shows an illustration of a simple three level hierarchy.
   The object of a hierarchy is to assess the impact of the elements of a higher level on those of a lower level or alternatively the contribution of elements in the lower level to the importance or fulfillment of the elements in the level above. This type of assessment is usually made by paired comparisons responding to an appropriately posed question eliciting the judgement. The mathematical definition of a hierarchy is given in Saaty's (2001) book.
2. **Priority discrimination and synthesis:** Setting priorities in a hierarchy requires that we perform measurements throughout the structure. We must then synthesize these measurements to obtain priorities for the

Figure 2. A three level hierarchy

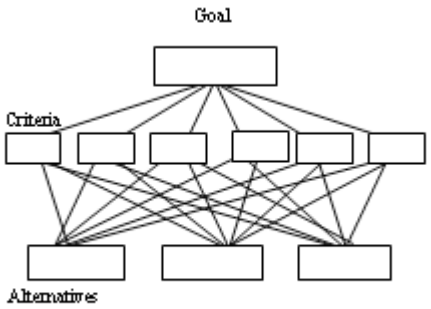


Table 1. The fundamental scale

Importance Intensity	Definition
1	Equal importance
3	Moderate importance
5	Strong importance
7	Very strong or demonstrated importance
9	Extreme importance
Reciprocals of above	If activity i has one of the above nonzero numbers assigned to it when compared with activity j, then j has the reciprocal value when compared with i

bottom level alternatives. The AHP is based on ranking activities in terms of relative ratio scales. In the paired comparison approach of the AHP, one estimates ratios by using a fundamental scale of absolute numbers in comparing two alternatives with respect to an attribute and one uses the smaller value as the unit for that attribute.

To estimate the larger one as a multiple of that unit, assign to it an absolute number from the fundamental scale. This process is done for every pair. Thus, instead of assigning two numbers w_i and w_j and forming the ratio w_i/w_j we assign a single number drawn from the fundamental 1-9 scale to represent the ratio (w_i/w_j) : 1. The absolute number from the scale is an approximation to the ratio w_i/w_j. The derived scale tells us what the w_i and w_j are. This is a central observation about the relative measurement approach of the AHP and the need of a fundamental scale. The scale of absolute values for judgments is shown in Table 1 (Saaty, 1998).

Let W be a matrix (1) whose row elements are ratios of the measurements w_i of each of n items with respect to all others.

$$W = \begin{bmatrix} w_1/w_1 & \dots & \dots & w_1/w_n \\ w_2/w_1 & & & w_2/w_n \\ \dots & & & \dots \\ w_n/w_1 & \dots & \dots & w_n/w_n \end{bmatrix} \qquad (1)$$

A number in the matrix is a dominance judgment. A judgment of 1.0 means that two activities contribute equally to the objective or goal, a judgment of 3.0 means that experience and judgement slightly favor one activity over another or three times as much (if you are dealing with measurable), a judgment of 5.0 means that experience and judgement strongly favor one activity over another, a judgment of 7 means that activity is strongly favored over another; its dominance is demonstrated in practice and 9.0 means that the evidence favoring one activity over another is of the highest possible order of affirmation (nine times as much). The elements should be grouped into homogeneous clusters so that it is not necessary to use a number larger than 9. In this way, we can interpret all ratios as absolute numbers or dominance units.

3. **Logical consistency:** The AHP provides guidelines for a test of consistency of judgments to ensure that elements are grouped logically and ranked consistently according

to a logical criterion. In order to achieve logical consistency it helps to make sure items are grouped into clusters of similar ideas or objects are grouped according to some form of homogeneity.

An inconsistency ratio is calculated for each set of judgments. Inconsistency follows the transitive property, for example, if you were to say that A > B, and B > C, then say that C > A, you would have been inconsistent. The Inconsistency Index, not ratio, is calculated for each node (and its cluster of children), and multiplied by the priority of the node, and summed over the entire model. A similar calculation is done for the Inconsistency Index for random judgments. The *overall inconsistency ratio* is the ratio of these two weighted sums. It has been shown that for any matrix small perturbations in the entries imply similar perturbations in the eigenvalues; thus the eigenvalue problem for the inconsistent case is:

$$Aw = \lambda_{max}\, w \qquad (2)$$

where the vector *w* is the eigenvector corresponding to the maximum eigenvalue, λ_{max} which will be close to n (actually greater than or equal to n) and the other values of λ will be close to zero. The estimates of the weights for the activities can be found by normalizing the eigenvector corresponding to the largest eigenvalue in the previous matrix equation.

The closer λ_{max} is to n, the more consistent the judgments. Thus the difference, (λ_{max}-n), can be used as a measure of inconsistency (this difference will be zero for perfect consistency). Instead of using this difference directly, Saaty (2001) defined a consistency index (CI) as: (λ_{max}-n)/ (n-1) since it represents the average of the remaining eigenvalues. In order to derive an accurate interpretation of either the difference or the consistency index, Saaty simulated a very large number of random pairwise comparisons for different size matrices, calculating the consistency indices, and arriving at an average consistency index for random judgments for each size matrix (Saaty, 1990). He then defined the consistency ratio as the ratio of the consistency index for a particular set of judgments to the average consistency index for random comparisons for a matrix of the same size. Since a set of perfectly consistent judgments produces a consistency index of 0, the consistency ratio will also be zero. A consistency ratio of 1 indicates consistency akin to that which would be achieved if judgments were made at random rather than intelligently. This ratio is called the inconsistency ratio, since the larger the value, the more inconsistent the judgments.

The consistency of a hierarchy is obtained by first taking sums of products of each consistency index, with the composite priority of its criterion. Then the ratio is formed from this number with the sums of the products of the random consistency index for the order matrix with the composite priority of its criterion (Saaty, 2001).

In general, the ratio should be in the neighborhood of 0.10 in order not to cause concern for needed improvements in the judgments. Too great a departure from the perfectly consistent value indicates a need to improve the judgments or to restructure the hierarchy.

THE CASE STUDY

A preliminary case study considering a number of public and private hospitals was carried out in Chile.

Health Care Service in Chile

Chile is a unitary state with a democratic government. The population considered for 2000 is of 15,211,308 inhabitants, 85% of which live in urban areas. The Chilean Health Service is organized as a mixed system including public and private health care institutions. It combines a scheme of social security with a system of insurances of competitive character. Nevertheless, these two components share a source of financing that is the obligatory contribution of the wage-earning workers (7% of its taxable rent), with a fixed-limit amount. The public expenditure is around $220 per capita. The public sector offers 196 public hospitals where 20 are of high complexity. In the private sector there are 19 complex hospitals and 216 clinics or hospitals of low and intermediate complexity (Organización Panamericana de la Salud, 2002). If we refer to the quality perceived through surveys of opinion the public sector users show that the dissatisfaction areas are concentrated on the patient-service relationship, deficient environment, and shortness of technology, while the private sector is dissatisfied with high expectations of the system and long delays in waiting rooms. The greatest challenge in the future consists of facing the demographic and epidemic changes. ICT and MT system development offers a crucial function to increase efficiency in health.

The Process

The system of Figure 1 involved three stages. The first stage is concerned with the identification of decisive factors and attributes that user/client (patient, physician, administrative staff, etc.) consider important in evaluating the quality of ICT and MT provision in health-related operations. This task will need to recognize the concerned participants (patient, physician, administrative staff, medical researchers, others.). Individually each participant will have different expectations about the health care system and possibly will desire many different characteristics to provide a health care system. Empirical data were collected from different types of hospitals and as a result, a large number of factors arose. The critical impacts of undertaking certain activities are identified. This practice allowed us to specify the criterion and to structure the problem situation. Once criteria and elements involved are identified, the next stage is to prioritize the different attributes by implementing a multiple criteria method. A comparison process is then carried out.

The main participants or agents are: the *patients* who demand prompt medical assistance, without delay, with precise, safe, and confidential information on their state of health. Moreover, they request updated information on therapeutic or preventive options; benefits and risks; efficiency of the services; and so forth. The *professionals* require the information on their patients, including that elaborated on by other professionals and corresponding to complementary tests, instantaneously and at the place of the attendance. In addition, they need to use management tools to deal with the information in order to reduce paper work. Furthermore, they need information that allows them to evaluate its effectiveness and to practice the clinical management that the administrator requests, that is, tools of revision and data processing on its own results and costs. The *research professionals* demand better access to the specialized bibliography, guides of clinical practice, protocols, and the opinion of other colleagues and the opportunity to value and to discuss this information to optimize their individual and collective practice. Finally, *administrative personnel* demand equipment and means for making their work more efficient (updated software for billing, sending test results, etc.).

For this study, potential and current ICT/MT system users were organized into four groups: (1) a group constituted by the *patients* who would demand a health care service, (2) another group conformed by the *clinical care professionals*

(physician, nurses, paramedics, etc.), who would make use of ICT to deliver a health care service, (3) a group represented by those users who develop *medical research*, collecting disease statistics and/or investigate new drugs and new devices; and (4) a group that is integrated by users who perform the *administrative* activities: billing, products distribution, and inventory control.

Data were collected from three categories of private and public health centers and hospitals which differ in resources, complexity, and infrastructure. The hospitals were classified according to their MT and ICT network infrastructure and resources availability. The three categories consisted of a group of hospitals with well provided ICT support, the other group constituted of hospitals that rely on scarce ICT support to perform its activities, and another group with intermediate ICT resources.

A team of experts consisted of representatives from each of the three categories of health center and hospitals, six from a metropolitan region and three rural hospitals. The group of participants from each hospital includes users of the four groups indicated previously (patients, clinic care professionals, medical research, administrative personnel).The total number of participants adds up to 480 and their ages ranged from 20 to 70 years old.

When criteria, factors, and the main representatives or agents are stated, a hierarchical structure incorporating quantitative and qualitative variables and their relationships identifying critical categories at each level Guidance was obtained through the judgments issued by the opinion poll.

This allows the implementation of an evaluation method to rank the different factors and elements considered in the hierarchy based on the agent's judgements. A comparison process is carried out derived from criteria and user preferences to prioritize MT and ICT system support to health-related activities. The experts' judgment was based on their own expertise and knowledge.

The final step involves applying the weights to the measured attributes of each activity to derive a ranking about the value of ICT support for each activity that would bring about a service improvement.

Structuring the Problem Situation

A three level hierarchical structure model has been designed. Each level has multiple nodes against which the alternatives on the next level are compared.

The first level is concerned with the global objective that needs to be obtained. For the situation studied, it is to identify the level of significance MT and ICT support is used in health-related activities.

The other levels and nodes represent the decision factors that have contributed to attain the goal. In this case the main agents are the users of the system and the main activities they perform.

The levels are represented as follows:

- Level 0 stand for the global objective indicated as, "ICT and MT system significance in Health care."
- Level 1 takes into consideration the implicated agents' perspective.
- Level 2 comprises the activities performed by the agents and would have an effect on each of them.
- Level 3 consists of the alternative ICT system that each activity depends on.

For this case, the alternative ICT systems were classified into four main groups labeled as: (1) fixed system (phone, fax, and office), (2) wireless system (wireless communications devices, mobile phones, radio communication devices), and (4) computer-assisted system (computer-assisted network communications).

Figure 3. Hierarchic structure

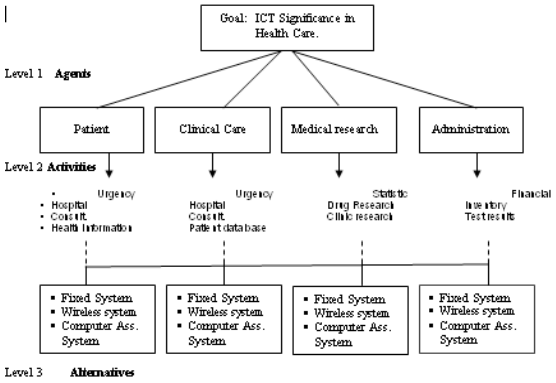


Figure 3 shows the basic and initial structure, which is a realistic simplification of a larger developed hierarchy.

An expert panel consisted of representatives from each of the main agents involved within each type of hospital considered. The main agents involved with the ICT support in health-related services were the patient; the personnel and professionals implicated in giving clinic assistance; the clinical researchers; and the participants involved in administrative tasks within the hospital. The categorization for health institution was made according to the ICT resources each hospital could rely on.

In agreement with the described basic hierarchic structure, a pairwise comparison was made in such a way that all the elements of a same level are compared and weighed to each other. The expert panel went through the hierarchical structure and derived a priority matrix for each level. The numbers in the matrix of (1) express the intensity dominance of the criterion in the column heading over the criterion in the row heading. The ratio scale of the matrix is reciprocal, the numbers which are symmetric with respect to the diagonal are inverses of one another, $a_{ij} = 1 / a_{ji.}$ In general, n (n-1) / 2 comparisons are needed if n is the number of elements being compared in the triangle above the diagonal of ones.

Table 2. The pairwise comparison process

Abbreviation	Definition
GOAL	ICT and MT system significance in health care.
Patient	Patient perspective criteria
Clinical care	Clinical perspective (physician, nurses, paramedics, etc.)
Medical research	Medical researcher group perspective
Administration	Administrative agent perspective
Urgency	Urgency-related activities to receive deliver health assistance
Hospitalization	Hospitalization and/or surgery requirement related activities
Consultant	Control and treatment related activities
Statistic	Statistic disease information and statistic records related activities
Drug research	Clinic and drug research and new devices investigation related activities
Financial	Administrative and financial activities
Inventory	Clinical needs distribution, supply, and inventory control actions related activities
Test results	Clinical and test result delivery, internal, and inter-institution communication activities
Fixed system	Includes fixed network communications (phones, fax, and office)
Wireless system	Wireless devices, mobile phones, radio communicationsdevices)
Computer-assisted system	Computer-assisted activities, Web, IP, browsing, and so forth

The judgements are entered into the matrix in response to the question: How much more important is one criterion on the left side of the matrix when compared with another at the top of the matrix to justify a fair decision selection? When a criterion is compared with itself, it is of equal importance and is assigned the value 1. Once all the pairwise comparisons of the group are completed, a scale of relative priorities is derived from them.

The final step is a weighing process that uses these priorities to synthesize the overall importance of the criteria and alternatives. This

procedure is repeated for all the elements of the structure, obtaining a ranking reflecting user perception. In addition, it was possible to detect inconsistencies when experts gave their judgments. Under such situations, it was necessary to review them until an acceptable index was obtained.

As an illustration, Table 3 shows the judgement and priorities that a single member of the expert team expressed with respect to ICT system importance from the perspective of the agents to develop their activity.

The numbers in Table 3 express the intensity dominance of the criterion in the column heading over the criterion in the row heading. From Table 3, the element in the second column and first row has an input of 0.5. It means that the expert considers that ICT is 2 times more important for a member of Clinic care group than for the patient. The value of 5 for the comparison of the ICT support for developing research activities versus the patient activities indicates that it is considered five times more important. The priorities are derived by applying the geometric mean and normalizing. The geometric average is the nth root of a product of n numbers.

To facilitate the calculations the method counts with the Expert choice (EC) software, which is a multi-objective decision support tool based on the AHP. This software is used to obtain the results (Saaty, 2003).

The Empirical Evaluation

Figure 4 shows it is possible to appreciate the overall prioritization results for the agents at level 1. It shows that globally, ICT support has a greater impact on supplying clinical care service. This service is concerned with the activities developed by the physician, nurses, and paramedics.

Taking into account that this is a global result for the present situation from Figure 4, it is possible to visualize that the support of the computer-assisted system (Internet, e-mails, Web, etc.) and the fixed network communications system have more significance for the participants, with a priority of 56.3% and 25.4 % respectively.

The comparison in usage and importance for fixed and mobile is shown in Figure 5, where it can be seen that more importance is given to administrative activities, and a fixed system plays an essential role.

Patient Perspective

The results indicate that the importance of an ICT system for patient health-related activities are mainly concerned with an urgency service requirement as shown in Figure 6. Therefore, the importance of wireless and fixed network systems to satisfy its demand has the highest priority.

However if we observe the comparison of the fixed system vs. the wireless system it can be clearly seen that wireless system has greater relative importance from the perspective of the patient as seen in Figure 7.

Table 3. Judgements and priorities of one of the members

Goal	Patient	Clinic Care	Administration	Medical research	**Priorities**
Patient	1	0.5	2	5	0.263
Clinic care		1	4	9	0.512
Administration			1	2	0.124
Medical research				1	0.101

Figure 4. Agents overall priority result and priorities for ICT alternatives

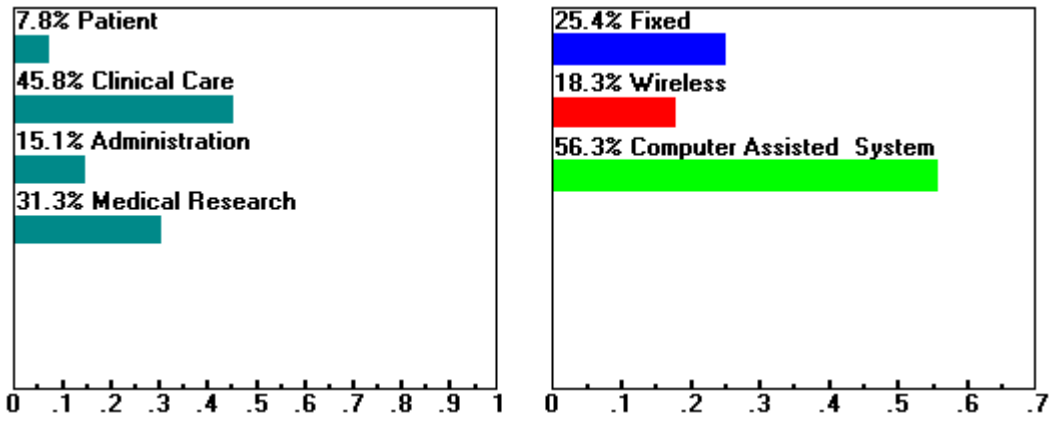


Figure 5. Fixed system vs. wireless system for each activity

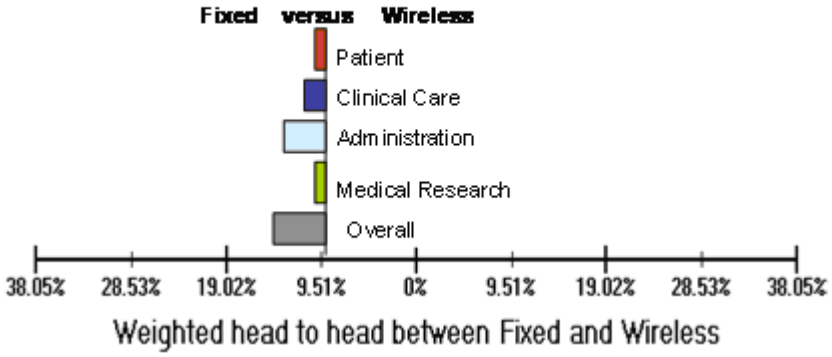


Figure 6. Priority results from the perspective of the patient

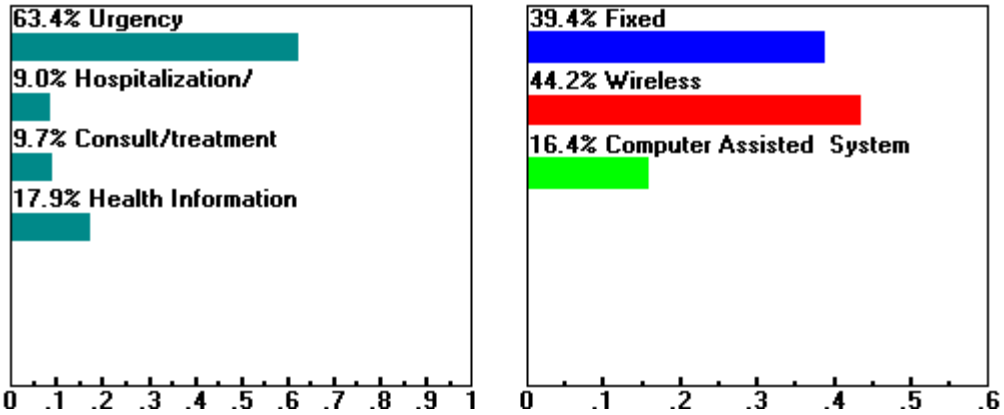


Figure 7. Fixed vs. wireless priority result for the patient's main activities

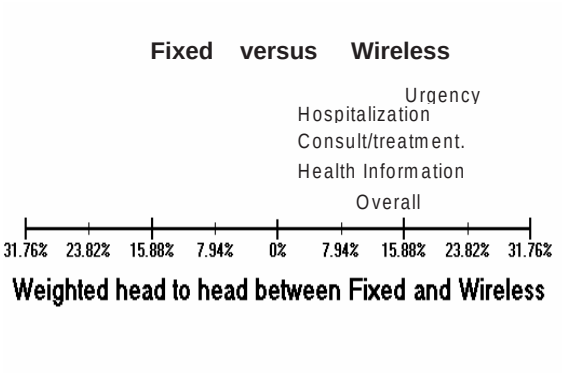
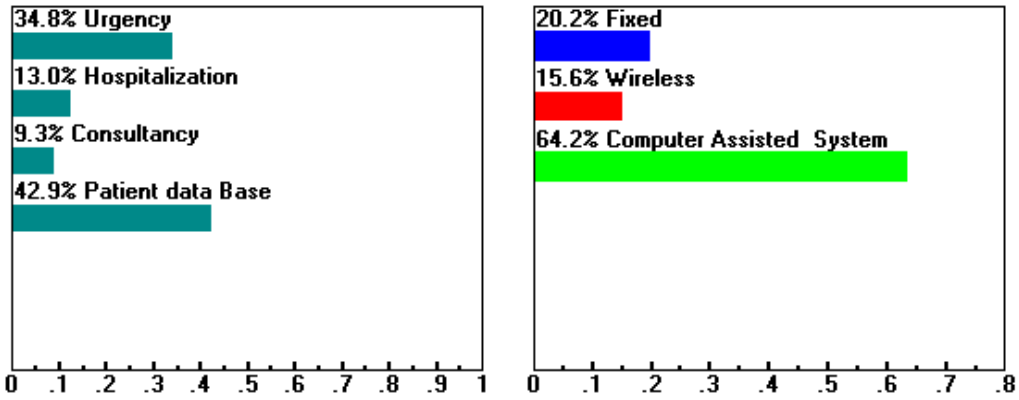


Figure 8. Priority results for clinical care perspective



Clinical Care Service Perspective

The overall result showed that the maximum importance regarding ICT and MT usage is to supply a clinical care service. The requirements related to urgency activities were the most relevant and there is a difference in the importance of the usage for providing service in hospitalization, control, and treatment activities. The different ICT systems are similar to the overall results as shown in Figure 8. The clinical care users found the computer-assisted system more useful for developing their tasks and did not consider the wireless system to be important.

However, a comparison between fixed technology and wireless technology shown in Figure 9 shows that MT is more important when facing urgent situations.

Medical Research Perspective

The medical research requirement showed a strong interaction with database applications implying a preference to work with computer-assisted support as shown in Figure 10.

Administration Perspective

From this perspective, the clinical activities such as delivering tests and exams results within the institution or externally has the highest priority. Therefore, the use of a fixed network system

Figure 9. Fixed system vs. wireless from clinical care perspective

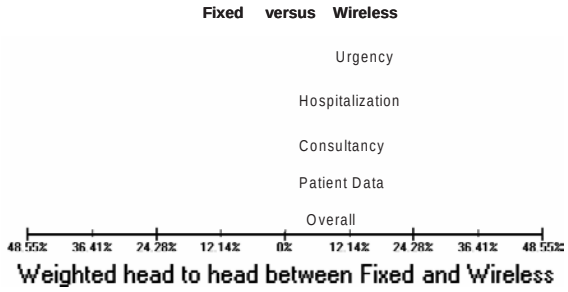
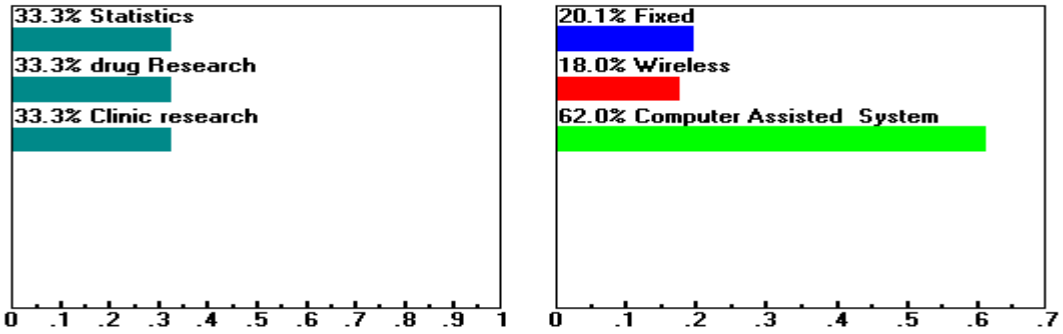


Figure 10. Preferences for medical research agents



(phone, fax, extensions, etc.) is of more importance.

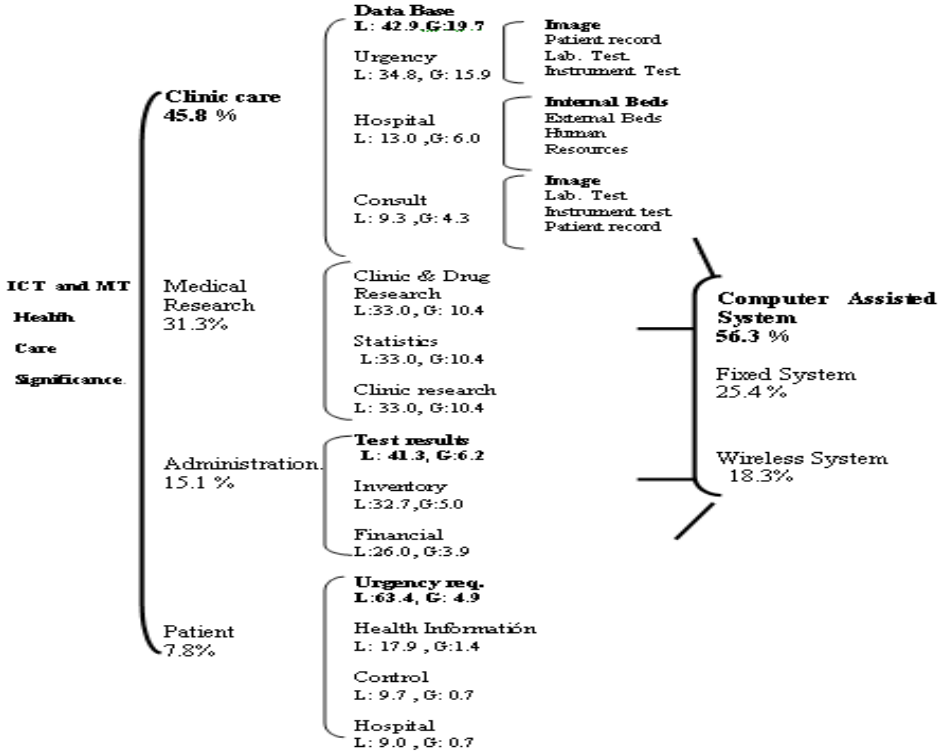
The Overall Results

The process incorporated data from the agents to achieve an overall result as shown in Table 4. From the patient perspective the results indicated that patient priority is mainly concerned with urgency service requirement (63.4%). At the present time, the importance of having access to wireless technology was ranked in first place while fixed network systems were the second priority. Nevertheless, gradient sensitivity for patients with urgency requirement show that wireless technology tends to increase.

From a clinical care perspective, the priority is for the support from computer-related systems. However, gradient sensitivity indicated an increasing priority for wireless systems while fixed system priority declines. From an administration perspective, the activities such as delivering test and exams results within the institution or externally has the highest priority. A strong usage of fixed network systems (phone, fax, extensions, etc.) is observed. From a medical research perspective, there is a strong interaction with database applications implying an increasing demand for computer-assisted support.

Figure 11 shows the hierarchy structure already ranked with priorities sorted. The local (L) priority refers to the percentage of contribution of that aspect to attain the requirement of the decision factor considered in a superior level. The global (G) priority refers to the contribution of that decision factor to the formulated Goal.

Figure 11. Hierarchy sorted according to local and global priority results



Result Analysis and Observations

One of the goals of every nation should be to have a healthy population. This is fundamental to arrive at other national goals such as quality of life, prosperous economy, and national security. ICT and MT offer a crucial function to achieve these goals in an efficient and economical manner. Our goal was to empirically investigate the MT and ICT provision and acceptance by Chilean hospitals and to provide a methodology to analyze user preferences and its effect. Although the model is based on user preferences, it is perceived that the user looks for satisfaction rather than optimization, since there are so many goals to obtain and it is not possible to satisfy all of them simultaneously. The results obtained show it is possible to be aware of the level of importance of the ICT and MT system for each agent involved and their trends.

As mobile devices have become consumer products, and become available to increasing number of users of all ages, consideration to incorporate its use for the health care service has begun to gain importance in this country. In this context, the user expectations of the ICT and MT system are to assist in developing new sources of knowledge and research, to support health care management, and to help increase efficiency and QoS by improving the processes that rely on ICT and MT systems.

From Table 4 the importance of computer-assisted systems to develop their activities for all the participants can be seen. Fixed network systems are of the highest priority and even though the appeal for wireless systems is increasing there is still some apprehension about its use and further work needs to be carried out in this field. However, from the patients group, older participant's concerns

Table 4. Agents priority results

Agents	Activities	%	ICT system	%
Clinic care 45.8%	Urgency service Hospital Consultant Patient data base	34.8 13.0 9.3 42.9	Fixed Wireless Comp. assisted	20.2 15.6 64.2
Medical research 31.3%	Statistics Drug research Clinic research	33.3 33.3 33.3	Fixed Wireless Comp. assisted	20.1 18.0 62.0
Administrative 15.1%	Financial Test results Inventory	26.0 41.3 32.7	Fixed Wireless Comp. assisted	45.1 13.8 41.2
Patient 7.8%	Urgency req. Hospitalization Consult, treatm. Health Inf.	63.4 9.0 9.7 17.9	Fixed Wireless Comp. assisted	39.4 44.2 16.4

were on obtaining information about their health status and urgency requirements (17.9% and 63.4% respectively). They also believed that the use of a wireless system was important (44.2%).

On the other hand, the results of the study indicated that mainly private hospitals are currently equipped with access to advanced networks and to develop e-business applications. From the physicians' point of view, they have a growing interest in adopting MT for customer applications and data records. Currently e-marketing applications in health care are low.

The application of AHP to the problem situation allows the integration of diverse judgements and preferences and therefore obtains an overall result.

Optional Actions Proposed

Based on the results, the next step is to propose optional actions and guidelines to follow with the purpose of managing efforts to obtain system improvement.

The options to follow are several and stick to the natural tendency to choose those factors that contribute with a greater relative weight to the objective. In this regard, three possible options for the result of the hierarchy are presented.

1. To consider each one of the decision factors of the first level with the same weight and the activities according to the ranking obtained. For the alternatives consider only the two high-priority ICT alternatives, concerning the particular agent to be developed as shown in Table 5.
2. Another option may be to develop all the agents proportional to the ranking and weights obtained and select two activities with higher priority. Then, choose the first rank alternative for each of the two activities considered.
3. Consider the whole structure and then pursue every activity in proportion and according to the ranking obtained.

These options could be related to the assignation of financial resources, governmental norms, technological resources, governmental support, and so forth.

CONCLUSION

The existence of competing goals in the health institution required a particular treatment, as the utilization of a scientific multi-criteria decision method. The AHP was beneficial for identifying the high-priority requirements of an ICT system in health-related activities, as well as, to be aware of MT acceptance and its importance in health care.

The process results identified the main ICT agents/users of health-related activities, the main

Table 5. Optional path 1

User / Agent	Activities	ICT Alternatives
Clinical service	1 Urgency 2 Hospital 3 Control & treatment	• Computer-assisted system • Fix (phone, fax, etc.)
Medical research	1 Clinic & Drug Research 2 Statistics	• Computer-assisted system • Fix (phone, fax, etc
Administration	1 Clinic activities (Test results) 2 Inventory 3 Financial	• Fix (phone, fax, etc.) • Computer-assisted system
Patient	1 Urgency 2 Health information 3 Control 4 Hospital	• Wireless • Fix (phone, fax, etc.)

activities they are involved in along with the ICT system support they require for each activity.

The AHP helped the experts and participants involved to identify the benefits of having an infrastructure of an adequate ICT network. From the results it can be seen that the user looks for satisfaction rather than optimization and they are willing to make a trade-off.

According to the resultant prioritization, efforts should be aimed at improving the QoS of the ICT system where they are most beneficial. It should be taken into account that the introduction of any new ICT could face problems not only in competing for financial support available but also possible interoperability problems. In this sense, the methodology helps to produce a distribution of the resources proportional to the users' requirement.

The attributes would be an improvement in clinical quality. The doctors, nurses, therapeutics, and other welfare personnel, provided with information to assist them, can offer a service of better quality in the clinical environment. Reduced costs: Health institutions can improve administrative efficiency and thus reduce medical costs. It can also achieve real savings in labor if the network is used for the execution of those manual tasks that are time consuming. This is particularly true in the traditional transactions among institutions such as derivations, claims, election, and even clinical data. Improved service to the client: The health institutions can use the ICT network system to provide faster ways for receiving/delivering health information through telephone aid links even between other organizations, to reduce the waiting period for hospital medical data and to avoid repetitive form filing.

This pilot study concludes that the combination of fixed and wireless networks can give the desired support to the patients when requiring information. The patients involved in the study gave a priority of 63.4% for urgency requirement and when selecting an alternative ICT system they revealed a 44.2% acceptance for wireless system, where a mobile device plays an important role for them. When comparing fixed vs. wireless systems they found it more important to use wireless technology for urgency requirements and for the rest of their activities they prefer a fixed system. For clinical care activities, computer-assisted technology support is preferable since professionals need relevant and timely information for better decisions. Comparing fixed vs. wireless systems for their normal activities they still preferred a reliable fixed system, however for urgency requirements the importance of having an efficient MT system is increasing.

When considering the factors for measuring quality in health care systems, the availability of the services and the need for ubiquitous access to integrated information are considered the most important.

The study revealed that mainly private hospitals have access to advanced network and Internet access; hence the technical basis for developing new applications is in position. Patient interest is online health information, e-health, and e-care services. Physicians would desire patients using MT and ICT applications where the support of MT is of increasing importance.

REFERENCES

Alexander, H. R., Biggers, J., Forman, E., & Schleicher, D. (1990). *Prioritization of civil tiltrotor technologies using the analytic hierarchy process.* Paper presented at the Third International Symposium on the Analytic Hierarchy Process, George Washington University, Washington, DC.

Alleyne, G. (1998). *Información en Salud para Todos.* En: Laerte PA, Castro E de. Biblioteca virtual en salud. Sao Paulo: OPS/OMS: pp 17-34.

Ammenwerth, E., Gräber, S., Herrmann, G., Bürkle, T., & König, J. (2003). Evaluation of health information systems problems and challenges. *International Journal of Medical Informatics, 71,* 125-135.

Birch, S., & Gafni, A. (2003). Inclusion of drugs in provincial drug benefit programs: Should "*reasonable decisions*" lead to uncontrolled growth in expenditures? *Canadian Medical Association Journal, 168,* 849-851.

Chan, A. T. S. (2000). WWW + smart card: Towards a mobile health care management system. *International Journal of Medical Informatics, 57,* 127-137.

Claxton, K., Sculpher, M., & Drummond, M. (2002). *A rational framework for decision making.* National Institute for Clinical Excellence (NICE). *Lancet, 360,* 711-715.

Clemens, R. T. (1998). *Making hard decisions: An introduction to decision analysis.* Duxbury Press Brooks/Cole publishing Company.

Colomer, M. J. (2002). El desafío es conseguir adaptar los hábitos y costumbres para afrontar la gestión clínica del presente Editorial. *Gestión Clínica y Sanitaria, 4,* 111-113.

Del Llano Señarís, J. E. (2003). Gestión Clínica y Sanitaria: ayudando a conciliar necesidad y escasez [Editorial]. *Gestión Clínica y Sanitaria, 5*(3-6), 2.

Geier, J. (2001, February 5). *Saving lives with roving LANs.* Retrieved from http://wireless.itworld.com/4246/NWW0205bgside/pfindex.html

Hikmet, N., & Chen, S. K. (2003). An investigation into low mail survey response rates of information technology users in health care organizations. *International Journal of Medical Informatics, 72,* 29-34.

Istepanian, R. S. H., Jovanov, E., & Zhang, Y. T. (2004). *M-Health: Beyond seamless mobility for global wireless healthcare connectivity* [Editorial]. *IEEE Transactions on Information Technology in Biomedicine, 8*(4), 405-412.

Istepanian, R. S. H., & Lacal, J. (2003, September 17-21). *Emerging mobile communication technologies for health: Some imperative notes on m-health.* In *Proceedings of the 25th. Annual International Conference of the IEEE Engineering in Medicine and Biology* (pp. 1414-1416). Cancun, Mexico.

Jan ten Duis, H., & Van der Werken, C. (2003). Trauma care systems in The Netherlands. *Injury—International Journal of the Care of the Injured, 34*(9), 722-727.

Jelekäinen, P. (2004). GSM–PKI solution enabling secure mobile communications. *International Journal of Medical Informatics, 73,* 317-320.

Khalifa, M., & Cheng, S. (2002). Adoption of Mobile Commerce: Role of Exposure, In *proceedings of the 35th Annual Hawaii International Conference on System Sciences* (HICSS'02)(Vol. 1, p. 46).

Mahmood, M. A., & Mann, G. J. (1993). Measuring the organizational impact of information technology investment: An exploratory study. *Journal of Management Information Systems, 10*, 1.

Oddershede, A., & Carrasco, R. A. (2006, November 5-8). *Analytic hierarchy process decision model for health institution selection: User perception.* Institute for Operations Research and the Management Sciences. Paper presented at the Informs Annual Meeting, Pittsburgh, PA.

Oddershede, A., Carrasco, R. A., & Ontiveros, B. (2006). Perception of wireless technology provision in health service using the analytic hierarchy process. *WSEAS Transactions on Communications, 5*(9), 1751-1757.

Oddershede, A., Carrasco, R. A., & Soto, I. (2005, October). Decision model for information and communications technology implications in health service: User perception. In *Proceedings of the SMDM 27th Annual Meeting, Society for Medical Decision Making Conference,* San Francisco, CA.

Oddershede, A., Soto I., & Carrasco, R. A. (2001, April). Analysis and prioritisation of Chilean mobile communication system. In *Proceedings of the International Conference on System Engineering, Communications and Information Technologies ICSECIT,* Punta Arenas, Chile.

Organización Panamericana de la salud. (2001). Marco general e institucional para el desarrollo de sistemas de información en servicios de Salud. Parte A. Organización Panamericana de la salud.

Organización Panamericana de la Salud, Programa de Organización y Gestión de Sistemas y Servicios de Salud**, "**Perfil del Sistema de Servicios de Salud, Chile**",**(1ra edición, marzo de (1999),(2da edición, enero de 2002)*,(Revisado, abril de 2002)

Pedersen, P., Nysveen, H., & Thorbjørnsen, H. (2003). *The adoption of mobile services: A cross service study. SNF-report no. 31/02*. Bergen, Norway: Foundation for Research in Economics and Business Administration.

Pederson, P. E. (2005). Adoption of mobile Internet services: An exploratory study of mobile commerce early adopters. *Journal of Organizational Computing and Electronic Commerce, 15*(3), 203-221.

Prados de Reyes, M., & Peña Yánez, M. C. (2002). *Sistemas de información hospitalarios: Organización y Gestión de Proyectos.* Ed.: Escuela Andaluza de Salud Pública, Granada.

Saaty, R. W. (2003). *Tutorial for building AHP hierarchical decision models.* Creative Decision Foundation.

Saaty, T. L. (1990). *Multicriteria decision making: The analytic hierarchy process, planning, priority setting, resource allocation.* RWS Publications.

Saaty, T. L. (2001). *Decision making for leaders.* Vol. II, AHP Series 315 pp., RWS Publications.

Saaty, T. L. (2006). *Fundamentals of decision making & priority theory* (2nd ed.). Vol. VI of the AHP series. RWS Publications.

Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making. *Medical Decision Making, 4,* 322-338.

Sosa, O El Nuevo Día, 10 de marzo de (2004), El País. *Boicot de aseguradoras a la contratación directa.*

Strasser, S., & Davis, R. M. (1991). *Measuring patients satisfaction for improved patient services.* Ann Arbor, MI: Health Care Administration Press.

Suh, C.-K., Suh, E.-H., & Baek, K. C. (1994). Prioritizing telecommunication for long range R&D planning. IEEE *Transactions on Engineering Management, 41*(3).

Weinstein, M. C. (2001). Toy E. L., Sandbergea, et al. Modelling for health care and other policy decisions: Uses, roles and validity. *Value Health, 4,* 348-361.

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, and J. Day, pp. 345-364, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.12
Relevance of Mobile Computing in the Field of Medicine

Henrique M. G. Martins
University of Cambridge, UK

Matthew R. Jones
University of Cambridge, UK

ABSTRACT

Mobile information and communication technologies (MICTs) are widely promoted as increasing the efficiency of work practices in many business sectors, including healthcare. There are numerous types of mobile computing devices available that provide users with capabilities that can be applied in a wide range of different work settings. Case studies of the use of different MICT devices by doctors in different hospital settings indicate that while some doctors easily adopt MICT devices and find them a helpful tool, others encounter problems with their usage and, in fact, a majority do not use MICTs at all. This chapter deals with identification of five factors influencing the uptake of MICTs in clinical work practices and proposes a framework for analysing their interactions with the aim of increasing its uptake in medicine.

BACKGROUND

Mobile information and communication technologies (MICTs) are widely promoted as increasing the efficiency of work practices in many business sectors, including healthcare. There are numerous types of *mobile computing devices* available that provide users with capabilities that can be applied in a wide range of different *work settings.* Case studies of the use of different MICT devices by doctors in different hospital settings indicate that while some doctors easily adopt MICT devices and find them a helpful tool, others encounter problems with their usage and, in fact, a majority do not use MICTs at all. This chapter deals with identification of five factors influencing the uptake of MICTs in clinical work practices and proposes a framework for analysing their interactions with the aim of increasing its uptake in medicine.

INTRODUCTION

In most business sectors, including healthcare, it is widely claimed that the use of mobile ICTs, either alone or in combination with existing desktop resources, has the potential to achieve significant increases in the efficiency of work practices (Kelly, 2001; Cox, 2002; Davis, 2002). In recent years, however, a growing variety of *mobile computing devices* have become available—including laptop personal computers (PCs), tablet PCs, handheld PCs/personal digital assistants (PDAs), and PDA-phones/smartphones—which differ significantly in terms of characteristics such as screen size, computing power, weight, or input mechanisms, some of which have been shown to have an influence on ease of use and usage patterns (Dryer, Eisbach, & Ark, 1999; Martins & Jones, 2005). The devices may also vary in whether or not they are connected wirelessly to an existing network. This may be significant since, although when unconnected they are able to provide mobile computing power and support asynchronous communication, for uses requiring synchronous communication or real-time collaboration, wireless capability needs to be in place.

It is not just devices that vary, but also the work settings in which they are used. In the healthcare sector, there are a variety of departments in which hospital doctors' work, often organised according to particular clinical specialities. These departments may be spatially contained (e.g., in wards or intensive care units), or clinicians in certain specialities such as genetics, metabolic conditions, or psychiatric support may work across a whole hospital. Less frequently, hospital doctors may work outside the hospital (e.g., accompanying acutely ill patients in transit from one location to another). In addition to potentially working in different physical settings, hospital doctors—like most highly skilled professionals—engage in several different types of activities over the course of their working day. These spatial and temporal dimensions of the organisation of doctors' work practices have been shown to influence how they use pen-and-paper and desktop ICTs (Westbrook, Gosling, & Coiera, 2004; Martins, Nightingale, & Jones, 2005).

Achieving the expected benefits from the use of MICTs, therefore, depends not simply on the provision of MICTs per se, but upon the appropriate matching of device characteristics and work settings. This chapter reports on research on the relationship between different types of MICT devices (laptop PCs on a cart or trolley, standard desktop PCs mounted on a cart or trolley, tablet PCs, and handheld/PDAs) and different clinical work practice[1] situations, and how this affects doctors' usage of MICTs.

RESEARCHING MOBILE COMPUTING IN HEALTHCARE

Two broad approaches may be used to study why and how doctors use (or do not use) MICTs in their clinical work practices: surveys across a large number of sites, or detailed studies in particular settings. This chapter largely focuses on the latter approach, presenting case studies of MICT usage at hospitals with different MICT devices and clinical settings.

Data were collected through multiple methods including interviews with doctors and hospital IT staff, observation, questionnaires, and analysis of usage logs for specific systems.

Case Descriptions

Case A: Paediatric Intensive Care Unit (PICU) with Handheld Computers

The PICU at a leading UK hospital had 12 beds, all located in a single ward. The layout of the unit comprised an open-plan central area with six beds and a nursing station (with two desktop computer) and a number of individual patient rooms and doctors' offices (with three desktop computers).

The distance from the nursing station to any bed was not more than 10-12 metres.

The unit was staffed by about 7-10 doctors working on a shift pattern. The senior doctors, who had their own offices in the unit, were relatively permanent, while the junior doctors shared a common doctors' office and rotated between different departments. As a result, a particular doctor might be away from the unit for periods as long as 3-4 weeks. The department also sent doctors to outside locations to assist in the transfer of acutely ill patients to the PICU.

Desktop computers had been in use in the unit for a number of years, providing access to some basic patient demographic information, and ordering and reporting laboratory results. The unit provided a handheld computer for collecting data on ward rounds, which could be synchronized with the desktop PC in the doctors' office to update the departments' database. The handheld was never used consistently except for a highly IT-savvy doctor. Instead doctors would record data on pieces of paper during the morning ward round and then key in the data on the desktop PC later in the day. The reasons for this appeared to mostly relate to problems with synchronising the handheld and the time required to use it during the ward round.

The unit had previously supplied handheld computers (without wireless connection)—providing drug and medical reference information and medical calculation applications—to some of the doctors. Despite this and the head of department's enthusiasm for handheld devices, only half of the 12 doctors interviewed used handheld devices in their work. Most of the doctors used them predominantly for arranging schedules and dairy appointments, whilst discussing a particular patient, and less frequently during ward rounds. Those that did not use a handheld device argued that they did not do so because patients tended to be on the ward for long stays and they could remember patient data and could access online information in their nearby offices if necessary. Some doctors also argued that due to the small screen, the technique required to input and access information on the handheld was less usable, if not difficult. This, according to them, was a reason for not carrying digital patient information, which they would have otherwise found useful. Interestingly, three female doctors said that since they had no pockets in which to carry the device, they only used it in the doctors' office. More than one doctor commented on the inconsistent way that colleagues used the handhelds, which restricted the possibility of it substituting for existing paper forms such as the unit job[2] list or pieces of paper used at shift handovers.

Those doctors using the handhelds found them helpful for storing information about drugs and infusions and telephone lists, and three of them stored short medical reference notes they created. One of the doctors also commented that he would use his handheld differently when working alone to record his own jobs from the shared list and set an alarm to remind him to carry them out, especially at night. Another enthusiast was one of the doctors involved in the retrieval of acutely ill children and who used his handheld extensively for these activities. Having customized most data to his needs, he said that he felt nervous about working without his handheld.

CASE B: Emergency Department with Handheld Computers

The emergency department (ED) of this U.S. hospital was seeing an average of 65,000 patients yearly and was staffed with about 45 physicians (attending and residents) working three regular eight-hours shifts. The area was a large, open space with four bays and several individual patient rooms. There were about 44 desktops available, four of which were located inside the individual rooms. The area had been rebuilt less than two years before, and the decision to install desktops was based on a fear that mobile computers (tablet PCs or mounted PCs) would be stolen or dam-

aged easily due to the high patient throughput and staff turnover.

The department had issued stand-alone handhelds to all residents and handhelds/PDAs were made available to all senior doctors to increase doctors' access to specific departmental information (clinical pathways, protocols) and as a tool for personal organization (for example, call schedules, contact information, procedural information). Out of the 30 doctors surveyed, all had a handheld and two-thirds used it more than once a day. These frequent users, predominantly 'junior' doctors, valued the devices, arguing that they made work more efficient, saved the doctor from having to remember formulae, and allowed easier access to drug and medical reference information. Some of the senior doctors said that they used the handhelds less frequently, as they were familiar with the commonly used drugs as well as frequent pathologies from long experience. Another reason for less frequent use was the fear devices could be lost in a busy environment and the fact that local practices often diverged from those suggested on handheld-based applications. It therefore made more sense to discuss cases with colleagues and obtain information from them. Only a few doctors reported using the devices to access hospital documents/guidelines, and none for accessing patient data, the Internet, or to send/receive clinical-work-related e-mails. The main reasons for this were said to be that it was hard to access data on the handhelds, especially compared to the widely available desktops. Department protocols, although available on handhelds, were also normally accessed via desktop PCs because doctors found the large screens more convenient than the small handheld ones and they could simultaneously access online resources (for medical and drug reference as well as for calculations).

CASE C: Breast Unit with Tablet PCs

This UK hospital breast unit comprised a multidisciplinary team of surgeons, radiologists, and physicians. Apart from short inpatient stays for surgery, the majority of contact with patients was via outpatient appointments at the unit. There would be about 5-6 clinic sessions per week when about 20 patients would be seen. Three physicians, two or three surgeons, and two radiologists would typically staff the unit. On a first visit patients would normally be examined by a physician and see a radiologist. On the second visit, physicians and occasionally surgeons would discuss results and the treatment plan. The unit had seven consultation rooms which had no desks, a large reporting room with three desktop computers where all team members (doctors, nurses, care managers) congregated during outpatient sessions, a waiting area, and a few offices for senior doctors.

Nine months previously a new clinical information system specifically designed for the unit was introduced. It was accessible via desktops in doctors' offices and the reporting room. The unit had also installed a wireless network and made available four tablet PCs for usage during sessions, with the intention that physicians especially would use them to record patient data on the first visit.

Observation during several outpatient sessions revealed that only the radiologists used the tablet PCs in the reporting room as stationary devices (permanently plugged to a power supply and with a mouse installed). Physicians would take a few sheets of paper into the consultation room and then enter their handwritten notes via the three desktop PCs available in the reporting rooms. Surgeons rarely used any computers, but would instead dictate their comments and provide these recordings to secretaries for entry into the application.

Amongst the reasons offered by the physicians for not using the tablet PCs were that they would still need to write certain details down on paper

(e.g., the patient's past history), as the application covered only part of their data recording needs and patient records were still largely paper based. Another issue was that it was more efficient for them to walk the short distance from the consultation rooms to the reporting room and enter the data on a comfortable and familiar device, rather than to feel awkward using a tablet PC to input data in front of a patient. This was especially the case with the text-based data, required by the nature of the speciality and patient history, which was felt as particularly difficult to enter using the active pen onscreen mechanism only. In the consultation rooms some doctors felt that having nowhere to place the device besides the chairs or the examination bed created a risk of it being dropped. Lastly, one of the physicians stated that she had not been made aware she could use the devices, and she thought that since radiologists were using them in the reporting room, they were for their exclusive use. The radiologists, for their part, predominantly used the tablet PCs because they did not have enough PCs in the area of the communal room where they worked. The tablets could also be brought closer to the X-ray viewing boxes. The senior radiologist was an enthusiastic user, who took the tablet PC home with him and had loaded personal work files and customized it to his own use. Occasionally he would also use it as a laptop in other unit locations.

CASE D: Renal Unit with Mounted PC and Tablet PC

The renal unit of this UK hospital had 64 beds. Most inpatients were located in three wards on one floor of the hospital building, although there were usually a further 6-10 patients located in other wards around the hospital. Some of the senior doctors' offices were some floors away from the ward, and the outpatient consultation areas were in another nearby building. The doctors' office (mostly used by less senior doctors) was located in the middle of the central ward and had three desktop PCs. The mounted PC on a trolley/cart would also usually be left here when not in use.

A wireless network had been installed in the wards that allowed access to the department's e-prescription/lab reporting/requesting application and to the British National Formulary (BNF) Web site. Tablet PCs were shared with nurses, although the two battery-chargers were located in the doctors' office. In addition to the doctors' office, desktop PCs were available in the consultant offices, outpatient consultation room, nursing stations, and some ward clerks' offices. These provided access to the departmental application as well as other hospital applications and full access to the Internet.

Two medical teams of about three to six doctors (comprising one or two consultants, one or two registrars, and one or two house officers) covered the wards each day. On most mornings team members would get together for a "consultant ward round," while on other days these would occur without the consultant. These ward rounds would start on the unit wards and then proceed to other floors of the hospital where their inpatients might be located. No doctor was seen picking up a tablet PC to visit an individual patient, although some said that they did so occasionally. Doctors were also not seen to use tablet PCs or the mounted PC except for morning ward rounds. Teams would use either one tablet PC or the mounted PC for each ward round, never both together, and occasionally on ward rounds without a consultant, neither would be used. The mounted PC was never taken outside the unit, but the team had started to take the tablet PC when visiting patients at a ward located in another floor of the hospital when wireless coverage was extended there. There were also a number of patients that the team would visit during the round at other hospital locations where mobile computing and the department application could not be used.

All doctors liked accessing the department application on the mobile computers for the ward round, although the need to change the tablet PC

frequently because the battery had died, or loss of wireless connection with the mounted PC in some of the most distant areas of the ward, made some doctors consider using nearby desktops. Convenience of access and having information readily available in one location (as opposed to team members needing to search for different pieces of paper) was said to improve decision making, especially compared with when teams visited patients outside the unit.

The hospital had no full electronic medical record (EMR), so teams would conduct their rounds with the files of patients' notes as well as mobile computers. This meant that, in addition to whichever mobile computer was in use, the team also needed a trolley to carry the sometimes voluminous notes. While with tablet PC this was not a big issue, it made the team reluctant to use the mounted PC if there were less than two junior doctors in addition to the consultant. Doctors' opinions were divided as to what they considered to be the better device for ward rounds. The mounted PC was seen as more cumbersome and as slowing down the rhythm of the round, because it could not be easily brought to the patient's bedside (it either stayed in the corridor or away from the immediate bedside area). Although the big screen of the mounted PC was seen as an advantage for information sharing and discussion, the mobility of the tablet PC between users was regarded as able to compensate for this. Such sharing of information and its collective use was not always observed or commented as unproblematic. Retaining control over the mobile computer or using it to show data on display as a backup for stronger argumentation in discussions occasionally seemed to reveal and was described as attitudes of power exertion.

Case E: Medicine Department with Laptops on Trolleys and Handhelds/ PDAs

The medicine department of the Veterans Affairs Medical Center, Washington, DC, had five medical teams (each one composed of one attending, one resident, two interns, and one or two medical students) covering about 72 inpatient beds on two consecutive floors. On each floor the ward had a central area (with several desktop computers) from which six corridors of patient rooms and two or three doctors' rooms (with five desktop PCs) led off. Service stairs connected the floors, but the lift (elevator) was located away from the centre of the wards. Normally teams had patients on both floors, and doctors identified this layout as affecting their usage of the Wi-Fi-enabled laptops made available to them, as there were attached to a food trolley and this apparatus was only movable as a whole. The laptops had been available for more than five years and provided full access to Windows software, two hospital systems (a full EMR and a drug administration system mostly used by the nurses), the Internet, and a resource-rich hospital intranet.

About one-quarter of the doctors claimed that they never used the mobile devices, and only about one-fifth of the doctors were high-frequency users, using the laptops more than five times a day for clinical work. All doctors used the laptops to check patient data, but only two for medical calculations and one to send clinical-related e-mails. No doctor reported using the laptops for checking hospital guidelines.

About four senior doctors had recently been give a PDA to try accessing the EMR wirelessly, although the handheld application was not as comprehensive and could not be used, for example, to prescribe or to retrieve certain patient data. Only one of the four senior doctors occasionally tried to use her PDA regularly.

Doctors only used the laptops during team ward rounds and did not use them when visiting

patient's rooms during the course of the day. A number of reasons were given for this, such as fear of losing/damaging the device, and that it was cumbersome to carry just for a quick visit to a patient room, especially as such visits often involved walking from one end of the department to the other or between floors. A number of doctors also said that they did not think they were allowed to use the device for tasks other than ward rounds.

Three out of five medical teams in the department never used the laptops, while the others used them regularly (albeit only for ward rounds). This appeared to be related to differences in ward rounds practices, different perceptions of the function of ward rounds, and whether this could be supported by MICT. For example, while one of the teams fostered the use of the laptops to access the Internet to look at online resources for retrieving pertinent medical information during ward rounds, others would not favour this although they might use them extensively to access the EMR. Individual doctors' attitudes towards technology, but also regarding the importance of rich and timely data in clinical decision making, the effectiveness of existing practices, and their willingness to change practices, also appeared to play a role. Strong social influences such as the power relationships between grades of doctors, existing work routines, and team dynamics were also associated with differences in use of mobile devices, as were departmental policy not clarifying that devices could be shared amongst teams on different floors.

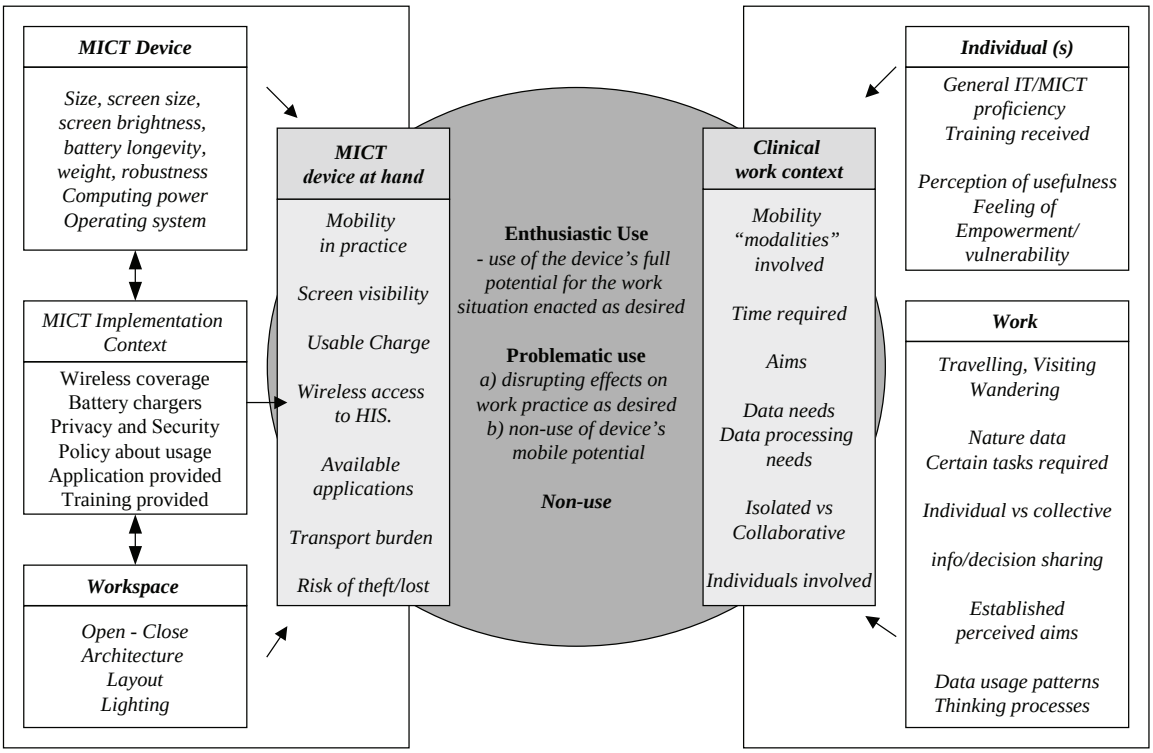
CASE F: Clinical Genetics Specialist with Handhelds

In this large New York State hospital, there was just one doctor providing clinical genetics services, supported by a nurse and some genetic counsellors. The doctor's main work involved diagnosis of outpatients, who he would then follow during their subsequent medical history, but he also often visited inpatients around the hospital. He had developed his own patient database, accessible via his handheld, to which colleagues contributed via desktop PCs. As he acknowledged difficulties with entering so much most data via his handheld, he included input strategies in the application design, like extensive drop-down menus and the option of entering full-customized sentences that he used more frequently. In addition he could always access patient data even though he was often away from his office with no access to a desktop PC or connection to the database in his department. He often received phone calls about patient care either in the hospital or at home. In both situations access to patient data and records of previous decisions was considered helpful, as was being able to record data immediately rather than when he next was in his office. The information on his handheld was shared with his secretary, who served as a "fixed" contact point. In addition to potential efficiency gains, he saw the possibility of accessing more complete information as a quality improvement, but also as a form of personal style regarding patients and families. The main reason, however, that made him keep using the handheld was for scheduled visits to inpatients around the hospital, where he regularly uses it by the bedside and then uses an infrared-enabled printer to produce a customized report to add to the hospital's paper record.

DISCUSSION

From these brief case descriptions, five themes may be identified that may be helpful in understanding the use (and non-use) of different MICT devices in medical work (the relationship between the categories and themes is shown in Figure 1).

Figure 1. Schematic representation of the relationship between a particular MICT device at hand and a specific clinical work context



MICT Devices Characteristics

MICT devices differ in their technical characteristics, and these alone can influence use. Screen size is perhaps the most obvious example of this, affecting viewing but also data entry on handheld and tablet computers. Where handhelds were involved, (small) screen size was shown to limit doctors' willingness to input large amounts of data (Case A) or retrieve information (Cases B and E). With tablet PCs (Cases C and D), this was no longer referred to as an issue for data retrieval, but when these were contrasted with larger screen alternatives (desktop or mounted PCs), tablet PC screens were considered less suitable (Case C) and a deterrent to information sharing and collaborative working (Case D).

Another significant characteristic is the size of the device itself, which has implications for the devices' *mobility in practice*. Handhelds are the most wearable MICT devices, but are still considered too big if the user does not have a suitable pocket to carry them in (Case A). When it comes to tablet PCs and PCs (or laptops) mounted on trolleys, size is again a distinguishing factor, as Cases D and E illustrate. Users had problems with mounted PCs/laptops on trolleys when their work involved a lot of walking. In Case D where a tablet PC alternative existed, it was often chosen due to size alone, and in Case E, it was clear that while the handheld would always be carried by the senior doctor into the patient's room, this was not the case with the laptop on the trolley. Size was also related to concerns about theft and damage, as was shown in Case B where the head of department provided doctors with handhelds because they could be personally owned and carried by professionals themselves.

Other characteristics such as screen brightness, battery longevity, weight, robustness, processing power, and operating system also influenced usage, both directly and in combination with other factors. This is in agreement with experimental studies (Dryer et al., 1999) that have shown that certain MICT device characteristics influence not just usage behaviours, but also users' attitudes towards the devices and towards other people using them.

Workplace Characteristics

As the cases illustrate, there is considerable variation in the characteristics of workplaces within hospitals. For example, while the size of some, such as the intensive care unit in Case A or the outpatient unit in Case C, was quite restricted and access tightly controlled, others such as the Accident and Emergency Department in Case B covered a larger area and were accessible to the general public. This had an impact on users' perceptions of the risk of theft or loss of devices. Similarly, the particular layout of some wards and the location of desktop PCs (Cases D and E) may create incentives to use mobile devices, while widely (Case B) or readily (Case C) available desktop PCs may discourage use of mobile devices.

Again, while some of these characteristics like physical layout or access to the settings may directly influence usage, others, such as the architecture of the building and even lighting conditions, may significantly affect MICT use when combined with certain device characteristics by influencing the convenience and comfort with which devices can be used (Figure 1).

MICT Implementation Context

In addition to the technical characteristics of the MICT devices and those of the workplace per se, adoption may be influenced by an organization's decisions about how to deploy mobile technology and for what purposes. These may relate closely with hardware, like the extent and reliability of wireless coverage (Case D) provided or the battery charging options. In Case E, the decision to use the standard laptop battery only (to reduce the weight of the trolley) meant that batteries would sometimes die, even when fully charged at the beginning of a ward round, if the device was used extensively or the round lasted longer than usual. Another type of decision related to policy measures to ensure confidentiality, privacy, and security (including theft of the devices themselves) which restricted the flexibility with which mobile devices could be used. A prime example of this was the laptops in Case E that were fixed to trolleys to prevent thefts. In the same case, the lack of policy on sharing of devices meant that teams did not use available devices when visiting other wards.

Another significant implementation issue concerns the applications and level of Internet access made available on the mobile devices. Applications obviously determine the particular data available and also how this is retrieved and manipulated. Some applications can overcome device limitations to a certain degree, as it was with the drop-down menus in Case F that compensated for the small handheld screen size. Others, however, may also highlight device limitations. In Case C, for example, the application required large text data fields to be entered, which was felt to be incompatible with data entry on tablet PCs, especially in full view of patients in the consultation rooms. Hardware-related training has been widely identified as a factor in the use of desktop technologies (Riley, Lorenzi, & Dewan, 2002), and it seems even more relevant with MICTs due to the new skills involved in data input on some devices and also their diversity.

Characteristics of the MICT Device at Hand

By *MICT device at hand,* we mean the mobile computing resource as a whole—hardware, software, and usage policies—as it presents itself to a particular context. For example, screen brightness in itself is not enough to ensure screen visibility; it also depends on the lighting conditions in a particular setting. *Mobility in practice* may also be seen as another instance of the contingent nature of the *device at hand.* Thus, while mounted PCs (Case D) or laptops on trolleys (Case E) were generally mobile when used on a single floor of the hospital, the exact same devices could not be used with some patients because of architectural barriers (stairs, narrow doorways).

Individual(s') Characteristics

Like MICT devices, individual doctors are not all the same. For example, age, seniority, and whether they are a surgeon or a physician may have an influence on users' attitudes to, and usage of, MICT devices (McLeod, Ebbert, & Lymp, 2003; Barret, Strayer, & Schubart, 2004). Most authors (Brenda & Gadd, 2001; McLeod et al., 2003), however, have suggested that such differences, for example with grade, are probably linked to work specificities and the roles enacted, rather than individual dispositions. The cases suggest, however, that there are a few individual characteristics that may directly influence usage in terms of *how* (rather than *what)* clinical work practices are carried out. For example individual doctors undertaking retrievals in Case A or working in Accident and Emergency in Case B experienced the same *MICT device at hand* in a consistent *work context,* but exhibited different usage patterns.

Doctors' general IT proficiency (keyboard skills, knowledge of applications, and ability to troubleshoot problems), as well as their familiarity with aspects particular to using certain mobile devices (e.g., using *graffiti* to input data on a handheld/PDA, or using active pens on a tablet PC), may also influence not only the adoption but also the ease with which devices are used in work situations. This was illustrated in Case A, where most doctors had difficulties with updating the database; only one, recognized to be more enthusiastic and IT proficient than the norm, was able to do this easily. Personal perceptions about the value of MICT devices compared to existing tools, such as paper or desktop PCs, were also found to vary significantly. For example, some doctors in training in Cases A and B valued the reminders provided by the MICT (with information about drugs, medical references, to-do lists), while others preferred to rely on their memory or notes on pieces of paper, especially where desktop terminals were nearby (Cases A and C).

Individual perceptions about their role and how this should be carried out also varied. In Case F, for example, the doctor's desire to be able to access every detail of his patients at all times meant that his handheld was seen not only as a useful tool, but one that actually improved efficiency and quality of care. Such perceptions could sometimes be shared across groups of doctors as was shown in Case E, where some teams were enthusiastic users, while others considered that the devices offered little or no advantage to their work and could even be a distraction.

A slightly different theme is that of doctors' feelings of empowerment, or conversely of vulnerability, when using mobile technologies. The concerns of physicians in Case C about using MICTs in front of patients may be contrasted with the positive feelings created by the personalized use of handhelds by the genetics specialist in Case F or the doctor involved in retrieval in Case A. There may also be a collective element to such feelings, as illustrated by the different responses of senior and junior doctors to the use of MICT during ward rounds in Case D.

Work Characteristics

Hospital doctors can potentially use MICT devices in a variety of work situations that vary along a number of dimensions over which they have more or less control. Work may vary, for example, in terms of their *modality of mobility.* Kristoffersen and Ljungberg (2000) distinguish three such "modalities"—travelling, visiting, and wandering—each of which has distinct characteristics: *travelling* is the process of going from one place to another, [often] in a vehicle; *visiting* is spending time in one place for a prolonged period before moving on to another place; and *wandering* is extensive local mobility in a bounded area.

Patient retrieval may be seen as an instance of travelling mobility (Case A), ward rounds as an instance of visiting, and junior doctors' response to requests at other times as an instance of wandering (Cases D and E).

These modalities are not necessarily mutually exclusive and there may be differences within them, but they are seen as affecting the suitability of certain types of MICT devices. Another aspect of mobility in work practices is highlighted by Luff and Heath (1998) who discuss *micro-mobility* as "the way in which the artefact may be mobilised and manipulated for various purposes around a relatively circumscribed, or 'at hand' domain." For example, a piece of paper may be easily passed from one person to another or read by more than one person at a time. This was illustrated in Case D, where the tablet PCs were handed around among the team at the patient's bedside, but the laptops on trolleys in Case E could not be so easily shared.

Work may also vary with respect to the *level* of *mobility* involved—predominantly stationary or predominantly mobile—and this seems to affect users' assessments of the value of mobile devices. This is clearest where doctors can "wear" handheld devices (Cases A, B, and E), but for devices like tablet PCs and especially mounted PCs, usage was low in both predominantly stationary activities (e.g., work in the doctors' office with occasional visits to patient rooms) and in highly mobile situations (for example, visiting patients on different floors of the hospital). It was in situations where there was an intermediate degree of mobility with significant periods of static use such as ward rounds (Cases D and E) that doctors seem to prefer these types of MICTs.

The individual and collective nature of the work practice seemed to influence usage of MICT devices. For example, handhelds were mostly considered suitable only for individual use due to small screen size and difficulties of communicating between different models. The team-based nature of some aspects of doctors' work, on the other hand, meant that the social influence of colleagues, especially those of senior grades, could encourage or discourage MICT use.

Perceptions about the objectives of MICT use clearly influenced how users individually (Case F) and collectively adopted them. In Case E, for example, three teams used the wireless laptops to access the EMR, but only one considered and encouraged their use to access online resources as well. Moreover, while the availability of nearby desktop PCs influences perceptions of the need for mobile access to data, this also appeared to depend on the type of data handled and the data analysis carried out. The genetics specialist in Case F, for example, felt that he needed a mobile device because he could be called on to make decisions about patients, based on complex data, at almost any time. Similarly, some senior doctors in Case E felt that the use of MICT devices on ward rounds was valuable to enable them to undertake more sophisticated analyses of patterns in patient data, rather than relying on the memory of the junior doctor presenting the case.

A Framework for Analysing MICT Use

From the above discussion we may identify two broad groups of influences on MICT use in clini-

cal settings, one of which relates primarily to the characteristics of the device, earlier described as the *MICT device at hand*, and the other relating to the particular work practices in which they are employed, which we will describe as the *clinical work context*. This is illustrated in Figure 1.

The cases do not provide sufficient evidence to suggest what particular combination of these influences will lead to enthusiastic use of MICTs, to problematic use of MICTs, or to non-use of MICTs, nor is it claimed that all possible influences have necessarily been identified. Figure 1 is presented, however, as an aid to conceptualisation of the types of influences that may need to be considered in understanding MICT use in clinical work practices, both at the macro level, in terms of addressing both technical and work practice issues, but also as an indication of potential influences (some of which appear to be distinctive to MICTs) within these two areas. The static representation of Figure 1 should also not obscure the dynamics of the balance of these influences in any setting—that is, a change in only one of the characteristics (e.g., loss of battery power or a new senior team member) may be sufficient to alter usage of MICT devices.

CONCLUSION AND FUTURE DIRECTIONS

In this chapter we have presented reports of MICT use in different hospital settings in four hospitals in two countries. Our own results in other countries and clinical settings suggest that it may be more broadly applicable, although it would be valuable if this could be confirmed by other studies. Perhaps the main contribution of this work is in drawing attention to the interaction between the technical and social influences on the use of a particular *MICT device at hand* in a specific *clinical work context.* A similar interplay may also be applicable in understanding the use of MICT in other business contexts. Future research is expected to focus more deeply on the relationships between individual factors and on assessing their relative influence on user behaviour. Research into clinical settings outside hospitals is another area of potential future work, especially as some of these, such as use in emergency vehicles, involve distinctive forms of mobility that may provide additional insights into the usage of MICTs. It is also planned to extend the framework to non-healthcare settings.

ACKNOWLEDGMENTS

Funding for this research was received from Fundação para a Ciência e Tecnologia, Lisbon (BD/8121/2002), and St. Edmunds College, Cambridge. The support of all those involved in obtaining access to the study sites as well as all staff at those sites is gratefully acknowledged.

REFERENCES

Barret, J. R., Strayer, S. M., & Schubart, J. R. (2004). Assessing medical residents' usage and perceived needs for personal digital assistants. *International Journal of Medical Informatics, 73*(1), 25-34.

Brenda, M., & Gadd, S. C. (2001). Introducing handheld computing into a residency program: Preliminary results from qualitative and quantitative inquiry. *Proceedings of the AMIA Symposium* (pp. 428-432).

Cox, J. (2002). *Networked mobile devices help improve patient care and diagnosis.* Retrieved from http://www.nwfusion.com/research/2002/1209sector.html

Davis, G. B. (2002). Anytime/anyplace computing and the future of knowledge work. *Communications of the ACM, 45*(12), 67-73.

Dryer, D. C., Eisbach, C., & Ark, W. S. (1999). At what cost pervasive? A social computing view of mobile computing systems. *IBM Systems Journal, 38*(4), 652-675.

Kelly, J. (2001). Going wireless. *Hospital Health Networks, 74*(11), 65-66, 68.

Kristoffersen, S., & Ljungberg, F. (2000). *Mobility: From stationary to mobile work. Planet Internet.* Lund: Studentlitteratur.

Luff, P., & Heath, C. (1998). Mobility in collaboration. *Proceedings of the ACM 1998 Conference on Computer Supported Cooperative Work,* Seattle, WA.

Martins, H. M. G., & Jones, M. R. (2005). What's so different about mobile information communication technologies (MICT) for clinical work practices: A review of selected pilot studies. *Health Informatics Journal, 11,* 123-134.

Martins, H. M. G., Nightingale, P., & Jones, M. R. (2005). Temporal and spatial organisation of doctors' computer usage in a UK hospital department. *Medical Informatics & the Internet in Medicine, 8*(2), 135-142.

McLeod, T. G., Ebbert, J. O., & Lymp, J. F. (2003). Survey assessment of personal digital assistant use among trainees and attending physicians. *Journal of the American Medical Association, 10*(6), 605-607.

Riley, R. T., Lorenzi, N. M., & Dewan, N. A. (2002). Barriers and resistance to informatics in behavioral health care. In N. A. Dewan, R. R. T. Lorenzi, & S. R. Bhattacharya (Eds.), *Behavorial healthcare informatics* (pp. 140-148). New York, Springer-Verlag.

Westbrook, J. I., Gosling, A. S., & Coiera, E. (2004). Do clinicians use online evidence to support patient care? A study of 55,000 clinicians. *Journal of the American Medical Informatics Association, 11*(2), 113-120.

ENDNOTES

1 We define clinical work practices as doctors' work directly related to patient care (e.g., prescribing a drug or viewing an X-ray for treatment plan decision making). Although equally interesting, doctors' activities such as teaching, research, or all those occurring "outside" hospital working hours are not considered in this chapter.

2 "Jobs" corresponds to discreet tasks to be carried out only by doctors, for example, drawing blood, changing an infusion, doing a small medical procedure. Due to the intensive care aspect of this department, these were numerous and new ones could be required throughout the day and night as the conditions of the patients evolved.

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 69-82, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.13
Integrating Mobile-Based Systems with Healthcare Databases

Yu Jiao
Oak Ridge National Laboratory, USA

Ali R. Hurson
Pennsylvania State University, USA

Thomas E. Potok
Oak Ridge National Laboratory, USA

Barbara G. Beckerman
Oak Ridge National Laboratory, USA

ABSTRACT

In this chapter, we discuss issues related to e-health and focus on two major challenges in distributed healthcare database management: database heterogeneity and user mobility. We designed and prototyped a mobile-agent-based mobile data-access system framework that can address these challenges. It applies a thesaurus-based hierarchical database federation to cope with database heterogeneity and utilizes the mobile-agent technology to respond to the complications introduced by user mobility and wireless networks. The functions provided by this system are described in detail and a performance evaluation is also provided.

INTRODUCTION

The integration of healthcare management and advances in computer science, especially those in the areas of information-system research, has begotten a new branch of science: e-health. E-health is becoming more and more widely recognized as an essential part for the future of both healthcare management and the health of our children. The 2001 President's Information

Technology Advisory Committee, in its report "Transforming Healthcare through Information Technology," noted that information technology "offers the potential to expand access to healthcare significantly, to improve its quality, to reduce its costs, and to transform the conduct of biomedical research"(p. 1). Although much has been done, reality has proven to us that there are still a great number of problems remaining to be taken care of. Health and human-services secretary Mike Leavitt told the Associated Press (2005) in an interview after hurricane Katrina, "There may not have been an experience that demonstrates, for me or the country, more powerfully the need for electronic health records…than Katrina" (p. 1). The article also pointed out that the "federal government's goal is to give most Americans computerized medical records within 10 years"(p. 1).

E-health embraces a broad range of topics, such as telemedicine, medical-record databases, health information systems, genomics, biotechnology, drug-treatment technologies, decision-support systems, and diagnosis aids, just to name a few. In this chapter, we focus on the topic of technologies that deal with integrating mobile-based systems with healthcare databases.

One of the major challenges in healthcare database integration is the fact that the lack of guidance from central authorities has, in many instances, led to incompatible healthcare database systems. Such circumstances have caused problems to arise in the smooth processing of patients between health service units, even within the same health authority (Svensson, 2002). For instance, electronic health record (EHR) systems have been used in practice for many years. However, they are often designed and deployed by different vendors and, thus, patients' information is collected and stored in disparate databases. Due to the lack of uniformity, these systems have very poor interoperability. Even though the wide deployment of networks has enabled us to connect these databases, a large amount of work still needs to be handled manually in order to exchange information between the databases.F

There are two potential solutions to the problems of interoperability and automated information processing: redesigning and reimplementing the existing databases or using a database federation. Redesigning and reimplementing existing databases require large capital investments, and are difficult to achieve. An alternative solution is to build a database federation in which problems caused by database heterogeneity are remedied by the use of a mediator: metadata. This approach is often referred to as the multidatabase solution (Bright, Hurson, & Pakzad, 1994).

The Internet and the client-server-based computing paradigm have enabled us to access distributed information remotely, where the data servers act primarily as an information repository, the user's workstation bears the brunt of the processing responsibility, and the client and server communicate through a well-formulated network infrastructure. Recently, the surge of portable devices and the wide deployment of wireless networks have ushered a new era of mobile computing. Users access information via wireless media and from lightweight and less powerful portable devices. This paradigm shift permits the exchange of information in real time without barriers of physical locations. This is particularly helpful in situations where emergency medical teams need to access patients' information as soon as possible at a disaster site (Potok, Phillips, Pollock, & Loebl, 2003). However, mobile computing has also brought upon several technical challenges. First, unlike workstations, portable devices usually have limited CPU (central processing unit) processing capability and limited battery capacity. Second, low bandwidth and intermittent wireless network connections are often debilitating to client-server applications that depend on reliable network communications.

The mobile-agent-based distributed system design paradigm can address the aforementioned limitations. Unlike the client-server-based compu-

tational model, which moves data to computation, mobile agents move computation to data. This allows mobile users to take advantage of the more powerful servers on the wired networks. In addition, mobile agents are intelligent and independent entities that posses decision-making capabilities. Once dispatched, they are able to fulfill tasks without the intervention of the agent owner. Network connectivity is only required at the time of an agent's submission and retraction. Therefore, the use of mobile agents alleviates constraints such as connectivity, bandwidth, energy, and so forth.

We proposed and developed a prototype of a novel mobile-agent-based mobile data-access system (MAMDAS) for heterogeneous healthcare database integration and information retrieval (Jiao & Hurson, 2004). The system adopts the summary-schemas model (SSM; Bright et al., 1994) as its underlying multidatabase organization model. Queries are carried out by mobile agents on behalf of users. Via a medical thesaurus, created by combining the Medical Subject Headings (MeSH) thesaurus (Chevy, 2000) and an English-language thesaurus WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), MAMDAS supports imprecise queries and provides functions for user education.

The purpose of this chapter is to provide details about the tools we developed for disparate healthcare database management and their potential applications. The rest of the chapter is organized as follows. First it provides the background, and then presents the design, functions, application, and performance evaluation of MAMDAS and a medical thesaurus *MEDTHES*. Finally, we summarize this chapter and discuss future trends.

BACKGROUND

In this section, we provide an overview of the current solutions to healthcare database management and introduce the two technologies on which we built our system: SSM and mobile-agent technology.

HEALTHCARE DATABASE SYSTEMS

The Veterans Health Administration (VHA) clinical information system began in 1982 as the Decentralized Hospital Computer Program (DHCP) and is now known as VistA (Veterans Health Information Systems and Technology Architecture; Hynes, Perrin, Rappaport, Stevens, & Demakis, 2004). VistA has evolved into a very rich healthcare information system that provides the information-technology framework for VHA's approximately 1,300 sites of care. VistA is built on a client-server architecture that ties together workstations and personal computers with nationally mandated and locally adapted software access methods. More specifically, VistA comprises more than 100 applications that clinicians access via the Computerized Patient Record System (CPRS) GUI (graphical user interface) to pull all the clinical information together from the underlying facility-based programming environment. CPRS provides a single interface for healthcare providers to review and update a patient's medical record. More than 100,000 VHA healthcare providers nationwide currently use CPRS. One important reason for VistA's success is the existence of a central authority. All VHA facilities are mandated to apply the same database-management system and unified access methods, which significantly eases the problem of interoperability among systems at different sites. Unfortunately, this uniformity is not a norm in today's healthcare databases. More often, we have to deal with heterogeneous databases that are designed and developed by different vendors.

IBM's DiscoverLink targets applications from the life-sciences industry (Hass et al., 2001). It is a fusion of two major components: Garlic (Carey et al., 1995) and DataJoiner (Cahmberlin, 1998).

Garlic is a federated database-management system prototype developed by IBM Research to integrate heterogeneous data. DataJoiner is an IBM federated database-management product for relational data sources based on DATABASE 2 (Cahmberlin). It is a mediator system that limits itself to metadata exchange and leaves the data in their original databases and format. When an application submits a query to the DiscoveryLink server, the server identifies the relevant data sources and develops a query execution plan for obtaining the requested data. The server communicates with a data source by means of a wrapper, a software module tailored to a particular family of data sources. The wrapper is responsible for mapping the information stored by the data source into DiscoveryLink's relational data model, informing the server about the data source's query-processing capability, mapping the query fragments submitted to the wrapper into requests that can be processed using the native query language of the data source, and issuing query requests and returning results. Since data sources may take one of the many formats—relational database, object-oriented database, or flat files such as XML (extensible markup language) files and text files—a wrapper is needed for each format. Thus, wrapper development is the key to the extensibility in DiscoveryLink.

The TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) project (Stevens et al., 2000), as its name suggests, aims to provide transparent access to disparate biological databases. TAMBIS includes a knowledge base of biological terminology (the biological concept model), a model of the underlying data sources (the source model), and a knowledge-driven user interface. The concept model provides the user with the concepts necessary to construct multiple-source queries, and the source model provides a description of the underlying sources and mappings between the terms used in the sources and the terms defined in the concept model. In other words, TAMBIS utilizes a domain-specific ontology for heterogeneous data-source integration. It is a novel and valid approach. However, the depth and quality of the TAMBIS ontology are difficult to evaluate because the ontology contents are not publicly available.

The PQL query language proposed by Mork, Shaker, Halevy, and Tarczy-Homoch (2002) intends to integrate genetic data distributed across the Internet. It is essentially a query language for semistructured data. It relies on metadata describing the entities and the relationships between entities in a federated schema. These metadata appear to be created manually. While providing a new query language, this approach also raises questions about the accuracy of the metadata and extensibility of the system.

The Query Integration System (QIS) of Marenco, Wang, Shepherd, Miller, and Nadkami (2004) is a database mediator framework that addresses robust data integration from continuously changing heterogeneous data sources in the biosciences. The QIS architecture is based on a set of distributed network-based servers, data-source servers, integration servers, and ontology servers that exchange metadata as well as mappings of both metadata and data elements to elements in an ontology. Metadata version difference determination coupled with the decomposition of stored queries is used as the basis for partial query recovery when the schema of data sources alters. The principal theme of this research is handling schema evolution.

We developed a prototype of a mobile-agent-based mobile data-access system that deals with heterogeneous healthcare data-source integration and information retrieval (Jiao & Hurson, 2004). Our work differs from the previously mentioned research in several ways. First, MAMDAS utilizes the summary-schemas model for multidatabase organization (Bright et al., 1994). The hierarchical structure of SSM enables automated metadata population and improves search efficiency. Second, supporting user mobility is an emerging demand and it has not yet received enough atten-

tion in healthcare information-system research. We proposed to apply the mobile-agent technology to cope with this issue. Third, existing biomedical thesauri often demonstrate poor interoperability and reusability due to their nonstandard designs. We modified the MeSH thesaurus (Chevy, 2000) so that it complies with the ANSI/NISO (American National Standard Institute/National Information Standards Organization) Z39.19 monolingual thesaurus-creation standard (NISO, 1994). In addition, most biomedical thesauri and ontologies are tailored to the needs of medical professionals and, thus, nonprofessionals often find them hard to use due to the lack of precise knowledge. We addressed this problem by augmenting MeSH terms with synonyms defined by a general English-lexicon thesaurus WordNet (Miller et al., 1990). Finally, MAMDAS can be coupled with thesauri or ontologies of different domains to provide an information-system infrastructure for various applications with minimal modification. In the following subsections, we briefly discuss the background information pertinent to the development of MAMDAS.

The Summary-Schemas Model

The SSM consists of three major components: a thesaurus, local nodes, and summary-schemas nodes. Figure 1 depicts the structure of the SSM. The thesaurus defines a set of standard terms that can be recognized by the system, namely, global terms, and the categories they belong to. Each physical database (local nodes) may have its own dialect of those terms, called local terms. In order to share information among databases that speak in different dialects, each physical database maintains local-global schema metadata that map each local term into a global term in the format of "local term: global term." Global terms are related through synonym, hypernym, and hyponym links. The thesaurus also uses a semantic-distance metric (SDM) to provide a quantitative measurement of semantic similarity between terms. This feature allows for fine-grained semantic-based information retrieval.

The cylinders and the ovals in Figure 1 represent local nodes and summary-schemas nodes, respectively. A local node is a physical database containing real data. A summary-schemas node is a logical database that contains metadata called summary schema, which store global terms and lists of locations where each global term can be found. The summary schema represents the schemas of the summary-schema node's children in a more abstract manner; it contains the hypernyms of the input data. As a result, fewer terms are used to describe the information than the union of the terms in the input schemas.

Figure 2 shows an example of the automated schema-abstraction process of four local terms, *human face*, *ear*, *navel*, and *belly button*, under

Figure 1. A summary-schemas model with M local nodes and N levels

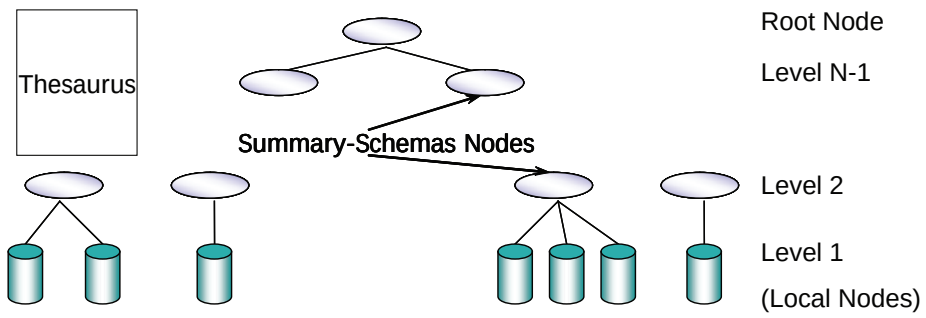
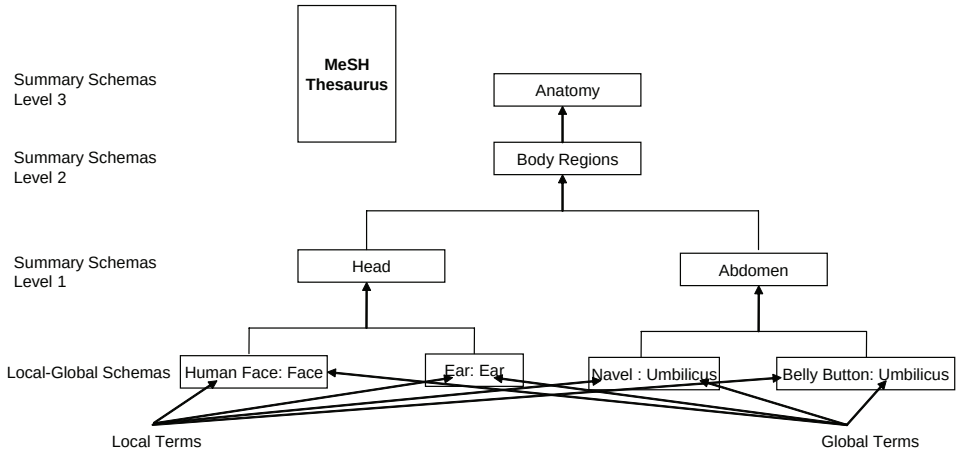


Figure 2. An example of the schema-summarization process



the guidance of the MeSH thesaurus in a bottom-up fashion. First, all local terms are mapped into global terms that are terms defined in MeSH. In the current prototype, this step is done by local database administrators manually. At SSM Level 1, the least common ancestors (immediate hypernyms) of the global terms are automatically identified by searching through the MeSH hierarchy: *Head* is the immediate hypernym of *face* and *ear*. Similarly, *abdomen* is the hypernym of *umbilicus*. At Summary Schemas Level 2, *head* and *abdomen* are further abstracted into *body regions*. Finally, at Level 3, *body region* is found to be a hyponym of 1 of the 15 categories defined in MeSH: *anatomy*.

The SSM is a tightly coupled federated database solution and the administrator is responsible for determining the logical structure of it. In other words, when a node joins or leaves the system, the administrator is notified and changes to the SSM are made accordingly. Note that once the logical structure is determined, the schema-population process is automated and does not require the administrator's attention.

The major contributions of the SSM include preservation of the local autonomy, high expandability and scalability, short response time, and the resolution of imprecise queries. Because of the unique advantages of the SSM, we chose it as our underlying multidatabase organization model.

The Mobile-Agent Technology

An agent is a computer program that acts autonomously on behalf of a person or organization (Lange & Oshima, 1998). A mobile agent is an agent that can move through the heterogeneous network autonomously, migrate from host to host, and interact with other agents (Gray, Kotz, Cybenko, & Rus, 2000). Agent-based distributed application design is gaining prevalence, not because it is an application-specific solution—any application can be realized as efficiently using a combination of traditional techniques. It is more because of the fact that it provides a single framework that allows a wide range of distributed applications to be implemented easily, efficiently, and robustly. Mobile agents have many advantageous properties (Lange & Oshima) and we only highlight some of them here:

- **Support disconnected operations:** Mobile agents can roam the network and fulfill their tasks without the owner's intervention. Thus, the owner only needs to maintain the physical connection during submission and retraction of the agent. This asset makes mobile agents desirable in the mobile computing environment where intermittent network connection is often inevitable.
- **Balance workload:** By migrating from the mobile device to the core network, the agents can take full advantage of the high bandwidth of the wired portion of the network and the high computation capability of servers and workstations. This feature enables mobile devices that have limited resources to provide functions beyond their original capability.
- **Reduce network traffic:** Mobile agents' migration capability allows them to handle tasks locally instead of passing messages between the involved databases. Therefore, fewer messages are needed in accomplishing a task. Consequently, this reduces the chance of message losses and the overhead of retransmission.

Contemporary mobile-agent system implementations fall into two main groups: Java-based and non-Java-based. We argue that Java-based agent systems are better in that the Java language's platform-independent features make it ideal for distributed application design. We chose the IBM Aglet Workbench SDK 2.0 (*IBM Aglets Workbench*, 1996) as the MAMDAS' implementation tool.

DESIGN, FUNCTIONS, APPLICATION, AND PERFORMANCE EVALUATION OF MAMDAS AND MEDTHES

Mobile-Agent-Based Mobile Data-Access System

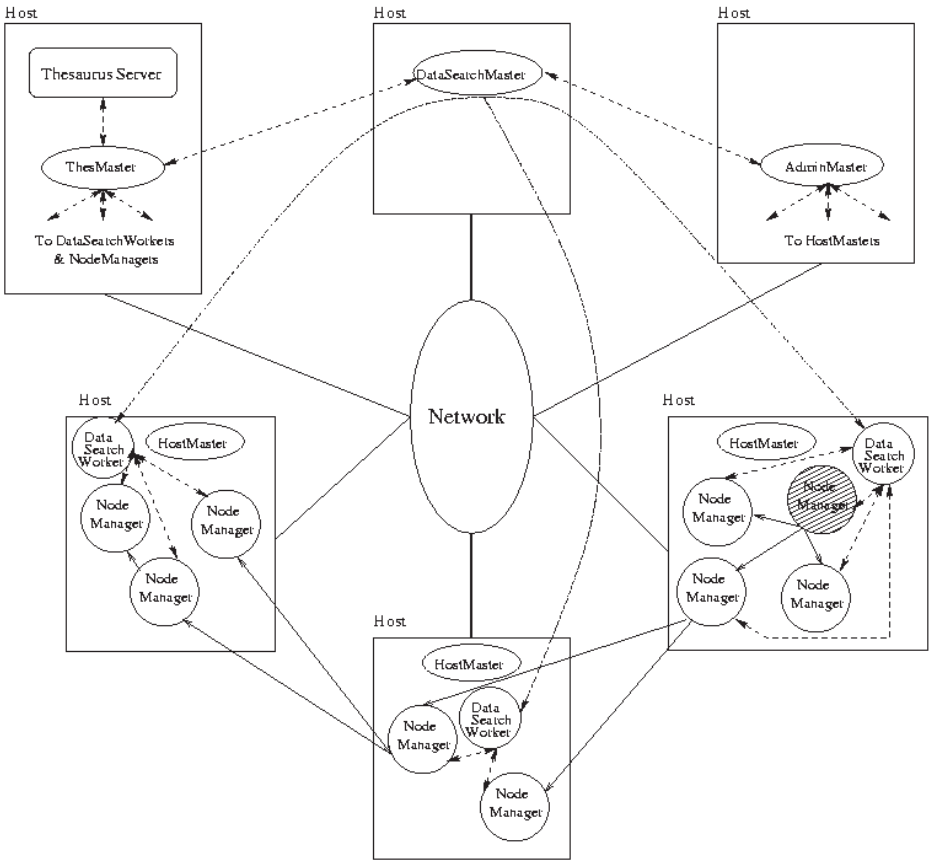
MAMDAS consists of four major logical components: the host, the administrator, the thesaurus, and the user (Jiao & Hurson, 2004). Figure 3 illustrates the overall architecture of MAMDAS.

The MAMDAS can accommodate an arbitrary number of hosts. A HostMaster agent resides on each host. A host can maintain any number and any type of nodes (local nodes or summary-schemas nodes) based on its resource availability. Each NodeManager agent monitors and manipulates a node. The HostMaster agent is in charge of all the NodeManager agents on that host. Nodes are logically organized into a summary-schemas hierarchy. The system administrators have full control over the structure of the hierarchy. They can construct the structure by using the graphical tools provided by the AdminMaster agent.

In Figure 3, the solid lines depict a possible summary-schemas hierarchy with the darkened node as the root and the arrows indicating the hierarchical relation. The ThesMaster agent acts as a mediator between the thesaurus server and other agents. The dashed lines with arrows indicate the communication between the agents. The DataSearchMaster agent provides a query interface, the data-search window, to the user. It generates a DataSearchWorker agent for each query. The three dash-dot-dot lines depict the scenario that three DataSearchWorker agents are dispatched to different hosts and work concurrently.

Once the administrator decides the summary-schemas hierarchy, commands will be sent out to each involved NodeManager agent to build the structure. NodeManagers at the lower levels ex-

Figure 3. An overview of the MAMDAS system architecture



port their schemas to their parents. Parent nodes contact the thesaurus and generate an abstract version of their children's schemas. When this process reaches the root, the MAMDAS is ready to accept queries.

The user can start querying by launching the DataSearchMaster on his or her own device, which can be a computer attached to the network or a mobile device. The DataSearchMaster sends out two UserMessengers (not shown in the figure): one to the AdminMaster and one to the ThesMaster. The UserMessengers will return to the DataSearchMaster with the summary-schemas hierarchy and the category information. The DataSearchMaster then creates a data-search window that shows the user the summary-schemas hierarchy and the tree structure of the category. The user can enter the keyword, specify the preferred semantic distance, choose a category, and select a node to start the search. After the user clicks on the "Submit" button, the DataSearchMaster packs the inputs, creates a DataSearchWorker, and passes the inputs to it as parameters. Since the DataSearchMaster creates a DataSearchWorker to

Figure 4. The search algorithm

```
1     Set all child nodes to be unmarked;
2     WHILE (NOT (all term(s) are examined OR all
      child node(s) are marked))

          IF (term is of interest)
3             Mark all the child nodes that
      contain this term;
4         ELSE
5             CONTINUE;
6         END IF
7     END WHILE
8     IF (no marked child node)

            Go to the parent node of the
current node and repeat the search
       algorithm (if a summary schema term
of the parent node only exists on the
    current node, we can skip this term);
9     ELSE
10        Create a DataSearchSlave for each marked
      child node;
11        Dispatch the slaves to the destinations
      and repeat the search algorithm;
12    END IF
13
```

handle each query, the user can submit multiple queries concurrently.

Once dispatched, the DataSearchWorker can intelligently and independently accomplish the search task by making local decisions without the owner's interference. During the query execution, the DataSearchWorker may generate DataSearchSlaves by cloning itself. The slaves can then work in parallel and report results to their creator. Figure 4 describes the search algorithm.

One of the major advantages of the MAMDAS framework is that it supports database heterogeneity and geographical distribution transparency. It provides the users with a uniform access interface. This property of MAMDAS significantly eases the use of the system and makes it possible for users with limited computer skills to benefit from it.

A Medical Thesaurus: MEDTHES

The quality of the thesaurus is critical to the effectiveness of MAMDAS because it provides semantic-similarity measures to assist users in performing imprecise queries in which the query term is different than the indexing term of a document. The proliferation of biomedical research and the public demand of e-healthcare systems have stimulated the development of biomedical thesauri. Several examples include MeSH (Chevy, 2000), the Unified Medical Language System (UMLS; McCray & Nelson, 1995), and the Systematized Nomenclature of Medicine (SNOMED; Spackman, Campbell, & Cote, 1997). While the existing medical thesauri have helped immensely in information categorization, indexing, and retrieval, two major problems remain:

- Their designs do not follow any international or national thesaurus standard and therefore they could result in poor interoperability and reusability.
- They do not provide information regarding the semantic similarities among terms and,

thus, the users are required to possess precise knowledge of the controlled vocabulary in order to make effective use of the thesaurus.

In order to alleviate these problems, we implemented a new medical thesaurus MEDTHES based on the medical thesaurus MeSH (Chevy, 2000) and the English-language thesaurus WordNet (Miller et al., 1990). It can be used as either a stand-alone thesaurus or an integral part of MAMDAS. In this subsection, we (a) briefly outline the ANSI/NISO standard for thesauri construction, (b) describe the two thesauri that have served as the foundation of MEDTHES, MeSH, and WordNet, (c) explain the concept of semantic similarity, (d) present the implementation of MEDTHES, (e) demonstrate the functions provided by MEDTHES as a stand-alone thesaurus, and (f) show the integration of MEDTHES with MAMDAS.

The ANSI/NISO Z39.19 Standard

The ANSI/NISO Z39.19 standard (NISO, 1994), entitled *American National Standard Guidelines for the Construction, Format, and Management of Monolingual Thesauri*, was developed by NISO and approved by ANSI. It provides guidelines for the design and use of thesauri, including rules for term selection, thesaurus structure, relation definitions, and thesaurus maintenance. Three types of semantic relationships between terms are distinguished in this standard: equivalence, hierarchical, and related. The equivalence relation establishes the link between synonyms, the hierarchical relationship provides links between terms that reflect general concepts (broader terms) and those that represent more specific information (narrower terms), and the related relationship exists among terms that have similar meanings or are often used in the same context but do not have hierarchical relationships. The design of MEDTHES follows this standard.

MeSH

The MeSH (Chevy, 2000) thesaurus is the standardized vocabulary developed by the National Library of Medicine for indexing, cataloging, and searching the medical literature. Currently, it contains approximately 22,000 terms (called descriptors) that describe the biomedical concepts used in health-related databases such as MEDLINE (*MEDLINE*, 2005), which is an online bibliographic database of medicine, nursing, health services, and so forth. All descriptors in MeSH are organized into 15 categories. Each category is then further divided into more specific subcategories. Within each category, descriptors are organized in a hierarchical fashion of up to 11 levels. In addition to the hierarchical structure, MeSH uses "Entry Term" or "See" references to indicate semantic relations such as synonyms, near synonyms, and related concepts of some terms.

Although MeSH is comprehensive and well maintained, it has several drawbacks. First, the synonymous relationship is not clearly listed and not differentiated from the related-term relation in MeSH. Second, the design of MeSH does not follow the ANSI thesaurus standard, which may result in poor interoperability and reusability. Third, MeSH is tailored to the needs of medical professionals. Nonprofessionals often find it hard to perform queries due to the lack of precise knowledge. For instance, a nonprofessional would use search terms such as *navel* and *belly button* instead of the official term *umbilicus* when submitting a query. Unfortunately, the query will fail because these terms are not defined in MeSH. We addressed this problem by augmenting MeSH with the well-defined synonyms found in WordNet, which we will discuss next.

WordNet

WordNet is an online thesaurus that models the lexical knowledge of the English language (Miller et al., 1990). It organizes English nouns, verbs,

adjectives, and adverbs into synonym sets, called synsets. In other words, a synset is a list of synonymous terms. Each term in WordNet may have one or more meanings, and each meaning has a synset. Different synsets are connected through hierarchical relationships.

In summary, WordNet is comprehensive and designed with the goal to include every English word; it makes a number of fine-grained distinctions among word meanings. Thus, we decided to take advantage of the well-defined synonyms of WordNet and use them to complement the MeSH thesaurus.

Semantic Similarity

Synonyms and related terms obtained from a thesaurus are often used in query expansion for the purpose of improving the effectiveness of information retrieval (Shiri, Revie, & Chowdhury, 2002). However, in order to improve the quality of document ranking, a more fine-grained measure is needed to describe the degree of semantic similarity, or more generally, the relatedness between two lexically expressed concepts (Budanitsky & Hirst, 2001). Naturally, semantic distance is the inverse of semantic similarity. For example, the semantic distance between synonyms can be defined as zero, and that between antonyms can be defined as infinite.

If a thesaurus provides functions that calculate the semantic similarity between terms, the users can perform fine-tuned queries by limiting the scope of the search via the constraint of semantic distance between the keyword and the search results. The user can indicate how closely the returned terms should be related to the keyword (searched term) by selecting preferred semantic-distance values.

Two main categories of algorithms for computing the semantic distance between terms organized in a hierarchical structure (e.g., WordNet) have been proposed in the literature: distance-based approaches and information-content-based approaches. The general idea behind the distance-based algorithms (Leacock & Chodorow, 1998; Rada, Mili, Bicknell, & Blettner, 1989; Wu & Palmer, 1994) is to find the shortest path between two terms based on the number of edges, and then translate this distance into semantic distance. Information-content-based approaches (Jiang & Conrath, 1997; Rada et al.) are inspired by the perception that pairs of words that share many common contexts are semantically related. Thus, the basic idea of these methods is to quantify the frequency of the co-occurrences of words within various contexts.

In order to avoid the potential bias introduced by context selection, we chose to implement three distance-based algorithms in the MEDTHES prototype: the edge-counting algorithm (Rada et al., 1989), the Leacock and Chodorow (1998) algorithm, and the Wu and Palmer (1994) algorithm.

The Edge-Counting Algorithm

In the edge-counting algorithm, the semantic distance is defined as the number of edges (nodes) along the shortest path between any two terms.

The Leacock and Chodorow Algorithm

The relatedness measure proposed by Leacock and Chodorow (1998) also relies on the shortest path between two terms, t_1 and t_2. The relatedness between two terms, t_1 and t_2, is calculated as follows.

$$relatedness(t_1, t_2) = -\log \frac{len(t_1, t_2)}{2D} \qquad (1)$$

where relatedness (t_1, t_2) is the similarity of terms t_1 and t_2, $len(t_1,t_2)$ is the length of the shortest path between two terms (using edge counting), and D is the maximum depth of the structure. Semantic distance is the inverse of relatedness (t_1, t_2), that is,

$$\frac{1}{relatedness(t_1, t_2)}.$$

The Wu and Palmer Algorithm

The Wu and Palmer (1994) algorithm uses the term *score* to define how two terms are related to each other. It measures the score by considering the depth of the two terms t_1 and t_2 in the tree structure, along with the depth of the LCA (least common ancestor). The formula used to calculate the score is shown in Equation 2.

$$score(t_1, t_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \qquad (2)$$

where N_1 is the length of the shortest path from t_1 to the LCA, N_2 is the length of the shortest path from t_2 to the LCA, and N_3 is the length of the shortest path from the LCA to the root. The range of relatedness is $0 < score(t_1,t_2) <= 1$. The $score(t_1,t_2)$ is 1 if t_1 and t_2 are the same. Semantic distance is the inverse of $score(t_1,t_2)$, that is,

$$\frac{1}{score(t_1, t_2)}.$$

MEDTHES Design

The taxonomy defined in MeSH is the foundation of MEDTHES. However, several major changes to MeSH have been made: (a) the semantic relations of MeSH were reconstructed according to the ANSI standard, (b) the synonym set of each entry in MeSH was enriched by synonyms extracted from WordNet, and (c) three algorithms of semantic-distance calculation were implemented in order to provide users with fine-grained control over the query results.

MEDTHES adopts the three standard relationships suggested by the ANSI/NISO standard for thesaurus construction: the equivalence relationship, hierarchical relationship, and associative relationship. Terms in MeSH are arranged hierarchically in a tree structure—top down from general to more specific. The broader term (BT) and narrower term (NT) relations can be easily extracted from this hierarchical structure. A program, MeSHFileParser, was developed to automatically parse such information.

In MeSH, synonyms and related terms (RTs) are not clearly differentiated. The definitions of synonyms are neither accurate nor complete. As a result, MeSH is not suitable to be used directly to obtain synonyms. Since the well-defined synonyms are one of the major strengths of WordNet, it was used as a reference when adding synonyms to MEDTHES. A term is selected from the synonym set as the preferred term, which means that term is used for (UF) indexing other terms in the same set. The reverse relation is use, which means that if a keyword is a nonpreferred term, it is substituted with the preferred term before searching.

A term may exist in one or more categories in MeSH. In order to establish a link between a term and the category it belongs to, an additional relationship, subject categories (SCs), was also defined. Table 1 summarizes the relationships used in MEDTHES.

Main Functions of MEDTHES

As shown in Figure 5, functions featured in MEDTHES include carrying out an imprecise search, calculating semantic distance, browsing MEDTHES, adding new terms, updating existing terms, and resetting MEDTHES. In this subsection, we demonstrate the usage of each of these functions.

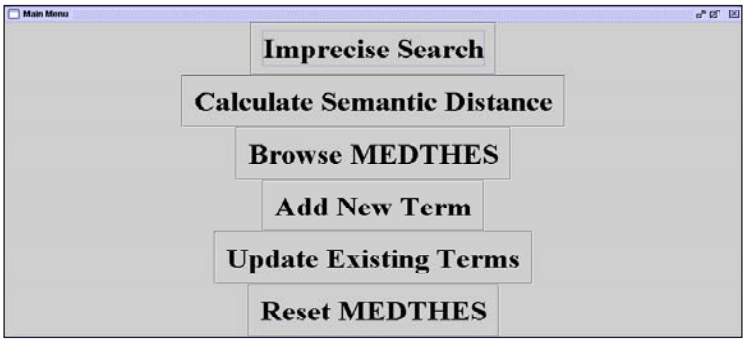
The imprecise-search function returns terms that are within a certain semantic distance to the keyword under a specific category. In the example shown in Figure 6, the category is Anatomy, the keyword is Brain, the semantic distance is 2.0, and the semantic-distance method selected is edge counting. Figure 7 shows the results and lists all the terms that have a semantic distance, with respect to the term Brain, as less than or equal to 2.0 according to the edge-counting algorithm. The BT, NT, UF, and SC of these terms are also listed.

Calculating the semantic distance is the function that can be used to find the semantic distance between any two terms. It also provides the ability

Table 1. Relationship definitions in MEDTHES

ANSI/NISO Relationship	MEDTHES Representation	Abbreviation
Equivalence	Use	USE
	Used For	UF
Hierarchical	Broader Term	BT
	Narrower Term	NT
Associative	Related Term	RT
	Subject Category	SC

Figure 5. Main menu



to compare the results generated by three different semantic-distance methods. Figure 8 shows an example demonstrating the semantic distances between two terms, Ankle and Toes, calculated by using different algorithms. The results are shown in Figure 9. This function endows the users with the opportunity of becoming familiar with the typical range of values of the various algorithms to aid in better selecting a good value when defining imprecise queries.

Browsing MEDTHES is the function that enables users to learn the content of MEDTHES including the tree structure and the alphabetical list of terms. Figure 10 shows a browser window.

Updating existing terms and adding new terms are functions that give users the freedom of customizing MEDTHES. Updating existing terms is a function that enables users to designate relationships among terms with their own knowledge instead of using the predefined interterm relationships. It helps users to modify their copies of MEDTHES according to their needs, such as changing the relationships among the terms and moving a term from one category to another.

Adding a new term is a function that can be used to add new terms to MEDTHES and establish relationships between the added terms and existing terms. A new term can be entered as the synonym of an existing term, in which case the new term is added to the synonym set of the existing term, and all other relations, for example, BT, NT, and RT are automatically established. A

Figure 6. Imprecise search

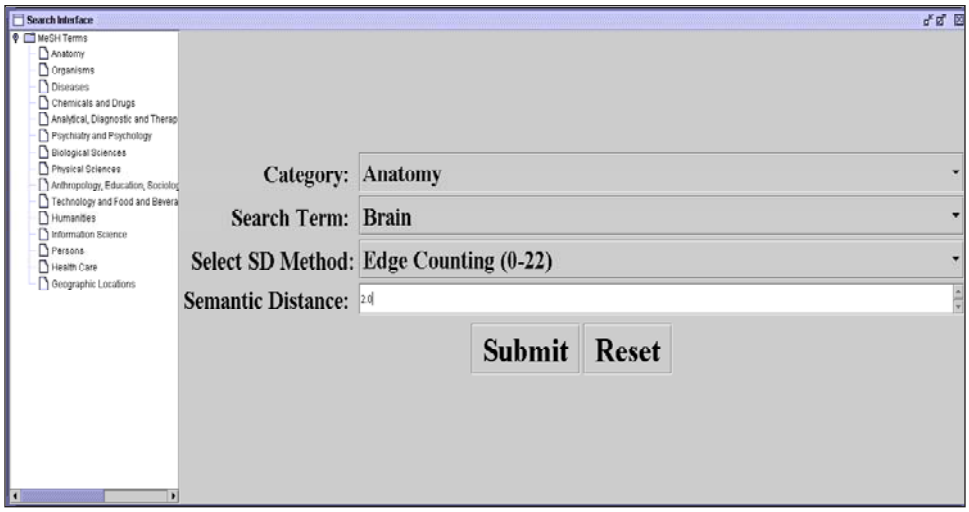


Figure 7. Result of the imprecise search

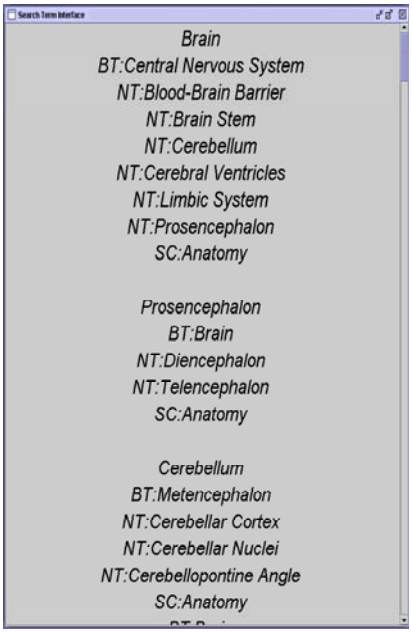


Figure 8. Calculate semantic distance

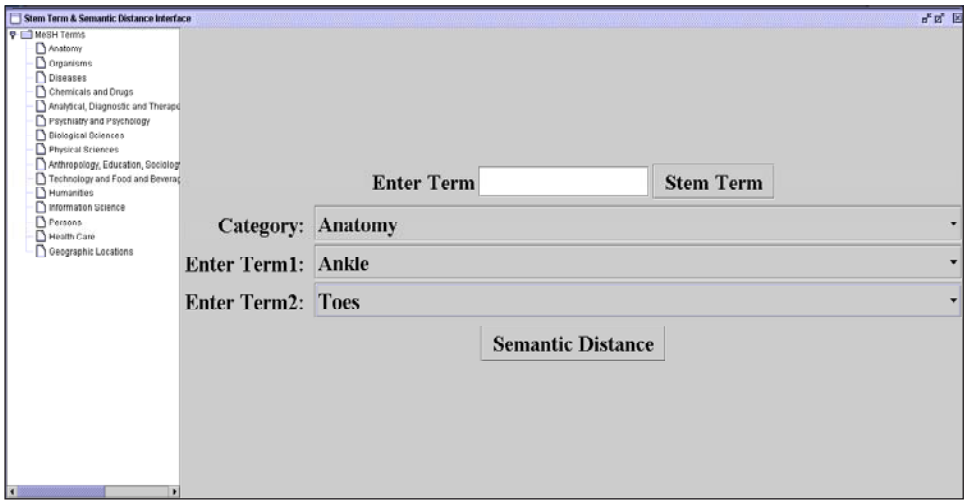


Figure 9. Results of semantic distances

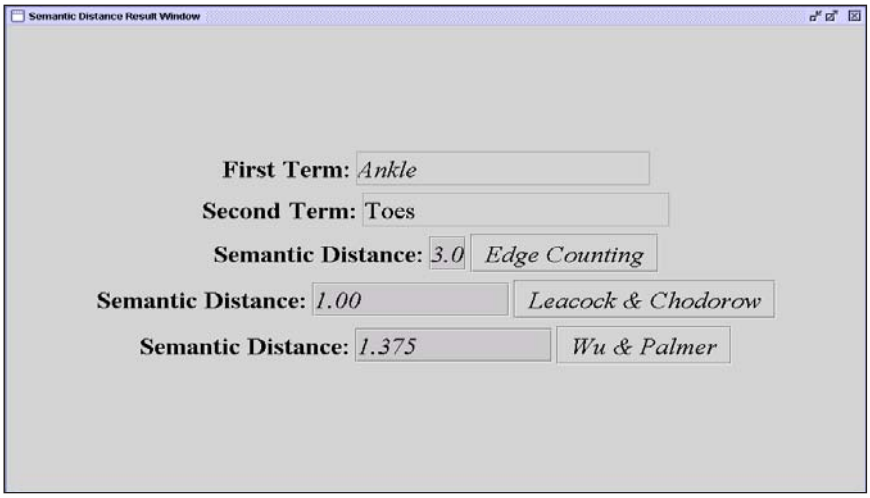


Figure 10. MEDTHES browser

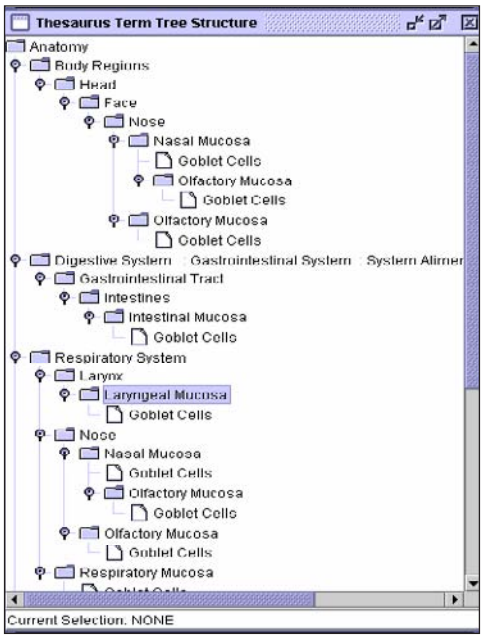


Figure 11. Add new term

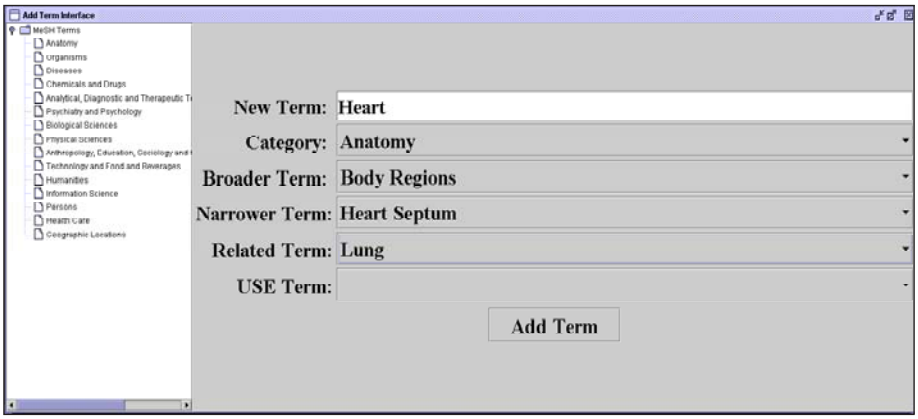
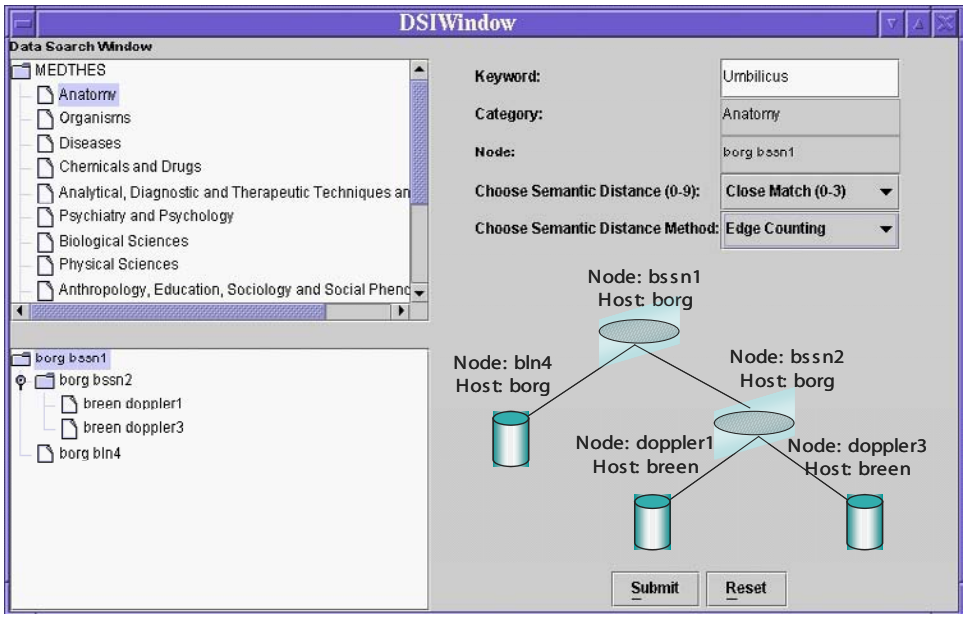


Figure 12. MAMDAS data-search GUI



new term and its relations can also be manually added, if desired. Figure 11 illustrates how to manually add a new term.

Integrate MEDTHES with MAMDAS

When MAMDAS is used in different application domains, the only modification required is to change the thesaurus in a plug-and-play fashion. In other words, we should choose a thesaurus that is the most appropriate for that domain. MAMDAS can work with any thesauri that follow the ANSI/NISO Z39.19 standard. In our study, we integrated MEDTHES with MAMDAS in order to resolve queries in the biomedical domain. Figure 12 shows an example of the MAMDAS data-search GUI.

The top-left window shows the content of MEDTHES, and the lower left window contains the current multidatabase hierarchy. The tree structure shown on the right hand side is an equivalent representation of the multidatabase hierarchy shown on the left. A query can be sent to any node in this hierarchy. In this example, the search keyword is umbilicus in the anatomy category, and the search starts at the root of the multidatabase hierarchy. In other words, every node (data source) in the hierarchy should be searched. The user wishes the search engine to return all terms that have a semantic distance of less than or equal to 3 with respect to the keyword according to the edge-counting algorithm. The search results are shown in Figure 13: Three terms, navel, umbilicus, and belly button, which satisfy the user-specified semantic distance, were found from three different data sources. Although the terms navel and belly button are not the exact lexical match of umbilicus, they are returned because they are defined as synonyms of umbilicus in MEDTHES. This example demonstrates that MEDTHES can be easily and successfully incorporated into MAMDAS and provides semantic-related information.

Figure 13. Search results

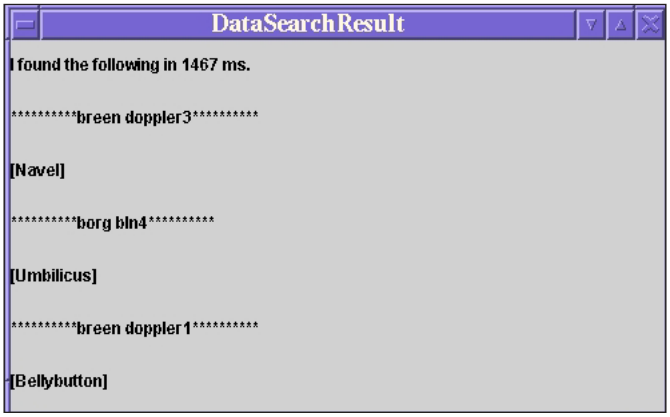
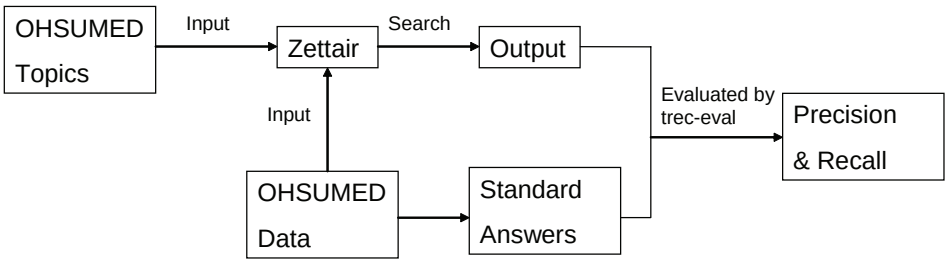


Figure 14. Experiment setup



Performance Evaluation

The effectiveness of information retrieval is assessed by the ability of the system to retrieve relevant documents while at the same time suppressing the retrieval of irrelevant documents. Thesauri-aided query expansion is one frequently used technique in information retrieval that improves the effectiveness of retrieval by adding terms related to the search keyword. We evaluated the practicality and effectiveness of MEDTHES by using it as the reference thesaurus in several query-expansion experiments.

Experiment Setup

We chose to use the OHSUMED(Oregon Health Sciences University's MEDLINE data collection) test collection, a large interactive test collection for information-retrieval evaluation, created by Hersh, Buckley, Leone, and Hickman (1994) at the Oregon Health Sciences University. It is a subset of the MEDLINE medical-domain abstracts and consists of 348,566 articles derived from a subset of 270 medical journals over the period from 1987 to 1991. The documents in the test collection are manually indexed by professional indexers by using the MeSH thesaurus. A set of 106 queries and corresponding relevance judgments are provided.

The queries were generated by actual physicians in the field of patient care and have at least one definitely relevant document. Each query contains a brief statement about the patient, followed by the actual information needed. OHSUMED has been widely used with many information-retrieval systems to evaluate the performance improvement of query expansion in the medical domain. It was also included in the TREC9 (Text Retrieval Conference, Track 9) Filtering Track (Robertson & Hull, 2001).

The data search engine used in our experiments is Zettair. It is a compact text search engine designed and developed by the Search Engine Group at RMIT University (*Zettair Search Engine,* 2003). It has been designed for simplicity as well as speed and flexibility. Users can use Zettair to index and search HTML (hypertext markup language) or TREC data collections. While working with TREC data, it takes a TREC topic file and an indexed data file as input, and it generates the search result in a format that can be analyzed by the trec-eval package.

The trec-eval package is a program that evaluates the documents retrieved by a search engine using performance metrics such as precision and recall. Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved. Recall is defined as the ratio of the number of relevant documents retrieved to the total number of relevant documents in the database. The trec-eval package also contains several measures derived from the precision and recall metrics. The measures that have been commonly used to compare different experiment runs are the recall-precision curve and the mean average precision. We used these two measures in our study.

Four main components are included in the experiments (Figure 14): (a) OHSUMED test topics and query-expansion techniques, (b) the OHSUMED test data collection that provides a set of documents, a set of topics, and standard answers, (c) the Zettair search engine that serves as the full-text search and retrieval engine, and (d) the trec-eval package that assesses query results and calculates the precision and recall.

The Zettair search engine takes OHSUMED topics (queries), either expanded or without expansion, and documents from the OHSUMED test collection as input. There are four types of query expansion: synonym expansion (Run 2), narrower term expansion (Run 3), synonym and narrower term expansion (Run 4), and broader term expansion (Run 5). In our experiments, we used the synonyms, narrower terms, and broader

Table 2. Five experiment runs

Mode	Name	Description
Run 1	Baseline	Use initial queries without expansion.
Run 2	Synonym Expansion	Initial queries are extended with synonyms.
Run 3	Narrower Term Expansion	Initial queries are extended with narrower terms (one level).
Run 4	Synonym and Narrower Term Expansion	Initial queries are extended with a combination of synonyms and narrower terms (one level).
Run 5	Broader Term Expansion	Initial queries are extended with broader terms (one level).

Figure 15. Eleven-point interpolated precision and recall

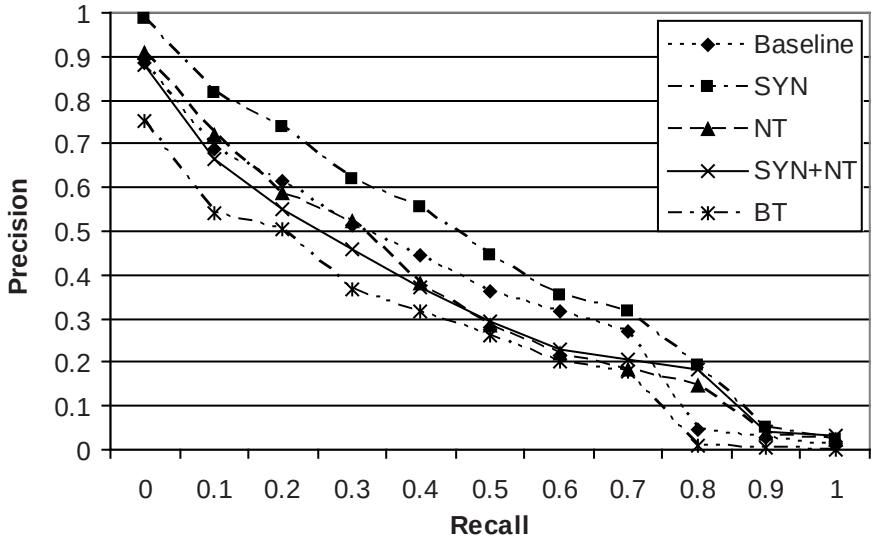
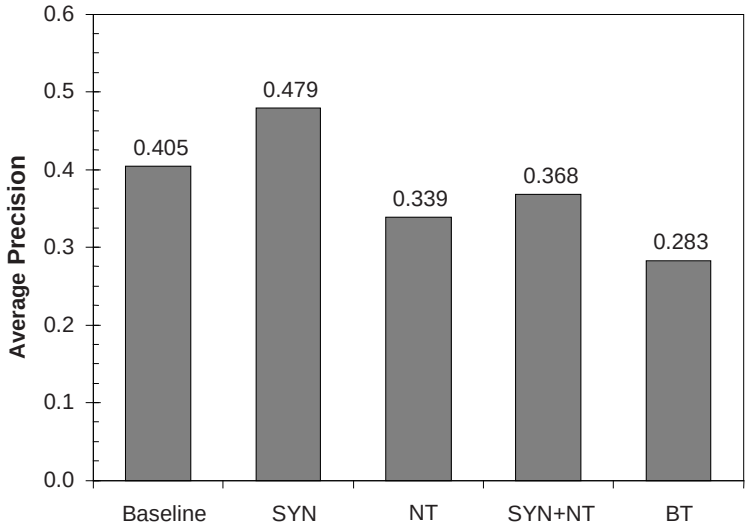


Figure 16. Average precision



terms defined in MEDTHES. The output from Zettair is compared with the standard answers provided by OHSUMED test collections by using the trec-eval package. Precision and recall are then generated automatically. Since the title portion of each topic resembles a typical user query, we used the titles in all the experiments. Our experiments included five distinct modes of tests. Table 2 describes each of them.

Experimental Results and Analysis

Figure 15 plots the 11-point interpolated precision and recall curves and Figure 16 shows the average precision of each run. The results show that only synonym expansion (Run 2) can improve the performance. However, narrower term expansion (Run 3), synonym and narrower term expansion (Run 4), and broader term expansion (Run 5) worsens the performance compared to the baseline due to the introduced noise. The general trends demonstrated by our findings are consistent with the results reported by Hersh, Price, and Donohoe (2000), who used the same data collection. However, we observe two major differences. First, the synonym expansion was done manually in the research of Hersh et al., whereas it was automatically added by using WordNet as a reference in our experiments. Second, synonym expansion achieved good average precision: In our work, synonym expansion increases the average precision by 7.4%. In contrast, the results reported by Hersh et al. show that adding synonyms to the query actually degrades the average precision. This evidence has led us to conclude that by combining the strength of MeSH and WordNet, we can enhance the effectiveness of information retrieval without requiring experts' involvement in MeSH modifications.

CONCLUSION AND FUTURE TRENDS

This study provides an innovative solution, MAMDAS, to integrating a mobile-based system with healthcare databases. It utilizes the summary-schemas model to address the difficulties of heterogeneous data-source integration and exploits the unique characteristics of the mobile-agent paradigm to handle problems in a wireless computing environment.

In order to promote interoperability and reusability, improve the effectiveness of information retrieval, and accommodate clients who are not medical professionals, we redesigned the MeSH thesaurus in accordance to the ANSI standard and augmented it with synonyms from a well-known English-language thesaurus, WordNet. We named our implementation MEDTHES. In addition, we incorporated three semantic-distance calculation algorithms into MEDTHES in order to support imprecise queries. Other than providing the basic search function, like many other thesauri, MEDTHES also empowers the users with functionalities such as adding new terms and updating existing terms for thesaurus customization.

MEDTHES can be used as a stand-alone thesaurus or as an integral part of MAMDAS. We demonstrate the integration of MEDTHES and MAMDAS via an example. We further quantitatively evaluated the performance of MEDTHES with regard to improving the effectiveness of information retrieval by using it as the reference for query expansion. Experimental results obtained from the standard biomedical test data collection, OHSUMED, show that by combining the strength of MeSH and WordNet, MEDTHES improves the precision of information retrieval by using query expansion with synonyms. However, the results also indicate that query expansion with broader or narrower terms may worsen the performance.

The MAMDAS framework can serve as the core infrastructure of many large, distributed healthcare applications where real-time access to heterogeneous data sources is required. However, we must note that high speed and high precision are not the sole requirements of e-health systems. Information security and user privacy also play an important role in the future of healthcare system development. For example, computer-based population or community health records usually use patient records anonymously. These systems are particularly valuable in public health where one is trying to trace different types of health hazards, linked either to medical, environmental,

or social agents. There is certainly important ethical concern in relation to the composition of records and access to them.

The administrative simplification provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) require the Department of Health and Human Services (HHS) to establish national standards for electronic healthcare transactions and a national identifier for providers, health plans, and employers. It also addresses the security and privacy of health data. Adopting these standards will improve the efficiency and effectiveness of the national healthcare system by encouraging the widespread use of electronic data interchange in healthcare.

Traditionally, the AAA techniques are used to secure information systems: authentication, access control, and auditing. A user is authenticated before he or she is allowed to gain entrance to the system. The access-control unit of the system determines and enforces the access privileges associated with each user. System-access information is often kept in a log for further analysis, and this process is often referred to as auditing. As the information systems grow more complex, however, new problems that cannot be addressed by the traditional methods frequently emerge. Many of them are still open-ended questions. Breakthroughs in the security and privacy research field are crucial to the success of e-health systems and appropriate steps must be taken in search for solutions.

REFERENCES

Associated Press. (2005). *Hurricane highlights need for digital records.* Retrieved September 13, 2005, from http://www.msnbc.msn.com/id/9316246/#storyContinued

Bright, M. W., Hurson, A. R., & Pakzad, S. H. (1994). Automated resolution of semantic heterogeneity in multidatabases. *ACM Transactions on Databases Systems, 19*(2), 212-253.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Proceedings of Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 29-34.

Carey, M. J., Haas, L. M., Schwarz, P. M., Arya, M., Cody, W. F., Fagin, R., et al. (1995). Towards heterogeneous multimedia information systems: The garlic appraoch. *In Proceedings of the Fifth International Workshop on research Issues in Data Engineering: Distributed Object Management*, (pp. 124-131).

Chamberlin, D. (1998). *A complete guide to DB2 universal database.* San Francisco: Morgan Kaufmann Publishers.

Chevy, C. (2000). Historical notes: Medical subject headings. *Bull Med Library Association, 88*(3), 265-266.

Fedyukin, I. V., Reviakin, Y. G., Orlov, O., Doarn, C. R., Harnett, B. M., & Merrell, R. C. (2002). Experience in the application of Java technologies in telemedicine. *eHealth International, 1,* 3.

Gray, R. S., Kotz, D., Cybenko, G., & Rus, D. (2002). Mobile agents: Motivations and state-of-the-art systems. Technical report: TR2000-365, Darthmouth University.

Hass, L. M., Schwarz, P. M., Kodali, P., Kotlar, E., Rice, J. E., & Swope, W. C. (2001). DiscoveryLink: A system for integrated access to life science data sources. *IBM Systems Journal, 40*(2), 489-511.

Hersh, W., Buckley, C., Leone, T., & Hickman, D. (1994). OHSUMED: An interactive retrieval evaluation and new large text collection for research. *Proceedings of SIGIR'94* (pp. 192-201).

Hersh, W., Price, S., & Donohoe, L. (2000). Assessing thesaurus-based query expansion using

the UMLS metathesaurus. *Proceedings of the AMIA Symposium* (pp. 344-348).

Hynes, D. M., Perrin, R. A., Rappaport, S., Stevens, J. M., & Demakis, J. G. (2004). Information resources to support healthcare quality improvement in the veterans health administration. *Journal of American Medical Informatics Association, 11*(5), 344-350.

IBM Aglets Workbench. (1996). Retrieved January 5th, 2005, from http://www.trl.ibm.co.jp/aglets/index.html

Jiang, J., & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference on Research in Computational Linguistics* (pp. 19-33).

Jiao, Y., & Hurson, A. R. (2004). Application of mobile agents in mobile data access systems: A prototype. *Journal of Database Management, 15*(4), 1-24.

Lange, D., & Oshima, M. (1998). *Programming and developing Java mobile agents with aglets.* Reading, MA: Addison Wesley Longman, Inc.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: A lexical reference system and its application.* Cambridge, MA: MIT Press.

Marenco, L., Wang, T. Y., Shepherd G., Miller, P. L., & Nadkarni, P. (2004). QIS: A framework for biomedical database federation. *Journal of American Medical Informatics Association, 11*(6), 523-534.

McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods of Information in Medicine, 34*, 193-201.

MEDLINE. (2005). Retrieved January 5, 2005, from http://www.nlm.nih.gov/pubs/factsheets/medline.html

Miller, G. A., Beckwith, R. T., Fellbaum, C. D., Gross, D., & Miller, K. J. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235-244.

Mork, P., Shaker, R., Halevy, A., & Tarczy-Hornoch, P. (2002). PQL: A declarative query language over dynamic biological schemata. *Proceedings of the AMIA Fall Symposium* (pp. 533-537).

National Information Standards Organization (NISO). (1994). *National Information Standards Institute American national standard guidelines for the construction, format, and management of monolingual thesauri.* Bethesda, MD: NISO Press.

Potok, T., Phillips, L., Pollock, R., & Loebl, A. (2003). Suitability of agent technology for military command and control in the future combat system environment. *Proceedings of the 8th International Command and Control Research and Technology Symposium* (pp. 1-20).

President's Information Technology Advisory Committee. (2001). Transforming healthcare through information technology. In *Panel on transforming healthcare.* Arlington, VA: National Coordinating Office for Information Technology Research and Development (pp. 1-17).

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(1), 17-30.

Robertson, S., & Hull, D. A. (2001). The TREC-9 filtering track final report. *The Ninth Text Retrieval Conference (TREC-9),* 25-40.

Shiri, A. A., Revie, C., & Chowdhury, G. (2002). Thesaurus-assisted search term selection and query expansion: A review of user-centered studies. *Knowledge Organization, 29*(1), 1-19.

Spackman, K. A., Campbell, K. E., & Cote, R. A. (1997). SNOMED RT: A reference terminology

for healthcare. *Proceedings of AMIA Annual Fall Symposium* (pp. 640-644).

Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N. W., et al. (2000). TAMBIS: Transparent access to multiple bioinformatics information sources. *Bioinformatic, 16*(2), 184-186.

Svensson, P. (2002). eHealth application in healthcare management. *eHealth International, 1*, 1.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* (pp. 133-138).

Zettair Search Engine. (2003). Retrieved January 5, 2005, from http://www.seg.rmit.edu.au/zettair

This work was previously published in Web Mobile-Based Applications for Healthcare Management, edited by L. Al-Hakim, pp. 197-225, copyright 2007 by IRM Press (an imprint of IGI Global).

Chapter 4.14
Adoption of Mobile Technology in the Supply Chain:
An Exploratory Cross-Case Analysis

Bill Doolin
Auckland University of Technology, New Zealand

Eman Al Haj Ali
Higher Colleges of Technology, UAE

ABSTRACT

The increasing utilization of mobile commerce technologies in e-business raises the question of their use in supply chain integration and management. This article presents a multiple case study investigation of the adoption of mobile technology in the supply chain. A technology-organization-environment framework of the contextual influences on technological innovation adoption is used to inform an analysis of three companies' adoption and use of mobile data solutions for sales automation, freight tracking, and service support. Analysis of the three case studies found that the relative advantage of the technological innovation and the information intensity of the company were the most important factors influencing adoption. Other factors that appeared to influence adoption included the compatibility of the technology with the company's business approach, the presence of top management support, and the degree of organizational readiness. Environmental factors such as competition within the industry or business partner influence seemed less influential for these pioneers of mobile technology use in supply-side activities.

INTRODUCTION

Supply chain management (SCM) can be defined as "the process of managing relationships, information, and materials flow across enterprise borders to deliver enhanced customer service and economic value" (Mentzer et al., 2001, p. 10). Information technology (IT) is pervasive

in SCM, and with the development of e-commerce it is playing an increasingly strategic role as supply chain activities are conducted, linked, and integrated electronically (Bhatt & Emdad, 2001). Companies are seeking to gain competitive advantage and create responsiveness to markets by adopting IT that enables them to utilize and manage information and knowledge within and across the extended enterprise (Lau et al., 2006). Of relevance to this article is the relatively recent but rapid development of mobile commerce and its application to SCM.

Mobile commerce is the conduct of e-commerce through mobile or handheld computing devices (e.g., mobile phones, PDAs, and tablet PCs), using wireless technologies and telecommunication networks (Siau, Lim, & Shen, 2003). Such mobile technologies facilitate communication, Internet access, data exchange, and transactional capabilities largely independent of time and location. The result is increased real-time interaction between companies, employees, supply chain partners, and customers, enhancing operational efficiency and providing new opportunities for customer service (Shankar & O'Driscoll, 2002).

A number of studies have examined the potential for mobile commerce to be applied to SCM. Mobile technologies are envisaged to have the most impact in areas of SCM such as e-procurement; materials handling; warehousing; inventory management; logistics and fulfilment; asset tracking; sales and field force automation; and dispatch management. For example, it has been argued that mobile applications integrated with a company's enterprise systems can provide greater visibility into supply chain operations, leading to real-time order status information and more responsive service management (Kalakota, Robinson, & Gundepudi, 2003). When deployed to mobile employees such as sales representatives or technical field service teams, mobile technologies can automate data collection, deliver necessary information to employees wherever their location, and reduce the time needed to update data from the field for the rest of the company, resulting in improved workforce productivity, process efficiency, data accuracy, and service quality (Rangone & Renga, 2006).

The idea that mobile commerce can transform SCM is reflected in the development of concepts such as "untethered" (Shankar & O'Driscoll, 2002), "adaptive" (Kalakota et al., 2003), and "responsive" (Lau et al., 2006) supply chains. However, there are few empirical studies that focus on the adoption and implementation of mobile commerce in the supply chain activities of companies—those that do have tended to report on financially modest or relatively simple applications that support mobile activities (operational mobility) rather than the mobile transmission of data (transmission mobility) (Rangone & Renga, 2006). In contrast, this article examines the adoption of more complex mobile applications that support transmission mobility as well as operational mobility, and integrate with existing company information systems and have the potential to change operating procedures and activities.

Since the organizational adoption of mobile commerce technologies in the supply chain is not well understood, we use an exploratory case study approach to provide an analysis of three New Zealand companies' development and use of mobile data solutions. We draw on the IT innovation adoption literature to inform our analysis. The next section summarizes this literature and presents a conceptual framework based on technological, organizational, and environmental factors influencing the innovation adoption decision. We then outline the research method used in the study before presenting our analysis of the three case studies. The final part of the article synthesizes some conclusions from the cross-case comparison and discusses the implications for research and practice in this area.

ORGANIZATIONAL ADOPTION OF IT INNOVATIONS

There is a long-standing interest in the adoption of IT innovations in the study of information systems. In this article we are concerned with the primary adoption of an innovation by an organization, rather than its secondary adoption by individuals in the organization. By organizational adoption of an innovation we mean a process beginning with initial awareness and evaluation of a new technology or product, followed by a decision to purchase and implement the innovation, and finally its acceptance or assimilation within the organization (Frambach & Schillewaert, 2002).

Researchers have utilized a number of approaches in attempting to explain why organizations adopt IT-related innovations. Probably the most common approach used is one based around the identification of a set of contingency factors that collectively explain the innovation adoption decision or outcome (Fichman, 2004; Frambach & Schillewaert, 2002; Jeyaraj, Rottman, & Lacity, 2006). Many contingency or factor studies of IT innovation adoption tend to follow a "technology-organization-environment" model pioneered by DePietro, Wiarda, and Fleischer (1990). The number of empirical studies following this approach provides support for its usefulness and, following calls to extend this framework to other innovation domains (Chau & Tam, 1997; Thong, 1999; Zhu, Kraemer, & Xu, 2003), we have used it to organize our exploratory study of the contextual influences on the organizational adoption of mobile commerce technologies in the supply chain.

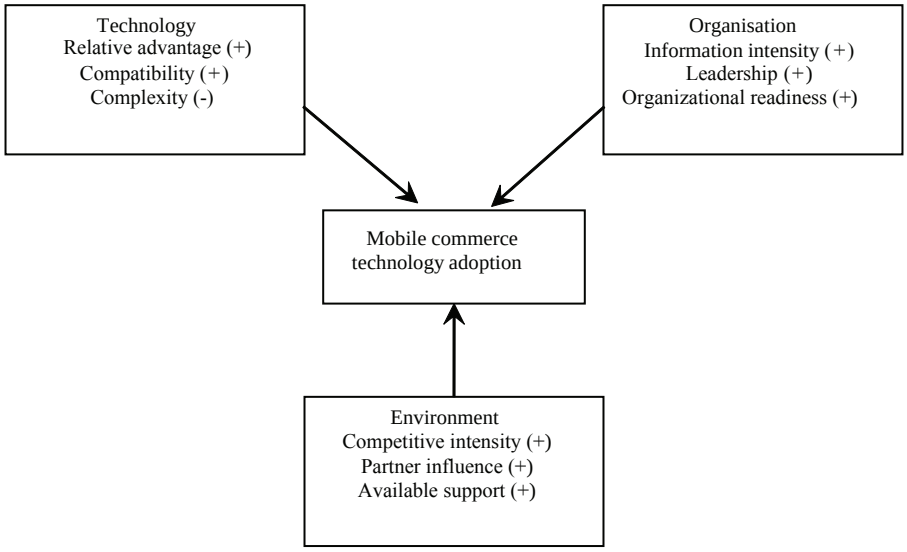
The technology-organization-environment model proposes that organizational innovation adoption is influenced by three elements of context: (1) the perceived attributes of the technological innovation, (2) organizational characteristics, and (3) environmental conditions. Prior studies of innovation adoption have identified a complex and rich group of potentially relevant factors within each of these three elements—too many for a single study to examine (Frambach & Schillewaert, 2002; Russell & Hoag, 2004). The adoption model we use in this study is shown in Figure 1. It includes three high-level factors for each contextual element, which we believe have an influence on organizational adoption of mobile commerce technologies in the supply chain. Each factor is discussed hereafter.

Technology Attributes

Tornatzky and Klein (1982) found that three perceived attributes of the technology or innovation itself were consistently associated with innovation adoption behaviors: the relative advantage of an innovation over its predecessor; its compatibility with the organization's needs and existing systems; and its complexity to understand and use (Rogers, 2003). Potential adopters typically evaluate the *relative advantage* of a technological innovation in terms of whether the costs of adoption are outweighed by the benefits likely to be received (Premkumar, Ramamurthy, & Nilakanta, 1994). This evaluation may be in economic or in more subjective terms; what is important is that an organization perceives the new technology as advantageous in comparison with existing or alternative technologies. Relative advantage may depend on how satisfied the organization is with their existing technological solution (Chau & Tam, 1997).

The more *compatible* a technological innovation, the less changes or adjustments needed and the lower the possible level of resistance to the technology when it is adopted (Teo, Tan, & Buk, 1997). Organizational compatibility involves the congruence of an innovation with organizational culture, values, and operating practices. For example, Flanagin (2000) found that some organizations adopted IT innovations at an early stage that they considered compatible with their perceived industry leadership or reputation. Technological compatibility reflects the abil-

Figure 1. Contextual influences on mobile commerce technology adoption ('+' and '-' indicate a positive or negative influence on adoption, respectively)



ity of the adopting organization to successfully integrate the new technology with its existing IT infrastructure and legacy systems (Dedrick & West, 2004; Premkumar et al., 1994). Finally, the *complexity* of a technological innovation, as well as the processes and activities involved in its adoption, may negatively influence adoption (Ramamurthy, Premkumar, & Crum, 1999; Russell & Hoag, 2004).

Organizational Characteristics

Factors related to the organizational context that may facilitate or inhibit adoption of an innovation are usually defined in terms of various characteristics of the organization, its employees, and available internal resources. Thong (1999) suggests that organizations that are more *information-intensive* in their products or services are more likely to adopt IT innovations based on their greater potential for strategic use for IT and perception of IT as a source of competitive advantage. This may be reflected in the centrality or strategic importance of IT to the organization's business and operations (Chwelos, Benbasat, & Dexter, 2001; Dedrick & West, 2004), the level of IT use or sophistication of the organization (Flanagin, 2000), or the organization's view of itself as innovative towards IT (Dedrick & West, 2004; Teo et al., 1997).

An organization's *leadership* may influence IT adoption decisions through senior management's willingness to innovate and explore the possibilities of new technologies (Thong, 1999). In particular, the commitment, involvement, and support of senior managers can provide direction, ensure adequate resources are made available, and signal the importance of the adoption (Premkumar & Potter, 1995; Ramamurthy et al., 1999; Russell & Hoag, 2004; Teo et al., 1997). Other members of the organization (often IT professionals) may act as internal champions for an innovation, raising awareness of the innovation and its benefits with managers and potential users (Premkumar & Potter, 1995; Russell & Hoag, 2004).

The *readiness* of an organization to adopt and introduce a technological innovation relates to the existence of adequate financial, human, and technical resources within the organization (Chau & Hui, 2001; Chwelos et al., 2001; Dedrick & West, 2004; Iacovou, Benbasat, & Dexter, 1995). Of particular importance is the level of internal technical expertise available to implement the new technology (Chau & Tam, 1997; Premkumar & Potter, 1995; Zhu et al., 2003). Adoption may depend also on the ability and confidence of employees to operate IT-related innovations (Thong, 1999).

Environmental Conditions

The environmental context constitutes the arena in which adopting organizations conduct their business (DePietro et al., 1990). The higher the *competitive intensity* in an industry, the stronger the pressure on an organization to adopt innovations in order to gain or maintain competitiveness (Chwelos et al., 2001; Ramamurthy et al., 1999). Competition also leads to environmental uncertainty, increasing the propensity for innovation adoption (Chau & Tam, 1997; Thong, 1999).

Business or trading *partner influence*, whether supportive or coercive, can also motivate an organization to adopt an innovation (Chau & Hui, 2001; Chwelos et al., 2001). Examples include external pressure from a trading partner (Iacovou et al., 1995), the presence of established trading relationships (Ramamurthy et al., 1999), and the readiness (or not) of business partners (Chwelos et al., 2001; Zhu et al., 2003).

The perceived level of *available support* from vendors (Chau & Hui, 2001), government (Damsgaard & Lyytinen, 2000) or third parties (Dedrick & West, 2004; Doolin, McLeod, McQueen, & Watton, 2003) for an IT innovation and its implementation is sometimes an important influence on an organization's adoption decision. Perceived support may also relate to infrastructural support for the use of an innovation. For example, a lack of standards may act as a barrier to the diffusion of a relatively complex IT innovation, such as electronic data interchange (EDI) (Damsgaard & Lyytinen, 2000).

METHOD

The research objective was to provide an empirical exploration of why organizations might adopt mobile commerce technologies in their supply chain activities. Because our understanding of this technological innovation context is relatively undeveloped and lacks a strong theoretical base, we used an exploratory case study approach (Benbasat, Goldstein, & Mead, 1987). Further, a case study approach facilitates our focus on the contextual conditions of mobile commerce technology adoption (Yin, 2003). We applied the preliminary model of mobile commerce technology adoption, shown in Figure 1, to three case studies of organizations that had adopted mobile data solutions in their supply chains.

Our primary source of data was semi-structured interviews conducted during 2004 with key informants in three New Zealand companies. The interviews were based on a common set of questions designed to elicit information on the company and its operations, its use of IT, the decision to adopt mobile technologies, the perceived benefits of the technology, factors facilitating or inhibiting adoption, the implementation process, and any implications of adoption for the company. The interviews were audio-taped and transcribed for qualitative data analysis. This involved both within-case and cross-case thematic analysis organized around the theoretical propositions identified previously (Yin, 2003). The interview data were supplemented with secondary data sources, including publicly available information on the companies and their activities. Table 1 summarizes the interviews and background details of the three case study companies.

Table 1. Case studies

	FoodCo	FreightCo	PowerCo
Business	Food manufacturing and marketing	Freight, logistics and warehousing	Electricity network and distribution
Company size	900 employees	1200 employees	280 employees
Turnover	NZ$220 million	NZ$890 million	NZ$870 million
IT team	4 employees	20 employees	20 employees
Application	Mobile sales automation	Mobile freight tracking	Mobile service support
Interviewees	IT Manager Systems administrator Commercial manager	IT manager Logistics manager Stock controller	IS manager Customer service manager

In the following sections we present our analysis of the three case studies. Each case is structured around a brief description of the company and the mobile data solution studied, followed by a discussion of the three types of contextual influence identified in the research model outlined in Figure 1: attributes of the technology itself, organizational characteristics, and wider environmental or industry conditions. Selected quotes from the interviews are used to illustrate the analysis.

MOBILE SALES AUTOMATION AT FOODCO

FoodCo is a New Zealand subsidiary of a multinational food company. It manufactures and distributes a range of product lines to a large retail customer base via mobile sales representatives. The company emphasizes speed and efficiency in order taking and fulfilment as essential to maintaining customer satisfaction. FoodCo has a small IT department for routine maintenance of the company's information systems. It was a pioneer in New Zealand in the use of barcode scanners to capture order information at the point of customer contact and the transmission of this data to its sales office, first by dial-up modem over a landline and then by car phone over a cellular phone network. In 1999, the company decided to upgrade its system and outsourced development of a customized mobile data solution used by the sales force via laptop computers. This system has been progressively updated since then both in terms of software and hardware. The major motivation for the adoption of a mobile sales automation technology was "to move key strokes out of the office into the field" (IT Manager).

Sales representatives now use battery-operated tablet PCs to download updated product information, customer information, sales promotions, territory management information, stock levels, and replenishment dates. Inputted order and invoicing information is transferred to the company's sales office where the information is processed via the company's enterprise resource planning (ERP) system and the required goods are dispatched as quickly as possible. Customer information and in-store negotiated promotion details can also be updated in real time. Other functionality includes a supermarket shelf management function and a sales effort screen, which provides information on sales targets and volumes and allows sales representatives to track their performance at product level. Data is transmitted over a general packet radio service (GPRS) wireless network, although the units also have built-in modems for use with a landline and infrared ports for use with mobile phones

if alternative data transmission mechanisms are needed.

Technology Attributes

FoodCo clearly perceives a relative advantage in their mobile data solution: "The benefits have certainly been there and pretty much delivered to our expectations" (Commercial Manager). The mobile data solution effectively automates the sales process, eliminating the paper work, which sales representatives were previously doing. Lightweight tablet PCs have replaced the "huge, big briefcases of paper" (Systems Administrator) previously carried by sales representatives. The added information and functionality provided by their mobile data solution enables FoodCo's sales representatives to undertake promotion management, conduct in-store deals, and manage customer relationships on a one-to-one, real-time basis. This was seen as enabling a shift in their role: "We see the [mobile] unit becoming even less an order entry unit and much more of a business management tool" (IT Manager).

The mobile data solution has enabled FoodCo to improve the efficiency of its order processing and logistics. Timely receipt of sales orders means that planning associated with warehouse picking and truck delivery loads can begin earlier: "We are becoming more and more focused in that area of getting that whole process more and more efficient. And having the orders coming in effectively within five minutes of them being taken into [the ERP system], ready to be picked, has been beneficial to us" (Systems Administrator). The mobile data solution is also considered to be a source of competitive advantage through the way that it integrates and synchronizes information regarding customers, products, and distribution, enabling the company to manage its key customer accounts more efficiently: "Historically we were very good at transactions and you've got good competitive advantage by being able to transact better than anybody else. But now it's not about transactions, it's about knowledge management" (IT Manager).

The current tablet PC technology is considered to be a significant improvement over previous units in terms of weight, screen size, and processing power. While some transmission and coverage issues had been experienced with the cellular network originally used to transmit the data, data is now transmitted over a GPRS wireless network selected because of its continuous availability, connection stability, high speed, and relatively cheap (data-driven) rates. Ironically, the speed and efficiency of the wireless transmission led to an unintended increase in projected data costs as sales representatives began transmitting data after every sales call (until reined in).

In terms of its compatibility, FoodCo's mobile data solution matched the business approach of the company in a number of ways. For example, the units allow sales representatives to manage customer relationships with key accounts in person rather than from head office. Similarly, sales representatives take a proactive role with small retailers: "It's all about presence in the marketplace and being there in front of them and actually influencing buying patterns" (Systems Administrator). The mobile data solution was also compatible with the IT infrastructure and approach used by FoodCo. The existence of the company's ERP system and the simultaneous roll-out of its sales and distribution modules provided the necessary complementary technology for the mobile data solution to function effectively.

Extensive training was required to up-skill the sales force in using both the mobile computer units and the extended range of functionality. The tradeoff of the more powerful, large-screened tablet PC units was their complexity, which made them more prone to breakdown and damage when dropped or mishandled. In addition, the mobile data solution project grew in size and complexity, creating some difficulties in coordination between

the various departments involved in its use: "I think the biggest thing was that it ended up bigger than it was ever planned to be … Sometimes what you find is that when you revisit it that a lot of the facility there isn't being used to its capability" (IT Manager).

Organizational Characteristics

The adoption of mobile technology for sales automation reflects both FoodCo's history of IT use (including sales automation) and its innovative attitude towards IT. FoodCo had been actively monitoring and developing the e-business side of its operations since 1999: "[FoodCo] has always been at the front of deploying that kind of technology to the market … We tend to pick up the new technologies quickly if we can see there's a clear business input" (Systems Administrator). The small IT department within FoodCo actively looks for ways to utilize new and innovative IT in the company's operations. However, the decision to explore new technological options in sales automation was a strategic one taken by FoodCo's senior management. According to the IT Manager, "That type of leadership has always been there … The current management is very, very supportive."

An unwillingness of some sales representatives to embrace the new technology initially slowed adoption and use of the mobile data solution within the company. Some lacked computer literacy, were reluctant to change established ways of doing things, or were reluctant to utilize the new functionality in front of customers in case they showed their inadequacy. As the Systems Administrator explained, "Some of our reps have been with the company for a long time … and putting a computer in front of them was terribly daunting." However, with time and training this barrier was overcome, with many of these representatives becoming advocates for using the new technology.

Environmental Conditions

FoodCo perceive themselves as leaders in their industry, particularly in gaining competitive advantage through the innovative use of IT for knowledge management. In relation to their use of mobile technology, "We were seen to be again, you know, market leading and out there doing things at the forefront basically" (Systems Administrator). FoodCo's largest customers, major supermarket chains, were beginning to move their suppliers to electronic ordering and invoicing, and FoodCo's significant investment in sales automation technology meant that they were well-perceived by these key customers. The proactive contact and support provided by FoodCo's GPRS wireless network provider was mentioned in our interviews as positively influencing the company's adoption of a wireless data solution.

MOBILE FREIGHT TRACKING AT FREIGHTCO

FreightCo is a supply chain logistics provider with operations in New Zealand, Australia, Asia, and the United States. The company offers a full range of logistics services, including managed warehousing, domestic distribution, and international freight operations, linked with IT and information systems. FreightCo operates a nationwide fleet of delivery vehicles in New Zealand servicing a large customer base. It coordinates its distribution operation through a centralized database supplied with real-time freight tracking data from delivery drivers in the field. FreightCo tends to outsource much its development work, with its IT team working on systems maintenance and IT innovations.

The original motivation for deploying a mobile freight tracking system was to "get even more satisfaction to the customers and get in that customer focus" (IT Manager). Drivers scan the barcode of each piece of freight on delivery us-

ing a lightweight handheld device with an inbuilt scanner. A consignment note, the date, time and location of delivery, the driver's identity, and the recipient's name is uploaded to the company's central database, where that information is made available via a Web site to customers, who can track the movement and status of their freight consignment in real time. The delivery information is also used as the basis for payment of the owner-drivers. New job information or updates flow back to the driver's handheld unit from FreightCo's administrative center. FreightCo was a pioneer in using systems such as this, transmitting data over a third-party operated trunk radio network via radio telephones in the delivery trucks since 1992. In 2004, FreightCo commenced transmitting data over a GPRS wireless network.

Technology Attributes

At FreightCo, the mobile data solution implemented for freight tracking removed the need for paperwork and reduced the administrative workload on the distribution fleet drivers, leading to considerable efficiency gains: "Basically we're piling through the freight, or the paperwork about the freight, in a much more efficient manner … The piles of paperwork that we would have had would have been enormous" (IT Manager). The automated system also decreases the chance of errors, improves the timeliness of information, and increases the speed at which information becomes available to customers: "[It] gave us the advantage of managing our network much better, in such a way that we knew where the freight was much better, we knew what our timing was, we knew we could monitor when things went wrong." (IT Manager)

FreightCo sees information and technology as central to its business of providing "intelligent" logistics solutions for its customers. It perceives technology to be the key differentiator in the logistics industry, and sees its ability to provide real-time information across the supply chain to customers as a competitive advantage: "It meant that we had much more to sell. I think we were already the premium provider out there, but it kept us the premium provider. Having been ahead of the technology, like we were, enabled us to continue to charge higher prices" (IT Manager).

The use of a GPRS wireless network for data transmission was seen by FreightCo as superior to the previous trunk radio network used, as it increased the amount of data that could be sent from a mobile unit at any one time (including, for example, customer signatures captured directly on the screen of the handheld devices) and also the overall data transmission capacity available to the company's distribution fleets. As the IT Manager observed: "[GPRS] was becoming a necessity … The more trucks we put on, the more delays we were getting with the data backing up and not coming through … [GPRS] seems to be unlimited."

The mobile data solution for freight tracking is compatible with FreightCo's business model and desire for technology leadership: "We've always had this fundamental business model of being the best … Although many companies may have said, 'Well, what's the benefit of … having the mobile communications today?', We didn't look at it like that" (IT Manager). Going mobile also allowed the company to cope with the huge growth that it experienced and continues to experience as a result of its business strategy.

Organizational Characteristics

As a company, FreightCo is proactive in keeping its IT capability ahead of the business in order to respond to new challenges in the business environment: "we wanted to take ideas to customers before they required it of us, so you know we wanted to be very forward thinking" (IT Manager). IT is essential in linking together and managing the company's range of logistics services. Expenditure on IT is high and the IT department actively seeks "innovative solutions

and ideas" (IT Manager). While adoption of the new mobile technology was initially IT-driven, FreightCo's management was quick to see the benefits and supported the innovation. As the company's IT Manager recounted: "We just had a belief that it would be better and we talked directly to the owners of the business and they thought it would be better and away we went."

Initially, the owner-driver contractors who comprise FreightCo's distribution fleets resisted accepting the new technology. The required expenditure on new technology may have been one reason for this, although FreightCo did subsidize half the cost of purchasing the handheld units: "There was a lot of resistance by the drivers … Resistance to change and technology. Yeah, they didn't want to do it" (IT Manager). However, when FreightCo more recently acquired a competitor's fleet, the newly arrived owner-drivers were generally receptive to using the new mobile data solution. The IT Manager suggested that this was because of the benefits to drivers were evident by then.

Environmental Conditions

The most important environmental influence on FreightCo's adoption of mobile technology was the competitive intensity of the logistics industry in which the company operates. As noted earlier, FreightCo's use of information provides them with a perceived competitive advantage: "We wanted to be ahead of the competition like we always are" (IT Manager). The availability and benefits of a supported GPRS network were acknowledged by FreightCo's IT Manager: "There's just going to be an exponential expansion … and you've got networks that are prepared to invest the money in it."

MOBILE SERVICE SUPPORT AT POWERCO

PowerCo is a large electricity distribution company that uses field crews from outsourced contractors to maintain and repair its electricity network. Good customer service in the form of reliable power supply is important to the company, so response times to the many emergency callouts the company experiences are critical. Around 2001, the company "identified the fact that we needed to get real time information back from the field, we needed to get more accurate information out to the field" (Customer Services Manager) in order to improve the response process. In 2003, after extensive piloting and field testing, PowerCo implemented a mobile data solution purchased from an overseas vendor and then customized for the company by predominantly outsourced developers (the company's in-house IT team works mostly on system maintenance).

When a fault is reported to PowerCo's call center or detected by the company's network management system, details are sent to a field crew's handheld PDA via a secure GPRS network using a Bluetooth, wireless-capable mobile phone as a modem. Crews can upload information on the job status, fault location, work required, and billing in real time from the field. Data is captured once and automatically updated on PowerCo's central information systems, including its customer relationship management (CRM) system and geographical information system (GIS). Customer contact representatives can access real-time information in order to accurately and quickly answer customer queries or claims. Service requests are logged against actual network assets and fault location data is uploaded from the field to the GIS, which facilitates monitoring, management, and long-term planning of PowerCo's networks.

Technology Attributes

The new mobile data solution was perceived as better than the previous system based on two-way radios and various paper-based forms, and its benefits matched PowerCo's expectations. Invoices are now created automatically from data relevant to a service request entered in the field, reducing the need for administrative data entry, decreasing costs and speeding up the invoicing process. Other benefits included a reduction in data duplication or redundancy, with a consequential decrease in the risk of errors in data entry: "So the main drive is reducing paper, data quality, and only capturing data once" (IS Manager). The efficiency of the emergency response process also improved markedly, with faster response times and more accurate information sent to and from field crews: "We were collecting data at the call center but it was never making it to the guys in the field … Now, everything gets passed through … so the sort of level of accuracy of information that the guys in the field are getting is much higher" (Customer Services Manager).

The information provided via the mobile data solution has enabled the call center to deal with customers' complaints efficiently and effectively, and to keep them informed of progress in a timely manner. Because information is updated from the field in real time and made accessible to the call center operators: "We know when they're [field crew] on-site. We know when they've restored power. We know that the job has been completed … We can follow up all the details … It's made a huge difference to us in terms of resolving customer complaints because all the information is actually there" (Customer Services Manager). This use of accurate, real-time information to maintain continuous power supply and improve customer service is consistent with PowerCo's role as a network provider of critical energy services.

Aspects of the complexity of the mobile data solution did become issues. For example, the limited battery life of the PDAs (which often stay docked in the field crews' vehicles in order to remain powered) and the range of the Bluetooth wireless connection between the PDA and the mobile phone modem (about 10 meters) effectively shape the crews' use of the technology. PowerCo's IS Manager described how aspects of the mobile data solution were designed to cope with crews periodically moving out of coverage. The crews are able to continue to work with the application off-line, updating the job status and then uploading the data when they come back within range. Screen layout and sequence on the PDAs was also modified to enhance the application's operability in field conditions.

In fact, the mobile data solution was deliberately developed in a way that accommodated the conditions and characteristics of field crews, who were consulted extensively. As the IS Manager recounted: "[The development company] supplied most of the developers and it was young people … [Their design] might be flashy but it's not always practical … [so] I arranged for them to go out with a field crew and their whole attitude changed. They suddenly started to think like the field crew and not just like a developer." Nevertheless, some aspects of the mobile data solution remain complex for the field crews to use: "The guys struggle a little bit with the GIS stuff and it's been quite a big learning curve for them, but they're getting there" (Customer Service Manager).

Organizational Characteristics

PowerCo has invested significantly in adopting new technology. It generates, on a daily basis, large volumes of multidimensional and interrelated asset, customer, financial, and operational data, which is compiled and displayed in a number of formats to allow users to select and drill into various areas for information. Business intelligence provides information analysis and distribution, data visualization, and spatial analysis for decision making and planning: "We're … an IT focused

[company] and we believe in IT solutions too. And it was most definitely a business decision that we needed to, that we wanted to go down that track [in adopting mobile technology]" (Customer Service Manager).

PowerCo's IT team takes a reactive approach to IT solutions for the company, focusing on supporting business requirements rather than "pushing" technology: "We're really in there to try and understand the business needs before we even talk systems" (IS Manager). The impetus for the adoption of mobile technology was from top management: "It was top down. It was a benefit that our executives … saw. And so, like, everybody's using wireless despatching in field crews and we should actually also be using it" (IS Manager).

PowerCo uses outsourced field crews, which meant that the contractors had to be convinced to adopt and use the new mobile data solution, including taking responsibility for maintaining the mobile technology itself: "We've provided a certain number of the devices to start with but then from then on they've got to buy their own, they've got to support their own hardware, that type of thing. So we had to sell it into them as well" (Customer Services Manager). However, PowerCo provided them with training. Project team members would go into the field with the field crews, "holding their hands" as they used the mobile technology: "You have to break the habit of what they would normally do" (Customer Services Manager).

The field crews generally accepted and used the new mobile units, despite management's concern that the modern "white collar" technology might be perceived as out of place in the *blue collar* field environment and that the field crews would *struggle with it*. In fact, although it was technology that most of the crews had not experienced before, "They picked it up pretty quickly … I think we thought that we'd have more problems teaching them than sort of we did" (Customer Services Manager). The field crews who selected to participate in piloting the system actually refused to return the units at the end of the pilot, wanting to continue using them, and placing unforeseen demands on the company's resources as they continued supporting the pilot while developing the full mobile data solution.

Environmental Conditions

The outsourced contractors who supply the field crews are an important business partner for PowerCo. The contractors' senior management apparently recognized the potential benefits of using wireless technology for dispatching field crews, and that at some stage they would need to adopt it: "I think they were quite pleased that we made the choice to actually roll it out, that they didn't have to do something themselves … I think they were pretty supportive. They could see the end result should be beneficial for their business" (Customer Service Manager).

Maintaining "robust connections" between the handheld PDA units and the GPRS wireless network, remains problematic according to PowerCo's IS Manager. The company initially used wireless cards in the PDAs to access the GPRS network, but experienced a high level of disconnections, hence the shift to using dedicated mobile phones as modems. However, there were still problems with disconnections, which appeared to be related to the standard that handles communication between the GPRS network and the mobile application: "That standard is still a grey area. It's not just related to [our application]; we are also talking to other people in the industry and we've found that they lose a lot of connections … Bit annoying, but we working with [network and application providers] to resolve it" (IS Manager).

Support from the original application vendor also became an issue, as while the application worked satisfactorily on the original handheld units used, it did not necessarily do so on the latest technology purchased by the contractor

users: "We're having some problems with newer technology, getting it to be able to support the software … That's been another issue to stop us rolling it [the mobile data solution] out wider, because there's been changes of device and [the vendor] hasn't necessarily kept up with that side" (Customer Service Manager).

Table 2. Summary of innovation adoption findings

	FoodCo	FreightCo	PowerCo
Technology Attributes			
Relative advantage	• Information integration and synchronization a source of competitive advantage • "Manage the business within the supermarket rather than just take an order"	• Providing real-time information to customers is a competitive advantage • "The piles of paperwork … would have been enormous"	• "Reducing paper, [improving] data quality, and only capturing data once" • "It's made a huge difference to us in terms of resolving customer complaints"
Compatibility	• "It's all about presence in the marketplace"	• Freight tracking system is a good fit with the company's focus on customer service	• "The end result is that customers spend less time in the dark"
Complexity	• "It ended up bigger than it was ever planned to be"	• Not mentioned	• Aspects of the mobile data solution had to be modified for field conditions • "The guys struggle a little bit with the GIS stuff"
Organizational Characteristics			
Information intensity	• "We tend to pick up the new technologies quickly if we can see there's a clear business input"	• If a company's IT capability stays ahead of the business, the business will always be prepared for new challenges	• "We're an IT focused [company] and we believe in IT solutions"
Leadership	• "The current management is very, very supportive"	• Management were quick to see the benefits and supported the innovation	• Adoption of mobile technology was initially a top-down decision
Technical readiness	• IT team actively scans the technological environment	• IT team actively seeks "innovative solutions and ideas"	• "We're really in there to try and understand the business needs before we even talk systems"
User readiness	• "Putting a computer in front of [some of] them was terribly daunting"	• "There was a lot of resistance by the drivers … to change and technology"	• "It's just technology that they're not used to" • "You have to break the habit of what they would normally do"
Environmental Conditions			
Competitive intensity	• "We were seen to be again you know market leading"	• "We wanted to be ahead of the competition"	• Not mentioned
Partner influence	• Major supermarket chains were beginning to move their suppliers to electronic ordering and invoicing	• Not mentioned	• "The contractors … were quite pleased that we made the choice to actually roll it out, that they didn't have to do something themselves"
Available support	• Proactive support from wireless network provider	• Not mentioned	• Continually changing hardware technology requires vendor support for software compatibility • Mobile device to wireless network communication standard problematic

DISCUSSION

Table 2 summarizes the findings of our cross-case analysis of the adoption of mobile data solutions in the three case studies.

Perceived *relative advantage* appeared to be influential in all three companies' adoption and use of mobile data solutions. The benefits they achieved related to (1) administrative efficiency, in the form of paperwork reduction and time savings; (2) improved data accuracy and timeliness; (3) improved operational efficiency in supply chain operations; (4) enhanced roles for company users of the mobile technology; and (5) competitive advantage. The *compatibility* of the mobile data solution adopted with a focus on customer service observed in all three companies was also a common factor across the three cases. Complexity only appeared relevant in two of the case studies, where it was perceived to increase the level of user training required.

All three companies are *information-intensive* in that information processing is an important part of their business and that IT is integral in managing customer services. The importance of this factor was reflected in the history of IT use in the companies and their proactive and innovative attitude towards IT, and e-business in particular. *Leadership*, in the form of top management support for the innovation adoption, was also a common theme across all three case studies. Even where the initial awareness of the innovation was not management-driven, management adopted a supportive attitude to the business use of new technology. With respect to *organizational readiness*, an interesting distinction emerged between the positive influence of *technical readiness* and the negative influence of *user readiness*. While the role played by two of the companies' IT teams in actively seeking innovative uses for IT was a positive influence on adoption of mobile commerce technology, the lack of readiness of some intended users to embrace the new technology tended to slow adoption or increase the time and training needed.

Although we expected wider environmental or industry conditions to play an important role in shaping innovation adoption decisions in the three case studies, overall they seemed to play less of a role than technology attributes or organizational characteristics. This may reflect the pioneering status of the three companies in their respective industries in New Zealand with respect to the use of mobile commerce technology in the supply chain. Industry *competitive intensity* was reflected primarily in FoodCo's and FreightCo's desire to be market leaders through the use of IT. *Partner influence* also played some role, with some of FoodCo's major customers innovating with electronic transactions themselves, and PowerCo's sub-contractors providing support for the innovation based on their recognition of the benefits of the mobile dispatch technology. While *available support* was a factor in the adoption experience of these two cases, it did not seem to be a direct consideration in terms of the adoption decision itself. FoodCo received proactive support from its wireless network provider, while PowerCo found itself reliant on vendor support because of changing or problematic technology.

CONCLUSION

This article has presented an exploratory empirical study into why organizations adopt mobile commerce technologies in the supply chain. The evidence from the three case studies suggests that the innovation adoption model presented in the article is likely to be of interest to researchers in this area. However, further research could refine or expand the model in several ways. Larger scale survey research could be used to statistically confirm the model's propositions at a more general level. Studies in different organizational or industry settings and for different types of mobile

commerce innovations would potentially increase the applicability of the model. The contextual factors used in our model were selected for their perceived relevance to supply chain applications of mobile technology. Other potentially relevant factors could be explored. Finally, our exploratory case study approach does not enable us to reliably assess the degree of influence on the organizational innovation adoption process of the various factors in our model.

As Frambach and Schillewaert (2002) note, such models are also of use to practitioners, including both technology suppliers and organizational managers, in marketing innovations to organizations and in gaining acceptance of innovations within organizations. A prominent issue in the analysis of the three case studies was that adopting and implementing a mobile data solution involves more than automating existing processes. This is not a new finding with respect to IT innovations, but the mobility, localization, and immediacy aspects of mobile commerce technologies provide opportunities for process redesign, which imply new ways of doing things for users. Addressing the latter involves extensive and carefully thought out training, but also recognition that changing existing user behaviors may be necessary.

REFERENCES

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly, 11*(3), 369-386.

Bhatt, G. D., & Emdad, A. F. (2001). An analysis of the virtual value chain in electronic commerce. *Logistics Information Management, 14*(1/2), 78-84.

Chau, P. Y. K., & Hui, K. L. (2001). Determinants of small business EDI adoption: An empirical investigation. *Journal of Organizational Computing and Electronic Commerce, 11*(4), 229-252.

Chau, P. Y. K., & Tam, K. Y. (1997). Factors affecting the adoption of open systems: An exploratory study. *MIS Quarterly, 21*(1), 1-24.

Chwelos, P., Benbasat, I., & Dexter, A. S. (2001). Research report: empirical test of an EDI adoption model. *Information Systems Research, 12*(3), 304-321.

Damsgaard, J., & Lyytinen, K. (2000). The dynamics of factors explaining EDI diffusion in Hong Kong in the late 1990s. In J. Thong, P. Chau, & K. Y. Tam (Eds.), *Proceedings of the Fourth Pacific Asia Conference on Information Systems* (pp. 1061-1074). Hong Kong: AIS.

Dedrick, J., & West, J. (2004). An exploratory study into open source platform adoption. In *Proceedings of the 37th Hawaii International Conference on System Sciences* (pp. 1-10). Hawaii: IEEE.

DePietro, R., Wiarda, E., & Fleischer, M. (1990). The context for change: Organization, technology, and environment. In L. G. Tornatzky & M. Fleischer (Eds.), *The processes of technological innovation* (pp. 151-175). Lexington MA: Lexington Books.

Doolin, B., McLeod, L., McQueen, B., & Watton, M. (2003). Internet strategies for established retailers: Four New Zealand case studies. *Journal of Information Technology Cases and Applications, 5*(4), 3-20.

Fichman, R. G. (2004). Going beyond the dominant paradigm for information technology innovation research: Emerging concepts and methods. *Journal of the Association for Information Systems, 5*(8), 314-355.

Flanagin, A. J. (2000). Social pressures on organizational Website adoption. *Human Communication Research, 26*(4), 618-646.

Frambach, R. T., & Schillewaert, N. (2002). Organizational innovation adoption: A multi-level framework of determinants and opportunities for

future research. *Journal of Business Research, 55*(2), 163-176.

Iacovou, C. L., Benbasat, I., & Dexter, A. S. (1995). Electronic data interchange and small organizations: Adoption and impact of technology. *MIS Quarterly, 19*(4), 465-485.

Jeyaraj, A., Rottman, J. W., & Lacity, M. C. (2006). A review of the predictors, linkages, and biases in IT innovation adoption research. *Journal of Information Technology, 21*(1), 1-23.

Kalakota, R., Robinson, M., & Gundepudi, P. (2003). Mobile applications for adaptive supply chains: A landscape analysis. In E.-P. Lim & K. Siau (Eds.), *Advances in mobile commerce technologies* (pp. 298-311). Hershey, PA: Idea Group.

Lau, H. C. W., Lee, C. K. M., Ho, G. T. S., Ip, W. H., Chan, F. T. S., & Ip, R. W. L. (2006). M-commerce to support the implementation of a responsive supply chain network. *Supply Chain Management, 11*(2), 169-178.

Mentzer, J. T., DeWitt, W., Keebler, J. S., Min, S., Nix, N. W., Smith, C. D., et al. (2001). Defining supply chain management. *Journal of Business Logistics, 22*(2), 1-25.

Premkumar, G., & Potter, M. (1995). Adoption of computer aided software engineering (CASE) technology: An innovation adoption perspective. *The DATA BASE for Advances in Information Systems, 26*(2/3), 105-124.

Premkumar, G., Ramamurthy, K., & Nilakanta, S. (1994). Implementation of electronic data interchange: An innovation diffusion perspective. *Journal of Management Information Systems, 11*(2), 157-186.

Ramamurthy, K., Premkumar, G., & Crum, M. R. (1999). Organizational and interorganizational determinants of EDI diffusion and organizational performance: A causal model. *Journal of Organizational Computing and Electronic Commerce, 9*(4), 253-285.

Rangone, A., & Renga, F. M. (2006). B2e mobile Internet: An exploratory study of Italian applications. *Business Process Management Journal, 12*(3), 330-343.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: The Free Press.

Russell, D. M., & Hoag, A. M. (2004). People and information technology in the supply chain: Social and organizational influences on adoption. *International Journal of Physical Distribution & Logistics Management, 34*(2), 102-122.

Shankar, V., & O'Driscoll, T. (2002). How wireless networks are reshaping the supply chain. *Supply Chain Management Review, 6*(4), 44-51.

Siau, K., Lim, E.-P., & Shen, Z. (2003). Mobile commerce: Current states and future trends. In E.-P. Lim & K. Siau (Eds.), *Advances in mobile commerce technologies* (pp. 1-17). Hershey, PA: Idea Group.

Teo, T. S. H., Tan, M., & Buk, W. K. (1997). A contingency model of Internet adoption in Singapore. *International Journal of Electronic Commerce, 2*(2), 95-118.

Thong, J. Y. L. (1999). An integrated model of information systems adoption in small businesses. *Journal of Management Information Systems, 15*(4), 187-214.

Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption implementation: A meta-analysis of findings. *IEEE Transactions on Engineering Management, EM-29*(1), 28-45.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

Zhu, K., Kraemer, K., & Xu, S. (2003). Electronic business adoption by European firms: A cross-country assessment of the facilitators and inhibitors. *European Journal of Information Systems, 12*(4), 251-268.

This work was previously published in International Journal of E-Business Research, Vol. 4, Issue 4, edited by I. lee, , pp. 1-15, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.15
Enabling the Glass Pipeline:
The Infusion of Mobile Technology Applications in Supply Chain Management

Umar Ruhi
Wilfrid Laurier University, Canada

Ofir Turel
McMaster University, Canada

ABSTRACT

In recent years, the prospect of information exchange independent of time and place has been a compelling driver for organizations worldwide to adopt mobile technology applications in their various business practices. In particular, the application of mobile technology in Supply Chain Management has drawn widespread attention from researchers and practitioners who endorse adaptive and agile supply chain processes. This chapter discusses the applications of mobile technologies in various areas of supply chain management and the potential benefits of those technologies along the dimensions of reduced replenishment time and transactions and billing cycles. Among other discussions, the role of mobile procurement, inventory management, product identification, package tracking, sales force, and field service automation technologies is highlighted. To substantiate the basis for adopting mobile technologies for supply chain management, different market drivers for mobile applications are exemplified and applied to the three macro-level processes of supplier relationship management, internal supply chain management, and customer relationship management; a resulting typology of mobile supply chain management applications is presented.

INTRODUCTION

The nature of competition is shifting away from the classic struggle between companies. The new competition is supply chain vs. supply chain. (Taylor, 2003, p. 3)

In recent years, we have seen various organizations from different industries focus their competitive strategies on improving their supply networks rather than concentrating on directly

contending with specific companies. Companies such as Wal-Mart, Dell, and Proctor & Gamble not only have made significant headway in optimizing their own supply chains, they also essentially have redefined the way business is done in their particular industries. Their competitors have had to follow suit in order to maintain their own competitive position in the marketplace.

A major factor that has contributed to more efficient supply networks is the increasingly unhindered and efficient flow of information within and among supply chain partners. Several researchers and practitioners have commented on the importance of information flow in effective supply chains (Chopra & Meindl, 2003; Handfield & Nichols, 2002; Kalakota, Robinson & Gundepudi, 2003). Consequently, much has been said about the role of technology in enabling effective supply chains (Holten, Dreiling, Muehlen & Becker, 2002; Knolmayer, Mertens & Zeier, 2002; Poirier & Bauer, 2000).

Mobile technologies and applications offer an advanced level of efficient and effective communications among business partners in supply chains. These applications augment the static nature of their predecessor, e-commerce, phone, and fax-based technologies, by adding flexibility and spontaneity to extant business processes. Technologies in mobile procurement, inventory management, product identification, package tracking, sales force, and field service automation are expected to change the current landscape of Supply Chain Management (SCM). It is expected that mobile technologies will bridge the functionality gap in traditional Electronic Data Interchange (EDI), Enterprise Resource Planning (ERP) and Web-based SCM technologies by providing the end-to-end transparency that can help businesses perform better through improved supply chain planning and execution (Kalakota et al., 2003).

In this chapter, we provide a value proposition for mobile SCM technologies and applications. By highlighting the benefits of the latest mobile applications, this chapter aims to explicate the role of these technologies in transforming integrated and collaborative supply chains into adaptive supply networks. We start this discussion with our working definition of SCM, which will be the gate to our analysis of various technology applications. Following that, we discuss the current state of information technologies in SCM and subsequently rationalize the business drivers for implementing mobile SCM technologies. This is followed by an elaboration of a typology of mobile SCM technology applications. Our conclusion and ensuing inferences follow after a discussion on the future outlook for mobile SCM technologies vis-à-vis other SCM information systems.

SUPPLY CHAIN MANAGEMENT: A WORKING DEFINITION

There are as many definitions of SCM as there are publications, which is quite enormous within the supply management literature. Furthermore, the terminology used to describe the concept or idea behind SCM is interchangeably used in various contexts to refer to the same thing. For example, supply chains, supply networks, and supply webs often are used to describe the same idea—coordination and collaboration across business partners. Recently, however, there is an increasing tendency to use the terms *supply networks* and *supply webs* as opposed to the notion of *supply chains.* The advantage of using the former terms over the latter, is to emphasize that the links among business partners are not linear and sequential but are, instead, dynamic, interdependent, and flexible (Bovet & Martha, 2000; Murphy, 2000; Rayner, 2004).

In this chapter, we use the terms *supply chains* and *supply networks* interchangeably, with the proviso that the nature of relationships among business partners is, indeed, more than just linear and sequential. As highlighted in the introduction and for the purpose of this discussion, we adopt a definition of SCM that incorporates the manage-

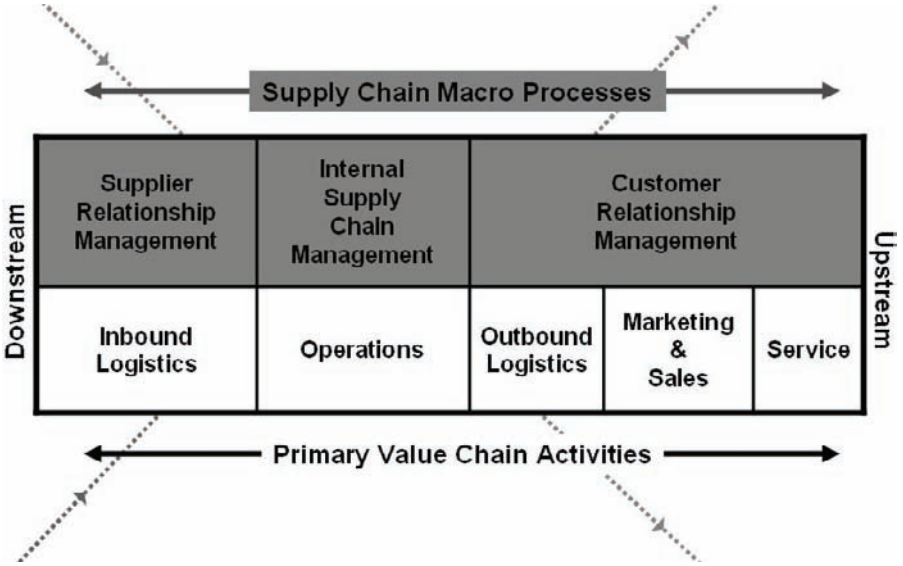
ment of information flow as the primary functional component—material flow and financial flow both upstream and downstream the supply chain. For a descriptive and formal characterization, we adopt Handfield and Nichols' (2002) definition of SCM. The authors define SCM as:

The integration and management of supply chain organizations and activities through cooperative organizational relationships, effective business processes, and high levels of information sharing to create high-performing value systems that provide member organizations a sustainable competitive advantage. (Handfield & Nichols, 2002, p. 8)

In addition to affirming the importance of high levels of information sharing, the definition prominently highlights the concept of a value system for sustainable competitive advantage.

The notion of a value system (also known as a value chain) is intertwined with that of a supply chain and needs some elaboration. Introduced by Michael Porter in his widely acclaimed book, *Competitive Advantage* (Porter, 1985), the idea of a value chain has been used to model a firm on the basis of its value-creating activities. The primary activities in the value chain include inbound logistics, operations, outbound logistics, marketing and sales, and service. Noticeably, it is several different business processes within the value-creating functions of a firm that constitute the various components in an organization's supply chain, as well. In fact, Chopra and Meindl (2003) classify all supply chain processes into three main macro processes; namely, Supplier relationship management (SRM), internal supply chain management (ISCM), and customer Relationship Management (CRM). Juxtaposing the value creating functions described by Porter (1985) with the three macro processes in a supply chain, it can be seen that inbound logistics and operations functions map diametrically to supplier relationship management and internal SCM, respectively, while the marketing, sales, and service functions map to the processes in customer relationship management. It is this inclusive framework of value-creating activities and supply chain macro processes that

Figure 1. Supply chain management functions and processes



forms the basis of our discussion throughout the rest of this chapter. Figure 1 summarizes these processes and functions in an inclusive conceptual model for SCM.

CURRENT STATE OF MOBILE SCM TECHNOLOGIES AND APPLICATIONS

As an affirmation of the prevalent adoption of wireless technologies, the Yankee Group predicts that close to 50% of large US enterprises will employ wide area wireless solutions, and that 3,000,000 mobile users will use these services (Yankee Group, 2004). In terms of technological maturity, third-generation (3G) technologies are emerging quickly, while more established technologies, such as Short Messaging Services (SMS) and the wireless Internet (e.g., Wireless Application Protocol [WAP]) are serving organizations to help fulfill their current business requirements. It is these existing and emerging technologies that act as the bearers of a large and ever increasing number of mobile SCM solutions. Furthermore, the diffusion of more advanced packet-switched data networks (e.g., GPRS, 1XRTT, etc.) is giving rise to innovative communication solutions. For example, Push-to-Talk services (P2T) that previously were available only to niche markets are now being deployed effortlessly over packet switched data networks and are being used to facilitate inexpensive interorganizational voice communications (Guy, 2003). At the same time, short-range wireless technologies, such as Radio Frequency Identification (RFID), Wireless Personal Networks (WPANs) such as Bluetooth, and Wireless Local Area Networks (WLANs) such as the 802.11, are becoming commonplace, forming the basis of wireless networking standards.

As anticipated, the previously mentioned technologies are finding their applications in various business processes in the organization's SCM, as well. A case in point is the recent adoption of RFID solutions by Wal-Mart, exemplifying the wireless services market trends and opening doors for next-generation logistics management (Emily, 2004). Additionally, WLAN technologies already are being used widely in many industries, such as the energy sector (Yankee Group, 2001). Decreasing deployment costs as well as communication and network costs are being proclaimed as the primary drivers for the growth of this market (Rao & Parikh, 2003).

To capitalize upon the opportunities presented by mobile technologies, vendors of traditional SCM systems, such as SAP and Oracle, and new ones, such as HighJump and @Par, are vying for a piece of the pie. Most of the mobile SCM technology vendors currently offer their solutions in bundled packages as part of their m-business or m-commerce technology suites. Such an offering (under a broader umbrella of m-commerce) is in line with the treatise of mobile SCM solutions by researchers (Burchett, 2000; Cousins & Varshney, 2001). M-commerce can be succinctly defined as "a layer of applications atop the mobile Internet" (Rulke & Iyer, 2003) or explicitly as "an extension of e-commerce in a mobile environment" (Dholakia, 2002). Consequently, Mobile Supply Chain Management (MSCM) can be regarded as a specific branch of m-commerce, and it can be characterized as a layer of mobile applications that enhances existing supply chain mechanisms while enabling efficient business processes. Just like their e-commerce counterparts, the incumbent technologies in MSCM are being used increasingly to support an evolving complex transactions landscape. Researchers classify the types of transactions supported through mobile technologies under the business-to-consumer (B2C), business-to-business (B2B), employee-to-employee (E2E), business-to-employee (B2E), and machine-to-machine (M2M) domains (Alanen & Autio, 2003). With this transactions landscape in mind, let us now examine the driving factors behind the adoption of mobile technologies in SCM.

Table 1. Drivers for information systems in SCM

Driver:	Applications:				
	Enterprise Resource Planning	Data Warehouses	Customer Relationship Management	Decision Support Systems	Mobile Technology Applications
Internal Integration	●	●	●	●	●
External Integration			●	●	●
Globalization	●	●	●	●	●
Data Information Management		●	●		●
New Business Processes	●		●		●
Replace Obsolete Systems	●		●		●
Strategic Cost Management	●	●	●	●	●

(Adapted from Handfield & Nichols, 2002)

DRIVERS FOR MOBILE TECHNOLOGIES IN SUPPLY CHAIN MANAGEMENT

The use of information systems in any business context always has been driven on the basis of available technological capabilities as well as on managerial vision toward fulfilling certain business requirements via those technologies. This is why we see the proliferation of different types of information systems, depending on different eras in which they are adopted. For example, whereas data and transaction processing systems were the order of the day in the 1950s, decision support systems started emerging in the 1970s, followed by e-business systems in the 1990s. The current trend is on the collaborative aspect of information systems.

Within specific business contexts of the current era, SCM has attracted myriad information systems (e.g., Materials Requirements Planning [MRP I] systems, Manufacturing Resource Planning [MRP II] systems, Enterprise Resource Planning [ERP] systems and Advanced Planning and Scheduling [APS] systems), which all have seen their peaks during the evolution of SCM as a discipline. Whereas, these systems all have a distinct logistics-oriented flare to them, SCM also utilizes cross-functional information systems, such as decision support systems and customer relationship management systems in day-to-day activities.

There have been several drivers that have led to the adoption of different types of information systems in SCM. Perhaps the most significant driver in the adoption of such systems has been the realization of an increasing need to internally integrate within and among different business functions as well as to integrate externally with supply chain business partners (Holten et al., 2002; Poirier & Bauer, 2000). Handfield and Nichols (2002) highlight additional drivers that have led to the adoption of various types of information systems in SCM. These additional drivers include

trends emerging from globalization, information management requirements, the need for new business processes, the desire to replace obsolete systems, and ensuring strategic cost management. Table 1 summarizes a mapping of these drivers to different types of information systems in SCM and presents our extension of this framework to assimilate the drivers for mobile technologies. An elaboration of the specific drivers for mobile technologies follows.

Internal and External Integration (Toward Pervasive Computing)

As highlighted earlier in this chapter, internal and external integrations are extremely important drivers in the adoption of information systems in today's business environment. Whereas, systems such as MRP I and MRP II lacked the functionalities that allowed effective internal integration among various supply chain functions, external integration with supply chain partners is a challenging problem for even modern-day ERP systems. It was only with the advent of the Internet and the connectivity accorded by it that the newer Web-based applications, including the likes of CRM systems, were made possible; these applications enabled more cross functional operations inside the firm and outside, to a certain extent.

The adoption of mobile technologies in SCM promises to take the integration phenomenon to the next level (i.e., pervasiveness). Gupta and Moitra (2004) characterize pervasive computing as saturating an environment with computing and communication capability, yet having those devices integrated into the environment such that they disappear. The same authors also consider mobility and wireless connectivity as an essential component in pervasive computing. Accordingly, the adoption of mobile applications in supply chains is driven partially by their capabilities to enhance internal and external integration. Internal integration at the systems level is enhanced through seamless network connectivity between the front end and back end systems through wireless local and personal area networks (WLAN and WPAN). Basic voice communication capabilities, like cellular technologies and push-to-talk (P2T), further allow employees to connect effortlessly to one another. Moreover, remote data access by employees increases the internal integration by allowing internal users to access organizational data from anywhere at anytime. Similarly, external integration is enhanced, as mobile applications enable remote access to relevant information for customers, retailers, and distributors through wireless wide area networks. This is particularly important with the increasing trend in large-scale organizations to establish supplier parks in geographic vicinities around their main manufacturing and fabrication plants (Moline, 2002). Furthermore, mobile applications also enable the organization to access external mobile data resources. For example, using a global positioning system (GPS) and wireless data services, an organization can view the location and status of a shipment arriving from one of its upstream suppliers and prepare for it accordingly. Late delivery of shipment (as well as occasional early delivery) requires careful planning, especially in case of time-sensitive goods, and mobile technologies and applications provide hitherto unknown possibilities in improving the efficiency of managing this supply chain.

Globalization

Globalization is another factor that is driving the infusion of mobile technology applications in supply chains. In remote locations, specifically where fixed landlines and other forms of communications are not available, mobile communications can provide a valuable means for voice and data transmission. Many researchers and visionaries have suggested that a wireless infrastructure can reduce the great divide between the developed and the underdeveloped countries, and it has the potential to facilitate and promote

business prosperity (Parker, 2000; Rice & Katz, 2003; Wareham, Levy & Shi, 2004). For example, the wireless infrastructure in China now has surpassed the fixed-line infrastructure in terms of penetration and coverage (GSM Association, 2004). This implies that businesses and people from other countries are more likely to engage in commercial transactions with firms located in China through mobile communication networks. The ascent of this mobile penetration, especially in developing countries, also can be attributed to the fact that once a mobile infrastructure (e.g., a mobile transmission tower) is set up, it does not require any additional work, such as installation and maintenance of landlines for communication. Furthermore, mobility is now regarded as a predominant characteristic of knowledge workers, and with the widespread availability and standardization of roaming capabilities, these knowledge workers are more likely to collaborate with other businesses through mobile technologies. Whereas, 15 years ago it was unthinkable to even call someone in a different part of the world, not to mention wireless connectivity with them while on the move, today, roaming agreements have made wireless connectivity around the globe a reality by connecting over 500 GSM/GPRS (Global System for Mobile/General Packet Radio Service) networks in almost 200 countries (GSM Association, 2004).

Also, with particular reference to SCM, there is an increasing tendency to outsource non-core functions to external service providers, such as third party logistics (3PL) companies. With contemporary provisions for boundary-less commercial exchanges under global economic forums, such as the North American Free Trade Agreement (NAFTA) and the European Union (EU), there is greater potential to offshore business functions to supply chain service providers in different countries. In fact, offshoring as a global phenomenon has been fueled by organizational attempts to gain competitive advantage through concentration on their core competencies while minimizing the costs of outsourcing at the same time (Bardhan & Kroll, 2003; Nair & Prasad, 2004).

Data Information Management

Mobile applications can improve the frequency and speed of communication (Gebauer, Shaw & Zhao, 2002). With data collection at the point of activity and the elimination of paper-based desktop-centric workflows, the velocity of transactions can be increased greatly. Proof of delivery and electronic signature capture technologies are being used in various businesses to enable more efficient and more streamlined information processes within the supply chain. Moreover, these processes, enabled through mobile technologies, allow real-time data access from the source, as opposed to batches of information being transmitted through various information systems at different times. Consequently, information synchronization errors that are attributed to batch processing can be reduced greatly.

While, on the one hand, real-time data transmission helps to increase the fulfillment velocity by making information instantaneously available throughout the supply chain, on the other hand, it facilitates enhanced visibility of supply chain processes by interacting directly with concerned parties through notification and alert mechanisms. The end result is reduced order-to-delivery time and more responsive service management.

New Business Processes

Mobile applications also are driven by the need to innovate operational business processes. Companies like Wal-Mart and McDonald's are well known for acquiring favorable market positions and competitive advantage through novel business processes. In fact, it has been conferred by many authors that it is the processes and not merely the underlying IT infrastructure that enables strategic advantage (Hurst, 2003). Hence, the need to gain and sustain competitive advantage

can drive companies to innovate and re-engineer their business processes using new technologies (Barney, 1995; Porter, 1980).

The well-known market leaders in SCM are just beginning to avail the opportunities afforded by mobile technologies, as well. For example, Wal-Mart's adoption of Radio Frequency Identification (RFID) and subsequent coercion of its suppliers to do the same will be indubitably accompanied by the establishment of new business processes. The workflows associated with tracking pallets and items from receiving to sales and the management of inventory status all will be affected. Among other benefits, RFID technology is being touted as an enabler for improved inventory accuracy, reduced receiving costs, lower safety stock levels, and reduced cycle count efforts. With mandates from companies such as Wal-Mart, other businesses in various industries also will be driven to explore such technologies in order to revamp their own business processes.

Replace Obsolete Systems

The adoption of mobile technology applications also is being driven by the need to replace obsolete systems and associated business processes. For example, mobile telemetry services (i.e., the use of telecommunication devices to automatically record measurements from a distance) can replace manual on-site data entry and other forms of continuous monitoring (Salz, 2003). With the proliferation of cellular networks, it makes sense to monitor remote resources and assets using wireless technologies.

The previous example is not a completely new technology phenomenon. Often, information systems from one industry find their way into other industries with a certain level of customization. Telemetry solutions have prevailed in the agricultural and military sectors for years. Feeding livestock and examining contamination levels in water are among the applications that use telemetry in an agricultural context. Similarly, remote surveillance using airborne and satellite-based cameras is an example of a telemetry application used by the military (Forrester Research, 1998).

Changing business contexts necessitate newer technology infrastructures. As illustrated in the example, wireless services and applications have the potential to steer more versatility in today's supply chains by offering features and functionalities that may have been tested in other business functions or in totally different industries. It is only a matter of time before more mobile technology applications emerge and change the current supply chain landscape.

Strategic Cost Management

As elaborated throughout this chapter, cost reduction is a major driver in adopting any new technology in SCM. Under compelling demands from various stakeholders, supply chain managers constantly are looking for ways to optimize operations with the objective of reducing operating costs. A feature that makes mobile technologies a viable contender for the supply chain applications market is their ability to account for financial information in real time. Technologies such as GPS (Global Positioning Systems), telemetry, and RFID can feed real-time data to static tethered information systems. Furthermore, by streamlining the order-to-cash process, mobile technologies can reduce the complexity in the overall supply chain execution (Kalakota et al., 2003).

Overall, it can be said that mobile technology applications are driven by a multitude of market forces, and the adoption of such applications is likely to have an impact on business functions across different industries. Mobile SCM technologies can provide processing efficiency in the form of time, cost, and quality of operations. Based on these various benefits that can be availed through the adoption of mobile SCM technology applications, organizations need a road map to implement these technologies in their various business functions and subsequently integrate these into a

seamless mobile infrastructure. The next section provides a typology of mobile SCM technology applications that can facilitate an organization's efforts in this area.

A TYPOLOGY OF MOBILE TECHNOLOGY APPLICATIONS IN SUPPLY CHAIN MANAGEMENT

In presenting our system of classification for mobile technology applications in SCM, we will discuss various mobile applications vis-à-vis their associated bearer technologies. Furthermore, to help us with our analysis, we will utilize the functions and processes from the inclusive SCM framework elaborated earlier in this chapter (see Figure 1). Table 2 at the end of this section illustrates the dimensions of our typology and the incumbent technology applications that are described herein.

Mobile Technology Applications in Supplier Relationship Management (Upstream Processes)

Three things in the life of a supply chain planner are for certain: death, taxes and reconciliation. (Hammer, 1997, p. 18)

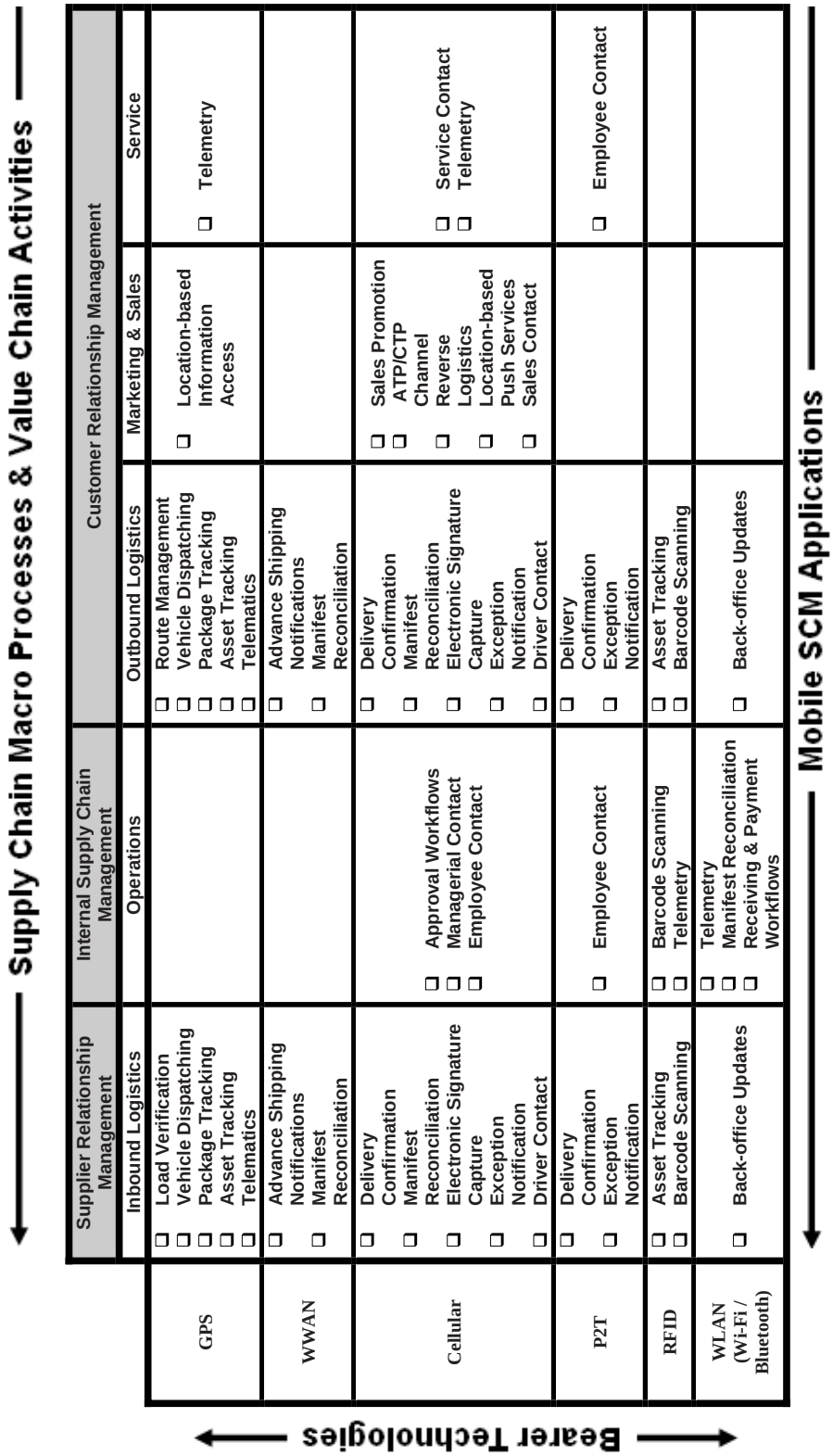
This statement by Michael Hammer (1997) epitomizes a classic problem in SCM. Traditional information systems fall short in their functionality in order to reconcile cash and inventory at various points in the supply chain. It is not surprising, then, that manifest reconciliation emerges as the most popular application in mobile SCM (see Table 2). An example in upstream supply chain processes is the enduring difficulty in reconciling purchase orders to truck manifests at check-in times and freight pickups. The solution to such problems lies in the integrated use of bearer technologies, such as RFIDs, WWANs, and WLANs. Using emerging technologies in smart bar coding, the receiver can seamlessly scan incoming shipments, note discrepancies between purchase orders and truck manifests, make relevant changes to purchase orders, and update back-office systems as well as supplier databases at the same time. Furthermore, the ability for the driver to connect instantaneously with the materials planner for the upstream supplier helps to resolve exceptions and system errors at the point of delivery. Together, these technologies allow increased transaction velocity in the supply chain and higher levels of supplier coordination versatility (Kalakota et al., 2003).

Also, mobile technologies based on GPS allow for greater inventory visibility throughout the supply chain. By pinpointing the location of delivery trucks and the status of delivery packages, upstream suppliers can coordinate shipment schedules with downstream customers, who, in turn, can communicate expectations to their own downstream business partners. Telematics, which is defined as the integration of wireless communications, vehicle monitoring systems, and location devices (GSM Association, 2004), is one such suite of applications that is known to enhance confidence in business functions, including SCM, through facilitating greater visibility and enabling greater control in the supply chain (Hanebeck & Tracey, 2003).

Mobile Technology Applications in Internal Supply Chain Management (Internal Processes)

Mobile technologies also offer significant advantages to internal business operations by facilitating express and streamlined workflows. Among others, workflow applications in business operations include document approval, expense reporting, payment, and purchase orders (Kalakota et al., 2003). According to a recent pilot study by Gebauer, Shaw, and Zhao (2002), the most significant benefits in wireless procurement services result from speeding the overall processing time of an

Table 2. A typology of mobile technology applications for SCM



approval request. It is estimated that close to half of the processing time of a purchasing request is due to managers being out of the office, and, from their study, the researchers conclude that wireless technologies can appreciably help manager approvers as well as finance and accounting approvers by providing support for delegation, communication, notification, and information access (Gebauer et al., 2002).

Another major category of mobile technologies that is redefining internal supply chain processes is wireless product identification. This suite of technologies is regarded as an enabler for handling efficiency, customisation, and information sharing (Karkkainen & Holmstrom, 2002). An example of wireless product identification technology is an RFID system that comprises electronic product codes stored on RFID tags. These tags can be read seamlessly through tactically placed RFID scanners, which, in turn, transmit inventory information to back-office systems through specific middleware. The advantage of such systems is that they can be used without requiring line of sight. The tag readers, hence, can be attached to forklifts, mounted in freight and shipment pathways, or built into stacking shelves. The idea in using such a technology is to eliminate the extra step in scanning pallets or items and to automate the process.

Other internal mobile applications that drive process efficiencies include direct machine-to-machine data exchanges through telemetry. As defined in the previous section, telemetry is the use of telecommunication devices (including wireless) to automatically record measurements from a distance. The automatic notification of inventory management system by an RFID reader when inventory gets depleted below a certain level would be an example of a telemetric application. Again, this type of application can facilitate efficient and streamlined process flows in warehouse and inventory management systems.

Finally, as mentioned earlier, wireless technologies usually are always strongly integrated with other SCM enterprise systems. Back-office updates resulting from real-time data capture at the source are almost never stored or used independently of these systems. The medium of transmission for updates between wireless devices and enterprise systems can be in the form of a WLAN, based on short-range Wi-Fi technology or the proximity-based Bluetooth technology.

Mobile Technology Applications in Customer Relationship Management (Downstream Processes)

Many technology applications in the outbound logistics function coincide with those in the inbound logistics function. This is because of the sophisticated omni-directional nature of supply networks of today, where the position of an organization might be that of an upstream supplier as well as a downstream customer at the same. Similar technologies can be utilized in both cases, albeit in different business contexts. For example, in the instance of delivering a package to a customer, the manifest reconciliation is still a useful application, except this time, the customer invoice will be reconciled with the bill of loading. Similarly, GPS bearer technologies can be used in conjunction with transportation management systems to determine dispatching routes and daily delivery schedules for outbound freight drivers and delivery personnel.

The marketing and sales function in customer relationship management can benefit greatly from information access provided through wireless handheld and pocket-pc-type devices. Using their handhelds, field sales employees can connect directly to back-office inventory management systems or enterprise resource planning systems in order to perform available-to-promise (ATP) checks or capable-to-promise (CTP) checks, respectively (May, 2001). They then can provide up-to-the-minute information to their customers. Not only do these types of applications result in lower costs, due to increased employee productiv-

ity and more streamlined workflows and faster decision-making, they also result in higher levels of customer satisfaction. Returns processing and reverse logistics presents yet another area where mobile technologies can help alleviate business pain spots. By allowing drivers and field service representatives to accept returns (due to product defects or a change of mind from the customer) and with the ability to dispatch pickup trucks in near vicinity, if need be, businesses can drastically increase customer satisfaction levels and operational productivity.

Lastly, with respect to mobile technology applications in the service function, telemetry applications have the potential to provide yet again an effective means of monitoring and controlling remote resources. Service levels can be monitored, and personnel can be assigned, based on the type of problem incurred. These applications can help to reduce employee time and costs associated with routine administration of assets in remote locations.

From the discussion in this section, it should be evident that the mobile supply chain applications described herein, along with their respective bearer technologies are indeed in harmony with the various drivers for mobile technologies described earlier. Table 2 presents our typology of mobile technology applications as a juxtaposition of bearer technologies and their functional scope in different supply chain activities. In the next sections, we discuss our future outlook for these mobile technology applications followed by our conclusions.

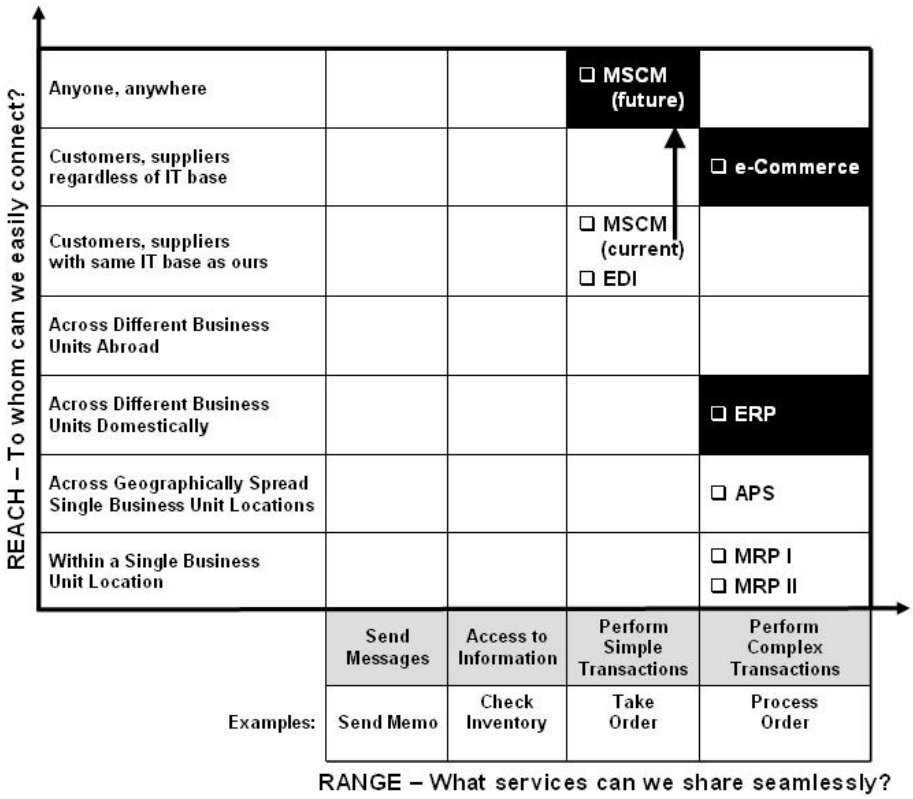
FUTURE OUTLOOK

Recent developments in new wireless technologies, more sophisticated end-user devices, and improved network coverage all have resulted in greater adoption rates for mobile applications (Alanen & Autio, 2003). However, in order to predict the future pathway of different types of mobile technology applications described in this chapter, we need to consider them vis-à-vis other technologies that have been used in SCM for some time, such as MRP I, MRP II, APS, ERP, EDI, and e-commerce systems. It is the combination of these functional and enabling technologies that constitutes the essential technology base for different supply chain environments. As elaborated throughout this chapter, mobile technology applications have the potential to assimilate in this current portfolio of technologies by improving process efficiencies and allowing streamlined access to these traditional back-end systems.

First, let us consider the technical interaction between current systems and how mobile technology applications are changing the current landscape. To reiterate our point from the previous section, mobile technologies currently are being used in conjunction with other systems, such as enterprise resource planning (ERP) systems, and it is our belief that this joint utilization will prevail at least in the short-term future. In order to explicate our position further and discuss our viewpoint for the future of mobile SCM technologies, we utilize the range/reach framework developed by Broadbent, Weill, and Clair (1999). The framework was used originally to explain the findings of an academic study that explored the functionalities of different types of information systems, and it lends well to our discussion of SCM systems. Figure 2 depicts our conceptual positioning of various SCM systems along the two dimensions.

As depicted in Figure 2, it can be seen that the current positioning of MSCM applications is such that these applications are bound to augment the reach and range of other enterprise systems (including ERP, MRP, and APS systems) that support various supply chain processes. The two types of technologies are complementary, in that although contemporary enterprise systems such as ERP are capable of performing complex transactions, they are limited to operations cross-functionally across the same organization. Wireless applica-

Figure 2. Positioning of various SCM systems in the range/reach framework (adapted from Broadbent et al., 1999)



tions can help overcome these spatial boundaries in order to allow communication, collaboration, and coordination across different businesses. Furthermore, with the ongoing standardization of communication protocols, the introduction of new transmission mechanisms (e.g., GSM/GPRS), and improved network coverage worldwide, wireless connectivity very soon can enable the anytime/anywhere business paradigm across global supply chain partners. Figure 2 illustrates the direction of this shift upwards from the current position of MSCM applications along the dimensions of range. However, it should be noted that, although this advancement in mobile technologies may enable more complex transactions to be executed than is possible today, the level of complexity cannot catch up to that of e-commerce and ERP systems. This is because of the nature of computing resource requirements for these complex systems. Hence, in the near to medium future, mobile technology applications in SCM will complement other more complex technologies, such as ERP systems and e-commerce systems. The shaded cells in Figure 2 illustrate this unison among the three systems.

Lastly, theorists and practitioners also envisage that mobile technologies will find their way into horizontal, function-specific services, only after they have been tried and tested in vertical, industry-specific applications (Alanen & Autio, 2003; Forrester Research, 2001).

CONCLUSION

The adoption of mobile technology applications for SCM is being driven by various business and technical factors. At the end of the day, managers at the helm of decision making are interested in increasing productivity by reducing process costs and time, increasing process responsiveness, and improving product and service delivery quality. Researchers and futurists contend that MSCM technology applications can turn that vision mentioned previously into a reality by enabling new processes in order to seamlessly connect into existing supply chain planning and execution systems. Through omni-directional, real-time transmission of information, instantaneous reconciliations, and elimination of non-value-added activities from the supply chain, these new mobile applications are enabling increased fulfillment velocity, improved inventory visibility, and higher levels of supplier coordination versatility in the supply network.

In this chapter, we have discussed various categories of mobile technology applications for SCM. A typology based on these applications and associated bearer technologies was presented to highlight various applications within the three supply chain macro processes of supplier relationship management, internal SCM, and customer relationship management. The current and predicted positioning of these technologies in the near to medium time frame shows that the business process changes that will be instigated by the introduction of mobile technologies will lead to gradual, albeit fundamental, transformations in the organization's operations.

It is hoped that the discussion of technology drivers and the typology of mobile applications will prove to be a useful conceptual vehicle for understanding mobile SCM technologies and aid practitioners in making a business case for adopting these technologies. Finally, organizations that recently have undertaken MSCM initiatives can provide useful test beds for the validation of these conceptual models. Case studies investigating the undertakings of these organizations and their experiences with different technologies can provide valuable insights for revising and improving the ideas presented in this chapter.

REFERENCES

Alanen, J., & Autio, E. (2003). Mobile business services: A strategic perspective. In M.B.E. Strader & T.J. Strader (Eds.), *Mobile commerce: Technology, theory, and applications* (pp. 162-184). Hershey, PA: Idea Group Publishing.

Bardhan, A.D., & Kroll, C.A. (2003). *The new wave of outsourcing.* Berkeley: University of California.

Barney, J. (1995). Firm resources and sustained competitive advantage. *Journal of Management, 17*, 99-120.

Bovet, D., & Martha, J. (2000). *Value nets: Breaking the supply chain to unlock hidden profits.* New York: John Wiley & Sons.

Broadbent, M., Weill, P., & Clair, D.S. (1999). The implication of information technology infrastructure for business process redesign. *MIS Quarterly, 23*, 159-182.

Burchett, C. (2000). Mobile virtual enterprises: The future of electronic business and consumer services. Paper presented at the *Acadmia/Industry Working Conference on Research Challenges (AIWoRC)*, Buffalo, NY.

Chopra, S., & Meindl, P. (2003). *Supply chain management: Strategy, planning, and operation*. Upper Saddle River, NJ: Pearson Education Inc.

Cousins, K., & Varshney, U. (2001). *A product location framework for mobile commerce environment. Proceedings of the International Conference on Mobile Computing and Networking,* Rome, Italy.

Dholakia, R.R., & Dholakia, N.D. (2002). Mobility and markets: Emerging outlines of m-commerce. *Journal of Business Research, Article in Press.*

Emily, K. (2004). Wal-Mart starts RFID test, promises privacy. *Forbes.com.* Retrieved April 4, 2004, from *http://www.forbes.com/reuters/newswire/2004/04/30/rtr1355059.html*

Forrester Research. (1998). *Telemetry's time is coming.* Cambridge, MA: Forrester Research.

Forrester Research. (2001). *Mobile data finds niche in risk-tolerant firms.* Cambridge, MA: Forrester Research.

Gebauer, J., Shaw, M., & Zhao, K. (2002). *The efficacy of mobile e-procurement: A pilot study. Proceedings of the Hawaii Conference on Systems Sciences*, Los Alamitos, California.

GSM Association. (2004a). *Mobile terms & acronyms.* Retrieved August 03, 2004, from *http://www.gsmworld.com/technology/glossary.shtml*

GSM Association. (2004b). *GSMA statistics Q1 04.* Dublin, Ireland: GSM Association.

Gupta, P., & Moitra, D. (2004). Evolving a pervasive IT infrastructure: A technology integration approach. *Personal and Ubiquitous Computing, 8*(1), 31-41.

Guy, A. (2003). *The evolution of push-to-talk represents a powerful carrier weapon.* Boston: The Yankee Group.

Hammer, M. (1997). *Beyond reengineering: How the processed-centered organization is changing our work and our lives.* New York: Harper Business.

Handfield, R.B., & Nichols, E.L. (2002). *Supply chain redesign: Transforming supply chains into integrated value systems.* Upper Saddle River, NJ: Financial Times Prentice Hall.

Hanebeck, H.-C.L., & Tracey, B. (2003). The role of location in supply chain management: How mobile communication enables supply chain best practice and allows companies to move to the next level. *International Journal of Mobile Communications, 1*(1/2), 148-166.

Holten, R., Dreiling, A., Muehlen, M.Z., & Becker, J. (2002). Enabling technologies for supply chain process management. *Proceedings of the Information Resources Management Association International Conference*, Seattle, Washington.

Hurst, S. (2003). IT doesn't matter—Business processes do: A critical analysis of Nicholas Carr's IT article in the *Harvard Business Review Library Journal, 128*(19), 78.

Kalakota, R., Robinson, M., & Gundepudi, P. (2003). Mobile applications for adaptive supply chains: A landscape analysis. In K. Siau, & E.-P. Lim (Eds.), *Advances in mobile commerce technologies.* Hershey, PA: Idea Group Inc.

Karkkainen, M., & Holmstrom, J. (2002). Wireless product identification: Enabler for handling efficiency, customisation and information sharing. *Supply Chain Management: An International Journal, 7*(4), 242-252.

Knolmayer, G., Mertens, P., & Zeier, A. (2002). *Supply chain management based on SAP systems.* Berlin, Germany: Springer-Verlag.

May, P. (2001). *Mobile commerce: Opportunities, applications, and technologies of wireless business.* Cambridge, UK: Cambridge University Press.

Moline, A. (2002). Supplier parks—The wave of the future? *Plants Sites & Parks, 29*(1), 18-19.

Murphy, J. (2000). Internet technology both forces and enables transformation of supply chains. *Global Logistics & Supply Chain Strategies,* March. Retrived June 8, 2004, from *http://www.glscs.com/archives/3.00.intro.htm?adcode=10*

Nair, K.G.K., & Prasad, P.N. (2004). Offshore outsourcing: A SWOT analysis of a state in

India. *Information Systems Management, 21*(3), 34-40.

Parker, E.B. (2000). Closing the digital divide in rural America. *Telecommunications Policy, 24*(4), 281-290.

Poirier, C.C., & Bauer, M.J. (2000). *E-supply chain: Using the Internet to revolutionize your business.* San Francisco: Berrett-Koehler Publishers Inc.

Porter, M. (1980). *Competitive strategy: Techniques for analysing industries and competitors.* New York: Free Press.

Porter, M.E. (1985). *Competitive advantage.* New York: The Free Press.

Rao, B., & Parikh, M.A. (2003). Wireless broadband drivers and their social implications. *Technology in Society, 25,* 477-489.

Rayner, B. (2004). More than a supply chain. *Electronics Supply & Manufacturing.* Retrieved June 8, 2004, from *http://www.my-esm.com/oped/showArticle.jhtml? articleID=21400478*

Rice, R.E., & Katz, J.E. (2003). Comparing Internet and mobile phone usage: Digital divides of usage, adoption, and dropouts. *Telecommunications Policy, 27*(8-9), 597-623.

Rulke A., Iyer, A., & Chiasson, G. (2003). The ecology of mobile commerce: Charting a course for success using value chain analysis. In B.E. Mennecke & T.J. Strader (Eds.), *Mobile commerce: Technology, theory and applications* (pp. 114-130). Hershey, PA: IRM Press.

Salz, P.A. (2003). New high-tech strategies aim to make supply-chain management smoother. *Wall Street Journal Europe,* November 2. Retrieved June 8, 2004, from *http://www.sensile.com/sentech/download/wsje_article.pdf*

Taylor, D.A. (2003). *Supply chains: A manager's guide*. Boston: Pearson Education Inc.

Wareham, J., Levy, A., & Shi, W. (2004). Wireless diffusion and mobile computing: Implications for the digital divide. *Telecommunications Policy, 28*(5-6), 439-457.

Yankee Group. (2001). Wireless technology in the energy industry. Boston: The Yankee Group.

Yankee Group. (2004). *The Yankee Group predictions for 2004.* Boston: The Yankee Group.

This work was previously published in Global Integrated Supply Chain Systems, edited by Y. Lan and B. Unhelkar, pp. 291-309, copyright 2006 by Information Science Publishing (an imprint of IGI Global).

Chapter 4.16
Mobile Automotive Cooperative Services (MACS): Systematic Development of Personalizable Interactive Mobile Automotive Services

Holger Hoffman
Technische Universität München, Germany

Jan Marco Leimeister
Technische Universität München, Germany

Helmut Krcmar
Technische Universität München, Germany

ABSTRACT

In this chapter we describe the systematic development and implementation of mobile services in the automotive sector. This includes a design framework that represents different requirements of automotive service engineering. The framework is used following a corresponding process model which combines iterative service development with classical prototyping. The framework and the process model are applied to a new mobile service MACS MyNews, a personalizeable, interactive news service, allowing the driver to be the editor and end user of his/her newscast at the same time. In order to design this service, we start with designing service scenarios. For these service scenarios a matching value-added network is derived, technologies for service provisioning are chosen, and a prototype is implemented. The service is then evaluated especially concerning driving safety. A final user evaluation helps the designers choose whether or not to include the service in series production before planning the service roll-out.

INTRODUCTION AND BACKGROUND

Mobile services in the automotive sector have been rather unsuccessful in Germany over the past years. Of all car manufacturers that offered services in this field only BMW and Fiat still offered mobile services in their cars in 2004/2005. The three main reasons for discontinuing mobile services are usually mentioned: (1) the costs for data transfer were too high (Frost & Sullivan, 2003), (2) the services offered did not fit adequately to the users' needs (Fuhr, 2001), and (3) mobile services were too focused on technology (Werder, 2005) and had hardly any economic aspects considered making it almost impossible to deliver viable and sustainable services.

But things are changing: The recent availability of new digital transmission channels such as the Universal Mobile Telecommunications System (UMTS) or digital audio broadcast (DAB), digital radio and the declining prices especially for cellular radio (i.e., mobile phone fees) almost eliminated the problem of transmission cost, leaving only two problems to solve. This is the starting point for the project MACS, a research project funded by the German Federal Ministry of Education and Research FKZ 01 HW 0207, and its central research question: "How can innovative mobile automotive services be systematically developed, structured and which steps have to be taken for deploying mobile services successfully?"

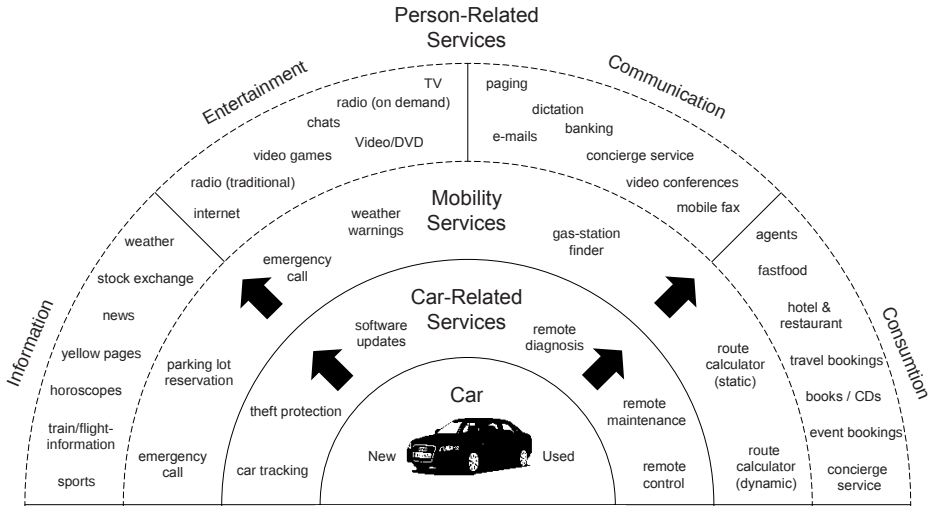
In order to address the problem of designing services that meet the end users' needs, the first step is to find and evaluate service scenarios that seem promising for new mobile services. Based on the scenarios found, the process of deriving a business model out of those scenarios has to be defined. Extracting common factors from these scenarios and models and finding common interfaces for mobile services are necessary for allowing a large and heterogeneous group of service providers to offer their products in the future. From an economic point of view there are several requirements a systematic design of mobile automotive services has to be able to deliver:

- Defining service scenarios and use cases which are comprehensible for the average car user and that are to be frequently used (in contrast to, e.g., emergency call services).
- Integrating mobile broadband services (DAB/UMTS/etc.) into mobile automotive services allowing data exchange at high data rates and thus enabling more powerful services in the years to come.

The following presented MACS design framework for developing mobile automotive services was designed to meet these criteria and to incorporate a wide variety of different solutions from different technical domains of mobile services. One of the main challenges for such a design framework which is unique to automotive mobile services, is the lifecycle mismatch between the car and the software in the car (Hartmann, 2004). While the average lifetime of a car is roughly 10 years, new technology and software comes to market every 2 to 3 years, thus making the manufacturing lifecycle complicated to manage (Mohan, 2006). The technological aspects, for problems imposed by the car as the service carrier, addressed by the MACS design framework thus are:

- Design and implementation of mobile services completely integrated into the car's infrastructure and operational concept
- Design of a modular infrastructure for mobile services, which is independent from individual car manufacturers' platforms in order to enable independent service providers to develop own services and reduce development costs
- Ensuring drivers' safety when using mobile services through responsible usage of different channels of user interaction (e.g., visual vs. audible content) in excess to complying with the current legal requirements

Figure 1. MACS "research radar" for mobile services, based on Ehmer (2002)



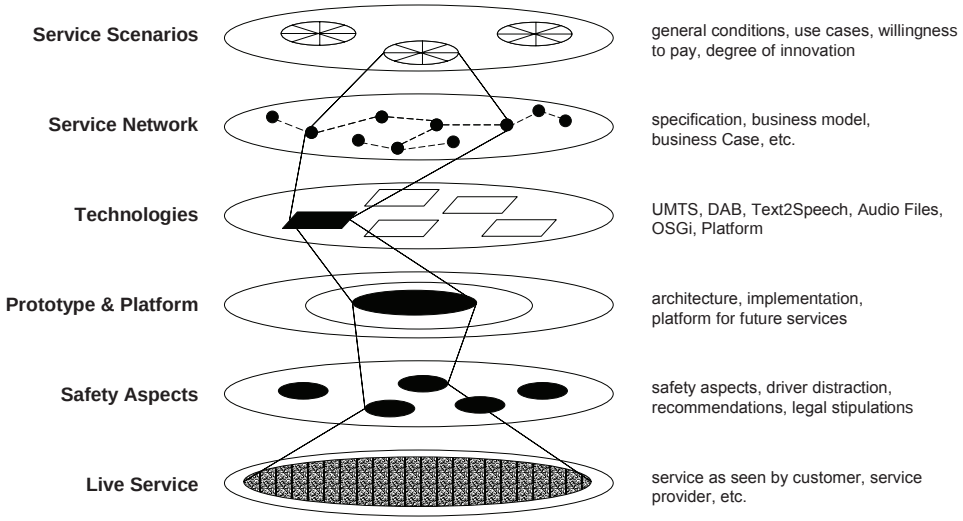
In order to be able to develop mobile automotive services successfully it is furthermore necessary to integrate both technical and economic views in one design framework. We will consequently present the MACS design framework which is composed of a set of guidelines for mobile service development as well as an example for a technical platform for service deployment in cars. To show the feasibility of that solution a specific mobile service—MACS MyNews—has been selected from a previously developed "research radar" of possible mobile services (see Figure 1) and it has been prototypically built according to the criteria derived from domain requirements.

THE MACS DESIGN FRAMEWORK

One of the main arguments in the past years when criticizing research in the field of mobile services was that most research conducted was too technical or technology-driven without offering usable solutions to existing needs or problems of drivers. Besides that there were hardly any viable business models for the developed mobile services.

In order to design attractive, usable, technically stable, and economically promising mobile services, both the economical as well as the technical perspective have to be taken into account. For that reason the MACS design framework (Figure 2) is divided into the following subsets or

Figure 2. MACS design framework



views, each describing one step in the life cycle of a mobile service:

- S**ervice scenario** which is targeted by the provider
- S**ervice network** of partners for service provision
- A selection of **technologies** used for the service
- P**rototype and platform** for service evaluation and assessing transferability to related services
- Consideration of **safety aspects** for ensuring the driver's safety
- Planning of **live service** deployment

No service whatsoever can be successful if it is not created according to the targeted users' needs. Similarly a service that offers useful functionality, but is not really usable for the customer will have to be redesigned. In order to enhance the chances of an economically promising service a targeted user group, partners needed for providing the service, and ways to successfully deploy the finished product have to be determined.

The "service scenarios" define the general set-up of the targeted environment and the use cases of mobile services. They are being studied in order to be able to determine more detailed information, like customer requirements or the willingness to pay for selected services. For finding the needed partners for the "service network" the service scenarios are being analyzed to determine the interaction between the different partners and how they should be organized in a network for value creation.

Since the provision of the mobile service as well as the service itself are technology centered the analysis of the economical aspects alone is insufficient for service design. The selection of

proper technologies, both fulfilling economic requirements and supplying a sustainable platform for the service, are essential. To ensure a viable selection of technologies and allow the creation of an infrastructure, which is able to compensate the lifecycle mismatch between car and technology, the process steps "technologies," "prototype & platform," and "safety aspects" are being examined. Initially the input coming in form of business cases is being matched with the available technology. Following this a prototype for demonstration and evaluation against safety guidelines is built. After being found safe the deployment of a "live service" is the final step in the whole process, merging the ideas generated and the prototypical implementations so far to a service offered to the public that should be usable, useful, technically stable, and economically sustainable.

Applying such a framework for a systematic iterative development of mobile automotive service confronts the developer with several challenges:

- Strategies for the service scenario have to be matched with the available technologies
- The service network design has to reflect the strategies for the service scenario
- People from different domains have to be able to work together, that is, understand each others' languages and rationales
- Functionality has to be built up iteratively and it has to be evaluated continuously against specific criteria and refined according to the results obtained

A solution for most of these aspects, which are not unknown in IS research, is presented, for example, by Hevner, March, Park, and Ram (2004) in form of the design science IS research process. At every point in the process of service design the previous step in the design framework defines the respective environment, composed of needs for which a solution is being searched and the setting in which the found solution has to work.

Whenever possible best practices already existing for a specific topic are being extracted from the domains' knowledge base and applied to the problem. Analogical to the "build and evaluate" loops found in iterative process models like the waterfall model (Royce, 1970) or in design science (Hevner et al., 2004), the assumptions being made in one process step of the model are evaluated in the following layer, such delivering input for a new "build" iteration. This is true in the form of additional requirements in the requirements elicitation phase conducted for each process (i.e., information flowing top-down) as well as for evaluation feedback for the process (i.e., information going bottom-up).

After a satisfactory solution has been found for the current layer the information gained in that process step is being applied to the preceding step (maybe causing a new iteration phase there), the change in knowledge, that is, the learning, is added to the respective domain's knowledge base.

APPLICATION OF THE MACS DESIGN FRAMEWORK

In the following section we will apply the MACS design framework to the exemplary service MACS MyNews. MACS MyNews was selected from a variety of services based on its high customers' benefit, the relative ease of price building for the service, the availability of technology needed for offering the service, and last but not least the very high degree of innovation of that service in the automotive sector. We will highlight the findings on each stage on the way towards a deployable mobile automotive service.

Employing only the standard methods of vertical or horizontal prototyping for mobile services built in multiple iterative steps are inefficient or hardly useable. Floyd (1983) describes vertical

prototyping as implementing system functions in their intended final form, only including selected functions, while in horizontal prototyping the functions are completely available, with part of their effect being omitted or simulated.

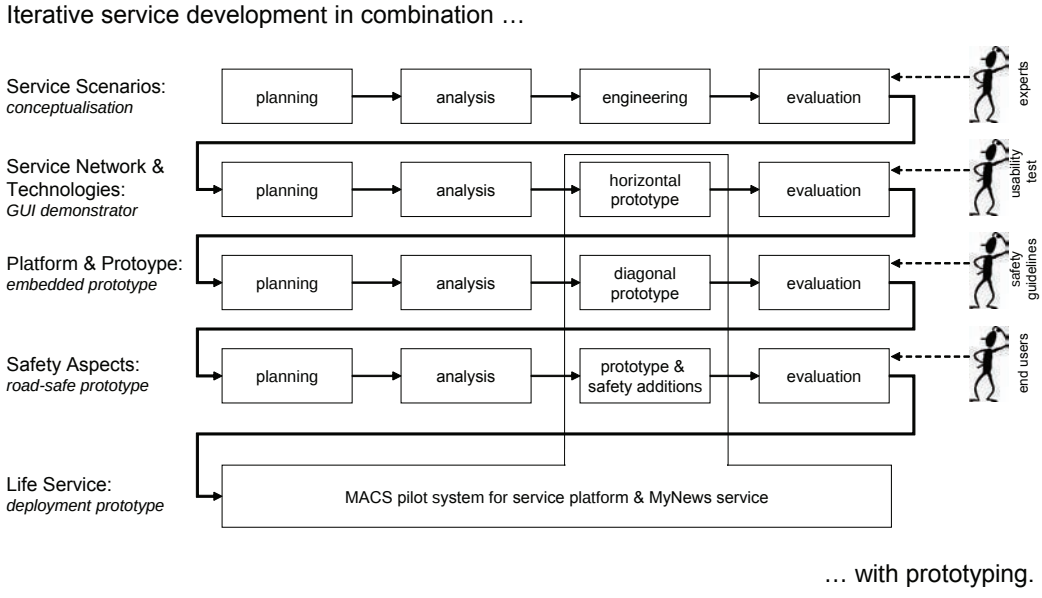
In the case of mobile services, a horizontal approach forbids the integration into the car's infrastructure. Evaluating mobile services in their target environment would be impossible. A tight integration using a vertical approach on the other hand does not allow to test the handling of the actual service, an evaluation of the safety implications is impossible.

A solution for this dilemma is to combine the two prototyping methods: functionality that is special or even unique to the service, like the ability to present personalizeable news or interact with the system via voice commands, is being implemented (vertical approach). Most of it in a form that is similar to what the user would experience while using the final service, but not necessarily in a way that would be used for technical realization for the final product (horizontal approach). This reduces the complexity of the approach from developing in a highly proprietary embedded system (i.e., in the car's infrastructure and on its infotainment unit) to presenting the user an approximation of the innovations important to the service in a way he or she would expect to see in his or her car.

Combining the iterative nature of the MACS design framework with its distinct phases described previously with the diagonal prototyping approach results in the MACS development model depicted in Figure 3.

The heart of the following process model is an iterative process adapted from the generic spiral process model (Boehm, 1988, p. 61; Wigand, 1998). It is combined with elements of prototyping. In opposition to the original spiral model, a

Figure 3. MACS development process model



stronger focus is placed on scenario discussions, the display of mock-ups and prototypes, and the active involvement of users. Requirements are collected and adapted within each iteration. Figure 3 shows the MACS process model that was used during the development of the MACY MyNews service.

Beginning with the planning phase, the activities for the respective iterations are scheduled. Afterwards the needed input for the tasks is analysed and the appropriate requirements are either deducted from the prior field studies and expert interviews (iteration 1) or simultaneously collected through user and expert involvement (iteration 2-4). The rendered part of the system is evaluated after the engineering phase is completed. Using the previously deducted general requirements (see above), the translation of the socio-technical and economic needs into system design is done iteration by iteration with the assistance of users and experts. After cycling through the phases a total of four times, the MACS MyNews service is operative since May 2006 and will be presented to the broader public by end of 2006.

MACS Service Scenarios

Defining the service scenario is the first step when planning the development and deployment of a new mobile service. It mainly consists of a thorough analysis of the target environment, including both the prospect customers as well as the competitors in the field. The desired outcome of this step is the definition of a service scenario in which the environmental factors and the proposed business model are evaluated. The service scenario is the starting point for selecting partners for future service provision and technologies for the service implementation.

The three elementary questions derived from the literature (Kotler & Keller, 2005) needed to identify a service scenario and players involved in the service network are the following:

- **Who** are my targeted customers (target group)
- **What** am I offering (value proposition)
- **How** will the service be delivered (service production and delivery, this will be addressed in the service network level)

"Who would be interested in offering a mobile service?" Despite the lack of success of in-car mobile services in the past years, researches results reveal both declining average prices for in-car mobile systems and rising numbers of subscribers for mobile services, ranging from enhanced navigation systems to digital entertainment (e.g., Lawrence, 2005). Furthermore Lawrence compiles figures from various research groups, that indicate an expected rise in subscriber numbers in the German market from below 5 million in 2004 to substantially over 15 million in 2009. A Roland Berger Strategy Consulting survey "How to hit a moving target" (Heidingsfelder, Kintz, Petry, Hensley, & Sedran, 2001) even projected 31 million mobile service subscribers by the end of the decade. Thus mobile services still could represent a promising way of generating new revenue streams in after-sales processes for both car manufacturers and independent service providers (Parnell, 2002), also offering an interesting strategy for value-added services for "premium" car manufacturers.

Parallel to identifying who would probably be interested in offering a new service an even more important issue is: "Who is interested in using or buying a mobile service?" For MACS we identified usage scenarios of mobile services that result from the everyday use of our cars. In the morning and evening commuters join professional drivers on the streets on their way to work or back home. In doing so, almost 2/3 of German job holders commute by car, 4 out of 5 when the distance to work is greater than 10 km. The time spent in the car varies, but more than half of the commuters need up to 30 minutes for one way,

for 1/5 the trip takes up to one hour and more (Statistisches Bundesamt Deutschland, 2005). Working on at least 200 days a year travel times easily top 8 whole days spent on the road for a majority of German commuters during one year. Similar situations are found globally, for example, average travel time of commuters in the U.S. is 25 minutes (Reschovsky, 2004; United States Census Bureau, 2003), so the following assumptions made for the German market will probably be rather similar for the U.S. and other markets.

The next question after knowing who could offer and who could use a service is: "What kind of service is useful for the target group?" Growing needs for information and time being a scarce and precious resource, providing interesting and "valuable" information for the driver while he/she is in the car appears to be promising. Drivers usually do not completely focus on their task of driving, but are distracted by other tasks like listening to music and so forth. Thus personalizable and interactive mobile information services could enable drivers to use time in their cars more efficiently—this is the setting for MACS MyNews.

MACS MyNews should be a solution for organizing the daily routine in the car more efficiently: a personalizable, interactive news service allowing people to be the editor, producer, and consumer of their very own newscast, thus minimizing redundancies and overlapping times. The vision of MACS MyNews is to completely liberate drivers from ordinary radio stations and their news program. MACS MyNews should therefore provide the user with news items which are up to date and ready for presentation all the time, not only every half or full hour as the radio stations' newscasts. MACS MyNews should enable every user to be the editor of "his/her" newscast, since they can not only specify the topics they are interested in, but also set the sequence in which the topics are being presented, assign weights to topics to define how much information they want in that category and more. News could be started when the user would like to hear the latest news, not just every full and half hour. The user could also interact with the news service, that is, if the user has to refuel his/her car he/she can pause the news; if the user missed a detail of a message he/she could listen to it again instantly, skipping forward and backward like with a CD or MP3 player is also possible.

Summarizing the service scenario obtained here is straightforward: MACS MyNews should be a service designed for commuters and professional drivers. It should be designed to replace the radio stations' news casts, offering to use the commuting times more efficiently through a tailored information supply.

MACS Service Network

After determining what kind of service should be supplied to which target group, the next step is the definition of a service network. This comprises the necessary partners needed to offer the service as it has been defined in the service scenario. The outcome of this process step is used in iterative loops together with the following selection of possible technologies to define a network of partners for providing a service according to the definition in the service scenario.

By asking what kind of service could be offered the partners that have to be integrated in the service network are the ones in charge of creating the value for the customer. In the case of a news service those partners will most likely supply the newscasts. Defining the targeted customer base allows the selection of a partner in the automotive sector that makes sense for service deployment: people driving a high end limousine might not very likely be interested in mobile games, but rather in a stock market live ticker. Asking how the service is going to be delivered finally should close the gap between creation and consumption of the service

It is very noteworthy that especially the first and last questions cannot be answered right away. In both cases certain technological conditions that

are being worked out in the following process step have to be considered to answer the questions correctly. The most pragmatic approach is to revise the assumptions and ideas for the service network and the technological aspects iteratively until a solution is found that offers a sound technological basis as well as a promising service.

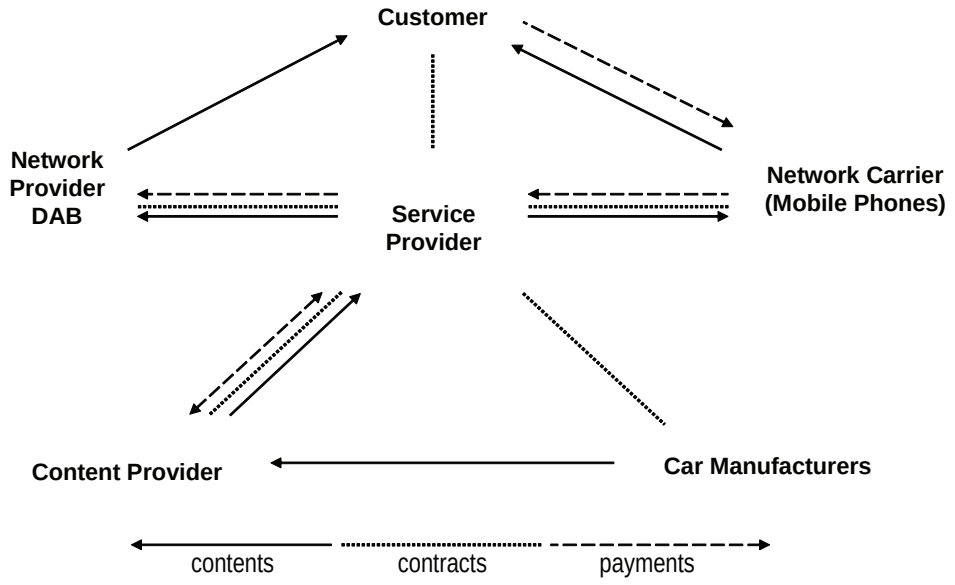
The following example illustrates how the service network for MACS MyNews was derived by analyzing the service scenario and technological implications: The news being used by the MACS MyNews service is delivered to commuters by various content providers. They are being selected, aggregated, and edited by the service provider and transmitted to the car where they are displayed. The main data channel to the customer is digital radio (DAB) for the news messages. Additional value-added services, such as more detailed information or multimedia files, can be downloaded on a pay-per-use basis via mobile phone, provided by a mobile network operator. The partners needed for offering the MyNews services thus are one or several content providers, a car manufacturer, network providers for digital radio, and cellular networks as well as one controlling instance that is in charge of service provisioning: the overall service provider.

The service provider plays a very central role and is in charge of:

- Selecting content providers
- Aggregating and editing their information (for use while driving)
- Selecting network providers for data transfer
- Ensuring the integration into the car's infrastructure
- Coordinating the efforts of the partners

The service provider is also the single point of contact for the customer subscribing to the MACS

Figure 4. Proposed MyNews service network



MyNews service, being responsible for customer management as well as offering a first-level support to the customer.

The topology chosen for the MyNews service network is a star-shape with the service provider as the central instance and single contractor for all other partners. The other partners are arranged in a way that represents their position for supplying the mobile service. The base is formed by the car manufacturer that offers the basic platform for service deployment and the content providers delivering the raw material for MACS MyNews. The middle layer is composed of the central service provider, aggregating and editing the contents before they are being transmitted to the customer, and the network providers needed for delivering the content to the customer. The customer as the end consumer of the service output is the tip of the MACS MyNews service network.

MACS Technologies

The development of the technical domain have to be made considering the economic considerations and, as mentioned for the previous step "service network," decisions or restrictions in one of the domains often have implications for the other. Using the described service scenario and integrating the service network partners the outcome of this process step are the technological fundamentals of the future service.

The logical separation of mobile services into three areas (creation/aggregation, transmission, and display of information) as seen in the service network is also applicable for technological considerations:

- The data formats are important for data creation or aggregation since all the information needed for supplying the service has to be representable by the format.
- The data format as well as the usage scenario has a great impact on the decision how the data should be delivered to the car. Along with different pricing models and costs for the transmission, the volume of data in combination with the data rate at which the information is being transmitted can be limiting factors for one or another channel. In the worst case a change of the assumptions made for the service scenario and service network is needed.
- The available amount of data and its quality influence the possible ways of user interaction in the car. While the information itself might be transferred in form of (compressed) text files that are being read to the user with a text-to-speech system it might be of more value for the customer to have higher quality audio or video files of pre-recorded text, which is slower in transmission.

The following example for MACS MyNews points out how the iterative decision process between the definition of the service and the selection of matching technologies. To reduce complexity only the technology used for content transmission is being presented here. MACS MyNews is supposed to deliver the latest news in a personalizeable form for commuters on their way to work or back home. That means:

- Personalized news selected are most likely only a subset of the news available
- Time until the first item is presented after start-up should be as short as possible
- The same is true for latency between news items
- A basic version of MACS MyNews is supposed to be provided free of charge

Those assumptions from the service scenario and a possible business case greatly affect the selection of the media format (audio/video or text) used in MACS MyNews and the channel used to transmit them. While audio contents provide a high sound quality, their production (i.e., a speaker recording them) is more expensive com-

pared to automated aggregation of texts, which are later synthesized by a text-to-speech engine. The transmission of audio content also is more expensive than the transmission of texts, since the same amount of information (i.e., a news item) uses a multiple of storage space, bandwidth, and time to transmit.

Even very highly encrypted audio files cannot be transferred over digital radio as it is available today at a reasonable rate. In the worst case, that is, when the desired information had just been transferred, the user would have to wait for up to 2 hours before that particular information would be retransmitted (based on the assumption of 2 hours of audio files, CSA-Celp encoded [61Mbit] broadcasted via 8kbit/s digital radio). This problem can be circumvented either by using text files, which are smaller and faster in transmission thus reducing the waiting time by a factor of 30, or by using broadband cellular services where the user queries for the information and receives only the news he/she is interested in. Since the MACS MyNews business case rests on the assumption that basic news is free for the user, the latter option cannot be realized as the user would have to pay the fees for data transmission over his/her cell phone. This means that in order to supply news in near real time after entering the car, text files have to be broadcasted.

The very brief and generic examination of the different possible technologies for data transmission and data representation for MACS MyNews alone shows the diversity of general technical conditions a service provider has to cope with in the design phase of the product. After the technologies promising to be the most viable have been chosen the next step is to evaluate that decision in form of a functional prototype in the next steps.

MACS Prototype and Platform

The preceding steps delivered the requirements for the new mobile service as well as possible partners in providing service and a selection of technologies that are to be used for providing the service. In order to be able to evaluate the service itself (e.g., user acceptance), the implementation (e.g., as to safety, usability, and perceived usefulness), and the practicability of the chosen technologies (e.g., technical proof of concepts), a prototype of the planned service has to be built. The outcome of the evaluation using this prototype will then be used for applying changes to the preceding levels in the MACS design framework as well as for the following considerations of safety aspects.

To be able to successfully implement mobile services in a car a service platform has to be available. Its two main purposes are to allow the easy integration of services into the car's infrastructure and thus allowing the developer to focus on the development of the services' business logic. In order to determine which services would be needed and how to structure the service platform as a framework for mobile services, an approach mixing domain analysis (i.e., automotive mobile services) (Aksit, Tekinerdogan, Marcelloni, & Bergmans, 1999) and "best practice" analysis (Boone, 1999), with the Siemens "Top Level Architecture" (TLA) in mind, proved to be useful and applicable.

The functionality of the MACS framework should be differentiated into "base services," each one grouping related methods logically and communicating among each other via the OSGi framework (Barr & Mata, 2000; OSGi Alliance, 2005). Those base services could represent either an interface to the car infrastructure, ensure the usage safety of the mobile services, or provide architectural building blocks on top of existing frameworks such as the OSGi framework (OSGi Alliance, 2005).

The OSGi framework offers many inbuilt services such as component life cycle management and dynamic service discovery of services, ensuring a robust end-to-end solution for embedded devices (Wong, 2001). Remotely upgrading the system or adding new features or services also poses no problem for OSGi-based solutions

(Palenchar, 2002). Furthermore the OSGi platform has a large user base and is used by car manufacturers such as Audi and BMW as well as tier one supplier such as Siemens VDO.

Since the base services mentioned previously are very specific to a car's infrastructure, they form an application programming interface (API) for the car's devices. Every car manufacturer would have to implement this API according to public and open specifications, allowing others to supply end-user software for the manufacturer's cars. Of course the embedded platform in a car may not be extended easily or quickly, thus the base services also enable the system designer to add or alter functionality even between platform life cycles (life-cycle mismatch, Hartmann, 2004).

For MACS MyNews several of the proposed base services have been implemented in order to allow realistic user evaluations. The user interface base service offers information output to a graphical display (i.e., textual output or graphical/video output) as well as the output of texts via a text-to-speech engine and different audio formats. On the input side the platform currently supports the (car manufacturer's specific) haptic input device as well as speech recognition. A universal cell phone adapter and a broadcast adapter allow exchanging data with services in the car or simply provide information on services in the car.

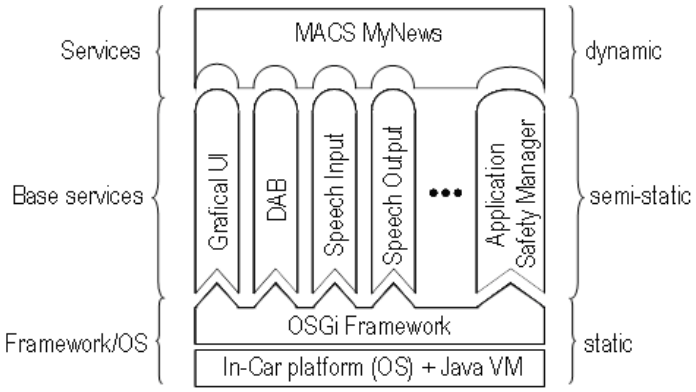
During the testing phases only the broadcast adapter is being used to receive data via digital radio broadcast. The technical specifications for unidirectional services (only using broadcast technology) and bidirectional services as worked out in the context of the DIAMOND project (Hallier, 2001, Hallier 2001a) served as foundation for the creation and definition of functionality merged in the broadcast adapter and the universal cell phone adapter. A very basic implementation of the *application safety manager* blocks detailed information from being shown on the display while driving.

An implemented a preliminary diagonal prototype (i.e., showing the unique features in the correct environment) is necessary for the following safety verifications for creating a service that is usable in a moving vehicle.

MACS Safety Aspects

The safe usage of mobile services also while driving is the single most important aspect for

Figure 5. MACS service platform architecture



mobile automotive services. The premise is never to distract the driver from the main task—driving. For that reason safety considerations already had an influence on the design of the service in the previous design steps, but also (and mainly) on the design of the mobile service platform in the car. One of the MACS project partners, the Institute of Ergonomics at the Darmstadt University of Technology, compiled a compendium of safety guidelines for mobile services from the current state of the art as found in the literature (e.g., in Becker, 1996; Tijerina, 2000; Tijerina, Parmer, & Goodman, 2000), as well as current laws and evaluated the MACS MyNews prototype against these guidelines.

In order to ensure driving safety for the automotive mobile services two main aspects have to be taken into account. First, the services are supposed to be usable in a safe way even while driving and second, the services have to comply with local legislation (Becker, 1999). The first step to achieve safety is to offer user interfaces that are qualified for usage while driving. This is why user input on the MACS platform is not limited to using a haptic input device, which means taking one hand off of the steering wheel, but allows the usage of a speech recognition engine, where the user uses voice commands (Färber & Färber, 1984). The same is also true for output devices: The MACS service platform not only offers the visual man-machine interface for the output of information but also a speech synthesis engine to output texts using the audio channel. The driver can thus focus on the street and does not have to look at the screen for information retrieval.

In order to assure a safe level of user interaction a special component, the application safety manager, controls the quantity as well as the format of information that may be delivered to users at any time. An internal rating function weights sensor information and incorporates the amount and the nature of information that is to be presented into the determination of how to present which information. Services using the MACS reference platform thus are flexible to avoid user distraction (e.g., limiting the output to only critical messages if the driver is under stress) and ensure that all services comply with

Figure 6. Taped road test usage scenario (TU Darmstadt)

current law (e.g., using text-to-speech instead of displaying written text while driving).

In order to be able to evaluate the risk potential of the prototypically implemented MACS MyNews using the service platform the Institute of Ergonomics conducted a road trial test, comparable to Wikman, Nieminen, & Summala (1998), with several test persons. The usage of MACS MyNews during those road trials were defined using different usage scenarios. On one hand the driving situation varied, that is, the road trials took place on motorways/freeways, on the highways, and on city roads. On the other hand the mode of user interaction was changed: The test persons were to operate the MyNews service using the haptic man-machine interface or using speech commands.

As a comparing parameter for the actual risk potential the car radio had to be operated in the same situations using only the haptic interface, as a speech recognition engine is not yet available for the radio. After each of those scenarios the subjective stress and distraction was evaluated using questionnaires the test persons had to fill out. In order to be able to objectively analyze the scenarios the general traffic situation (density of traffic, etc.), the man-machine interface (both input device and display) and the face of the driver were recorded by four video cameras mounted in the test vehicle. That enabled us to determine, for example, if the driver stayed in the lane, how often and for how long the driver looked at the screen in the middle of the dashboard, and how often and for how long the driver took his/her hands off of the steering wheel to use the haptic input device.

The evaluation of the road trials is not yet completed, but first evaluations of the questionnaires give strong evidence that the drivers felt comfortable using the service while driving and did not make any difference between using MACS MyNews and the familiar car radio as to a distraction from the driving task. Minor changes concerning details of the interface design have been extrapolated from that feedback and have been included in another iteration of the prototyping process step.

MACS Live Service: Some Preliminary Conclusions

The final step and conclusion of a successful development process of a mobile service is the deployment of a live service. In order to achieve a successful introduction into the market all process steps have to be re-evaluated quickly and checked for inconsistencies. These steps are currently (early 2006) being done.

The service scenario built at the very beginning evaluated the general conditions and pointed out usage scenarios and use cases. Since some time has passed since all that information was collected a review of the assumptions made has to ensure that they are still valid, for example, the targeted customer base is still available and willing to pay for the service, and so forth. A detailed business plan and an up-to-date *strengths weaknesses opportunities threats* (SWOT) analysis are mandatory for a successful and sustainable mobile service.

In order to be able to offer the service derived from the service scenario a service network was designed, also considering major technological implications. For all roles in the service network specific partners have to be found. If these partners cannot be found a start-up or spin-off company could be considered. In the current case the most likely candidate for such a spin-off is to fill the role of service providers themselves, which is about to be founded by a car manufacturer (maybe in corporation with content providers).

Since the field of technologies is very dynamic it might be possible to find better or cheaper solutions for the tasks to be solved for the service in the future. Examples for the fast changing environment are the availability and prices for mobile broadband connectivity using UMTS in Germany. While UMTS services seemed to be

far away in Germany back in the fall of 2003 (Dernbach, 2003), today more than half of the people in Germany can use the high speed mobile network (Teltarif, 2006). As of 2006, almost all of the German network providers offer flat fees for data services over their UMTS networks for about $50 per month and prices are about to decline in the future.

The service platform and the service itself exist in early 2006 as a prototype. Before the actual service can go live for the mass market the needed interfaces, also derived from the technologies that are coming into operation, have to be finally determined together with the car manufacturers. Since the target system, the car, is a safety-critical environment even more special precautions have to be taken in order to ensure a safe operation at all times. One approach to ensure that automotive software meets requirements, for example, response times, is proposed by Botaschanjan, Kof, Kühnel, and Spichkova (2005). In their development model traditional testing phases for software components are exchanged with software verification processes that are proving that critical functionalities are working properly and cannot be affected by other subsystems.

Driver safety is the most critical issue when offering a live service. The proposed service platform, however, if implemented correctly, liberates the service developer from coping with those problems. The platform is a fast and reliable way of controlling services' output to the driver at any time. Rules for the output can be adjusted to reflect the current legal situation at anytime, so for the deployment of the service those rules have to be defined according to the current jurisdiction. The safety guidelines worked out by the Institute of Ergonomics are a good starting point for re-evaluating the upcoming life service.

Last but not least: the service framework not only allows the developers to work with standardized and open interfaces, it also enables the service provider to deal with the lifecycle mismatch between the car's and the software's life cycles and keep the services offered up to date. Innovative technologies in the automotive sector, such as Wireless LAN or WiMax, will have to be supported by the framework as soon as possible in order to allow the adoption of services and the creation of novel services using the distinct properties offered by new technologies. In the case of Wireless LAN, services may be altered so it is possible to receive data additionally via W-LAN access points offering location-based information.

OUTLOOK

The application of the development framework and the development process initially proposed using MACS MyNews as an exemplary service proved the practicability and usefulness of the approach. Especially the interconnections between the different layers of the framework could thus be handled. The MACS MyNews service worked as expected, the results obtained at the various stations in the development were often presented to the public to not only receive evaluation feedback for the service itself, but also for checking for the development framework and the development process itself.

One of the main results learned through these presentations was that the users' evaluation of the usefulness of the service as well as the service usability is highly dependent on the setting in which the service is being presented. The most obvious example was one person who complained about the "unacceptable" quality of the text-to-speech engine when evaluating a GUI mock-up of the service on a PC, but complimented us on the very same component when he tried the service integrated in the car.

This lead us to the conclusion that the platform described in the development process for mobile services should be extended to offer a rapid prototyping environment enabling software developers to show services in the car very early

in the development process. Such an in-car GUI demonstrator also helps to communicate and coordinate the service development among engineers and management. It also enables the research of very new terrain of user interaction in the car, like an intelligent avatar based co driver system.

REFERENCES

Aksit, M., Tekinerdogan, B., Marcelloni, F., & Bergmans, L. (1999). Deriving frameworks from domain knowledge. In M. E. Fayad, D. C. Schmidt, & R. E. Johnson (Eds.), *Building application frameworks—Object-oriented foundations of framework design* (pp. 169-198). New York: John Wiley & Sons.

Barr, J., & Mata, R. (2000). OSGi: Spec basics, interface issues. *Electronic Engineering Times, 1144,* 112.

Becker, S. (1996). *Panel discussion on introduction of intelligent vehicles into society: Technical, mental and legal aspects. Mental models, expectable consumer behaviour and consequences for system design and testing.* Paper presented at the IEEE Intelligent Vehicles Symposium.

Becker, S. (1999). Konzeptionelle und experimentelle Analyse von Nutzerbedürfnissen im Entwicklungsprozess. In Bundesanstalt für Straßenwesen (Ed.), *Informations- und Assistenzsysteme im Auto benutzergerecht gestalten. Methoden für den Entwicklungsprozess.* (pp. 64-72). Bergisch Gladbach: Verlag für neue Wissenschaft.

Boehm, B. W. (1988). A spiral model of software development and enhancement. *IEEE: Computer, 21*(5), 61-72

Boone, J. (1999). Harvesting design. In M. E. Fayad, D. C. Schmidt, & R. E. Johnson (Eds.), *Building application frameworks—Object-oriented foundations of framework design* (pp. 199-210). New York: John Wiley & Sons.

Botaschanjan, J., Kof, L., Kühnel, C., & Spichkova, M. (2005, May). *Towards verified automotive software.* Paper presented at the 2nd International ICSE workshop on Software Engineering for Automotive Systems.

Dernbach, C. (2003). *UMTS-Start in Deutschland nicht in Sicht.* Retrieved April 20, 2006, from http://www.heise.de/newsticker/meldung/print/39973

Ehmer, M. (2002). Mobile Dienste im Auto—Die Perspektive der Automobilhersteller. In R. Reichwald (Ed.), *Mobile Kommunikation: Wertschöpfung, Technologies, neue Dienste* (pp. 459-472). Wiesbaden, Germany: Gabler.

Färber, B., & Färber, B. (1984). *Sprachausgaben im Fahrzeug. Handbuch für Anwender.* Frankfurt am Main: Forschungsvereinigung Automobiltechnik e.V.

Floyd, C. (1983). *A systematic look at prototyping.* Paper presented at the Approaches to Prototyping, Namur, Belgium.

Frost & Sullivan. (2003). *Customer attitudes and perceptions towards telematics in passenger vehicles market.*

Fuhr, A. (2001). *Die Telematik ist tot—es lebe die rollende Schnittstelle.* Paper presented at the Euroforum Jahrestagung "Telematik," Bonn.

Hallier, J., Betram, G., Koch, H., Perrault, O., Kuck, D., Korte, O., & Twietmeyer, H. (2001a). *DIAMOND: Technical specification for bi-directional services.*

Hallier, J., Kuck, D., Perrault, O., Rucine, P., Twietmeyer, H., Korte, O., Capra, L., Betram, G., Fernier, M., & Schulz-Hess, T. (2001). *DIAMOND: Technical specification for uni-directional Services*

Hartmann, J. (2004). Wo viel Licht ist, ist starker Schatten - Softwareentwicklung in der Auto-

mobilindustrie. *Automatisierungstechnik, 52*(8), 353-358.

Heidingsfelder, M., Kintz, E., Petry, R., Hensley, P., & Sedran, T. (2001). *Telematics: How to hit a moving target—A roadmap to success in the Telematics arena*. Detroit/Stuttgart/Tokyo: Roland Berger.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Kotler, P., & Keller, K. L. (2005). *Marketing management* (12 ed.). Upper Saddle River, NJ: Prentice Hall.

Lawrence, S. (2005, January). Wireless on wheels—The latest advances in telematics. *Technology Review,* 22-23.

Mohan, L. R. (2006). Driving down the fast lane: Increasing automotive opportunities the EMS provider way. *Frost & Sullivan Market Insight* Retrieved April 29, 2006, from http://www.frost.com/prod/servlet/market-insight-print.pag?docid=67150588

OSGi Alliance. (2005). *About the OSGi service platform.* Technical Whitepaper. San Ramon, CA: Author.

Palenchar, J. (2002). OSGi networks ready to roll. *TWICE*. retrieved January 15, 2005 from http://www.twice.com/article/CA198367.html

Parnell, K. (2002, Fall). Telematics drives the new automotive business model. *Xcell Journal.*

Reschovsky, C. (2004). *Journey to work: 2000.*

Royce, W. W. (1970). *Managing the development of large software systems.* Paper presented at the International Conference on Software Engineering, Monterey, CA.

Statistisches Bundesamt Deutschland. (2005). *Leben und Arbeiten in Deutschland - Ergebnisse des Mikrozensus 2004.* Wiesbaden, Germany.

Teltarif. (2006). *UMTS: Wo sind die neuen Netze schon verfügbar?* Retrieved April 30, 2006, from http://www.teltarif.de/i/umts-coverage.html

Tijerina, L. (2000, May). *Issues in the evaluation of drive distraction associated with in-vehicle information and telecommunication systems.* Retrieved May 5, 2006, from http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/PDF/3.PDF

Tijerina, L., Parmer, E. B., & Goodman, M. J. (2000). *Individual differences and in-vehicle distraction while driving: A test track study and psychometric evaluation.* Retrieved April 5, 2006, from http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/PDF/4.PDF

United States Census Bureau. (2003). *2003 American community survey.* Washington DC: Author.

Werder, H. (2005). *Verkehrstelematik als Element der Verkehrspolitik*. Paper presented at the its-ch, Olten, Switzerland.

Wikman, A.-S., Nieminen, T., & Summala, H. (1998). Driving experience and time-sharing during in-car tasks on roads of different width. *Ergonomics, 41*(3), 358-372.

Wigand, R., Picot, A., & Reichwald, R. (1998). *Information, organization and management: Expanding corporate boundaries.* Chichester.

Wong, W. (2001). Open services gateway initiative: OSGi links devices and clients. *Electronic Design,* p. 86.

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, and J. Day, pp. 199-214, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.17
Using the Railway Mobile Terminals in the Process of Validation and Vending Tickets[1]

Marko Horvat
Croatian Railways Ltd., Croatia

Mario Žagar
University of Zagreb, Croatia

EXECUTIVE SUMMARY

This article describes the functional and technical side of Railways Ltd. mobile terminals project. The advantage of mobile terminals lies in the greater efficiency of railway tickets vending, the control and real-time supervision of complete process of vending tickets in the country. Mobile terminals allow railway conductors to automatically vend and verify tickets. Also, information about each sold ticket is transmitted wirelessly via GSM/GRPS in real time or near real time. The information about sold tickets is received by the central server and stored in the main database. The data are available for analysis and report making.

ORGANIZATION BACKGROUND

Railways Ltd. is one of the railway companies in the Republic of Croatia. The company was founded in 1990 after the country gained its independence from Yugoslavia. However, the history of railway traffic in Croatia starts in the 19th century with the first railway line operating in 1860.

Today, the national railway network connects all major Croatian cities except Dubrovnik. As can be seen in Figure 1, due to geographic reasons, Croatian national railway network is quite spread out. Statistically, railroads are mostly mountainous with only one truly straight and fast line in the Slavonia lowland region that connects the capitol Zagreb and the city of Vinkovci.

The country of Croatia is located well geographically on the crossroads of Central, Eastern, and Southern Europe. There are three Pan-European Corridors running through Croatia forming the backbone of the railway infrastructure (see Figure 2). Croatia has direct railway lines to Slovenia, Hungary, Italy, Austria, Switzerland, Slovakia, France, Germany, Bosnia-Herzegovina, Serbia, and Montenegro. Also, there are indirect lines to almost all other European countries.

Public transport of passengers and freight in domestic and international railway traffic and construction and maintenance of railway infrastructure are the company's main, or core, business. Railways Ltd. employs more than 15,000 people located throughout the country and many more than 50,000 passengers travel by train in Croatia every day. As a large railway operator, Railways Ltd. is of obvious national interest. Railways Ltd. is a limited liability company in the state ownership. The company has a management board, supervisory board, 14 offices, and five regional subsidiaries. Railways Ltd.'s total annual budget is approximately 373 million Euros. Many capital investment projects, such as the mobile terminals project described in this case and the modernization of the railway infrastructure, are financed directly by International Monetary Fund (IMF) and European Bank for Reconstruction and Development (EBRD).

The implementation of the mobile terminals project is a part of the company's global modernization project. In the last five years, Railways Ltd. has gone through a series of different restructuring and modernization projects like upgrading basic railway infrastructure, dismissing unnecessary personnel, and implementing new information technologies. In the next few years, the company should be intensely privatized and divided into several smaller, more efficient, and specialized companies. The intent is to make the company completely self-sufficient and profitable.

Figure 1. Croatian railway network

Figure 2. Pan-European corridors

Table 1. Relevant information about the company

Type of business	(a) Public transport of passengers and freight in domestic and international railway traffic, (b) construction and maintenance of railway infrastructure
Budget	Annually approx. 373 million Euros + direct investments from IMF and ERBD
Organization	Management board, supervisory board, 14 offices, 5 regional subsidiaries, over 15,000 employees
Company's goal	Modernization, restructuring, privatization, improving the level of service

SETTING THE STAGE

Railways Ltd. has more than 35 years of experience in implementing and maintaining IT (Information Technology) systems. As a result, the company currently uses a mixture of information technologies for its operations. The significant acquisitions of new information systems have been carried out in three distinct periods, or waves, at the beginning of 1980s, 1990s, and in the last three years. The new systems were acquired approximately every 10 years, which was followed with intense schooling of personnel and subsequent project development. The company's corporate strategy is to seamlessly shift all existing information systems to the newest technologies using synergy and mutual integration/cooperation of all available resources. In this process, many existing systems will merge, and some will become redundant, but the main imperative is to continue to provide the current level of IT service while implementing brand new systems at the same time.

The oldest IT system in operation is Enterprise Resource Planning (ERP), which is responsible for such matters as Human Resources (HR), finance, analysis, and reporting. It runs on an IBM mainframe acquired 20 to 25 years ago. Of course, this IBM's DB2 database holds millions of data rows due to decades of operation and the company's significant business complexity. The limits of scalability were reached a long time ago, and even the limits of physical functionality could be reached any time soon. The second system in everyday use during the last 10 to 15 years is MAPPER. As a contemporary system at that point, it was acquired to relieve IBM DB2 and to improve on overall project development and management. Numerous core business projects have been developed successfully with MAPPER, like maintaining train schedules, train lists, cargo lists, ticketing, trafficking, and so forth. Many of these systems are critical for the functioning of a railway company.

In the last three years, great attempts have been made to move the entire business to the latest technologies like Microsoft's .NET Framework, Active Directory (AD), IIS Web server, SQL Server, and others. Other development platforms like Java + Oracle are in the focus, too. It is interesting to note that Railways Ltd. operates the second largest Windows Datacenter Server in Europe.

The mobile terminals project, described in this case, is at the forefront of the company's mobile computing development. The project was com-

missioned by the Railways Ltd. Office of Public Transport. The entire software development on the project was accomplished by the IT division of Railways Ltd. The Railways Ltd. Accounting Office also took part in the project.

The next few chapters will bring out important information about the project such as its cost and duration, the project team and its composition, the technology used, constraints and resources, implemented software, acquired hardware, and so forth.

CASE DESCRIPTION

The mobile terminals project supports Railways Ltd.'s core business — vending and validation of passenger tickets. However, the mobile terminals are used primarily to improve the overall quality of railway service. The reasons for implementation of this system are (a) faster data input, (b) machine ticket printout, (c) direct wireless connection to the central database for ticket validation (using the barcode), and (d) counterfeiting prevention and data protection (with database replication and synchronization). To accomplish these targets, the following conditions had to be fulfilled: (a) it is necessary to have a robust mobile device; (b) the transfer and synchronization of data has to be possible anywhere and anytime; (c) it is necessary to develop intuitive software user interface; and (d) user requirements have to be taken into account.

Depending on the project's development phase, the project team consisted of five to 15 persons in the roles of project leader, technical project leader, IT consultant, chief business process consultant, business process consultant, and software developer. Members of the project team came from three Railways Ltd. offices: Public Transport, Accounting, and IT. Because of the complexity of the railway business cases, the number of business consultants was equal or even greater than the number of technical personnel from the IT Office. During the project development, the project team had help from a consulting company specialized in mobile computing. The consulting company helped with a few algorithms and parts of the code that were incorporated into the project's software. Overall, the development team needed very little additional education, as they had strong experience in corporate computer software design and implementation using various object-oriented (OO) technologies. Only the project's mobile corporate architecture was truly a novel and previously untested approach that required the most attention from the development team. The project's users were more than 600 railway conductors and revisers from the Railways Ltd. Office of Public Transport.

It was chosen to use the iterative project development cycle and Microsoft Solution Framework (MSF) (Hansen, 2004; Thomsen, 2004) as a method of project management. MSF generally is similar to Rational Unified Process (RUP) (Booch, 1998; Hausmann, 2001; Heckel, 2001; Jacobson, 1998; Krutchen, 2000; Rumbaugh, 1998) and defines iterative project development with each iterative loop having the following development phases: envisioning, planning, developing, stabilizing, and deploying.

In order to speed up the software development and reduce errors and costs, it was decided early on in the project to use only off-the-shelf components. Such components are easily available, pre-tested, and often simpler to use than custom components. They have easily available support and a large pool of experts who can offer valuable help in crucial moments. It was also very important that all components were mutually integrated according to international IT standards. For all these reasons, the chosen software architecture was Microsoft .NET. All software was written in object-oriented languages (VB.NET and C#.NET) and database languages (SQL/SPL). Mobile terminals had the Windows CE 4.2 .NET operating system (Nienaltowski, Arslan, Meyer, n.d.), local SQL Server CE 2.0

database for storing data, and Conductor Compact Edition Railways Ltd. software application. The mobile application was written by using Microsoft .NET Compact Framework (.NET CF). The project's central server was Microsoft Windows 2000 Datacenter operating system. The server ran Microsoft SQL Server 2000 database and Microsoft Internet Information Server IIS 6.0 Web server. More detailed information about the project's architecture, software, and hardware can be found in the next chapters.

Table 2. Major concerns perceived before the start of the project development

Technical issues
- Nationwide GSM/GRPS signal coverage and how to seamlessly overcome interruptions in the wireless dataflow? - Will the mobile terminals be able to safely and securely store a large amount of data? - Is the wireless bandwidth wide enough for real-time data transfer and how to achieve it? - How to implement an efficient mobile user interface? Is it possible at all? - Are the mobile terminals enough robust and durable?
Non-technical issues
- How to efficiently organize Railways Ltd. offices on this project? - Will the users actually benefit from the project? - Will the project fulfill its goals?

Table 3. Project overview

Players involved	**Commissioned project**	Railways Ltd. Office of Public Transport
	Implemented project	Railways Ltd. IT Office
	Managed project	Railways Ltd. IT Office
	Project's users	Over 600 railway conductors and revisers from the Office of Public Transport
Team	5-15 people from Railways Ltd. offices of Public Transport, Accounting and IT Departments.	
Cost	3.5 million Euros	
Duration	2 years	
Current status	The development is over. The project is undergoing last testing and awaiting implementation.	

The biggest concerns or problems noticed before the project development started were related mainly to the used mobile technology and how well it would perform in the actual situation in the field. The second group of perceived concerns, the non-technical, was related to the project management and the users — will the company be able to successfully manage and carry out such a project, how well will 600 computer-illiterate people use the new devices, and how much will they actually help them. All major concerns perceived before the project started are listed in Table 2.

The entire project development from start to finish lasted two years. This was due mainly to the usage of new technologies and several changes to the project's use cases that occurred relatively late in the development. Total project cost was 3.5 million Euros. This included acquisition of the hardware, education of the development personnel and project's users, development of the software, consulting, and all other fees.

At the moment of writing, the project development has been finished successfully, and the project is undergoing final testing and awaiting implementation. The last chapter of this use case explains in detail the most valuable lessons learned during the development and how the most prominent concerns were solved.

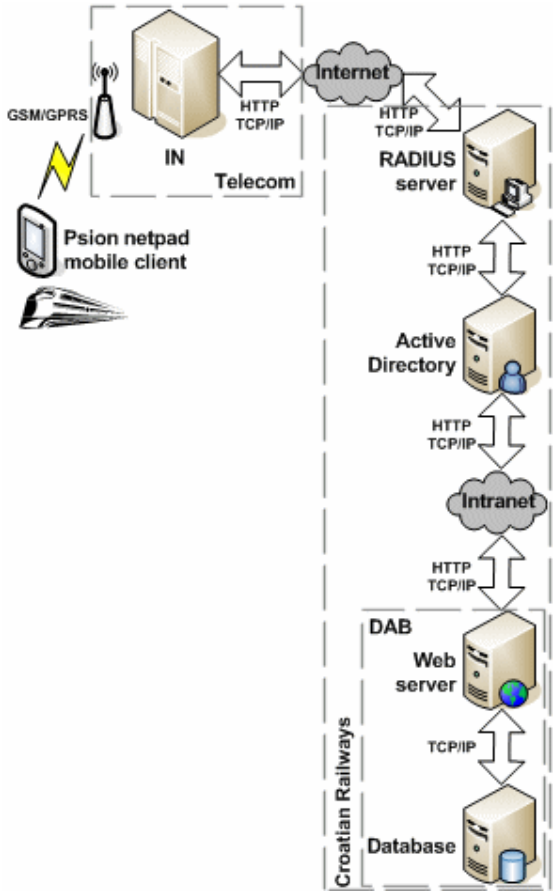
Architecture

The mobile terminals project has a multi-tiered (n-layered) client-server architecture (Chow, 1994; Socic, 1994; Wijegunaratne, 1994). The schema is shown in Figure 3. The mobile terminals and their accompanying mobile software represent the client side. Clients dial a specific number and connect wirelessly over GPRS to the infrastructure of mobile telephony provider (Cartwright, 2002; Chakravorty, 2002; Pratt, 2002).

The provider routes any IP or HTTP requests to the RADIUS (Remote Authentication Dial-In User Service) server used exclusively for the Railways Ltd. mobile terminals project. RADIUS server performs centralized connection authentication, authorization, and accounting for many types of network access, including wireless, authenticating switch, dial-up and virtual private network (VPN) remote access, and router-to-router connections. RADIUS server is the access point to the Railways Ltd. intranet network infrastructure, and together with Active Directory (AD), it authenticates a connecting wireless client. After the successful authentication, AD assigns the appropriate network group policy to the client. With the policy client is granted specific rights related to accessing and sending data in the Railways Ltd. intranet.

After the netpad mobile terminal is connected, it can start making requests to the project's Web server. Web server runs XML Web services for exchange of Extended Markup Language (XML) data files between netpad mobile terminals and server-side database. Instead of a NET-capable Web server that receives and sends HTTP requests, any IP server could be used. In this case, the exchange of data would be by sockets. The server side, or the server layer, consists of the mobile's provider infrastructure, RADIUS server, AD, Web server, and database server. In this architecture, security is guaranteed by several methods. The first one is the verification of login/password by the RADIUS server. Login and password are unique, administrated and manually defined by administrator personnel in Windows CE settings before netpad mobile terminal is put into service. Each netpad has its own SIM card, so it is possible to track completely the usage and network traffic of every mobile terminal. Also, the whole server architecture for mobile terminals located in Railways Ltd. is separated and inaccessible from other parts of company's intranet. In the future version of mobile terminal's software, it is planned to secure the entire project's network traffic by VPN (Virtual Private Network) standard. RADIUS server and other hardware for VPN support already exist.

Figure 3. Project's architecture



Hardware

A single mobile terminal set consists of two devices: Psion Teklogix netpad (5000 series) mobile computer (PDA) and Extech Bluetooth (S3500T series) mobile printer. A conductor wears both devices at the same time. The netpad (Figure 4) is placed in a leather holster and worn over the shoulder, and the mobile printer is strapped to his belt. Each netpad has Windows CE 4.2 .NET SP2 real-time operating system with color touch-screen with pen navigation, Bluetooth (Haartsen, 2000) and IrDa communication interfaces (IRDA, 1997), SecureDigital/MultiMediaCard (SD/MMC) slot (Praca, 2003; Praden, 2003), SIM card slot, inbuilt laser barcode scanner, and GSM/GPRS wireless connection (Kalden, 2000; Meirick, 2000; Meyer, 2000). Wi-Fi 802.11b module is optional. The netpad has Intel SA–1110 Strong ARM processor at 206 MHz, 64MB SDRAM, and 32MB flash ROM. LCD screen has 8-bit color depth (256 colors) and half VGA resolution of 640x240 pixels. Using a user interface picture can be dynamically rotated to portrait or landscape the mode. The size of the netpad is 215 x 85 x 25 mm, and its mass with standard battery is 620 g. Two battery packs are available: standard battery has 1800 mAh, and the larger (and heavier) has 4400 mAh. Both batteries supply 7.2 V and guarantee eight to 14 hours of operation. The netpad can have either tranreflective or transmissive screen. In the so-called netpad outdoor model, the screen is transreflective and, therefore, more readable in the direct sunlight. The backlight can be turned off. In the indoor model, graphics are less readable in the sunlight. The screen is transmissive, the colors are more vivid, and the backlight can be turned off. The netpad is certified according to IP67 (Ingress Protection) rating, which ensures its robustness. It was specifically designed for use within demanding mobile computing environments. Netpad is completely dust-proof. The netpad can sustain a 1.5-meter high fall on the concrete and is protected from temporary immersion in water one meter deep for 30 minutes. Also, apart from having a metal casing and being covered in the protective rubber, the netpad also has a very tough screen. That kind of ruggedness is highly needed, since the devices will be deployed in the field in tough conditions exposed to hits, drops, rain, and different weather conditions. The alternative to the durable netpad was the standard Pocket PC, but since it cannot withstand a great deal of physical abuse, it would be necessary to acquire additional

Figure 4. Psion Teklogix 5xxx netpad PDA

Figure 5. Extech S3500T Bluetooth printer

units to quickly substitute the one that has been broken. Therefore, the acquisition of netpads was judged to be the most prudent choice.

Extech Bluetooth printer is visible in Figure 5. This is a monochrome thermal printer. It has serial (RS232), IrDa, and Bluetooth communication interfaces. It also has Magnetic Stripe Reader as a factory-installed option. The printer's size is 152 x 120 x 57 mm, and mass with paper and battery pack is 580 g. Operating voltage is 5V, and with one battery charge, approximately six rolls of paper can be printed. All settings are configured with DIP-switches. In Mobile Terminals, the project netpad and its mobile printer communicate only with Bluetooth interface. When each mobile printer is turned on, it first automatically pairs itself via Bluetooth (Morrow, 2002) with netpad in its mobile set. Afterwards, the netpad freely can send data to the paired mobile printer using Bluetooth wireless connection.

Software

In order to fulfill their tasks outlined in the introduction, the netpad mobile terminals, which constitute the client's hardware, must have a mobile application called Conductor Compact Edition installed as the client's software.

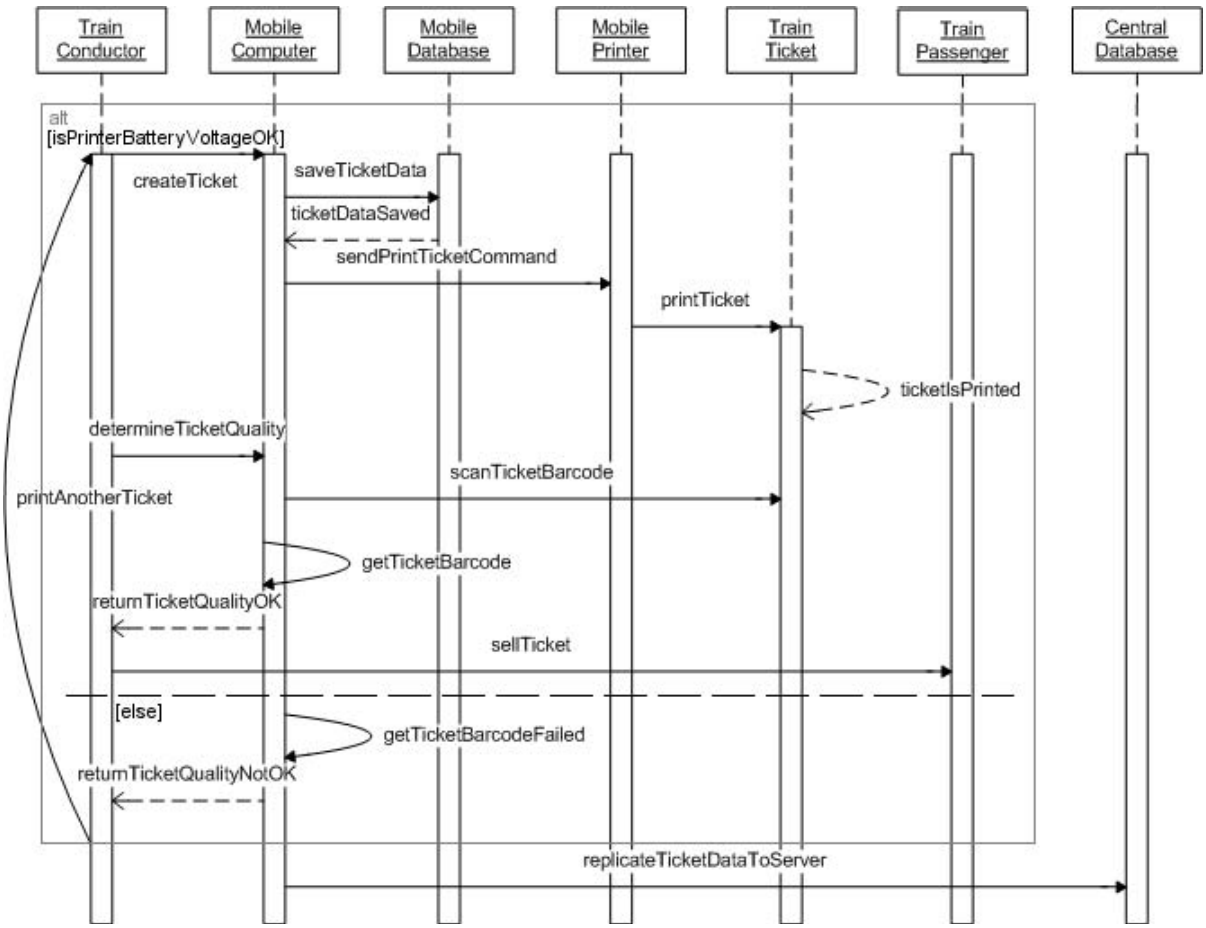
The application Conductor Compact Edition is designed for use on a Windows CE mobile computer with Bluetooth, GSM/GRPS wireless networking, and barcode reader. These components make one of the most capable all-in-one configurations in the mobile development today. This, in turn, means that the software developers can have a lot of IT tools at their disposal to accomplish the project's goals.

The application uses all the features of a large color touch-screen to enable faster and simpler data input. In accordance with the software requirements, Conductor Compact Edition allows maintaining encrypted data in a mobile database, replication between the mobile and the central databases, the transfer of data and synchronization of databases by fixed docking station, serial connection (RS232), IrDa interface, or GPRS wireless connection. Conductor Compact Edition loads on netpad startup, and the user is required to log on. The user enters his unique access code in the application's login screen. The data are checked against access codes stored locally (on the netpad) in a separate SQL Server CE database file, which

is encrypted and protected. This feature enables user login, even when GSM/GRPS connection is unavailable. In its business logic, Conductor Compact Edition uses graph representation of railways and the network of railway stations to calculate the route and distance with Floyd's algorithm (Floyd, 1962). The price of each passenger ticket is calculated using these data and mathematical algorithms with the business logic. The price of the ticket is proportional to the route length and a number of parameters, such as various fare reductions, class, single or two-way ticket, and so forth. Conductor Compact Edition executable is permanently stored on a single SD card, so it cannot be erased by loss of power on the netpad. The most important application's use case — ticket vending — is displayed in Figure 6. While the user is working with the application, it dynamically stores all the data about the vended tickets on the mounted SD card. The data are stored in a single SQL Server 2.0 CE database file.

The data located on the card are erased after each successful synchronization session when the central database is over. This is done because sev-

Figure 6. UML sequence diagram of "ticket vending" use case

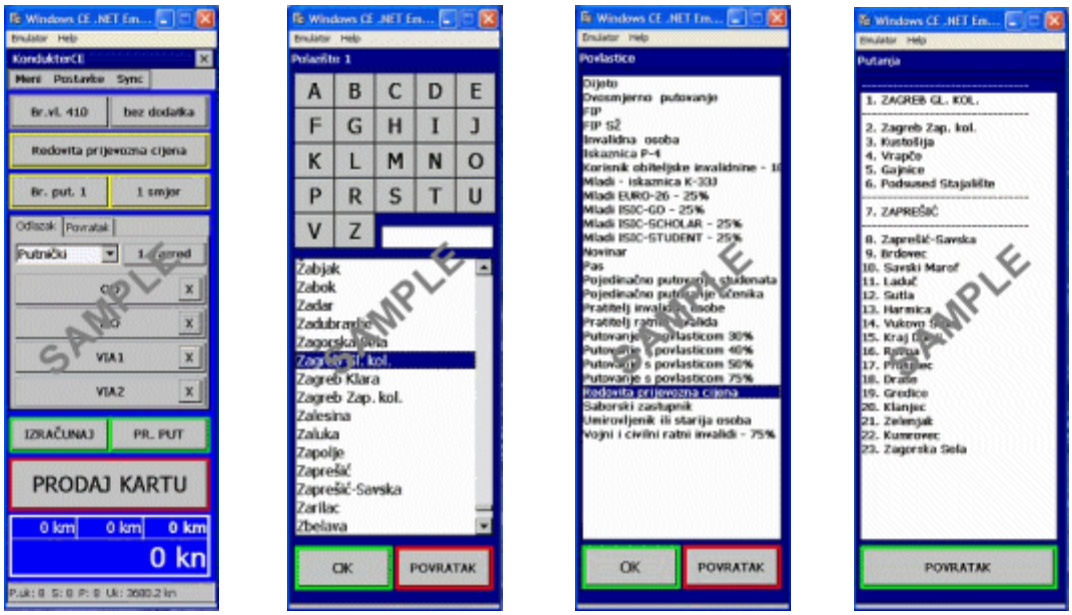


eral conductors can use a single mobile terminal. Therefore, permanently maintaining their data on the mobile device can be a security problem, but it also serves no logical purpose, since the data already have been replicated to the server. The project description dictates that the database synchronization and the data replication are done via GPRS in real time. If at any point GPRS connection isn't available, Conductor Compact Edition waits until GPRS connection is working, so it can send data to the server. GPRS can be unavailable due to a number of reasons (e.g., the mobile terminal can be in a tunnel or in a remote area without GSM/GPRS signal coverage). At the end of each working shift, the mobile terminal is returned to the supporting IT center; if the local database has not been synchronized already with the central database, then it is synchronized by using the netpad cradle and PC workstation. The PC is connected to the Internet with the telephone lines via a standard modem, ISDN, or DSL. Because lower-level software (Windows CE, netpad drivers, and API functions) handles the basic networking issues, Conductor Compact Edition application uses the same .NET functions for the data replication over wireless and cable networks. In other words, the application doesn't care how it is connected to the Internet, as long as it can communicate with the project's Web services and synchronize local and central databases. After a successful synchronization, the data are erased from the SD card as with GRPS synchronization.

Mobile Graphical User Interface

GUI (Graphical User Interface) of a mobile computer application that is intended to be used in field conditions outside the office is quite particular in its design and functionality. Although Pocket

Figure 7. Conductor compact edition graphical user interface



PC and Windows CE mobile devices have one of the richest and most capable user interfaces in the mobile world, using all of its complicated features would be a huge mistake.

Even though the elaborate user interfaces can show more graphical and textual data in a given display area and allow different types of user input, they obviously are not always the best choice. In moving environments, the varying lighting conditions and cramped areas complicating user interface only would thwart the user. For example, it is completely plausible that at some point, the pen stylus might be dropped or lost, so the user must be able to continue the work with his or her hand instead (i.e., slowly, but steadily). Because of this, Conductor Compact Edition application has only a few full-screen forms (windows). Some of them can be seen in Figure 7. Conveying the information with the standard message boxes has been avoided completely, and the usage of overlapping forms has been reduced to the very minimum. Critical messages are displayed in the currently displayed form or in a custom message box. Usage of any more complex user interface controls, such as dropdown menus, checkboxes, combo boxes, lists, list views, and tree views, also has been avoided completely. Only textboxes, standard buttons, toggle buttons (as a replacement to the checkboxes), and lists were used. All parts of the user interface have increased the width, height, and font size to make sure that they all can be read clearly and used in out-of-office conditions. The colors are used to relate intuitively the basic information (i.e., red borders around buttons indicate command termination and green borders continuation).

Printing and Validating Tickets

One of the most important capabilities of mobile terminals is printing and validation of tickets. Therefore, it was important to ensure good ticket printing quality and to pay attention to possible counterfeiting. Various use cases that can occur in everyday situations were analyzed. It was decided that tickets printed by mobile terminals should have two levels of prevention against forgery. First, tickets would be printed on a special paper that cannot be photocopied. A conductor would be able to determine with a simple visual examination whether a ticket is genuine or not. Second, each ticket bears a barcode containing the information needed for its verification. Importantly, each printed ticket has a unique identification number and a fare's data stamp encoded in the barcode with all other important fare data. Since a large quantity of data has to be encoded in the barcode, it was decided to use a high density Code 128 standard (Bushnell, 1998; Meyers, 1998).

Data are encrypted (scrambled) in the barcode with a custom mathematical algorithm. The data encoding algorithm was improved throughout the project's development and judged to be adequately safe. The barcode would be horizontally centered in the upper part of the ticket, so it was clearly visible and scanned. This way, it was less likely that during handling the barcode would be smudged, painted over, or torn off from the ticket altogether. The appearance of a single test ticket printed with the terminal set is shown in Figure

Figure 8. Printed test ticket

8. Commands to the mobile printer were issued in formatted escape codes and transmitted via Bluetooth along with the ticket data.

Since Extech S3005T is a thermal printer, the quality of printed text greatly depends on its battery power level. As the printer's battery output falls, so does the quality of its printouts, and, at some point, tickets will become unusable. Therefore, it is very important to determine the quality of a printed ticket before it is handed over to the passenger, and as outlined in the UML diagram in Figure 6, it is sometimes necessary to bend the business processes a little and to accommodate for the limitations of the project's hardware.

The project's system tests have yielded encouraging results concerning ticketing. By using mobile terminals, the time required to vend a single ticket has been reduced three to five times, which is visible mostly when several tickets are vended in a row. All users have pointed out that their workload has been reduced significantly. As can be seen in Table 4, a conductor can type in a single ticket in 30 to 45 seconds or even less, if several identical tickets have to be produced. Then, it takes another 10 to 15 seconds for a ticket to be printed. Parallel to these tasks, if the GSM/GPRS signal is present, the mobile computer wirelessly transmits the vended ticket's data. Information about one vended ticket is transmitted to the server in 25 to 50 seconds. As can be seen, all these tasks are measured in seconds, and they are quickly finished, but when 150 to 200 tickets have to be vended per hour, which is expected to happen, this certainly will be a more significant task for the system than it is designed to handle.

CURRENT CHALLENGES/ PROBLEMS FACING THE ORGANIZATION

As was already mentioned, the project development has been finished successfully, and the project is awaiting implementation, after which the actual benefits or drawbacks will become visible. However, from the test runs, it is already evident that the mobile terminals greatly improve the speed and efficiency of the railway conductors and revisers.

The mobile terminals project is not the only project of its kind in the world, but it has several special and advanced characteristics that distinguish it from the others. First, this is a nationwide project with more than 600 users and a large data turnaround. Second, the unique technical features are (a) database synchronization via GSM/GPRS, (b) ticket validation via custom barcode and GSM/GPRS, and (c) printing on Window CE-based proprietary device via Bluetooth. Third, only off-the-shelf components have been used, reducing the project's price and time to market.

Table 4. Ticket vending speed

Task	Duration
Vend a single ticket	30-45 sec.
Vend several similar tickets	10-15 sec. per ticket; after the first ticket
Vend several different tickets	20-45 sec. per ticket; after the first ticket
Print a single ticket	10-15 sec.
Synchronize data for a single ticket	25-50 sec.

Table 5. Major issues observed after the end of the project development

Technical issues
- Technical problems are solvable. Majority of technical problems can be solved directly while only a few problems, or obstacles, must be avoided using different technical solutions. - It is possible to construct efficient real-time GSM/GPRS data transfer on a national level with custom software mechanisms to overcome signal gaps. - Practical and intuitive mobile user interface is possible and even computer inproficient personnel can start to use it with little or no training. Over time the users will completely familiarize themselves with the interface and learn to use it by memory. - Mobile terminals are durable devices but attention has to be paid to maintaining their power supply. Software can be far less robust than the hardware.
Non-technical issues
- Organization of a similar project in a large company is a major undertaking. Maybe this has been the greatest challenge during the entire project development. Any project introducing new technologies into business areas that haven't been computerized before can face the same problems. The problems are numerous and they are mostly coming from various inefficiencies in the company's organization. It is of a paramount importance to obtain all business workflows and use-cases as early in the project development as possible and not to change them later. - Users will benefit from any computer application that automates their jobs. Given the proper climate and incentives personnel will be happy to use the application especially when they understand how well it helps them. - Acquiring needed hardware and developing required software is not enough to make a project successful. Complex projects with a lot of users spread over a large area often must have numerous support staff that is even larger than the development team. Keeping the project up and running is a large and ongoing task. - The project has achieved all its perceived goals. All that remains is for the management to make a corporate decision and introduce the mobile terminals into everyday business.

Table 6. Problems that the organization has faced and challenges that it will face in the future

Past challenges and obstacles
- How to best define an information technology project of this size and how to optimally define its scope? How to incorporate such project in the existing organizational structure? - How to efficiently run the project? In other words, what is the best way to plan, implement and manage the project in the given organizational situation? - What is the optimal composition of the project team? Who are the players involved and what is their expertise? - What are the obstacles in the development of the project's software? How many off-the-shelve technology solutions can be used, and how many must be custom developed?
Future challenges and obstacles
- Are there any unforeseen difficulties that will arise during the project's implementation? Has the project team done good enough job, and are the information technologies mature enough so implementation difficulties will not threaten the project? - After the project has been successfully completed how big positive impact it will have on the organization's business? What will be the project's overall consequences on the company as a whole?

In the future, some of the planned improvements to the project are two-way communication (messaging and information-on-demand) between the mobile terminals' users and the control, enabling reservation of seats on the train, data analysis with data warehousing (DW), and data mining and printing of Railways Ltd.'s timetable.

All significant project issues noticed after the project development was finished and the lessons learned can be found in Table 5.

Finally, Table 6 lists all the questions and problems that the organization faced before the project was completed, as well as the challenges that the project and the organization will face in the future.

REFERENCES

Booch, G., Rumbaugh, J., & Jacobson, I. (1998). *The unified modeling language guide*. Reading, MA: Addison Wesley.

Bushnell, R., & Meyers, R. (1998). *Getting started with bar codes: A systematic guide*. Arlington, MA: Cutter Information Corp.

Chakravorty, R., Cartwright, J., & Pratt, I. (2002). Practical experience with TCP over GPRS. In *Proceedings of IEEE GLOBECOM*.

Floyd, R. W. (1962). Algorithm 97. Shortest path. *Communications of the ACM, 5*(6), 345.

Haartsen, J. (2000). The Bluetooth radio system. *IEEE Personal Communications, 7*(1), 28-36.

Hansen, J. E., & Thomsen, C. (2004). *Enterprise development with visual studio .NET, UML, and MSF*. Berkeley, CA: Apress.

Hausmann, J. H., & Heckel, R. (2001). Use cases as views: A formal approach to requirements engineering in the unified process. *GI Jahrestagung*, (1), 595-599.

IRDA. (1997). *IrDA object exchange protocol (IrOBEX)*. Retrieved from http://www.irda.org/

Kalden, R., Meirick, I., & Meyer, M. (2000). Wireless Internet access based on GPRS. *IEEE Personal Communications, 7*, 8-18.

Krutchen, P. (2000). *The rational unified process: An introduction* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Morrow, R. (2002). *Bluetooth: Operation and use*. New York: McGraw-Hill Professional.

Nienaltowski, P., Arslan, V., & Meyer, B. (n.d.). Concurrent object-oriented programming on .NET. In *IEEE Software Proceedings, 150*(5).

Praca, D., & Praden, A.-M. (2003). Smart cards and smart objects communication protocols: Looking to the future. In *Proceedings of the 2nd Gemplus Developer Conference*, Montpellier, France.

Wijegunaratne, I., Socic, M., & Chow, C. (1994). An architecture for client/server application software. *Australian Computer Journal, 26*(2), 30-41.

ENDNOTE

[1] Organization name, data, trademark products, or services referred to in this case are fictitious. All facts, data, and figures in this case are free and publicly available.

This work was previously published in the Journal of Cases on Information Technology, edited by M. Khosrow-Pour, Volume 8, Issue 2, pp. 30-44, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.18
An Evaluation of U.S. City Government Wireless Networks for Mobile Internet Access

Ben Coaker
Whiting-Turner Contracting Company, USA

Candace Deans
University of Richmond, USA

ABSTRACT

The purpose of this chapter is to provide guidelines for city governments considering implementing large-scale wireless networks to provide Internet access for their citizens and businesses. Case studies of cities in the United States that have implemented wireless networks will be evaluated in the context of opportunities and potential challenges. Some key considerations discussed in this chapter involve free versus fee-based models, security considerations, conflicts with local telecommunications companies, and network support. Opportunities to benefit police and emergency services are examined in terms of potential benefits as well as considerations of security in mission critical situations. Strategy guidelines will be presented as a means for providing structure to this decision-making process.

INTRODUCTION

The spectrum used for wireless technology is under FCC control, but is not charged a usage or license fee. This simple fact has many cities looking at the opportunity of providing wireless access to local residents and businesses. Some municipalities have a business model that provides this service for free, while others are considering the options for charging some type of usage fee. Many issues come into play as city government administrators and boards evaluate the opportunities and potential problems associated with taking on the responsibility of implementing and maintaining a wireless network that provides residents with new conveniences and flexibility afforded by mobile Internet use. There is also potential benefit to businesses to enhance visibility and marketing efforts at a cost less than telecom-

munications companies or other profit-oriented businesses might provide.

Evaluation of long-term benefits for the municipality is essential and yet difficult to evaluate given the fast changes in technological developments. Should city governments even get into this business of mobile Internet access? What are the decision variables? This chapter will help simplify the decision process by providing analysis of city case studies and the current state-of-the-art in terms of benefits and potential drawbacks.

The major objective of this chapter is to evaluate the options for U.S. city governments considering the implementation of large-scale wireless networks to provide mobile Internet access to residents and businesses. Several cities within the United States have already successfully implemented these systems for their residents. Case studies of cities that have implemented wireless networks will be examined to provide insights into the opportunities and potential challenges that are unique to government entities. Revenue generation is a key issue in the overall strategy and decision process. This chapter will provide a discussion of the pros and cons of implementing wireless networks in the context of unique issues faced by city governments.

BACKGROUND

Currently, there is very little research and literature addressing the issues of wireless Internet implementation for government entities. The technology is relatively new in this environment and cities have only recently begun experimentation in this arena. Most of the literature currently available is in trade publications and press releases that address current issues of interest to business and government leaders. The focus of this chapter is on city governments in the United States. Similar trends and issues are emerging in other parts of the world as well.

A major issue surrounding municipal ownership of wireless networks involves competition with the telecommunications companies. Most telephone and cable companies oppose community efforts to offer wireless Internet and in many cases have lobbied to stop municipal Internet zones (DeGraff, 2005). Those favoring municipalities in these efforts believe legislation will ultimately impact America's ability to compete globally. It is essential that all Americans have access to Internet services in order to benefit from the higher standard of living afforded by the Internet. Those without access will be left behind in the rapidly evolving marketplace. According to the National Telecommunications and Information Administration, 40% of Americans do not have dial-up access to the Internet at home and 80% do not have high speed access. One out of four Americans do not use the Internet at all. This places the U.S. behind many countries in Europe and Asia. Those without Internet access cannot benefit from online advertisements for job openings and other available information on the Internet. Those who lack Internet access are typically low income, minority, less educated, and unemployed. Affordable or free Internet access could provide these members of society with the benefits of participating in an Internet-based society (DeGraff, 2005).

Income is a major factor in this divide between the Internet haves and the have-nots. Charges run on average $40 to $60 a month and prices continue to rise. Cable and phone companies claim that municipal Internet service is unfair competition. They argue that cities could still provide access to community centers, schools, and libraries through their services. Public interest groups argue that communities have a right to build their own Internet networks to offer more services at a lower cost. School districts could save considerably on current costs for high-speed Internet access (DeGraff, 2005).

Some argue that cities could better spend their money on teaching literacy and computer

skills since these are the issues at the root of the digital divide. They argue that broadband Internet access is different than power, telephone, and other "utilities". The costs are astronomical and failures are also high. Building a redundant competitive broadband system may deter the private sector from building new systems or upgrading systems already in place. The private sector can provide higher quality and more secure systems (Effros, 2005).

As more cities investigate this alternative, the debate will likely become more heated as both sides move to protect what they perceive as their best interests. Closing the digital divide is the ultimate goal but finding the best solution remains the challenge for all parties involved.

TYPES OF WIRELESS TECHNOLOGY

There are two main types of wireless technology that are available to municipalities that are interested in implementing a large-scale wireless network. These available technologies are local area networks (LANs) using Wi-Fi technology or metropolitan area networks (MANs) using Wi-Max technology. Each of these technologies has benefits as well as weaknesses associated with them.

Wi-Fi technology utilizes the IEEE 802.11 standard. This standard operates on three different levels. 802.11b was the first standard released and provides transfer speeds up to 11 mbps and operates in the 2.4 GHz range. 802.11a was released several years later and increased transfer speeds up to 54 mbps; however 802.11a operates in the more expensive 5.0 GHz range. Recently, 802.11g has been introduced as a cost effective alternative. 802.11g operates in the 2.4 GHz range keeping it relatively cheap, but still provides transfer speeds up to 54 mbps (Senn, 2004). While Wi-Fi technology is relatively inexpensive, the main weakness of Wi-Fi is the fact that the signal starts to degrade after approximately 100 feet. This gives Wi-Fi an effective operating area of about 31,000 ft^2, which is a little more than one city block. Using Wi-Fi technology, the municipality will have to put one node on approximately every corner in the covered area. Not every node will need to be hard wired to an Internet server, but a large portion will. Wi-Fi performance starts to deteriorate as more and more users access the system. In order to counter the problem of lesser service, several Wi-Fi nodes will need to tie directly into an Internet server, giving the wireless network several access points needed to cope with the anticipated volume.

A new development in wireless technology is the use of MANs, using Wi-Max technology. Wi-Max technology falls under the IEEE 802.16 specifications. Wi-Max technology operates in much the same way cellular technology operates. Instead of having an effective operating distance of 100 feet as in Wi-Fi, Wi-Max can operate up to 30 miles under ideal conditions. Wi-Max operates in one of two ways. The first is through line of site from one tower to another. A steady stream of data is beamed from one tower directly to another, up to 30 miles under ideal conditions. The effective distance can be affected depending on weather conditions and other obstacles that may be in the way. The second way Wi-Max operates is through non-line-of-sight, similar to the way Wi-Fi works. When Wi-Max is not relying on line of sight, its effective distance is cut to about a 5 mile radius, which translates to approximately 25 mi^2. Under this scenario, a city could cover its entire corporate limits with four to six towers strategically placed to provide maximum coverage. Also, since the line of sight aspect of Wi-Max can carry such high capacities, very few towers (possibly only one) will need to be hardwired into an Internet server.

When deciding whether to use Wi-Fi or Wi-Max technologies for its wireless Internet network, each municipality has to make some basic decisions which will dictate how to proceed.

First, the local government needs to determine if the existing network of cell phone towers would provide adequate Wi-Max coverage should the cell phone companies allow the municipality to attach a Wi-Max transmitter atop the towers. Second, if towers are not available, or coverage would not be adequate, what would the local reaction be to putting up more towers? Would more towers cause an eyesore to the existing layout of the covered area? Third, would it be more expensive to erect Wi-Max transmitter towers or place a Wi-Fi transmitter node on every city block corner?

CITY GOVERNMENT WIRELESS NETWORK INITIATIVES: CASE STUDIES

The following case studies represent first initiatives by cities in the United States to implement city-run wireless networks. Each city faces different opportunities and challenges but the solutions are similar. The main issue involves whether to offer the service for free and if not what to charge for the service. In some cases, cities have found sponsors to help defray the costs while in other cases the citizens cover the entire costs through monthly fees. The challenge is to find the long-term solution that best meets the needs of the particular situation at hand. The lessons learned by these cities will provide valuable insights for other cities evaluating the potential for their own citizens.

Philadelphia, Pennsylvania

According to the 2000 census, there are approximately 1.5 million people living in the Philadelphia county area. Philadelphia also saw 25.5 million visitors (business and leisure) in 2004, (www.gophila.com). Therefore, on any given day there are more than 2 million people within the Philadelphia metropolitan area. These population facts accompanied by the fact that Philadelphia is a major business center and tourist destination along the east coast, makes this city a perfect test bed for low cost wireless Internet access. Philadelphia has decided to install a city-wide wireless network based on current Wi-Fi 802.11b standards. In July of 2004, Mayor John F. Street appointed an independent council to research, recommend, and coordinate the implementation of a low cost wireless Internet network. This council, the Wireless Philadelphia Executive Committee (Wireless Philadelphia, 2005), is acting in an advisory/advocacy role in the process of developing a wireless community network by instituting community outreach programs, communications with the press, and participation in meetings and conferences (www.phila.gov/wireless).

With the creation of Wireless Philadelphia, the city now has a full-time organization whose sole purpose is to oversee the implementation of a wireless network throughout the city and coordinate all the resources needed to maintain this network. The sole purpose of Philadelphia Wireless can be broken down into three main functions: provide a forum, recommend policy, and identify barriers along with strategies to overcome them.

According to its Web site, www.phila.gov/wireless, Philadelphia Wireless will provide a forum that will enable potential future users to provide feedback on various issues of concern. This forum also allows Philadelphia Wireless to explore any and all emerging technologies that may enhance and improve performance on the wireless network once it is in place or before implementation is complete. Another aspect of the forum to Philadelphia Wireless is the promotion of third-party development of research, development, and use of mobile mesh networks. Inclusion of a third party, or multiple parties, will help defray the cost of the wireless network as well as the burden of maintaining and servicing this new network. By using the open forum to explore concerns, new technologies, and partner-

ing scenarios, Philadelphia Wireless allows for a higher volume of information to be examined as well as multiple viewpoints to be presented. With all the information gathered from the forum, Philadelphia Wireless will then be able to make recommendations on policy.

The second function of Philadelphia Wireless is to provide recommendations on different policy areas. These policy areas include fee structure, roles and responsibilities of personnel, level of service, and privacy and security issues. Some recommendations of Philadelphia Wireless already include a fee structure that provides some level of free wireless Internet access to everyone within Philadelphia city limits with a tiered system of payments for more advanced access. Preliminary fees for advanced access are $10/month for qualified residents and less than $20/month for non-qualifying residents and businesses. Since Philadelphia is a large city with infrastructure in place, wireless nodes will be placed on city streetlights and traffic lights. Locating the nodes in these places will allow for connection to a pre-existing source of power as well as a grid network that is already in place. According to a press release by Philadelphia Wireless, no city money will be used to install and/or maintain the wireless network. Third-party personnel will place and connect the nodes as well as maintain and service the nodes as appropriate. The level of service provided initially will be free access in some parks and public spaces (Philadelphia Wireless Press Release dated October 3, 2005) with an opportunity to expand service areas as more nodes are installed. Privacy and security issues will be handled by the individual service providers and the end users. Each service provider and end user will be responsible for protecting their data and abiding by all applicable laws.

The third function to Philadelphia Wireless will be to identify barriers and develop strategies to overcome these barriers. One such barrier is the resistance from telecommunication companies. As more and more cities are starting to look into the benefits and opportunities of providing free and/or low cost wireless Internet access to the population, telecommunication companies are providing more and more resistance through political lobbying techniques. For example, in November of 2004, Pennsylvania Governor Ed Rendell signed a law that disallowed any local government from providing broadband Internet access "within the service territory of a local exchange telecommunications company". The law also states that the only way a municipality can deploy its own network is if the government first offered the local telecommunication company the opportunity to deploy its own network. This law was largely backed by telecommunications giant, Verizon (Chen, 2005). Philadelphia was allowed to circumvent this roadblock by brokering a deal with the governor and Verizon.

The most important fact of Philadelphia's initiative to provide wireless Internet access to its population is that no money will come from city coffers for installation, maintenance, and upgrades of the wireless network. According to a press release from Philadelphia Wireless, (October 3, 2005) Philadelphia Wireless and Earthlink have entered into the final stages of contract negotiations for Earthlink to "…develop and build the nation's largest municipal Wi-Fi broadband network." According to the press release, the agreement between Earthlink and Philadelphia Wireless specifies that Earthlink will "…finance, build, and manage the wireless network, and provide revenue-sharing fees to support Wireless Philadelphia." While providing a great service to the population such as free and/or low cost wireless Internet access is important, more important is the need to relieve financial stress on the municipality. In today's economic times, everybody is trying to reduce budgets and cut spending. Therefore, finding a means to increase services to city residents and visitors without spending more money is the goal of every

local and state government in the United States. According to the press release (October 3, 2005), the agreement between Philadelphia Wireless and Earthlink specifies that Earthlink will provide a 135mi^2 city-wide Wi-Fi mesh network that shall provide the following:

- Low cost wireless high-speed Internet access;
- Hot spot access providers will posses roaming capabilities;
- Free access in some parks and public spaces;
- Occasional users and visitors will be able to obtain daily and weekly access;
- Small businesses will be able to connect to the wireless network and use it as a T-1 alternative;
- Future expansion of wireless network into different areas of Philadelphia; and
- Future implementation of emerging technologies as they become economically viable.

According to Philadelphia Wireless's Web site, www.wierlessphiladelphia.org, "Philadelphia's goal is to become the number one wireless city in the world and intends to set the standard by which wireless accessibility is measured." This statement, along with the recent progress made by Philadelphia Wireless, shows the aggressive nature that Philadelphia is pursuing in wireless technology in order to provide a better experience for its citizens and visitors. Estimates of installing Wi-Fi transmitters in Philadelphia are 10 units per day with approximately 8 to 16 units needed per square mile.

Alexandria, Virginia

Alexandria, Virginia is another locality that is pioneering the municipality supplied wireless Internet access. However, Alexandria's plan differs from Philadelphia's plan in two important ways. Alexandria is planning on offering wireless Internet access at no charge and only in a small portion of the town as well as all the public libraries. Alexandria's wireless plan, appropriately named Wireless Alexandria, is to provide free wireless Internet access in a small portion of Old Town Alexandria, specifically the eight block zone from Washington street to the Potomac River along King Street (Gowen & MacMillan, 2005). This portion of Alexandria is a vibrant area that attracts tourists and residents alike due to the large number of shops and restaurants in the vicinity. There are several positives as well as negatives associated with a wireless network of this size.

Wireless Alexandria has all the usual benefits that come with a municipal government providing wireless Internet access. The wireless footprint provided can easily be enlarged due to the fact that the local government owns several buildings in the area as well as all the street and traffic light poles on which to mount the wireless receivers. Emergency officials can utilize the wireless network while within the area receiving the wireless signal. Tourists and residents alike can check and send e-mail or surf the Internet while enjoying a nice day in an outdoor café or while taking in the sites alongside the Potomac River. Due to the fact that the network is confined to a small area, Alexandria has been able to implement the network at a relatively low cost of approximately less than $14,000 initial investment and about $650 per month for the T1 Internet connection (Fifer, 2005). One other benefit unique to the situation in Alexandria is the high tourism traffic in this area. Parking control can be greatly enhanced due to the wireless network. For instance, instead of having traffic officers concentrate on reading meters all day, these same meters can be monitored automatically in a central location, and when the meter expires a notification can be sent to any officer in the vicinity to go and ticket the parking meter violation. Another possible usage for Alexandria's wireless network would be for trash cans to

sense when they are getting full and notifying the sanitation department automatically (Walsh, 2005). This section of Alexandria has many points of interest, whether it is shops, restaurants, or historical landmarks. A free wireless network will allow users to learn about the history while actually standing in front of the point of interest, or peruse the dinner menus of several different restaurants while walking down the street. The most important benefit of Alexandria's wireless network is the way the network is set up; there is no competition between the telecommunication companies and the local government providing free wireless Internet access.

While there are many benefits to Alexandria's wireless network, there are also some negative aspects as well, many common to free networks. First, the wireless network that Alexandria has put into place is an unsecured network. An unsecured network is one that has no encryption software in place guarding the information that is being sent back and forth along the network. Alexandria is making a concerted effort to let users know that Wireless Alexandria is an unsecured network and no financial transactions or any type of communications with sensitive information should be conducted while connected to Alexandria's wireless network. Alexandria's wireless Internet is set up so that the only reliable signal will be available out doors along the eight block area between Washington Street and the Potomac River, along King Street. Finally, since this is a free service, users are not going to be able to get customer support 24 hours a day.

Alexandria is avoiding the wrath of the large telecommunication companies that some other cities are coming into contact with as they try to implement wireless technology for public use. However, it should be pointed out that in large cities such as Philadelphia, where the local government is trying to install a service that will take the place of the private Internet service provider, Alexandria is not trying to replace the services provided by the telecommunication companies. Alexandria's main purpose for the wireless network is to provide services for emergency personnel and other city services in a strategically located area. The fact that the residents and tourists are able to "piggy back" onto this wireless network is merely a beneficial secondary opportunity. The fact that this network is available for outdoor use only, is an unsecured network, and lacks 24-hour customer support can almost guarantee that the telecommunication companies in Alexandria will see no noticeable drop in customers due to the implementation of a free wireless Internet access zone in Old Town Alexandria.

New Orleans, Louisiana

In late August of 2005, Hurricane Katrina slammed ashore near the outfall of the Mississippi River as a category 3 hurricane. This storm caused considerable damage far and wide; however the city that received the most attention was New Orleans. The storm surge from the hurricane caused New Orleans' extensive levee system to fail and the waters of the Gulf of Mexico came flooding into the city streets. What followed over the course of the next week or so can only be described as chaos. Emergency services broke down, communication networks throughout the city were non-existent, and many lives were lost. During the confusion, the difficulty in getting accurate information in a timely manner caused many delays for much needed emergency supplies and services. The response time was further delayed due to lack of communication in the next few weeks. Some of this lack of communication was due to the fact that phone and power lines were down over a very substantial area. Another source for lack of communication was the fact that information coming out of New Orleans was so disjointed, that it was very difficult to put the big picture together because there was no way to organize the information and distribute it to the responsible parties quickly enough. Once the storm waters subsided and the rebuilding efforts got under

way, New Orleans officials began to investigate a means for speeding the recovery process and giving incentives for displaced residents to return to the streets of New Orleans.

On November 29, 2005, city of New Orleans officials announced plans to implement a free wireless Internet access network initially in the central business district and the French Quarter. New Orleans is in a unique situation right now. Hurricane Katrina has caused massive damage to the public and private utilities as well as the structures in the New Orleans area. Many other cities that have tried to implement a wireless Internet network have done so as a single project, not as a large-scale rebuilding effort. New Orleans has an opportunity to not only replace the infrastructure that was damaged or destroyed by Hurricane Katrina, but upgrade the infrastructure to a higher level of service. The purpose behind New Orleans' wireless Internet network is two-fold. First, the wireless network will aid in the recovery effort by giving the police, emergency services, and aid workers a communications network to utilize in order to make the rebuilding process more efficient. By using the wireless network, city officials can assess damage and log their findings while still on site. Aid workers will be able to communicate with their parent organization and order more supplies efficiently using the wireless Internet network. Police and emergency services will be able to use the wireless Internet network to access databases on current activity as well as coordinate responses to deal with any situation more efficiently than ever before. Second, the wireless network will hopefully jump start the economy in New Orleans. With free wireless access, more tourists are likely to come back to New Orleans quicker than if the network was not in place. Residents will have an incentive to come back home and try to rebuild what they lost in the storm. But the most important result from the city of New Orleans implementation of a free wireless Internet network is the fact that this kind of initiative has never before been deployed by a city trying to recover from a natural disaster. By moving forward and implementing a city-owned free wireless Internet access network, the city of New Orleans is showing that they are not only trying to replace what was lost to Hurricane Katrina, but city officials are trying to make New Orleans a better place than it was before the storm ravaged this world famous city. However, not everyone's reaction to New Orleans' announcement was positive. Just hours after the city of New Orleans made the announcement to implement a city-owned free wireless Internet network, BellSouth officials withdrew a building they had donated for the police to use as their new headquarters (Krim, 2005).

Cities all over the United States are starting to look into the opportunities available stemming from a free or low cost wireless Internet network. Obviously there are several benefits that can come out of a city taking on this kind of initiative; however, the biggest roadblock is coming from the telecommunications companies. When Philadelphia implemented their wireless Internet network, the governor had to get involved and state legislation was created just for that instance. The BellSouth building damaged by Hurricane Katrina only had a flooded basement. Since Katrina, New Orleans headquarters police forces have been scattered throughout the city staying wherever there is space, such as hotels, precinct stations, and other makeshift locations. The idea was to renovate the BellSouth building and use that as the new police headquarters building. This is just one example of the many negative responses that other telecommunication companies have had whenever a city decides it wants to implement a city-owned, free, or low-cost wireless Internet network.

Other Cities

The three previous case studies represent the most prevalent examples of city government implementations of wireless networks in the U.S.

Other cities have implemented similar wireless systems as those discussed. The opportunities and challenges fit similar trends as the three cities examined. Scottsburg, Indiana is an example of a town that was losing jobs to nearby Louisville because of its lack of affordable high-speed Internet access. The mayor addressed the issue head on by providing Wi-Fi access to the city's 6,000 residents who pay $35 a month for the city-run wireless service. Chaska, Minnesota offers a broadband network that costs its citizens $16 a month and provides for more effective communication with the police as an added benefit. St. Cloud, Florida deployed wireless Internet in order to attract more people to its downtown area. Plans are in place to expand free Internet access to the entire city rolling the costs into the general budget (DeGraff, 2005). Although the reasons and opportunities vary from city to city, the advantages are outweighing the costs as more cities decide to move in this direction.

Similar Applications

Large-scale wireless networks have been implemented by entities besides cities. One example is universities that are rolling out campus-wide wireless networks. One university in particular is the University of Richmond. The University of Richmond deployed a wireless Internet network with the intent to provide the freedom for students and professors the ability to teach and learn wherever they choose. The University of Richmond's wireless network uses the 802.11g IEEE standards, coupled with a 10 gigabyte wired network in order to supply users with fast reliable service. The network consists of 560 wireless access points and covers all the buildings on campus as well as many outdoor areas. In order to supply the bandwidth needed for an academic setting, each wireless access point is hard wired to a "data cabinet" which is then connected to the University's Internet backbone. In order to provide security to wireless users, the University of Richmond is using the Wi-Fi Protected Access (WPA) security standard. The security provided as well as ample bandwidth truly makes the University of Richmond's wireless network a success story as well as a good model for other academic institutions and small localities. For more information see http://is.richmond.edu/network/wireless.htm.

STRATEGY FRAMEWORK

When a city government decides to look at the possibility of providing wireless access to a certain area, there are some critical decisions that need to be made before pursuing any type of implementation. First and foremost, the city needs to research state laws to ensure that there is nothing on the books that would not allow a local government to provide a wireless network in direct competition with the area telecommunication companies. Case in point is the situation in Philadelphia, Pennsylvania, where Verizon lobbied the Governor of Pennsylvania to enact a law that would make it illegal for Philadelphia to provide wireless Internet access without first giving the local telecommunications company an opportunity to do the same (see the Philadelphia case study earlier in the chapter). Philadelphia was able to overcome this road block because the governor made a deal with Verizon to make Philadelphia an exception. If there are no legal barriers already in place then the municipality should make every effort to get state and federal government official's support before the telecommunication companies can lobby and change their minds. Once the decision has been made to provide wireless Internet access, there are two more key decisions the municipality must entertain that will dictate future actions. The first key decision is whether to provide wireless Internet to a small portion of the city or provide access throughout

the city limits. The second key decision is whether to provide wireless Internet access for free or for some type of usage fee.

When deciding to provide the network for free or for a nominal user fee, the municipality must weigh such factors as maintenance, upgrades, implementing new technologies, and financial strain or benefit on the local government's budget. If the network is provided for free, then there will need to be some source of income other than user fees in order to pay for the deployment, maintenance, and future upgrades for the wireless network. Some cities are able to do this by creating a non-profit organization that runs the wireless network (i.e., Wireless Philadelphia) from funds through grants, donations, and other sources. If the municipality decides to offer wireless Internet access for some type of user fee, there is a large opportunity to leverage the wireless network into a revenue generating operation. The technology is already available and inexpensive, while the installation of the network can be done quickly with sections of the network becoming operational as soon as transmitters are put into place.

The good thing about wireless technology is that it is immediately operational in its effective area as soon as the transmitter is installed. Because of this fact, the municipality does not have to decide on whether to roll out full-scale city-wide access or small pockets of wireless Internet access at the onset of the initiative. Several small wireless Internet access points can be set up throughout the city initially in public spaces such as parks and libraries and then expanded as demand increases and funds become available. Even if the plan is to implement wireless Internet access city wide, this can be done quickly.

Opportunities

The ability to access the Internet anywhere in the city carries great opportunities that cities are eager to tap into. Wireless access city-wide can benefit police and emergency service workers (Malykhina, 2005). Police can use the wireless access to hook up to databases that store information on known criminals, sex offenders, and get up-to-date information on traffic situations as well as other cruiser locations. Firefighters and other emergency personnel can use wireless access to retrieve patient information from local hospitals to better serve their patients upon arrival at the location. Using the wireless network to enhance emergency services is a benefit of the wireless network, however there still needs to be a backup plan in the case of a disaster. The wireless network depends on a power source in order to be operative. In the case of a disaster, power may not be available and therefore an alternate means of communication needs to be readily available. Cities providing wireless access can use this as an incentive for businesses to come back to the downtown areas and revitalize areas that have seen an egress of residents in the past few decades. Another opportunity would be for tourist attractions to make their Web sites interactive for tourists on the property. Since municipalities already own the street lamps and all the government office buildings in a downtown area, they will be able to avoid the leasing fees that most telecommunication companies have to pay in order to put up transmitters and receivers (Malykhina, 2005). Local businesses within the wireless zone will be able to use this feature in their marketing plan in order to attract more customers or get customers to stay longer. Along with the further development of wireless technology, more and more cities will be looking into this service as options become faster, cheaper, and more reliable. However, even with all these opportunities, there are still several roadblocks that need to be overcome in order to consider wireless access a viable project to undertake.

Challenges

Whenever a governmental entity takes on a new project of any scale there are several challenges

it must overcome. The major challenge for municipalities looking into the prospect of providing free or low-cost wireless Internet access is the reaction of the telecommunication companies. Some telecom companies fear that with more and more cities looking into the ability to provide wireless access, their customers will continue to dwindle. Others, however, feel that government-provided wireless access will be of such quality that a large population will continue to use the telecom's services (Reardon, 2005). Cities providing wireless access to its residents are faced with the daunting task of building this network from the ground up. Starting from scratch requires solutions to such problems as:

- Level of customer support to provide;
- Maintenance of the system and who will provide this service;
- How to pay for the system (bill residents, bill businesses, increase local taxes, get a major corporation to sponsor, or some combination of these options);
- Implementation delays of needed upgrades or initial roll-out of the system due to the cumbersome process of city government operations;
- Interference with neighboring and/or pre-existing networks;
- Providing the high quality the customers of telecommunication companies have come to expect while still offering the service for free or at a low cost; and
- Having enough capacity once the wireless signal reaches the router with the actual wired connections.

Each municipality will deal with these challenges in different ways, but these are the issues that will play a big role in the decision of whether to move forward with implementation.

Another challenge that each municipality must overcome is network security. Wireless information security is vital to the continued use of wireless technology. If data transferred wirelessly is not protected properly, then the data can be stolen or corrupted. Knowing the intended uses of your wireless network will allow you to map out a proper plan to provide adequate security. The level of security should depend on the intended use of the network. For municipal-owned networks, the level of security provided is a function of costs. If the network is intended to replace personal, private networks, then the municipality should be prepared to provide adequate security to allow sensitive data to be transmitted as well as financial transactions to be performed. If the network is intended to be a supplemental network, then users should be made aware of the intended level of security to be provided. In order to provide security, the municipality must first determine the intentions of the network.

REVENUE GENERATION

One of the big issues in this debate is whether to provide Internet access as a free service to everyone or as a means to generate revenue. This decision needs to be made at the onset of the project. A free service may attract residents and businesses while additional revenue may be necessary to provide value-added services. If the municipality decides it wants to use the wireless network to generate revenue, then the equation is fairly simple. The installation cost is low when divided out on a per capita basis. Therefore, the city will be able to charge an equally low fee in order to recoup its costs, cover the network's overhead, and generate revenue at some point in the future. If the municipality wants to provide the network for free to attract business then the situation becomes more complicated. Many cities that are providing or getting ready to provide free wireless access have found some companies to sponsor the initiative. Companies such as Cisco Systems, Dell, IBM, Microsoft, and SAP have done this (Malykhina, 2005). Once the system is

in place, these cities need to find a way to fund the maintenance, customer support, and other overhead for the network. These miscellaneous expenses can be put into the budget in a variety of places and funded by the increased (hopefully) income from sales and other taxes that the wireless network will bring in by providing a more appealing environment for certain types of shoppers, businesses, and tech savvy residents.

FUTURE TRENDS

Wireless technologies will continue to evolve offering more alternatives and opportunities to cities and private companies. One of these trends is open mesh network technology. Mesh technology allows nodes in the network to send data over multiple routes to multiple neighbors. This results in redundant connections with many paths for transport of data. Data that is sent directly to a neighbor node in the LAN without accessing the Internet travels at faster speeds. Mesh technology also allows for decentralization. There is no need for a central server or central administration for the network. The best path is chosen by each node based on signal strength and shortest distance. This technology is suitable for larger outdoors LANs (Watkins, 2005).

Technological advancements will continue to impact the decision-making process and in some cases change the course of events in unexpected ways. It is essential to keep abreast of technological trends and remain adaptable as new advances provide opportunities that were not before possible.

An emerging technology that will affect future wireless networks is the advancement of Wi-Max to construct wireless MANs. The ability of Wi-Max to operate in two separate modes makes it a perfect candidate for future wireless uses. As mentioned before, Wi-Max can operate through line of sight from tower to tower, as well as an access point in much the same way as Wi-Fi, but with a much larger broadcast area. These two modes of operation will enable Wi-Max to be a viable alternative. The fact that it can carry more data than a Wi-Fi signal will alleviate the bandwidth problem, and become a viable alternative to expensive T1 lines. An indication that Wi-Max is a technology for the future is the fact that it has standards for both fixed wireless (desktop and laptop computers), as well as mobile devices (cell phones, PDAs, and blackberries).

As businesses continue to become more "mobile" in their operations and work environments, these trends will flow into homes and impact personal expectations. As the world becomes more connected and mobile, expectations for ordinary citizens and demand for mobile Internet access will increase. Cities who are already leaders in this arena will likely reap the benefits of being early movers in the midst of complex technological advancements. Decisions made today will influence future opportunities and direction regardless of the outcome of this debate.

CONCLUSIONS

Given the emerging trends as discussed in this chapter, it has become essential that city governments develop a strategy for wireless network deployment. The issues should be addressed head on and discussed openly. The decision of whether to move or not move in the direction of implementing a city-run wireless network is not the key issue. Competition today dictates that city governments be pro-active in this arena and at least debate the issues and determine a plan of action. It is no longer acceptable to be unaware of these trends.

The mobile wireless Internet revolution is only in its infancy. Much experimentation and trial and error is underway and much remains to be determined in terms of long-term direction and outcomes. Decisions made today that will impact access to the Internet by all sectors of society

will have lasting and important implications not only for businesses and governments but for all citizens as participants in the evolving global digital future.

REFERENCES

Chen, M. (2005). *Philly to defy Telecom giants, set up public wireless network*. Newstandardnews.net. Retrieved December 10, 2005, from www.newstandardnews.net/content/index.cfm/items/1658

DeGraff K. (2005). Community wireless networks: Why not? *Mobile Government,* June, 7-9.

Effors, S. (2005). Community wireless networks: Why? *Mobile Government,* June, 19-21.

Fifer, C. (2005). *City launches "Wireless Alexandria"*. Alexandriava.gov. Retrieved December 15, 2005, from www.alexandriava.gov/fyi_alexandria/sept_05/fyi_alexandria4.html

Gowen, A., & MacMillan, R. (2005). The Web is in the air. *Washington Post*. Retrieved December 15, 2005, from www.washingtonpost.com/wp-dyn/content/article/2005/06/09/AR2005060901770_pf.html

Krim, J. (2005). Angry BellSouth withdrew donation, New Orleans says. *Washington Post*. Retrieved December 21, 2005, from www.washingtonpost.com/wp-dyn/content/article/2005/12/02/AR2005120201853_pf.html

Malykhina, E. (2005). Square off over Wi-Fi in the town square. *Information Week, 26*(September), 47-53.

Reardon, M. (2005). *The citywide Wi-Fi reality check*. News.com. Retrieved December 10, 2005, from http://news.com.com/The+citywide+Wi-Fi+reality+check/2100-7351_3-5722150.html?tag=st.num

Senn, J. A. (2004). *Information technology, principals, practices, opportunities*. Upper Saddle River, NJ: Pearson Education, Inc.

Walsh, T. (2005). When trash talks: Embracing wireless technology. *Government Computing News*. Retrieved December 15, 2005, from www.gcn.com/vol1 no1/daily-updates/35466-1.html

Watkins, S. (2005). Open mesh. *Mobile Government,* June, 27-28.

Wireless Philadelphia. (2005). *Wireless Philadelphia enters final contract discussions with Earthlink*. Wirelessphiladelphia.org. Retrieved December 10, 2005, from www.wirelessphiladelphia.org

This work was previously published in Mobile Government: An Emerging Direction in E-Government, edited by I. Kushchu, pp. 357-374, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.19
The Prospects of Mobile Government in Jordan:
An Evaluation of Different Delivery Platforms

Ala M. Abu-Samaha
Amman University, Jordan

Yara Abdel Samad
Ministry of Information & Communication Technologies, Jordan

ABSTRACT

This chapter aims to assess the viability of mobile governmental services in Jordan as a precursor to embracing mobile government as a complementing medium of communication. Reflecting on Jordan's experience with electronic governmental services, it is evident to say that the first wave of electronic governmental services was delivered through the Web as the sole communication channel. Despite the success of a number of governmental entities to utilise such a communication channel, the penetration of the Internet in the Jordanian society is very low which dampens such limited cases of success. Currently, the e-government initiative is considering mobile phones for the future waves of its electronic service delivery on a multi-channel platform. This chapter articulates the concerns and issues surrounding the viability of mobile government in terms of availability of bandwidth and reach. The chapter will provide a number of statistics and other qualitative reviews concerning previous experiences in the Jordanian electronic government initiative.

INTRODUCTION

It is evident to say that the past few years have witnessed the emergence of many communication channels and media based on information and communication technologies and networks as an alternative to the more "traditional" face-to-face, phone and fax modes of business conduct. The most prominent of these channels and communication media were the Internet and the

Mobile/Cellular network. The Internet started as a distributed system for information and knowledge interchange, and evolved to become the medium for several crucial applications, such as e-commerce, e-banking, e-government, e-business, and so forth, that share business processes and connect different organisations. Abu-Samaha (2005) points out that contents delivered via such channels, that is, digital libraries, electronic markets, virtual stores, electronic business, and electronic commerce, are primarily perceived as electronic replacements to the traditional modes of conducting trade, business, and transactions; that is, converting such transactions into a stream of electrons representing data, products, services, and payments. Such move from physical, face-to-face to electronic modes and channels is expected to provide organisations with many benefits; some of these include: expanding market reach (global availability, small compete with large organisations), generating visibility at very low cost, strengthening business relationships (EDI and B2B using XML), offering new services online; reducing cost (through paperless inter- and intra-business activities; that is, exchanging e-mails to support conducting daily activities within and across organizational boundaries), shortening time to market and expediting time to respond to changing market demands, improving customer loyalty, real-time training and conferencing; personalization of goods; enabling employees to carry out tasks internally and externally, reducing cost of creating, processing, distributing, storing, and retrieving paper-based activities, reducing inventories and overheads, saving time to look for resources; obtain useful expertise from the Internet, savings in communication costs; promoting current and future products and/or services, and disseminating information (Abu-Samaha, 2005; Amor, 2002, p. 17; Lawrence, Corbitt, Tidwell, Fisher, & Lawrence, 1998; Simpson & Swatman, 1998; Turban, Kuy Lee, & King, 1999, p. 15). Conducting business electronically is believed by many professionals and academics to be the "most promising" innovation of the future. *Turban and Potter (2001)* explain that a highly sought business strategy is "a multifaceted concept, ranging from electronic transfer of funds between buyers and suppliers, to Internet-based marketing, to intranet- and extranet-based information networks for both inter- and intra-organizational support".

Electronic government can be perceived as an implementation of electronic business/commerce (EB/EC) within governmental domains from an operational point of view. Though e-government concept possesses a number of distinguishing features when compared to electronic business, such as strategic and operational reasons for investment, expected benefits and outcomes, and targeted audience (citizens, businesses, and other governmental and non-governmental entities). Ginige and Murugesan (2001) propose a classification of Web-based applications, where these applications are grouped into seven categories (Table 1).

Based on Table 1, it can be said that electronic governmental systems/services can provide a variety of services, information and application to its beneficiaries—that is, information, interaction, transaction, collaboration, and Web portals. As well, it is evident to note that electronic government applications and services are one of those applications/services that do not fit exactly into a specific category/class; on the contrary, they seem to span many different categories/classes.

Reflecting on Jordan's experience with electronic governmental services, it is evident to say that the first wave of electronic governmental services was delivered through the Web as the sole communication channel. Despite the success of a number of governmental entities to utilise such a communication channel, the penetration of the Internet in the Jordanian society is very low which dampens such experiences of success. Currently, the e-government initiative is considering mobile phones for the future waves of its electronic service delivery on a multi-channel platform using mobile/cellular phones/network

Table 1. Categories of Web-based applications (Ginige & Murugesan, 2001)

Category	Examples
Informational	Online newspapers, product catalogues, newsletters, service manuals, online classifieds, online electronic books.
Interactive (user-provided information or customised access)	Registration forms, customised information presentation, online games.
Transactional	Electronic shopping, ordering goods and services, online banking.
Collaborative work environments	Distributed authoring systems, collaborative design environment tools.
Online communities, marketplaces	Chat groups, recommender systems that recommend products or services, online marketplaces, online auctions.
Web portals	Electronic shopping malls, online intermediaries.

as a complementary medium of communication and delivery.

This chapter will articulate the concerns and issues surrounding the viability of mobile government in terms of availability of bandwidth and reach. The chapter will provide a number of statistics and other qualitative reviews concerning previous experiences in the Jordanian electronic government initiative to establish different future scenarios in terms of viability and willingness to incorporate mobile government. The chapter is structured into a number of sections—each with it own domain of interest. The second section introduces Jordan's Strategic Initiative REACH (**R**egulatory Framework, **E**nabling Environment (Infrastructure), **A**dvancement Programs, **C**apital & Finance, and **H**uman Resource Development) and a synopsis of the Jordanian telecommunication industry/market. The third section provides a detailed description of the Jordanian electronic government initiative including its aims, constituent ingredients, and stakeholders. The fourth section provides a number of quantitative and qualitative assessments of the first wave of Jordan's electronic governmental services, while the final section provides the emergence of mobile phones and networks as an alternative or a supplementary medium of communication to the established electronic services. The chapter provides a number of conclusions and recommendations for the future waves of the e-government initiative.

BACKGROUND TO REACH AND TRC

Background to REACH

Jordan's strategic IT initiative (REACH) was intended to lay out a clear plan of action to bolster

the country's nascent IT sector and maximize its ability to compete in local, regional, and global markets. Jordan's strategic initiative came to life as a response from the local IT industry to His Majesty King Abdullah the 2nd's directive to the private sector to formulate a realistic strategy and action plan that would launch Jordan's Information Technology sector. The result was a comprehensive IT plan that was called REACH. REACH 1.0 was available to the public in March 2000; REACH 2.0 followed this in January 2001 and REACH 3.0 in 2003. REACH stands for **R**egulatory framework, **E**nabling environment infrastructure, **A**dvancement of national IT programmes, **C**apital and finance, and **H**uman resource development. These five areas of concern are perceived to be the most vital and important to the success of such an initiative (REACH 1.0, 2000).

The long-term goal of the Jordanian strategic initiative is to position Jordan favourably within the knowledge economy. The Jordanian strategic initiative is foreseen to be led by the private sector in partnership with the government. The government role is perceived to be of a supporting nature in legal and national senses. On the short to medium terms, the Jordanian strategic initiative (REACH), aimed by the end of 2004 to create 30,000 IT and IT-related jobs, generate a revenue of $550 million per year in export, and attract $150 million in foreign direct investment (REACH 1.0, 2000). Twenty thousand of these jobs were expected to be directly related to IT, ranging from software/system development to IT consultancy. And, 10,000 jobs will be indirectly related to IT as supporting jobs, ranging from lawyers to intellectual right property experts (REACH 1.0, 2000). The figures of INT@J (The Information Technology Association of Jordan) for the year 2001 showed that at least 10,000 Jordanians are currently recruited in the IT services and software sector. The reported annual revenue from local and foreign sales in 2001 was estimated at $106,586 million. The total size of the industry is estimated to be $176,959 million (REACH 2.0, 2001). INT@J's estimates showed that this number would rise to $270 million by 2002 in both domestic and export revenue. Six areas of concern are identified by the strategic IT initiative to realise the projected figures and to aid in achieving the stated objectives and aims. Here follows a summary of the most significant facts of each area of concern (REACH 1.0, 2000; REACH 2.0, 2001):

1. IT Industry Development
2. Regulatory Framework Strengthening
3. Human Resource Development
4. Government Support
5. Capital and Financing
6. Infrastructure Improvement

The strategic initiative has called upon the Government of Jordan to build on the recent successful liberalization and reform efforts to establish Jordan as the most competitive country in the Middle East in terms of IT policies and regulatory systems. The IT plan has identified a number of actions, these being: to reduce indirect taxes on all IT-related products; to streamline customs clearances procedure; to continue and formalise policy of no censorship of IT media and products; to adopt more competitive taxation policies; to enhance access to Investment Promotion Incentives (IPI); to remove constraints to employee's stock ownership plans; to sign information technology and customs valuation regulations of the World Trade Organisation; to develop Electronic Commerce legislation; to enforce intellectual property rights; and to amend labor law (REACH 1.0, 2000). The REACH initiative identified 25 laws, bills, and articles of urgent need for amendment or ratification by the Upper and Lower Houses of Representatives of the Jordanian House of Parliament. The Government of Jordan has succeeded in getting 11 of these 25 pieces of legislation to be ratified amongst of which labor law, electronic commerce legislation, electronic signature legislation, private corporation, and stock option law.

According to these laws, computer print-outs and e-mails electronically signed can be held as legal evidence in a court of justice.

The Government of Jordan is the largest employer and the largest consumer of IT products in the Jordanian local market. The action plan is designed to focus government support efforts in appropriate areas that will stimulate private sector IT development while improving the delivery of government services. The plan identifies the following actions: to establish a high-level body for the Jordanian software and IT services industry; to initiate electronic government initiatives; to focus export and investment promotion efforts to the IT sector; and to develop and implement an IT incubator program (REACH 1.0, 2000). The high price of both hardware and telephone calls and the low quality of telecommunication services are perceived to be one of the major reasons for the lack of IT proliferation in both households and businesses. The IT plan had identified a number of actions to establish Jordan as a regional leader, these being: to provide high-speed lines to software developers and IT service companies on a priority basis; to provide competitive pricing on high-speed telecommunication connections for software developers and other IT service firms; and to plan and develop information technology park (REACH 1.0, 2000).

Infrastructure Liberalization and the Establishment of TRC

National infrastructure in the form of computers and public and private networks plays a pivotal role in the realization of any strategic IT initiative. Tapscott (1996) compares the significance of computers and networks of the digital economy to steel, automobiles, and roads of the industrial economy. Tapscott (1996, p. 15) indicates that "Just as the highway system and electrical power grid were the infrastructure for the industrial economy, so our information networks will be the highways for the new economy. Without a state-of-the-art electronic infrastructure throughout organizations, no country can succeed". In terms of national infrastructure, Jordan has one telecommunication company, fully owned by the private sector. As well as four mobile network operators mostly owned by the private sector and an extended number of Internet Service Providers (ISPs) all of which are owned and operated by the private sector. The Jordanian government to meet the increasing demand for high quality, high bandwidth Internet connectivity has issued 21 more ISP licenses (Arab Advisors Press Room, 2005; Telecommunications Regulatory Commission, 2005a).

The de-regulation of the telecommunication market in Jordan started as early as 1992, which led to the establishment of the Telecommunications Regulatory Commission (TRC) in the year 1995 (Telecommunications Regulatory Commission, 2005b). The vision of TRC is "A telecommunications environment that is competitive, advanced, regulated and available to all" (Telecommunications Regulatory Commission, 2005a). While, the mission statement of TRC is "To ensure the availability of advanced and high quality Information and Communications Technology (ICT) services to all users at just, reasonable, and affordable prices by working with all stakeholders in an independent, open, and transparent manner to create a regulatory environment that promotes fairness, competition, and investment, thus assuring fulfilment of the Kingdom's long-term ICT needs" (Telecommunications Regulatory Commission, 2005a).

TRC's scope includes the following services/products: Public Switched Telephony Network (PSTN), Public Mobile Telephony (Cellular), Public Mobile Telecommunications, Radio Trunking, Paging, Data Communications Services, Global Mobile Personal Communications by Satellite (GMPCS), and Pre-paid Cards Services (Telecommunications Regulatory Commission, 2005a).

Regarding Public Switched Telephony Network (PSTN), TRC regulates the service providers (Jordan Telecom (JT)) who operate and manage a fixed public telecom network that provide local, national, and international fixed telephony services and leases lines and BATELCO Jordan who were granted a class license in May 2005 to provide PSTN services in the near future (Telecommunications Regulatory Commission, 2005a). Regarding Public Mobile Telephony (Cellular), Jordan Mobile Telephone Services (Fastlink), which is partly owned by Motorola Co., has been providing this service since 1995 through a countrywide GSM900 cellular network. Moreover, JT has been granted a license to provide this service through an affiliate (MobileCom) to compete with Fastlink providing such service since September 15, 2000. The two companies had dual exclusivity (duopoly) for providing GSM900 public mobile telephony service until the end of 2003 when a license to operate public mobile telecommunications service was granted to (Umniah) on the 9th of August 2004 (Telecommunications Regulatory Commission, 2005a).

The New Generation Telecommunication Company (Xpress) is the sole provider of Radio Trunking service where a license was granted from Telecommunication Regulatory Commission (TRC) on April 6, 2003, and the company launched its services commercially in June 2004. This service provided by Motorola's iDEN technology which allows push-to-talk radio access in addition to full access to mobile telephony services including short messaging services (SMS) and data services. In addition to the main services, operator services (24 hours) numeric, e-mail notification, Internet paging, and voice-mail service. The number of Radio Trunking subscribers reached 7,300 subscribers by mid-year 2004 (Telecommunications Regulatory Commission, 2005a).

The provision of the data communication service is fully liberalized in Jordan. The data communication service has developed substantially in Jordan over the last few years. The service is licensed through a class license of a 10-year duration. Prices are completely liberalized while quality and standards of service are underscored in the license. There are 21 data communications licensees as of June 30, 2004. The total number of subscribers increased from 3,146 in 1996 to 91,566 in 2003. The subscribers' penetration rate for the year 2003 equals 1.7% while users penetration is 8.1%. The estimated number of users until the end of 2003 is 444,000 users (Telecommunications Regulatory Commission, 2005a).

Regarding Global Mobile Personal Communications by Satellite (GMPCS), this service is provided by the Thuraya Satellite Telecommunications Company (Thuraya), as a system operator, and by the Middle East Communications Co., as a service provider for Thuraya. The provision of the service, whether as system operator or service provider, is fully liberalized through class licenses. The number of subscribers is 2,300 until the end of 2003 (Telecommunications Regulatory Commission, 2005a).

Regarding Pre-paid Cards Services, this service started in Jordan in 2001. There are four providers for this service: Jomotel, TeleCard, Weinak, and Swiftel. The number of sold cards reached (137,597) from all different categories by mid-year 2004, with a growth rate of (2%) when compared with the first quarter in the same year, where the number of sold cards reached (134,849) (Telecommunications Regulatory Commission, 2005a). Most network accesses are provided via dial-up connectivity, while most recently ISPs have introduced ISDN (Integrated Services Digital Network) and ASDL (Asymmetric Digital Subscriber Line) access to the Internet, which is expected to increase bandwidth, and fastens access to the net. On the other hand, mobile network operators have introduced Internet services over mobile phones (Arab Advisors Press Room, 2005; Telecommunications Regulatory Commission, 2005a).

Despite limited affordability of personal computers and network access, Int@j estimated that Jordan had 42,000 Internet users, 50% of whom surfed the Internet via Internet Cafes (REACH 1.0, 2000), while Abu-Ghazaleh & Co. Consulting (2005) estimated the number of Internet users in Jordan in the year 2004 at 111, 054 users up from 62,242 Internet users in 2002—a penetration rate of almost 2% of the population. Sixty-seven percent of who used the Internet via pre-paid Internet cards while only 1% used the Internet via a leased line. This can be attributed to "relatively expensive [Broadband connection]…why dial-up connection is popular among Citizens but still it has its limitations: high cost, limited time and low speed" (MoICT, 2005a).

E-GOVERNMENT INITIATIVE

Electronic government can be defined in many different forms. The adopted definition of e-government in the Jordanian initiative is "the ability to submit [governmental] transactions on-line and make payments electronically where they are required" (MoICT, 2000). Such a definition indicates the need to coordinate information from a variety of different governmental sources presented in an easily navigable format (MoICT, 2000). The use of electronic services and channels to provide governmental transactions has been considered to be a powerful tool for the improvement of internal managerial efficiency and the quality of public service delivery to citizens as well as enhancement of public participation (Moon, 2003). Lawrence et al. (1998, p. 8) show that "the Internet and intranets give businesses the opportunities to improve their internal business processes and customer interfaces to create a sustainable competitive advantage". The eEurope 2002 Action Plan indicates that "eGovernment could transform old public organisation and provide faster, more responsive services. It can increase efficiency, cut costs, increase transparency and speed up standard administrative processes for citizens and business" (eEurope, 2000). The overall objectives of e-government can be summarised as: improve the quality of governmental service delivery, increase transparency, improve responsiveness, save time, money, and other resources, and create a positive spin offs (MoICT, 2000). On the other hand, mobile government can be seen as a supporter of one-stop government services, where one-stop government refers to "the integration of public services from a citizen's- or customer of public services-point of view. Online one-stop government allows citizens to have 24-hour access to public services from their home or even on the move" (Tambouris, 2001).

It is highly believed that information and communication technologies (ICT) can provide a more pro-active action to handle the causes of the chronic decline of public trust in governments (Moon, 2003). Moon (2003) indicates that "IT appears to offer a useful opportunity to government to enhance public trust and citizen satisfaction by improving procedural transparency, cost-efficiency, effectiveness, and policy participation. IT provides positive opportunities as well as many challenges to governments and the public". A conclusion emphasised by Vélez-Rivera, Díaz, Fernandez-Sein, Rodríguez-Martínez, Núñez, & Rivera-Vega (2005), "Electronic government systems have an unprecedented potential to improve the responsiveness of governments to the needs of the people that they are designed to serve. To this day, this potential is barely beginning to be exploited".

The electronic government of Jordan was conceived and established as a national program in the year 2000 (Jordan eGovernment Initiative, 2003). Jordan's e-government vision is to be a major contributor to Jordan's economic and social development by providing access to government e-services and information for everyone in the kingdom irrespective of location, economic status, IT ability, and education (MoICT, 2000).

A number of challenges to e-government in Jordan has been encountered; these can be summarised as: low level of Internet penetration (1.9% Internet users of the Jordanian population), infrastructure constraints (high cost and inadequate), digital divide, privacy and security concerns, limited IT skills, limited public sector reform efforts, lack of an enabling legal framework, and lack of awareness (MoICT, 2000). To overcome those challenges, eGovernment in Jordan was based on five major building blocks: electronic services, the technology infrastructure to enable the delivery of electronic services, the regulatory framework to provide the legal coverage and acceptance of electronic services, the educational reform and skill development to ensure effective and efficient services, and the organizational reform to ensure inter-government cooperation and coordination to required by e-government (MoICT, 2000).

The Ministry of Information and Communications Technology (MoICT) was assigned to take the lead role in implementing the e-government pProgram. MoICT's mission towards e-government is to provide support and capability to coordinate the management, implementation, interoperability and benefits of the National e-Government Initiative for the Government of Jordan. The program objectives can be summarised as follows (Jordan eGovernment Initiative, 2003):

- Develop and support the e-government strategy to be implemented across government entities.
- Participate in the planning and coordination of a sustainable national portfolio of e-government initiatives.
- Maintain technological integration and interoperability of e-government initiatives, and encourage the re-usability of application components, to achieve consistency among ministries/departments for technical solutions.
- Plan and implement security policies and a secure network environment for e-government initiatives.
- Promote and monitor a systematic method of planning, developing, and implementing e-government initiatives.
- Promote and monitor organizational transformation (change management) at the ministry/department/organizational level necessary to establish effective e-government.
- Educate Government of Jordan employees and transfer the knowledge in order to have consistency in the level of skills and competencies among the GoJ employees.
- Establish a common understanding of e-government program across government and to the public.
- Deliver successfully e-government initiatives and projects that are managed by dedicated project managers.
- Provide analysis and information on the status of e-government initiatives and projects to sponsors and major stakeholders (to maintain buy-in).

Over the past five years, the e-government program has been involved in developing and implementing major e-government initiatives. This included scanning and analysing the business operations at each of the first wave government departments (Income Tax, Sales Tax, Drivers and Vehicle Licensing, Lands and Survey, and Borders and Residency). Subsequently, drivers and vehicle licensing, income tax e-services, and land registry e-service were launched in January 2006. Furthermore, a Secure Government Network (SGN) providing connectivity for Internet and e-mail services to 18 government entities was implemented and hosted at an Operation Centre established by the e-government program. An e-Government Contact Centre was also established. Currently, the contact centre provides technical

support to SGN administrators while it is envisioned to provide business and technical support to different categories of e-government services' users. A comprehensive information security roadmap for the Government of Jordan was produced and the e-Government Information Portal was launched in November 2006. The bilingual (Arabic/English) Information Portal will provide a single official access point on the Internet to government information required by different categories of users, including citizens, businesses, and government entities and employees.

GOVERNMENT SERVICES DELIVERY CHANNELS ASSESSMENT (2000-2005)

This section presents the findings of a number of surveys conducted by the MoICT to evaluate the e-government initiative as Jordanian governmental institutions have invested both in technology and personnel to provide electronic governmental services to both citizens and businesses. This post-implementation review will aid in conceiving future scenarios for both Jordan's citizens and institutions to embrace alternative communication medium, that is, to embracing mobile government as a medium of communication.

The survey of national outlook and expectations about the e-government program plays a critical role within the e-government initiative as it is an important source of information about the priorities of the main stakeholders (key national figures, citizens, companies, and government employees) and their expectations related to strategic direction of the program. At the same time, the survey is also concerned with issues directly related to the expectations of stakeholders of about how the selection, implementation, and operation of new e-services will look like, which allows to set a stage for set a stage for the evaluation of access and delivery channels of electronic governmental services.

The survey included 21 interviews with key national figures, 142 focus groups with government employees, 53 focus groups with companies, 48 focus groups with citizens, 395 questionnaire surveys of government employees, 254 questionnaire surveys of companies, and 409 questionnaire surveys of citizens. Arthur Business Consulting carried out the assessment under the supervision of the e-government program team. The findings of these interviews, focus groups, and questionnaires can be summarized as follows:

- Implementation of a new set of e-services will face many issues of which resistance to change, technical weaknesses, and budgetary constraints are perceived as the largest threats and obstacles. The new e-services are expected to bring benefits like time-savings, simplification of processes, and fewer number of mistakes.
- Internet and phone should be considered as primary channels for implementation of new e-services.
- Effective project and change management as well as communication are singled out as the most important success factors.
- Shared services are perceived as good idea that can however meet with many practical implementation problems. (MoICT, 2005b)

As mentioned earlier, the MoICT survey included a questionnaire survey of 395 government employees, 254 companies' representatives, and 409 citizens. This questionnaire included a number of companies and governmental employees covering a number of industries and governmental entities. The demographics of these participants are shown in Figures 1 and 2. Where Figure 1 shows the demographics of companies representatives split into two groups small/medium and large organizations. Where any organization with less than 250 employees is considered as a small/medium enterprise (SME) and any organization with

Figure 1. Companies questionnaire survey (MoICT, 2005b)

Industry	Grand total
Agriculture, Hunting, Forestry and Fishing	**2**
Manufacturing	**36**
Construction	**7**
Wholesale & Retail Trade, Restaurants and Hotels	**60**
Transport, Storage & Communications	**36**
Finance, Insurance, Real Estate and Business Services	**65**
Public, Community, Social and Personal Services	**48**
Grand total	**254**

Figure 2. Citizen questionnaire survey (MoICT, 2005b)

	Age				Grand total
Education	**15-29**	**30-44**	**45-59**	**60+**	
Primary or lower	13	19	11	6	**49**
Secondary or intermediate diploma	46	99	21	15	**181**
Bachelor or above	51	86	24	18	**179**
Grand total	**110**	**204**	**56**	**39**	**409**

greater then 250 employees is considered a large organisation. While Figure 2 shows the demographics of governmental employees grouped in terms of education (primary or lower, secondary or intermediate diploma and bachelor or above) and location (central, north, and south).

The assessment of the 21 key national figures interviews shows agreement on the importance of the e-government initiative. In terms of communication channels, the interviewees were unanimous that the Internet should be selected as the primary contact channel of both citizens and businesses in terms of governmental relationship/contact whether via personal computers, kiosk, or local school computer center. However, a need to maintain traditional methods of contact should be supplemented by phone (mobile and fixed) while the more traditional channels such as fax or traditional post were rather deemed obsolete and not worth investing in. Hence most interviewees

stressed that the traditional methods of contact with the government must be kept as an alternative to indirect channels (MoICT, 2005b).

Regarding the focus groups, there is a feeling that the e-government program will bring numerous advantages, like improving the quality of services provided by the government. Furthermore, the Internet was ranked as the most preferable channel for e-services while telephone comes second regarding number of users and frequency of usage (MoICT, 2005b).

Regarding the questionnaire, traditional face-to-face contact remains the most popular way of contacting the government but alternative solutions are already in use. Eighty-two percent of companies' representatives and 87% of citizens contact the government in a traditional way (face-to-face). Thirty-nine percent of respondents used the Internet, 55% used the telephone or post, more than 60% sent documents via fax. Frequency of usage of the alternative channels among citizens is much lower, usage of the Internet accounts for 14% only, and the most popular channel—voice—has been used by 36% of respondents (MoICT, 2005b). Figure 3 shows the distribution of preferred communication method where the first bar refers to business while the second bar refers to citizens.

On the other hand, the Internet is the most preferred by respondents for getting and providing information. Figure 4 shows the distribution of

Figure 3. Preferred way to contacting the governmental agency (MoICT, 2005b)

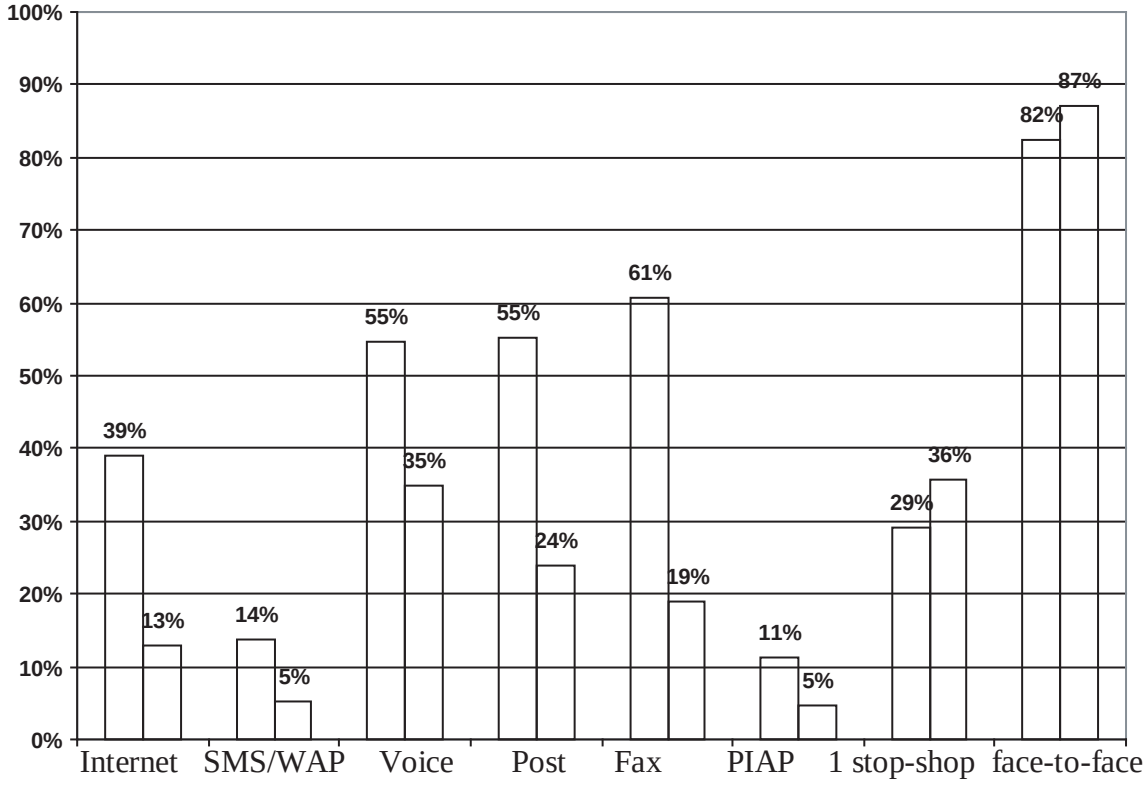
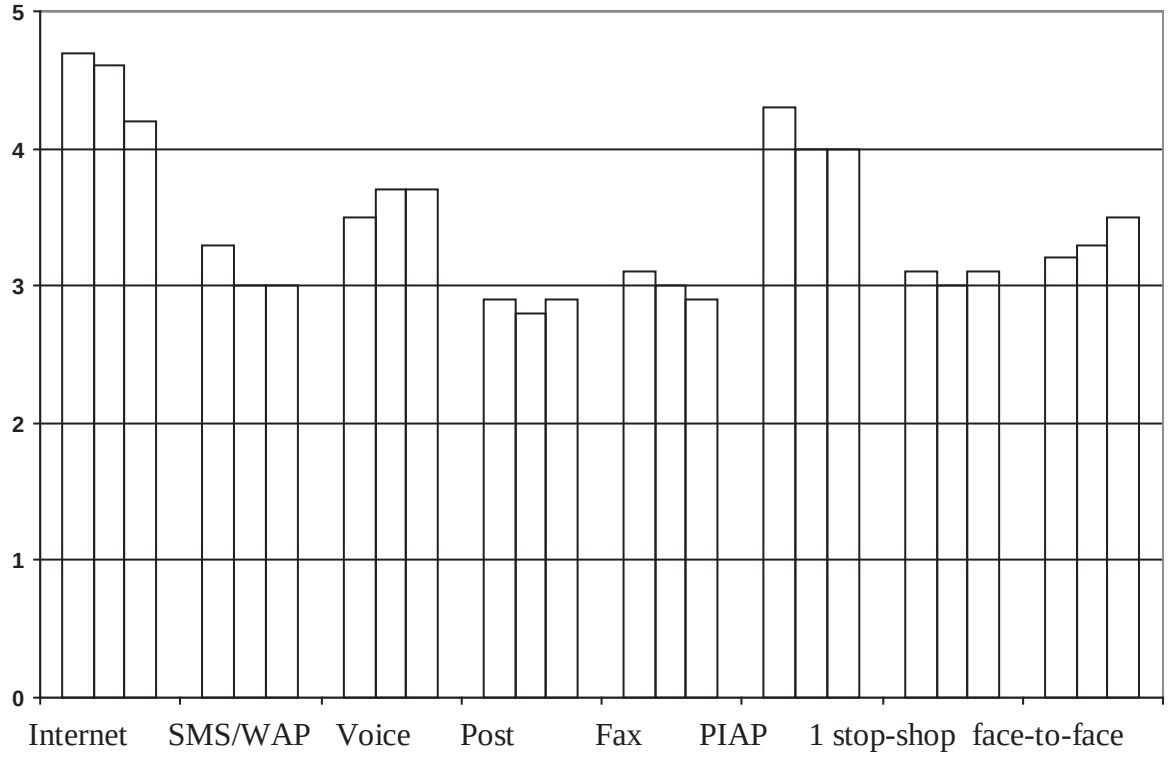


Figure 4. Preferred information distribution channel (MoICT, 2005b)



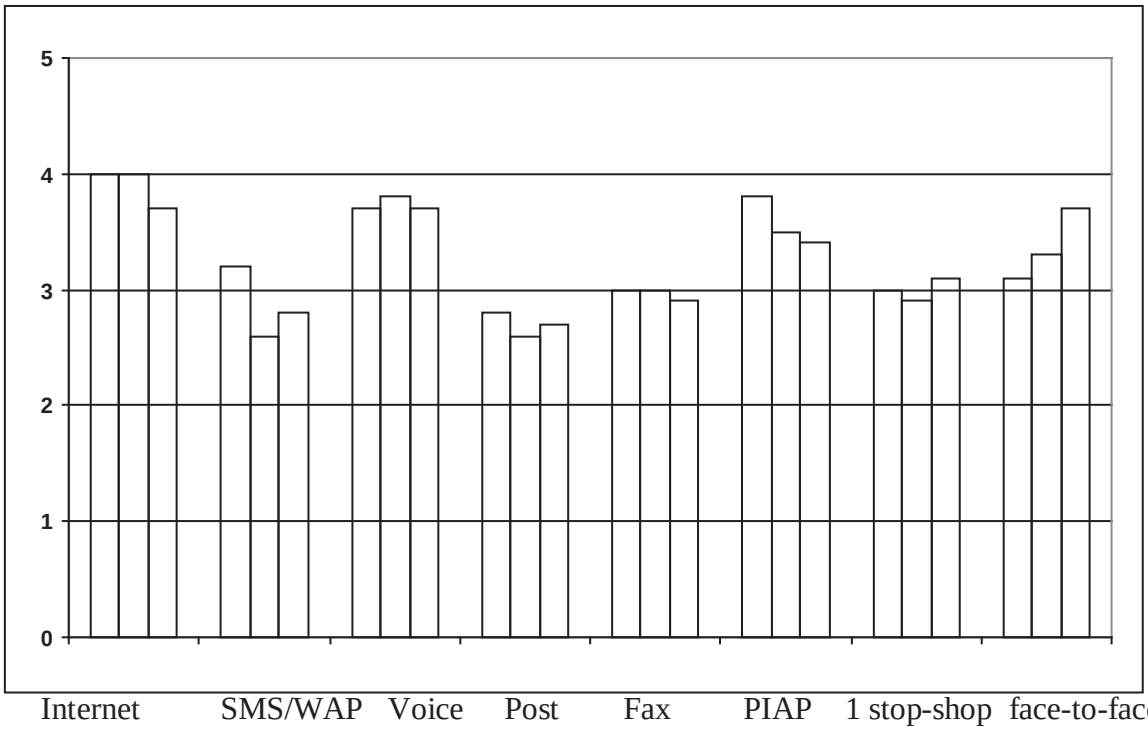
preferred information distribution channel where the first bar refers to governmental entities, the second bar refers to business, and the third bar refers to citizens.

The Internet got the highest rating from all groups (government employees, companies, and citizens) in terms of the most preferred interactive channel, while the telephone got a high evaluation from all groups of respondents. On the other hand, citizens strongly prefer traditional face-to-face contact for interactive communication with service providers. On the other hand, the Internet and the telephone are widely accepted channels for interactive communication between services' providers and services recipients. Figure 5 shows the distribution of a preferred information distribution channel where the first bar refers to governmental entities, the second bar refers to business, and the third bar refers to citizens.

MOBILITY TELEPHONY

Jordan has embarked upon a progressive reform of its telecommunications and postal sectors. This process commenced in 1995. Progress has been made in establishing some measure of competition in specific markets, and enabling regulatory structures have been established through the TRC (Telecommunication Regulatory Commission). In

Figure 5. Preferred Interactive communication channel (MoICT, 2005b)



certain instances, however, prevailing legislation, commercial agreements and the WTO (World Trade Organization) and other international obligations demand that further reform measures must be implemented. In others, the perceived needs of the market, the Jordanian economy as a whole, and social development factors are the drivers of further action (MoICT, 2003).

Many local and regional policy makers and planners consider Jordan as one of the regional pioneers in the full liberalization of the communications market in the Middle East. The Jordanian legislators and regulators have embarked upon a liberalization of the fixed Public Switch Terrestrial Network (PSTN) and International Long Distance (ILD) services and started the licensing process of new operators. By September 2005, the Telecommunication Regulatory Commission (TRC) has granted a total of 29 licenses, Jordan Telecom was the sole operator with an individual license until May 2005, when BATELCO-Jordan was granted its own individual license in addition to the class license it had before. The remaining 26 licenses are all class licenses. There are only two categories of licenses according to TRC: the individual licenses and the class licenses. The individual license "is for the operation of telecommunication networks and/or offering telecommunication services using scarce resources, such as radio spectrum, public rights of way and numbering"

(Abu-Ghazaleh & Co. Consulting, 2005). While a class license is "for any other licensed services, which would include public telecommunications networks operators and public telecommunication services providers not using scarce resources" (Abu-Ghazaleh & Co. Consulting, 2005). A total of five new licenses were granted in 2005; one individual license was granted to BATELCO-Jordan, while the other four were class licenses granted to "Sirat Telecom Technology", "LaSilkee Virtual Connection Company Ltd.", "Pella", and "Jordan Bell Telecom". On October 21, 2004, the Council of Ministers in Jordan approved the TRC's proposed licensing program, which has fully opened the fixed telecommunications sub-sector to competition as of January 1, 2005 (Arab Advisors Press Room, 2005).

Jordan's cellular subscribers grew at a rate of 46.8% between 2000 and 2004. Reduced rates, per second billing, extended validities and special offers, are expected to introduce the market with a healthy growth rate of over 46% in 2005. Between 2006 and 2009, the Arab Advisors Group projects the Jordanian cellular market to grow at a rate of 10% to exceed 3.43 million subscribers by 2009, a penetration rate of more than 57% (Arab Advisors Press Room, 2005).

The figures of the Department of Statistics and Ministry of Information and Communication Technology show that telephony penetration in Jordan is high, especially when compared with other communication medium like the Internet. The Department of Statistics indicates that the percentage of Jordanian households who own a telephone line is 54.4%, while percentage of Jordanian households who own a mobile telephone is 47.4% (Department of Statistics, 2005). While MoICT figures indicate that 87.4% of households have either a fixed or mobile phone (MoICT, 2005a).

The MoICT's figures indicate that almost all households with a monthly income above 350 Jordanian Dinars (500 US$) have a fixed or mobile phone (MoICT, 2005a). Moreover, the number of mobile phones rose with monthly income where a majority of households with over JD350 per month have more than one mobile phone (MoICT, 2005a).

The number of telephone subscribers, fixed lines, has gone down to 623,000 in 2004 from 629,000 in 2002, while at the same period of time, the number of mobile subscribers has gone up to 1,624,000 in 2004 from 1,219,000 in 2002 (Department of Statistics, 2005). And, Abu-Ghazaleh & Co. Consulting (2005) estimates the penetration rate of cellular/Mobile phone at 30.9% in 2004 with a total number of mobile owner population of 1,801,100.

THE PROSPECTS OF MOBILE TELEPHONY

Despite the continuous liberalization of the Jordanian Telecommunication market and the existence of a number of competitive service providers whether wireless or wirefull, the unaffordable prices of both hardware and telecommunication devices as well as the high cost of telephone calls are perceived as the major constraints on the proliferation of ICT in Jordan. The Jordanian IT strategic plan (REACH) has called upon both the private and the public sector to provide preferential access to the high-speed lines and permit private up- and downlinks, as well as to provide competitive pricing on high-speed lines. The current IT infrastructure of Jordan can be described as primitive in comparison to leading countries in the IT and Telecommunication sectors. For Jordan to become a regional leader in the Middle East, Jordan needs to work out its local infrastructure both in the technical terms as well as in the economic terms.

There are two major indicators to assess any communication channel or medium: reach and richness. While Evans and Wurster (1997) define

reach as "the number of people, at home or at work, exchanging information", they define *richness* in terms of bandwidth, customisation, and interactivity. *Bandwidth* refers to the "amount of information that can be moved from sender to receiver in a given time"; *customisation* refers to the "degree to which information can be customised"; and *interactivity* refers to "dialogue" (Evans & Wurster, 1997). Evans and Wurster (1997) explain that this trade off between reach and richness shapes "how companies communicate, collaborate, and conduct transactions internally and with customers, suppliers and distributors".

In terms of reach, as explained earlier the penetration of the Internet, whether at home or at the workplace, is very minimal in Jordan, estimated at 2% of the population in 2004, up from 0.7% in 2000. Furthermore, the affordability of a personal computer coupled with the high tariff of telecommunication has created a great barrier to overcome. While the penetration of the mobile/cellular phone is much higher, currently estimated at 47.4% of the population (Department of Statistics, 2005) and expected to reach 57% of the population by the end of 2009 (Arab Advisors Press Room, 2005). Furthermore, the figures of MoICT show that 39% of businesses and 13% of citizens used the Internet to contact with governmental entities, while 55% of businesses and 35% of citizens used the telephone to contact the governmental entities (MoICT, 2005b). These figures indicate the higher percentage of phone contacts compared to Internet-based contacts. Furthermore, MoICT (2005c) confirms such a conclusion stating that this low penetration of the Internet in Jordan is concentrated in few large urban areas, while the technical infrastructure "is insufficient to support high-speed connection on a mass scale". On the other hand, the increasing competition fuelled by the de-regulation of the telecommunication industry "results in decreasing prices for final users, higher quality, new investments and development of new services...Fixed line phone penetration is high and spreads over the region; mobile telephony is rapidly growing and is expected to develop further in coming years" (MoICT, 2005c).

In terms of richness, currently the Internet can provide a far larger amount of information that is easily navigable in many formats while the current wireless infrastructure is incapable of competing with such advantage. The display of the mobile phone is still cumbersome and very much limited in space.

Furthermore, the figures of the MoICT show that the Internet is the most preferred channel by businesses, citizens, and government employees for interactive communication and information distribution (MoICT, 2005b). These indicate that the Internet is preferred to the phone in terms of interactivity, bandwidth, and more importantly customization and personalisation. This makes the Internet a better option when it comes to richness. On the other hand, face-to-face is considered by the majority of citizens and businesses as the preferred way to contacting governmental agencies. The penetration of mobile phones is much higher in Jordan when compared to personal computers which make such a medium more accessible and widespread when compared to PCs. This section articulates the concerns and issues surrounding the viability of mobile government in terms of availability of bandwidth and reach. The section includes statistical data and interviews with top strategy makers at the Ministry of Information and Communication and Technology in Jordan to establish Jordan's willingness to incorporate m-government on the technical, human, organisational, infrastructural, and legislative levels.

The high penetration of mobile phones in Jordan makes the mobile service provision a much better opportunity especially when compared with the low penetration figures of the Internet. Though the limited display space and the cumbersome interface makes such a communication medium unfavourable for service providers. Until such

issues are resolved in the near future, the usage of mobile phone would be limited to alerts and as a payment gateways for micro-payments with an exchange value of less than 10US$ (Amor, 2002).

Based on the surveys conducted in 2005 by Arthur Business Consulting, the belief at the Ministry of Communication and Information Technology is that the Web seems to be an appropriate channel for most of the governmental e-services where e-mail is usually suggested as a channel supplementing Web since the majority of electronic services usually require either large amounts of data or legal enforcement. On the other hand, SMS channel can be used for notification about the status of the transaction (MoICT, 2005a). Though, the figures of the MoICT indicate that the majority of companies' employees use fixed telephone lines and mobile phones for business purposes while the Internet comes in second place using either broadband Internet connection or dial-up (MoICT, 2005a).

In order to assess the viability of m-governmental services in Jordan, the authors carried out two major interviews with two prominent figures in the e-government program. They are e-government Program Director (Mr. Khaldoun Naffa) and e-government Program Chief Technology Officer and Head of Operations (Mr. Hasan Hourani). These interviews aimed to assess the viability of mobile government from a strategic point of view. The interviews lasted for one hour each both held on Thursday, November 24, 2005, at the e-government program headquarters in the Ministry of Information and Communication Technology.

The interviews started by assessing the past experiences on electronic governmental services. Both interviewees agreed on the utter importance of electronic governmental services and explained concerns regarding many implementational/operational obstacles. These obstacles included: education, awareness, preparedness, public sector reform, organisational and technical change management, and transformation management. The interviewees agreed on the successful implementation so far in terms of systems, technology, and infrastructure but lacking attention toward the softer human aspects of the change process. The interviewees indicated that the e-government program is perceived as a tool of public sector reform in terms of becoming more customer centric, that is, improve governmental entities/employees performance, increase cost effectiveness, and increase transparency.

In terms of successes so far, the interviewees pointed out that the internal surveys show a fluctuation in usability of electronic governmental services. This fluctuation in service usability can be attributed to many factors, mainly: over sensitivity of users toward information confidentiality and security, lack of proper change management during the transition process, the need for more than one sponsorship promotion and support, loss of key staff during the change process, and ulterior/personal motives of both enthusiasts and resistors to change.

Regarding mobile government, the interviewees indicated that it should be used to provide a complementary delivery channel to the already established Internet/Web-based services delivery to make use of the high penetration (reach) of such a medium. Though the interviewees raised concerns regarding richness of such medium, these concerns addressed the following issues: limited bandwidth and capacity in terms of intensive data services, lack of maturity in terms of devices, limited data presentation/display capacity, challenges of security/confidentiality of private data over public networks, and lack of display capacity of the device itself.

Regarding the liberalization and de-regulation of the local telecommunication market, the interviewees indicated that this pushes the whole initiative toward providing alternative delivery option in addition to the existing Web-based de-

livery. The interviewee's vision for the future is that mobile lines will take over fixed lines based on increasing cost reduction of call tariffs, increasing bandwidth, and mobility of the device. Though, one of the most neglected areas of the liberalization effort is legislation, where a new legal road map is needed to cater for the differences in delivery mode and contents. The interviewees indicated the need for electronic legislative enablement to be led by involved stakeholders.

CONCLUSIONS

This chapter has presented and described the different aspects of the Jordan's Electronic Government Program based on the published material of the Jordanian governmental as well as non-governmental organisations in Jordan. The chapter presented the findings of a number of quantitative and qualitative assessment projects to the past few years of electronic services provided by a number of governmental entities. In addition, the result of a limited number of interviews with key decision makers was provided in order to augment the results of the carried surveys.

Jordan's local IT and Telecommunication sector can be described as evolving rather than developed. Currently, Jordan has one national telecommunication operator, one licensed PSTN provider, four mobile phone operators, and an extended number of Internet service providers. Accessibility to both private and public networks is provided via dial-up connections, which is regarded as slow and expensive. More liberalization from the government and more participation from the public sector both locally and internationally are needed to enable Jordan to become a regional leader. While current figures of IT usage and proliferation in Jordan shows a great advancement in the past five years, several issues are needed to be taken into account like the affordability of computers and phone calls as well as the different laws, by-laws, and pieces of legislation needed to be endorsed by the House of Parliament.

Lack of advanced and secure technical infrastructure, lack of high-volume of Internet users, and limited use of credit cards in Jordanian society remain the main reasons why most organizations and individuals in Jordan refrain from using the Internet to exchange products/services and funds online. To justify investment in online technologies and processes, high velocity of Internet traffic is needed. REACH 2.0 (2001) shows that despite the sharp fall in subscription fees, affordability of personal computers remains the main hindrance to engaging in online activities for individuals. These issues need to be resolved on national and regional levels rather than on national level only. Regarding payments methods, whether physical or digital, the limited spread of credit cards makes the use of online experiences to exchange digitized and physical products and services for funds a true nightmare. It is becoming evident that more creative mechanisms should be used to overcome such a hurdle, like cash exchange at delivery time or using pre-paid cash cards or even using automatic teller machine (ATM)/debit cards as an alternative to credit cards.

The chapter provided an analysis of the possibility of using mobile governmental services as supplementary and as an alternative to Web-based services. The analysis relayed on validating the richness and reach of the mobile network in comparison to the Internet. The analysis showed that while mobile networks enjoy a higher level of penetration, these networks and mobile devices provide an inferior service in terms of display, interactivity, and customisation specially when compared to Web-based services. Generally speaking, in order for mobile phones to surpass the current position used mainly for alert services and micro-payments, the mobile phone networks need to provide more richness coupled with higher bandwidth.

REFERENCES

Abu-Ghazaleh & Co. Consulting. (2005). *Market brief on telecommunications sector in Jordan.* Retrieved February 1, 2005, from http://commercecan.ic.gc.ca/scdt/bizmap/interface2.nsf/vDownload/ISA_2665/$file/X_8404815.DOC

Abu-Samaha, A. (2005). Strategic and operational values of e-commerce investments in Jordanian SMEs. In S. Kamel (Ed.), *Electronic business in developing countries: Opportunities and challenges* (Chapter 16) (pp. 315-335). Hershey, PA: Idea Group.

Amor, D. (2002). *The e-business (r)evolution: Living and working in an interconnected world.* New York: Hewlett-Packard Books.

Arab Advisors Press Room. (2005). Retrieved November 12, 2005, from http://www.arabadvisors.com/Pressers/presser-181005.htm

Department of Statistics. (2005). Retrieved November 12, 2005, from www.dos.gov.jo

eEurope (2002). *An information society for all action plan.* Retrieved February 1, 2005, from http://europa.eu.int/comm./information_society/eeurope/documentation/index_en.htm

Evans, P. B., & Wurster, T. S. (1997). Strategy and the new economics of information. *Harvard Business Review*, September-October, 71-83.

Ginige, A., & Murugesan, S. (2001). The essence of Web engineering: Managing the diversity and complexity of Web application development. *IEEE Multimedia, 8*(2), 22-25.

Jordan eGovernment Initiative. (2003). *The e-government status update.* Retrieved December 1, 2005, from http://moict.gov.jo/MoICT/downloads/e-Gov_Intaj_6102003_v1.0.ppt

Lawrence, E., Corbitt, B., Tidwell, A., Fisher, J., & Lawrence, J. (1998). *Internet commerce: Digital models for business.* Brisbane: John Wiley.

Ministry of Information Communication and Technology, MoICT. (2000). *Launching e-government in Jordan: Readiness and approach.* Retrieved December 1, 2005, from http://www.MoICT.gov.jo

Ministry of Information Communication and Technology, MoICT. (2003). *Statement of government policy.* Retrieved December 1, 2005, from http://www.MoICT.gov.jo

Ministry of Information Communication and Technology, MoICT. (2005a). *Universal service/access ICT policy and fund mechanism: Market research study.*

Ministry of Information and Communications Technology, MoICT. (2005b). *National outlook and expectations.* Retrieved December 1, 2005, from http://www.MoICT.gov.jo

Ministry of Information and Communications Technology, MoICT. (2005c). *Detailed study and recommendations on the access and delivery channels for e-services implementation.*

Moon, G. J. (2003). Can IT help government to restore public trust? Declining public trust and potential prospects of IT in the public sector. In the *Proceedings of the 36th Hawaii International Conference on System Sciences.*

REACH 1.0. (2000). *IT forum.* Retrieved December 1, 2005, from http://www.reach.jo/Downloads/R1/R1_report.pdf

REACH 2.0. (2001). *INTAJ.* Retrieved December 1, 2005, from http://www.reach.jo/Downloads/R2/R2_report.pdf

Simpson, P., & Swatman, P. (1998). Small business Internet commerce experience: A longitudinal study. In the *Proceeding of the 11th Bled International Electronic Commerce Conference,* Bled, Slovenia, June 8-10 (pp. 295-309).

Tambouris, E. (2001). An integrated platform for realising online one-stop government: The

e-GOV project. In the *Proceedings of the DEXA International Workshop "On the Way to Electronic Government"* (pp. 359-363). Los Alamitos, CA: IEEE Computer Society.

Tapscott, D. (1996). *The digital economy: Promise and peril in the age of networked intelligence*. New York: McGraw-Hill.

Telecommunications Regulatory Commission (TRC). (2005a). Retrieved November 12, 2005, from http://www.trc.gov.jo/Static_English/market.shtm

Telecommunications Regulatory Commission (TRC). (2005b). Retrieved November 12, 2005, from http://www.trc.gov.jo/Static_English/telecomss.shtm

Turban, E., Kuy Lee, J. ,& King, D. (1999). *Electronic commerce: A managerial perspective*. New Jersey: Prentice Hall Business Publishing.

Turban, E., Jr., & Potter, R. (2001). Introduction to information technology. USA: Von Hoffmann Press.

Vélez-Rivera, B., Díaz, W., Fernandez-Sein, R., Rodríguez-Martínez, M., Núñez, M., & Rivera-Vega, P. (2005). Multidisciplinary e-government research and education as a catalyst for effective information technology transfer to regional governments. *DGO*, 133-134.

This work was previously published in Mobile Government: An Emerging Direction in E-Government, edited by I. Kushchu, pp. 268-290, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.20
Usability Driven Open Platform for Mobile Government (USE–ME.GOV)

Paul Moore Olmstead
Atos Research and Innovation, Spain

Gertraud Peinel
Fraunhofer FIT, Germany

Dirk Tilsner
EDISOFT, Portugal

Witold Abramowicz
The Poznan University of Economics, Poland

Andrzej Bassara
The Poznan University of Economics, Poland

Agata Filipowska
The Poznan University of Economics, Poland

Marek Wiśniewski
The Poznan University of Economics, Poland

Pawel Žebrowski
The Poznan University of Economics, Poland

ABSTRACT

This chapter introduces the USE-ME.GOV project that supports and encourages the authorities with the access to new e-government services at any time and anywhere through the use of mobile communications and Semantic Web technologies. The USE-ME.GOV system addresses openness,

interoperability, usability, and security scientific goals, and throughout the chapter the methodology and main outcomes are described.

MOTIVATION AND GOALS

IST initiatives for improving services to citizens and businesses are increasingly being promoted and implemented by individual authorities and organizations. Even smaller towns operate their own Web site with access to general public information, whereas larger cities and institutions generally offer a wider range of more sophisticated electronic (Web-based) services.

However, the richness and quality of these services can vary significantly. In particular, small authorities, for example, in rural areas, have limited financial, technical, and human resources in order to implement and deploy electronic services with the same quality as large organizations (Leenes & Svensson, 2002). This aspect becomes even more critical for the deployment of mobile services because of a higher complexity of service implementation, the required organizational changes as well as higher costs for commercial exploitation due to the complexity of the value chain.

Authorities are usually organized in departments, each with their own responsibilities, tasks, structure, and customers. Unfortunately, the IT infrastructure and equipment, as well as the corresponding technical background knowledge, are often different in each department. Mobile operators or portals are searching for content to promote their new mobile technologies and approach public organizations to deliver services on Internet and wireless networks. Once contracted, one department connects to a particular mobile operator and "somebody" implements a proprietary bridge to one specific operator interface. This bridge can normally not be reused for other applications or other mobile operators.

Authorities are now actively searching for mobile solutions to implement regulations and recommendations from state, national, and European bodies calling for e-government, e-governance, and of course m-government. But due to a lack of adequate technical background, monetary shortcuts, legal restrictions on innovative partnerships and business plans, and less experience in mobile markets and their interdependencies, the authorities are hesitant about investing time and money in stand-alone proprietary solutions that require major investments.

The deployment of an open service platform, that can be shared by networked authorities and institutions (e.g., on a regional scale) in terms of technical resources as well as commercial exploitation, would harmonize the quality of public services and overcome related *divide* phenomena. On the other hand, resource sharing on the basis of attractive business models would also provide the conditions for cost-efficient m-government services especially in geographical areas with low Internet penetration.

Therefore, the USE-ME.GOV project aims to provide an open and interoperable platform that can be shared by different local authorities and diverse organizational units. This sharing of the platform means that the cost of ownership of such a solution is reduced. Also, by emphasizing the openness of the platform, any involved party is given the choice of providing services in any suitable or available technology.

Hence, the USE-ME.GOV project's key objective is to support authorities entering the mobile market with an open source platform that allows:

- The sharing of common modules with other departments or other authorities (for example, subscription, alerting components);
- Making development and operation more secure through open source transparency;

Figure 1. Current state of authorities' mobile interface architecture

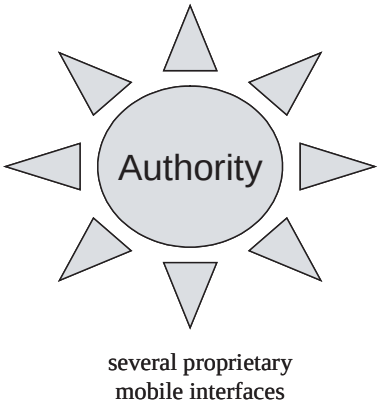
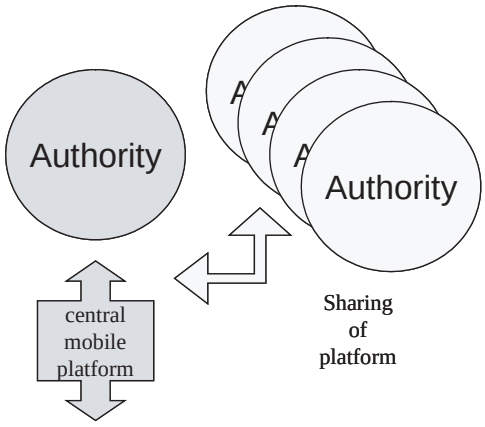


Figure 2. "Dream" situation while using the interoperable platform



- Bringing on board other mobile operators independent of their respective interfaces; and
- Being able to estimate costs, efforts, outcomes, and benefits in advance.

This proposal gives an overview of the USE-ME.GOV project, its current state and findings including planned services, the obstacles that have been experienced, and the technological design and implementation work in progress.

USABILITY REQUIREMENTS

Groundbreaking topic-specific mobile applications focus mainly on the challenges related to mobility itself, mobile assistance, and location-based services (Abramowicz, Bassara, Filipowska, Wisniewski, & Zebrowski, 2006). In all these areas, usability is seen as a key challenge (Barton, Zhai, & Cousins, 2006).

For the design of the open service platform and pilot services, a usability-driven approach was followed, as also indicated by the name of the project. The concept of usability is many-fold and encompasses the following fields of research and application:

- **Enlarged access to public information services.** In order to ensure broad access by a significant part of the population, the platform has to provide openness and interoperability with regard to the interconnection with different networks, the integration of external content providers and public authorities providing their services, and must further consider diverging mobile device characteristics and capabilities.
- **Intuitive and user-friendly mobile interfaces.** Services are designed taking into account heterogeneous user characteristics, addressing the common needs of the citizens with different educational or even cultural background, age, and interests, allowing for easy access to and search of information considering location, context, and user interests.
- **Deployment of services.** The concept of usability also implies that mobile services

must be easy to deploy for the authorities, not depending on expensive software-hardware products or demanding technological skills for their configuration, maintenance, and continuous update of service content.

- **Economical sustainability.** The participation in the platform is thought to be open to all interested providers of public mobile information services including small authorities and organizations that have limited financial capabilities to deploy mobile services on an individual basis. The framework for exploitation takes into account the diverse needs and interests of public and private providers of services and information.

One of the major problems developers must deal with is the small display and limited user input of mobile devices. User interfaces for mobile technology and ultra-portable devices like cellular phones, PDAs, computer tablets, and wearable devices tend to mimic user interfaces originally designed and established for desktop computers to date but these basic versions of the interfaces do not usually translate well to mobile situations.

There were two divergent streams of usability research in the USE-ME.GOV project. Pilot services designed for present day smartphones were created using the standard interactive methods available in these devices (basically keyboard and pen/finger). The usability of these services was then validated in field and laboratory testing with 10 users following the criteria detailed earlier. But also, in parallel, one of these pilot services (the City News Broadcasting Service) was redesigned for future smartphones with an additional voice input channel, hence turning its UI into a multimodal UI. The objective of this work was to identify and to give recommendations that can be applied to the design of multimodal 3G services on mobile devices. A comparative user test with 20 subjects was conducted to assess and compare the usability of the tactile only and the multimodal UI.

Needs and Benefits for Public Mobile Services

The need for and usefulness of mobile services provided by public administrations is mainly seen to be from the perspective of multi-channel service delivery. While the debate on mobile government is ongoing, research on the benefits and applicability of mobile services for the public administration is sparse and much of the available literature actually refers to pilot projects and implementations. The previously-mentioned project USE-ME.GOV therefore conducted a study on the particular needs and expectations of each of the authorities involved in the project:

- Bologna City Hall (Italy)
- Extramadura Association (Spain)
- Vila Nova de Cerveira (Portugal)
- Gdynia City Hall (Poland)

It was found that these organizations follow a variety of operational, economical as well as political and IST strategic goals (D2.2, 2004). Despite the existence of common needs, a diversity of requirements and potential benefits were extracted and synthesized:

- Mobile services are seen as a new and/or complementary dissemination channel and a means of access to public information. Public information is of various types:
  - General public information;
  - Time-critical information (emergencies, traffic); and
  - Notifications according to user-specific interests.

Key Benefits

- o Dissemination of information to a larger number of people (mobile access) in a very short time.
- o Increased accessibility, transparency, and citizen satisfaction.
- o Improved public perception of the town or city.

- The mobile channel provides an efficient means of communication between the authority and the citizen The most evident application of this concept is in the sending of notifications, bulletins, and so forth, having to do with specific cases and processes such as, for example, requests for certificates and other documents issued by the public authority). Even though mobile services cannot be expected to eliminate entirely the need for personal attendance, they can substantially simplify process-related correspondence and provide instant and accurate information to the citizen whether serving as the primary channel or as a complement to other channels.

Key Benefits

- o Reduction of average service processing time, especially for correspondence concerning simple notifications.
- o Ubiquitous and instant contact.
- o Reduction of costs.
- o More time freed up and which can be dedicated to particular cases.
- o Citizen and private-user satisfaction.

- Mobile services can serve as a stimulus for the participation of the citizen in local community affairs and be applied as a channel for the submission of complaints, suggestions, and so forth, accessible to the public. This kind of service also encompasses the communication between the authority and the citizen during the follow-up of the complaint/suggestion.

Key Benefits

- o Early detection of problems reported by the citizens.
- o Greater accessibility.
- o More transparency.
- o Increased participation of citizens in community affairs.
- o Higher levels of citizen satisfaction.
- o Ubiquitous and instant contact.

- Within the context of general public information services, mobile services can also be used as vehicle for promotion of local (cultural, fairs) events. The promotional effect would be particularly useful for local businesses with limited financial and organizational capabilities to announce their presence at events such as local fairs.

Key Benefits

- o Dissemination of information to a larger number of people (mobile access) in a very short time.
- o Reduction of costs.
- o Contribution to sustainability.
- o Promotional support to local businesses.
- o Improved image of city, town, region, …

The results of the study show that mobile services have a very high potential and can bring substantial benefits. It should be noted that, generally speaking, the improvement of the quality of public services for citizens and private businesses are probably the most important ex-

pected benefits, whereas the apparent potential for increased service efficiency and economy, stemming from the streamlining of internal administrative work processes, is recognized, but not seen as the key driver for the adoption of mobile service delivery.

It was further concluded that the authorities who participated in the study had a clear understanding of how the targeted mobile services fit into the organization's individual IST strategy and which particular benefits could be obtained. On the other hand, a lack of experience as well as missing (defined, tested, proven) business models make it difficult to achieve a reliable assessment of the relation between costs and benefits and financial sustainability. Obviously this kind of uncertainty is often perceived by the authorities as a risk and ultimately can cause an organization to refrain from mobile service provision.

Even though the potential for increased organizational efficiency and productivity is realized, the impact on administrative workflow sometimes turns out to be a barrier to adoption. As a matter of fact, mobile service provision cannot stay disconnected from underlying workflows for service provision and require a certain amount of process re-organization and re-engineering. The bottlenecks and problems of current workflows and processes are generally known. However, the impact of introducing particular mobile services can be quite significant and complex, and organizational resistance to change and the need for modifications to established norms and administrative procedures must also be considered.

Usability in the Multimodal Services

By combining different output modalities (text, graphics, sound), information transmission and comprehension can be both easier and more effective. By allowing the use of different input modalities (phone keypad, virtual keyboard, pen, handwriting, spoken commands), users can easily adapt to different circumstances (stationary use, noisy environment, use while walking). Finally, context-awareness mechanisms can also be considered since they can make input tasks easier by providing the service with required information (e.g., user's location, environmental information, time) and hence alleviating user's effort.

Dealing with heterogeneity means that active (user-initiated) and passive (system-initiated) adaptation mechanisms should be considered for filtering and presenting relevant information to users. However, intelligent systems raise specific usability issues related to the need for users to control their system. Future research must explore acceptable guidance strategies that will help users decide how to set their preferences. It is also important to identify the procedures by which an intelligent system could maintain users awareness of its current setting, allowing them to easily understand how and why it behaves differently in different circumstances.

Following this study, the main considerations to take into account for integrating speech interaction into a graphical and tactile UI can be summarized as follows:

- Natural language tends to flatten functions hierarchies, hence the navigation structure must be deeply reconsidered to take into account the power of speech to combine multiple information in a single sentence.
- Visual prompts and feedbacks must be carefully integrated into the graphical design so that users can guess the most efficient verbal phrases and get control on the interaction.
- It may be necessary to create advanced graphical interactive components with tactile interaction that mimic the phrasing and chunking capabilities of speech to keep both modalities equivalent.

However, all these considerations are general and do not guarantee success. As the user tests showed, users interacting with a PDA or a smartphone are willing to naturally speak with

sentences once they realize it is possible, but their first expectation seems to have a system that only understands keywords or simple phrases. Moreover, it appeared that the visual prompts we chose were not efficient enough. Further research is clearly needed to find better ways to guide users toward the appropriate and effective ways to utter their requests and commands. A few directions are indicated by the test results:

- Creating an area of the screen dedicated to user guidance. Visual prompt and feedback will appear only in this area. Such an approach could help users to know which information to exploit in order to know what to say but we anticipate difficulties with it due to the reduced space of the screen of mobile devices.
- A dialogue based on speech input/speech output instead of speech input/visual output: based on previous studies (cf., Hone & Barber, 2001), it may be argued that such an approach should help users to naturally speak to systems more rapidly and, with appropriate prompts, reduce the number of errors. However, such a solution might not fit some mobility constraints (disturbance caused to other people with an audio output, privacy concerns, …).
- In any case, improved strategies to deal with dialogue errors are necessary. In particular, rather than expecting users to produce unambiguous requests and commands, ambiguity should be expected and a specific dialogue should be designed to deal with it. Also, users should be prompted to shift to another input modality when more than two errors in succession occur.

Another lesson learned from our user test is that even with appropriate prompts—whether audio or visual—users would probably deviate in uttering spoken commands from the lexical and syntactical rules exhibited with these prompts. The conclusion is that designers of multimodal UI should create flexible grammars and sufficiently wide vocabulary. Since it is difficult to anticipate all the required words and phrases from the beginning, an iterative design process should be adopted where pre-design studies (Yankelovich, 1998) and Wizard-of-Oz tests (Fraser & Gilbert, 1991) can help in identifying them before any program development begins. We have also learned that it should be preferable to make any word or label visible on the screen understandable by the system.

Most importantly, technological improvements are required to reduce the number of recognition errors. Even with an iterative design process, it seems difficult to expect today, with the available market products (such as Nuance) and under real use conditions, recognition rates higher than 80%. Since other available modalities on small devices exhibit a better reliability, which also means a better ease-of-learning and ease-of-use, a majority of users could prefer them to voice input.

One could question the need to expect these improvements. After all, if users learn very quickly and found easy to use the current user interfaces of smartphones and PDA, why providing them with the ability to use voice as an input channel? The reasons why we still believe necessary to look for multimodal UI are:

- There exists a great variability among users about which input modality is preferable whatever the situation. As our test revealed, some users clearly seem to prefer voice input even if it is less reliable than other available modalities while other users seem to prefer pen-based interaction, whatever the speech recognition performances.
- Each modality fits particularly well some specific task characteristics. For instance, voice input is the preferred input modality to select objects that are not visible on screen. Other studies have shown that gestures are preferred to select points or geographical

areas on a map (Oviatt, DeAngeli, & Kuhn, 1997).

- Even if a specific modality is preferred, users may encounter situations where they need to shift to another modality. This is especially true when an user faces repetitive errors with a specific word or in selecting a specific area of a touchscreen, for instance.
- More generally, multimodal UI enhances the users' adaptive capabilities. Depending on the situation characteristics, users may prefer to use voice or the pen. For instance, when moving, they should prefer to use voice. Being stopped in a quiet place, they should prefer to use the pen.

All these reasons lead us to believe that research on multimodal UI should continue and produce technological, methodological, and design progresses so as to increase their usability and acceptance.

GENERAL ARCHITECTURE

The analysis of the user requirements leads to the design tasks that are initiated with the architectural design. Collected, formalized, and classified requirements are analyzed to give functional designers a framework and technologies to work upon. This is not an easy task, as it needs to satisfy both the functional requirements and provide a basis for the detailed system design. One the one hand, the system must fully comply with the requirements. But, on the other, system designers must feel comfortable with this basic architecture so that during the entire design process they are not unduly constrained by technology or architectural decisions. These constraints could arise from the limitations of both the chosen technologies and/or the sketched architectural solution. Although there is no ideal solution where such constraints would not appear, the architectural designers should aim at minimizing that gap.

In the area of m-government, according to the main European Union's initiatives in the field, such as IDABC (former IDA Framework), systems' analysts should conform to the service orientation paradigm. This paradigm describes a single unit of functionality as a service. The service providers form a specific marketplace, on which the services can be searched, utilized, joined together, and ran. Therefore, the main focus is placed on services, and therefore, on service orientation. In a more technical fashion, we can talk about service-oriented architectures (SOA) that are a roadmap, and serve as the set of recommendations on how the specific applications need to be constructed.

The application of SOA opens up other "ideological" advantages:

- **Openness.** The ability of the system to stay open for multiple non-proprietary technologies and frameworks. The more open the system is, the more extendible, and thus usable, it is.
- **Interoperability.** The ability to provide technical, organizational, and semantic means for data and information interchange and utilization. In the area of e-government, this is of steadily increasing importance, and it has been one of the major focus points for the technical and research partners in the USE-ME.GOV project.

Interoperability

Interoperability has been an issue since the first information systems lost their homogenous nature. Multiplicity led to the emergence of "islands" of objects that were originally thought as cooperative units. However, the operability on the "island" level is still only a possibility, and these units are unable to operate within external frameworks. Here the problem of interoperability arises.

There is no coherent and agreed definition of interoperability. Mainly because the concept is

broad enough to be comprehended from many different perspectives. The common pitfall is to think of interoperability in terms of only technical issues, whereas all its aspects should be considered. These additional aspects could be grouped into the semantic and organizational domains.

On the other hand, some misunderstanding of the concept of interoperability led to multiple and inconsistent definitions. The definitions vary from extremely simple:

the ability of two or more systems or components to exchange information and to use the information that has been exchanged (IEEE, 1990),

the ability of software and hardware on multiple machines from multiple vendors to communicate[1]*;*

to more complex:

capability to provide successful communication between end users across a mixed environment of different domains, networks, facilities, equipment, etc., from different manufacturers and(or) providers. In this context the communication is meant between end users or between an end user and a service provider[2].

Even so, the European Commission (EC) (2003) has its own way to define interoperability. It is not a real definition but a neat comparison:

Interoperability is like a chain that allows information and computer systems to be joined up both within organizations and then across organizational boundaries with other organizations, administrations, enterprises or citizens.

The purpose of interoperability is to share and reuse information in a way that the exchanged content will be understandable by the applications. Furthermore, the sharing of information shall not interfere with internal organizational code.

Technical Interoperability

Weiser's (1993) original research on pervasive computing was driven by the vision of interoperable miniaturized devices. Multiple kinds of devices, software, and operating systems and network access mechanisms all stand behind the idea of interoperability. All these pre-requisites form the technical aspect of interoperability. Weiser's vision is now regarded as a roadmap for the contemporary research communities.

Technical interoperability in m-government is a complex term because it involves interoperability issues on five distinct levels:

- Representation languages
- Data formats
- Operating systems
- Transmission protocols
- Heterogeneous hardware

To help understand the problem, we may ask the question: where has the technical interoperability succeeded? There are obviously domains in which interoperability was not only the key issue, but also the main necessity. In those areas, it has succeeded. The postal code is an example—unified, standardized, commonly used, and simple. The example may seem trivial, but moving on to telephone numbers things get little more complex. Finally, the latest benefactor from the interoperability revolution is the Internet. It almost seems unreal, but the compliance to agreed standards has been achieved. The backbone of the Internet lies in the hands of technical protocols, namely TCP/IP and HTTP. To some extent, the success of the Internet is based on technical interoperability. However, not all of the interoperability aspects are covered in the Internet which is clearly a driving force behind the initiatives devoted to the promotion of coherent conceptualization.

But in many spheres technical interoperability is still failing. The realm of information systems is tremendously diverse, and in the foreseeable

future, this diversity means that, for the most part, problems in interoperability will continue to plague us.

To bring us closer to the ultimate goal of real interoperability different approaches are being considered. Service-oriented architecture and all the accompanying standards with Web services as a backbone bring the promise of application level integration.

Technical interoperability is, in specific cases, achievable. The problem is of a different nature—how to convince or impose standards and technologies. In the case of simple homogeneous systems architecture the situation is trivial. However, as the complexity grows the interoperability problem escalates. This requires that all participating parties conform to the technologies that have been agreed upon.

Semantic Interoperability

Semantic interoperability is to ensure that the precise meaning of exchanged information is understandable by any other application not initially developed for this purpose (EC, 2003).

When dealing with semantic interoperability one has to consider problems of a different kind. That are of a structural or semantic character. The structural problems are due to the variety of model representations, whereas semantic problems have to do with incoherency in meaning. The idea is to build a framework which will form an internally compact solution. Thus, a new application will be able to understand and use the information. This should be considered as a shortcoming of the EC's definition of semantic interoperability. New players should not only gain access to the information but also be able to utilize it.

Knowledge representation has been a subject of many new trends over the past decades. Currently, the leading of these are based on ontology theories. In fact, the semantic aspect of interoperability is all about the ontology and how it deals with knowledge:

- representation;
- management; and
- utilization.

Semantic aspects therefore are vital to achieve the overall objective of interoperability and are presented in more detail in the *Service Repository* section.

Application level integration promised by service-oriented architecture requires mechanisms to compose exposed WSDL endpoints; these mechanisms are known as service composition. Currently, the OWL-S[3] initiative defines the processes behind Web services composition.

Organizational Interoperability

Organizational changes are not easy to be implemented. The time of simple, hierarchical, and unified organizational structures seems to be gone. The organizational structures of Europe's administrations and enterprises vary significantly. At the same time, they tend to have closed structures that do not easily allow for operating with external units. Therefore, even providing technical and semantic interoperability is not enough, as often organizational constraints will not permit interoperability. This implies re-organization of internal structures.

Sharing and reuse of administration-specific information, which is the whole point of interoperability, should not, however, unduly interfere with internal organization structures. Of course, the pre-requisite is the appropriate preparation of these structures. Though re-engineering of the authorities' processes is often one of the stated goals of their IT projects, developers cannot expect them to very actively participate in the technical design of the business processes. The main challenge, therefore, is to define and execute process re-organization in such a way that platforms such as USE-ME.GOV do not require intense technical participation of the authorities in the course of the

product life cycle but, in the end, the functional requirements of the authorities are met.

Functional and Architectural Partitioning

Platform requirements analysis released as an object model (D5.1.1, 2004) should be the initiator for the subsequent analysis and design activities. Analysis and design activities should be split into functional and architectural partitioning according to the Unified Process (Larman 2005; Maciaszek, 2005) as the most common and acceptable methodology. An object-oriented approach results in the Unified Modelling Language being used as a modelling language for communication purposes as well as documentation of work conducted.

Functional partitioning starts with an analysis of the functional requirements of the platform. Stated requirements should be analyzed both syntactically and semantically to derive direct mappings to the objects. For non-trivial object identification, sequence diagrams are constructed and refined according to the outcomes of the analysis model. The initial functional partitioning activities are conducted in three subsequent phases in order to demonstrate the legitimacy of the achieved results. For each identified functional area, detailed sequence diagrams should be constructed to define objects' operations and detect all possible design imprecisions.

Architectural partitioning refers to the analysis of the current, most appropriate technologies for the domains involved and the subsequent architecture elaboration.

The main factor behind the architectural partitioning is to design the architecture in a such way that to ensure that both stable and extensible application systems can feasibly be provided. The stable application system should include all the functionalities that are indispensable for all other system's parts and therefore, constitute a core system's functionalities. Extensible parts of the system should provide for all those functionalities that can be joined up, shared, and utilized among all other participating entities.

As well, the project team should approach the issues of openness and interoperability of the proposed solution with an exhaustive attention. As already stated, technical aspects require that the platform utilize and accept common standards of the most influential standardization bodies, notably the W3C[4] and OASIS[5] organizations. Architectural usability aspects require that other defined entities have the possibility to co-exist with the system being developed, so that the content can be shared or added and services or any other defined functionality can be provided. Openness and interoperability of the system should permit all parties to have a wide choice of technologies that they are allowed to use.

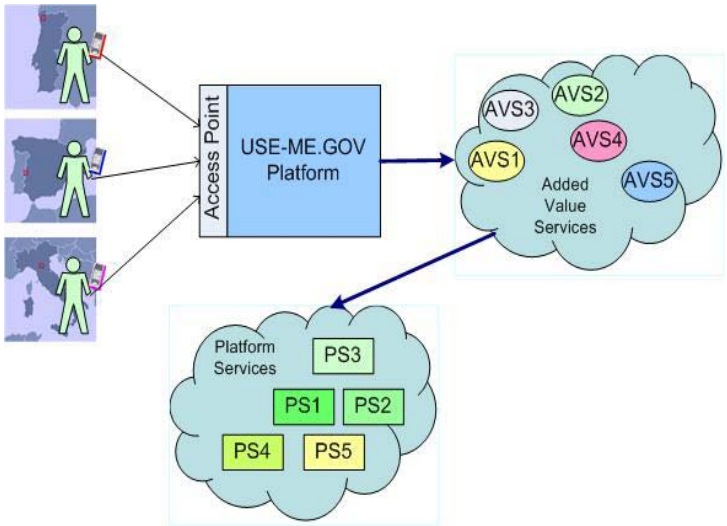
The architectural requirement for the division of application systems and the project's technical goal for open and interoperable platform are able to be fulfilled only by one architectural framework. Service-oriented architecture, most notably its latest version of Web services architecture (WSA, 2004) has proved to be a huge success among research bodies.

To ensure that the SOA is reflected in the architecture, the system under construction should, where possible, map the concepts that are defined in WSA to the context of the system being defined. Not all defined concepts will necessarily be used, of course, but the most relevant ones will probably form the main part of the architecture.

Web services architecture can be applied from four different viewpoints, thus its reference consists of four different models:

- Message-oriented model that deals with a message as a focus point;
- Service-oriented model that focuses on a service;
- Resource-oriented model that deals with resources; and

Figure 3. USE-ME.GOV general architecture



- Policy model that defines the modelling of the constraints to the resources, services, and agents.

In USE-ME.GOV system, we have, in particular, utilized the service-oriented model as a focal point of our system.

Platform Design

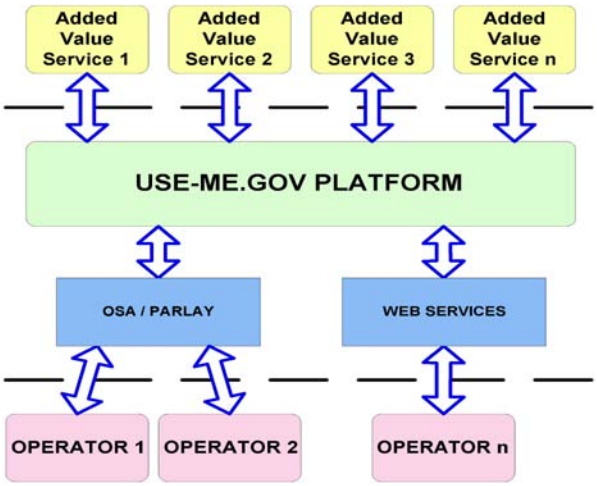
The USE-ME.GOV system is designed to allow the delivery of content and e-services to users who use a variety of mobile devices with different capabilities and connecting by various communication channels. These services constitute an added value (from now on added-value services—AVS) and are not an integral part of the USE-ME.GOV platform.

AVS are intended to be delivered by third parties and may be created using virtually any technology and deployed on any machine as long as its functionality is accessible via the designed interface. For interoperability issues, open and commonly accepted technologies are used. AVSs use Web services (conformance to WS-I Basic Profile) for remote procedure calls and electronic documents interchange and use WML and xHTML over HTTP for content delivery.

The user is not allowed to invoke AVS functionality directly. The USE-ME.GOV platform takes care of finding appropriate service, dispatching request from user to AVS and forwarding responses to users and contacting users on behalf of services.

The USE-ME.GOV platform consists of two separate application systems that are deployed on J2EE servlet container—the core platform and service repository. The core platform serves as a single point of contact for users and is also responsible for management of users and terminals. The user does not need to be aware of the AVSs available or their location. The core platform is

Figure 4. USE-ME.GOV platform interfaces



responsible for forwarding messages to the proper location. This task is achieved with the help of the service repository.

The service repository serves as a central registry of available services. Every service which wants to be discovered must register its description within the repository. The description (description language is defined by meta-protocol of service types) contains functional and non-functional features of service encoded in semantically rich format. These descriptions allows for easy finding of relevant services as well as their automated execution.

The USE-ME.GOV system also contains platform services—services provided either by the platform operator or third parties which extend the functionality of the USE-ME.GOV platform. Their functionalities must also be exposed as a Web service, and they also must be registered within the service repository. They differ from AVSs in that there are not directly accessible by users. Sample platform services included in the platform installation are:

- Context aggregation service
- Context provision service
- Localization service
- Content aggregation service
- Content provision service

Developing spatially-aware information systems for dynamic, location-specific information in mobile environments has been a challenge on it own (Harsha & Joel, 2005). Such solutions can be plugged-in to the USE-ME.GOV architecture according to the openness and interoperability capabilities.

Currently for external connections to mobile operators, there is no accepted standard to which all operators comply but the most common standard is OSA/ParlayX. For this reason, the interfaces for connecting the USE-ME.GOV

platform to the national mobile operators have been developed conforming to the OSA/ParlayX standard. In those cases where operators did not have a ParlayX platform, specific connectors had to be developed.

The core platform includes the following modules:

- **HTTP Server Adapter:** It is an adapter to the application server. HTTP server allows for interactive connection with the platform. It should forward every request to an HTTPAccessPoint and return the response generated by an AVService.
- **MO Messaging Adapter:** It is an adapter to mobile operator messaging capabilities. It allows sending and receiving SMS and MMS messages. It also allows to be notified when a message arrives and to check message delivery status. Initially, UseMeGov will provide implementations for interacting with a ParlayX interface. If the network operator does not have such a platform, special adapters will need to be implemented to access network messaging functionality.
- **Communication:** This subsystem is responsible for managing the communication with user terminal. The communication is established between user terminal and specific access point.
- **Communication Channel:** Manages communication channels registered in the platform. Communication channel which possesses specific characteristics is used to describe access point capabilities.
- **Billing:** This subsystem records data necessary for billing purposes. Client subsystem decides which data post to billing subsystem. Due to the variety of the billing needs of each operator, this subsystem will be modeled as a controller which dispatches billing events to some registered listeners.
- **Platform Management:** This subsystem allows access to the platform administrative capabilities like subsystems monitoring, status reporting, and resources management.
- **Operations:** Start stop platform, manage users, terminals and access points, subsystems status management, configuration management, performance management, and so forth.
- **Terminal:** Terminal package deals with all significant issues related to terminal management in the UseMe.Gov platform. Main objectives of the package is terminal data persistency, terminal identification, and terminal properties management. This package handles also with the mobile operator data that are crucial to platform operation.
- **User:** User package provides necessary infrastructure for the management of user-related persistent data. It performs four main tasks: manages user authentication data, manages user subscription status to external services, provides infrastructure for user profile management, and enables indirect access to terminal package.
- **Properties:** Property package deals with provision of mechanisms for properties management. Properties could be either user preferences, terminal preferences, or any profile-related attributes. It is designed in order to provide terminal and user subsystems with a capability to manage properties sets.

WS Interfaces:

- **Messaging Service:** This Web service exposes basic messaging functionality
- **User Management Service:** This Web service exposes basic user management functionality. It allows the suspension an reactivation of a user subscription in the platform.

SERVICE REPOSITORY

Many functional requirements for the USE-ME.GOV system are formulated around the concept of automatic service discovery and execution. It is assumed that it should be easy to create new services that are hosted on the platform or may be used by the platform (or its services). These services include services which are directly accessible by the end user—AVServices (added value services), and services that extend platform capabilities such as:

- Content provision services
- Content aggregations service
- Localization services

Services may be created and provided by virtually anyone, which means that the major problem is platform awareness of their existence and the ability to communicate with them. Therefore, there exists a strong demand for a mechanism for the exchange of service offers and requests, allowing client applications to dynamically locate services that satisfy their requirements.

For this reason, every technology that allows for creation of SOA should be equipped with this kind of mechanism, which may be generally divided in two categories: discovery and lookup.

- The lookup mechanism requires the existence of a central repository (or repositories), which stores all necessary information on available services. This information should be sufficient to successfully establish a connection between the service requester and the service provider. In the most common scenario, the service provider announces (registers) its services in the central repository providing all the necessary information such as: service name, address, terms of usage, interface (or many others depending on the target technology). The service requester knows the location of service repository as well as the protocol of communication with it. The requester may then retrieve services that fulfill certain criteria. The sample implementation of this mechanism are: CORBA Trading Object Service (Wohlever, Fay-Wolfe, Thuraisingham, Freedman, & Maurer, 1999) and Universal Description, Discovery and Integration (UDDI).
- In contrast, the discovery mechanism does not require a central repository. A client request is usually multicasted (or a search agent is used) to all available or potential service providers. In response, service providers which host services that fulfill the client's criteria respond with the necessary information. The sample implementation of such an approach is service location protocol (Hagen, 2001).

From among the many potential technologies that allow for development of service-oriented systems, Web services became the backbone of inter-component communication in the USE-ME.GOV system. The choice has been made based on the requirements of interoperability (highlighted in the previous section) which are most fully complied with by this technology.

In Web services, the most important element is UDDI, which is a standard platform and API for publishing and discovering information about Web services. UDDI's approach is based on a distributed registry of organizations and descriptions of respective provided services, implemented in a common XML format.

The main component of UDDI is the registry, which corresponds to an XML repository containing information about organizations and their services. Conceptually, information about an organization stored in a UDDI registry consists of three components: "white pages" including address, contact information, and identifiers; "yellow pages" that describe categorizations based on standard taxonomies; and "green pages" including references to Web services specifications. These

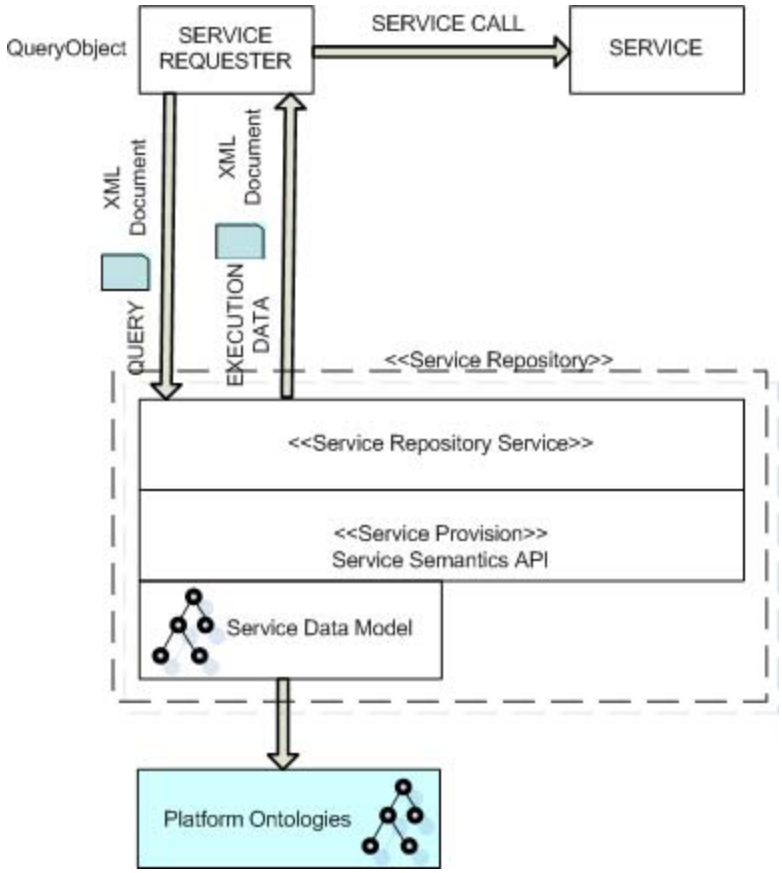
three conceptual components are, in practice, implemented in the XML format through four basic elements containing information about the organization itself (*businessEntity* element), offered services (*businessService*), service access (*bindingTemplate*), and service specification (*tModel*) (UDDI).

The UDDI is however not sufficient for dynamic execution of services, while it lacks means for the negotiation of message semantics (semantic interoperability is not supported; WSA assumes that the semantics are negotiated outside the framework). The semantics of the service is a kind of agreement between parties that provides a coherent vision of the behaviour of the service. In other words, it is an agreement on effect of service's invocation.

This agreement may be reached if the parties involved agree on a common vocabulary, which in the platform's service repository is encoded as a set of ontologies. Sample ontologies include the definition of (D5.2.4, 2005):

- content
- type of content
- processes along with necessary parameters
- types of services

Figure 5. Process of service execution



The general mechanism is very similar to all lookups mechanism except that it uses ontologies to compare queries with service descriptions and the fact that the result of the query may be directly used for services execution (without any additional information).

The main scenario is as follows:

1. The service requester connects to the service repository. The service repository itself is exposed as a Web service and so the client application may be constructed in any language.
2. The service requester formulates a query which uses concepts from ontologies provided by the platform. A sample query may expressed as follows:

   content: news
   contentType: text
   deliveryRegion: sectorA

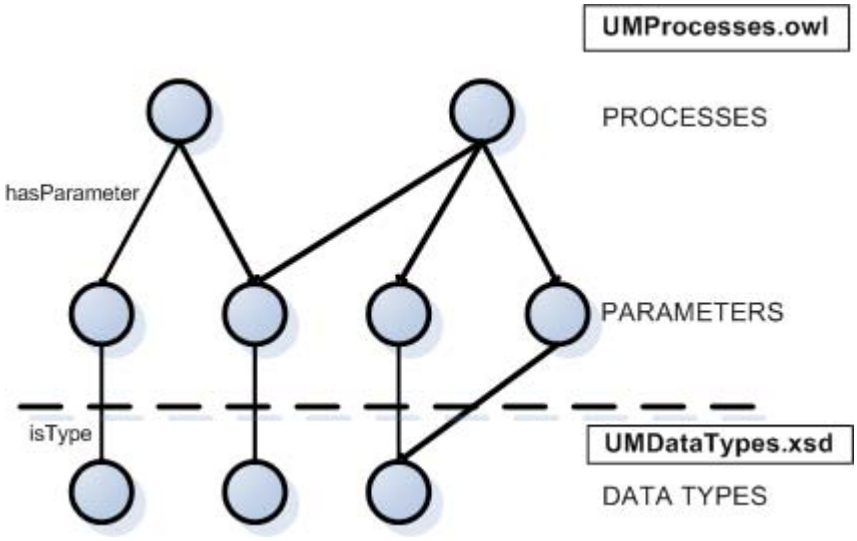
   Queries are always considered to be a conjunction of criteria. In this case, the service requester should be provided with all the services which provide textual news in a region called sectorA.
3. This query is compared against all registered services which are described by a language which is a rough modification of OWL-S.
4. The requester is provided with a set of services that satisfies the criteria. Each service may be automatically executed while the service description contains all necessary information.

The service repository utilizes this description mechanism, which is based on a solution known as OWL-S. The description of each service consists of:

- **ServiceProfile:** Contains a non-functional description of the service, which includes information on the service provider, optional rating, charging information as well as possibly some internationalization information if applicable. ServiceProfile may be easily extended with so-called provision models. A sample extension includes spatial model, which allows defining (using both geographical coordinates and region names) the spatial range of service.
- **ServiceModel:** Defines the behavior of the service in terms of processes that may be executed as well as all parameters that must be passed to the service in order to achieve the agreed functionality.
- **ServiceGrounding:** Contains all information that is sufficient to execute the service,

Figure 6. The architecture of service description

Figure 7. Platform ontologies



that is, to achieve the functionality defined by ServiceModel.

Bearing in mind that both service descriptions and queries are expressed using concepts (individuals) from ontologies, the matching process may now be conducted on a semantic level instead of using string character comparison. In the previous example, the query for news will return all kinds of news.

The most important change with respect to UDDI is in the area of services execution. Every process that may be executed contains the list of input and output parameters. For instance, the simple process of computing a quotient requires two input parameters: dividend and divider and one output parameter result_of_division. As opposed to the Web service performing the division operation, it may be defined as an: operation which has as an input two numbers and one number as a result. Either statement itself is not sufficient to execute such a service. In the first case, one knows (assuming that the semantics of "process of computing a quotient", "dividend", etc., are agreed on) the exact semantic of the operation but is unable to execute the service as long as the technical invocation details are unknown. In the second example, however, technical details are known but the semantics of the operation cannot be determined. Therefore, the service caller after querying the service is provided with full semantic description of the service along with the technical details of invocation.

USE-ME.GOV APPLICATIONS

Added value services are intended to be delivered by third parties and may be created using virtually any technology and deployed on any machine

as long as its functionality is accessible via the designed interface. The main goal of the USE-ME.GOV project was to develop the platform so that it can be used by any partner that wants to provide end services to citizens.

For the test purposes, four diverse added value services were developed and validated. Each of the selected services present some of the aspects of the designed solution. In this sense, AVS are treated as applications of the USE-ME.GOV platform. In this section, we would like to present one of these services—health care service.

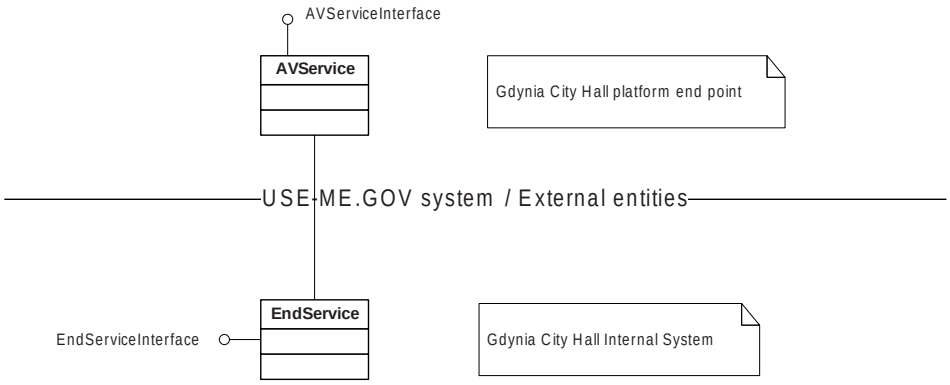
The main objective of the health care service is to provide information about health care prevention programmes and initiatives for citizens (especially young families with babies and elderly people). Nowadays in Gdynia (Poland) where the service is to be implemented information about health care centres and health care prevention programmes is published on the Web and as well as in the bulletin of the City Mayor. This should be enriched with the functionality of providing citizens with an opportunity to request and make appointments at a health care centre (according to needs—medical speciality and time preferences).

The health care service aims at providing citizens with an easier way to get an appointment, discover information about health care programmes and health care centres using mobile devices.

The health care service is designed to be used with Gdynia's specific internal system and end service. AVS is to be deployed in the public authority (in the case of the pilot—in Gdynia) and is stable between installations. Stability means that even if the place of the deployment changes, the component stays the same. It is a stand-alone application that encapsulates all service business logic and functionality. In general, end services are used in situations where some IT systems already exist within the authority or within the unit the authority collaborates with. Typical situations include databases with user information or information on authorities activities. The existence of the end service means that some IT systems already exist by the time USE-ME.GOV platform is deployed, or these systems are going to be introduced in the near future.

The solution for this approach is depicted in Figure 8.

Figure 8. Health care general architecture



The pilot service implemented and deployed in Gdynia includes the AVS part of the health care service architecture and an additional part of the end service.

Any platform for e-government service provisioning should not impose any technological choices on the services that are being provisioned by the service providers. By showing the architecture of the health care service as a typical AVS, it was intended to show how the USE-ME.GOV platform deals with that challenge. The USE-ME.GOV platform does not impose technological choices on health care services, or for that matter on any other future AVS. The internal construction of AVS or end services are transparent from USE-ME.GOV's point of view.

RELATED WORK

Years of research and development in the domain of Web services have resulted in the existence of many platforms from various vendors that allow for deployment and execution of Web services. At the same time in the past few years, there has been a huge increase in the use of mobile services as well as in the mobile communication systems that support them. As a consequence of this, there has appeared the need for solutions that bring together theses two different worlds (Abramowicz, Bassara, Filipowska, Wisniewski, & Zebrowski, 2006). One of the main movements in this field is the existence of the Parlay Group that acts dynamically in a favor of the creation of some standardized interfaces to help opening mobile operators' networks to enterprises and content providers. As a result, there came into being several open specifications devoted to the interfaces between the mobile networks and the Internet. The existence of these APIs permits the linking up of IT applications with the capabilities of the telecommunications world as well as generating new possibilities for revenue streams. The API for the area of Web services is Parlay X (Parlay, 2005) which is a set of standardized Web services that provides developers with access to telecom functionalities available in an operator's domain such as short messaging, multimedia messaging, call control, terminal location, and so on. This specification has been incorporated into several existing commercial platforms, for instance, Oracle9iAS (Oracle, 2004), IBM WebSphere (IBM, 2005), or HP Mobile Services Delivery Platform (HP, 2003). According to the vendors, their solutions are robust, scalable, flexible, versatile, and easy to maintain and deploy. The drawback is that they are commercial products and so, in most cases, that also means that they are expensive. Unfortunately, to the best of our knowledge, developers do not have other choices, as there is no open solution that can be considered to be mature enough to be used as a platform, on the one hand, for orchestrating Web services and, on the other hand, for integrating with a mobile operator network.

CONCLUSIONS

With the introduction of the USE-ME.GOV project, we tried to make the provision of m-government applications much more convenient than ever before. In this chapter, we presented the general process that can be applied to the development of such platform. In particular, we focused on the usability design, its end users, and authorities aspects. Especially, much attention must be paid to the end users (local authorities) who need to be abstracted from the technical and development details that often obstruct their day-to-day duties.

The main challenges that are highlighted throughout the chapter are requirements studies and platform design. The requirements analysis is particularly important with usability studies becoming the essential aspect of users acceptance. The proper design with openness and interoperability guarantees the acceptance on one hand

and choice on the other. There is no danger of being a "lone island", allowing islands owner to choice whatever technology is suitable at the same time.

REFERENCES

Abramowicz, W., Bassara, A., Filipowska, A., Wisniewski, M., & Zebrowski, P. (2006). Mobility implications for m-government platform design. *Cybernetics and Systems, 37*(2-3), 119-135.

Barton, J. J., Zhai, S., & Cousins, S. B. (2006). *Mobile phones will become the primary personal computing devices.* WMCSA 2006 Workshop on Mobile Computing Systems and Applications, Washington, USA, April 6-7.

D2.2 (2004). *Deliverable 2.2: Service and Use Scenario Definition*. USE-ME.GOV Consortium. Retrieved November 14, 2006, from http://www.usemegov.org

D5.1.1 (2004). *Deliverable 5.1.1: System requirements*. USE-ME.GOV Consortium. Retrieved November 14, 2006, from http://www.usemegov.org

D5.2.4 (2005). *Deliverable 5.2.4: Meta-protocol of Service Types.* USE-ME.GOV Consortium. Retrieved November 14, 2006, from http://www.usemegov.org

European Commission. (2003). *Linking up Europe: The importance of interoperability for egovernment services.* Retrieved November 14, 2006, from http://ec.europa.eu/idabc/en/document/2036/5583

Fraser, N. M., & Gilbert, N. G. (1991). Simulating speech systems. *Computer Speech and Language, 5*, 81-99.

Hagen, S. (2001). *Guide to service location protocol.* San Jose, CA: Podbooks.Com Llc.

Harsha, T., & Joel, J. (2005). Developing spatially-aware content management systems for dynamic, location-specific information in mobile environments. In the *Proceedings of the 3rd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, Cologne, Germany, September 2 (pp. 14-22). New York: ACM Press.

Hone, K. S., & Barber, C. (2001). Designing habitual dialogues for speech-based interaction with computers. *International Journal of Human-Computer Studies, 54,* 637-662.

HP. (2003). *Mobile service delivery platform.* Retrieved November 14, 2006, from http://www.hp.com/

IBM. (2005). *Telecom Web services toolkit preview.* Retrieved November 14, 2006, from http://www.alphaworks.ibm.com/

IDABC. *Interoperable Delivery of European eGovernment Services to public Administrations, Businesses and Citizens.* Retrieved November 14, 2006, from http://europa.eu.int/idabc/

IEEE. (1990). *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries.* New York, NY: Institute of Electrical and Electronics Engineers.

Larman, C. (2005). *Applying UML and patterns* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Leenes, R. E., & Svensson, J. S. (2002). Size matters – Electronic service delivery by municipalities? In the *Proceedings: Electronic Government – First International Conference, EGOV 2002*, Aix-en-Provence, France, September 2-5 (pp. 150-156). Berlin Heidelberg: Springer.

Maciaszek, L. (2005). *Requirements analysis and system design* (2nd ed.). Harlow, England: Addison Wesley.

Oracle. (2004). *Oracle application server wireless 10g Parlay and Parlay X*. Retrieved November 14, 2006, from http://www.oracle.com/

Oviatt, S. L., DeAngeli, A., & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. In the *Proceedings of Conference on Human Factors in Computing Systems (CHI '97),* Atlanta, Georgia, USA, March 22-27 (pp. 415-422). New York: ACM Press.

Parlay. (2005). *Parlay X Web Services Specification, Version 2.0*. The Parlay Group. Retrieved November 14, 2006, from http://www.parlay.org/en/specifications/

UDDI. *Universal Description, Discovery and Integration*. Retrieved November 14, 2006, from http://www.uddi.org/

USE-ME.GOV Project Deliverables. (2004-2006). Retrieved November 14, 2006, from http://www.usemegov.org/

Weiser, M. (1993). Hot topics: Ubiquitous computing. *IEEE Computer*, October, 71-72.

Wohlever, S., Fay-Wolfe, V., Thuraisingham, B., Freedman, B., & Maurer, J. (1999). CORBA-based real-time trader service for adaptable command and control systems. *Second IEEE International Symposium on Object-Oriented Real-Time Distributed Computing*, Saint-Malo, France, May 2-5 (p. 64). Los Alamitos, CA: IEEE Computer Society.

WSA. (2004). *Web Services Architecture*. Retrieved November 14, 2006, from http://www.w3.org/TR/ws-arch/

Yankelovich, N. (1998). Using natural dialogs as the basis for speech interface design. In S. Luperfoy (Ed.), *Automated spoken dialog systems* (pp. 30-56). Cambridge, MA: MIT Press.

ADDITIONAL READING

D4.1.1 (2004). *Deliverable 4.1.1: Review of State of the Art in User Interface Design for Mobile Applications*. USE-ME.GOV Consortium. Retrieved November 14, 2006, from http://www.usemegov.org

D4.1.2 (2004). *Deliverable 4.1.2: Usability Requirements Definition for Selected Scenarios*. USE-ME.GOV Consortium. Retrieved November 14, 2006, from http://www.usemegov.org

D4.1.3 (2004). *Deliverable 4.1.3: Usability Driven Design and Mock-Up Evaluation*. USE-ME.GOV Consortium. Retrieved November 14, 2006, from http://www.usemegov.org

D6.1.2 (2004). *Deliverable 6.1.2: Pilot Services Requirements Specifications*. USE-ME.GOV Consortium. Retrieved November 14, 2006, from http://www.usemegov.org

ENDNOTES

1. http://www.hyperdictionary.com/computing/interoperability
2. http://www.anuit.it/conv0312/hebert03a/tsld004.htm
3. http://www.w3.org/Submission/OWL-S/
4. W3C—World Wide Web Consortium (http://www.w3.org/).
5. OASIS—Organization for the Advancement of Structured Information Standards (http://www.oasis-open.org).

This work was previously published in Mobile Government: An Emerging Direction in E-Government, edited by I. Kushchu, pp. 30-59, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.21
Mobile Computing for M–Commerce

Anastasis Sofokleous
Brunel University, UK

Marios C. Angelides
Brunel University, UK

Christos Schizas
University of Cyprus, Cyprus

INTRODUCTION

The ubiquitous nature of modern mobile computing has made "any information, any device, any network, anytime, anywhere" a well-known reality. Traditionally, mobile devices are smaller, and data transfer rates are much lower. However, mobile and wireless networks are becoming faster in terms of transfer rates, while mobile devices are becoming smaller, more compact, less power consuming, and, most importantly, user-friendly. As more new applications and services become available every day, the number of mobile device owners and users is increasing exponentially. Furthermore, content is targeted to user needs and preferences by making use of personal and location data. The user profile and location information is becoming increasingly a necessity.

The aim of this article is to present an overview of key mobile computing concepts, in particular, those of relevance to m-commerce. The following sections discuss the challenges of mobile computing and present issues on m-commerce. Finally, this article concludes with a discussion of future trends.

CHALLENGES OF MOBILE COMPUTING

Current mobile devices exhibit several constraints:

- Limited screen space: screens cannot be made physically bigger, as the devices must fit into hand or pocket to enable portability

(Brewster & Cryer, 1999)
- Unfriendly user interfaces
- Limited resources (memory, processing power, energy power, tracking)
- Variable connectivity performance and reliability
- Constantly changing environment
- Security

These constraints call for immediate development of mobile devices that can accommodate high quality, user-friendly ubiquitous access to information, based on the needs and preferences of mobile users. It also is important that these systems must be flexible enough to support execution of new mobile services and applications based on a local and personal profile of the mobile user.

In order to evaluate the challenges that arise in mobile computing, we need to consider the relationships between mobility, portability, human ergonomics, and cost. While the mobility refers to the ability to move or be moved easily, portability relates to the ability to move user data along with the users. A portable device is small and lightweight, a fact that precludes the use of traditional hard-drive and keyboard designs. The small size and its inherent portability, as well as easy access to information are the greatest assets of mobile devices (Newcomb et al., 2003). Although mobile devices were initially used for calendar and contact management, wireless connectivity has led to new uses, such as user location tracking on-the-move. The ability to change locations while connected to the Internet increases the volatility of some information. As volatility increases, the cost-benefit trade of points shift, calling for appropriate modifications in the design.

Wireless communications and mobile connectivity are overridden by bandwidth fluctuations, higher loss rates, more frequent and extended disconnections, and network failures that make Quality of Service (QoS) a continuous challenge. As a result, applications must adapt to a continuously changing QoS. Although mobile devices are designed to run light applications in a stand-alone mode, they still make use of wireless communication technologies such as Bluetooth, GPRS, and WiFi, which makes them useful in the new mobile world sphere, but they succumb to QoS limitations as a result of portability.

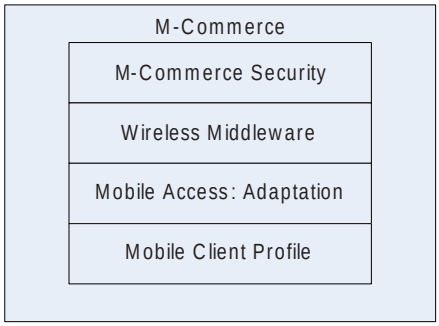
Mobility also is characterized by location transparency and dependency. A challenge for mobile computing is to factor out all the information intelligently and provide mechanisms to obtain configuration data appropriate to the current user location. In fact, in order to resolve a user's location, it is necessary to filter information through several layers: discovering the global position, translating the location, superimposing a map, identifying points of interest for the user and their relative range to that of the user. This suggests a multi-layer infrastructure. A number of location tracking services were developed in order to provide location information transparently to application developers who need to deploy location-aware applications.

M-COMMERCE

Mobile commerce is fast becoming the new trend for buying goods and services. As with e-commerce, it requires security for mobile transactions, middleware for content retrieval, and adaptation using client and device information.

The enormous effect of mobile commerce in our lives can be noticed by studying the effect of m-commerce on industries in a way that will exceed wire-line e-commerce as the method of preference for digital commerce transactions (e.g., financial services, mobile banking), telecommunications, retail and service, and information services (e.g., delivery of financial news and traffic updates). The global m-commerce market is likely to be worth a surprising US $200 billion in 2004 (More Magic Software, 2000). Report statistics confirm that in 2003, over a billion mobile phone users regarded it as a valuable communication

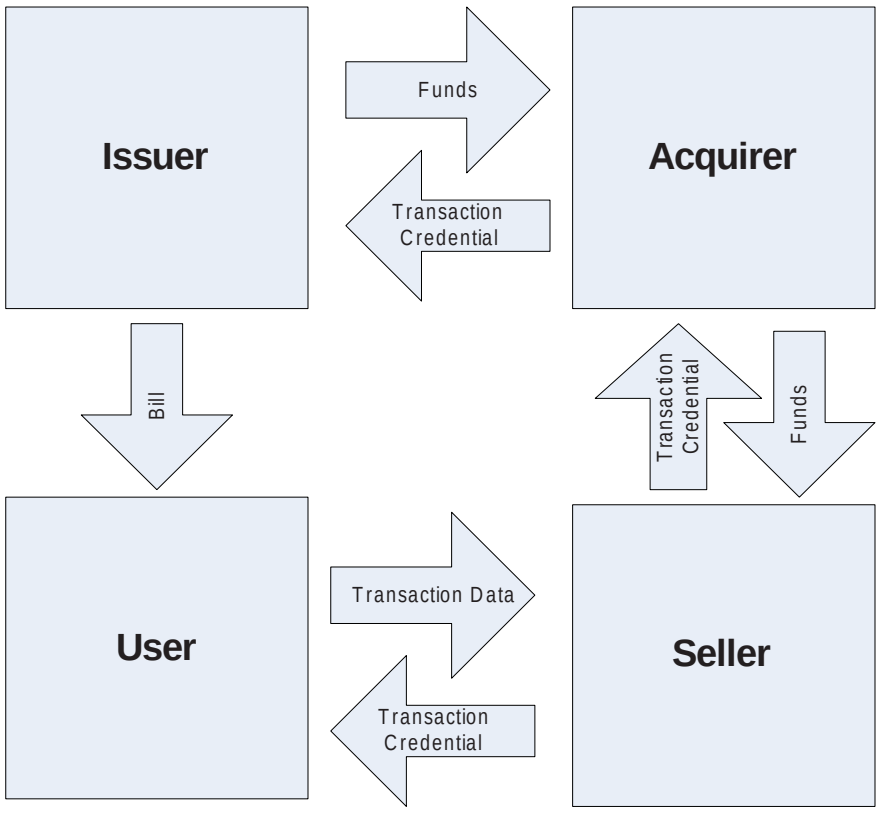
Figure 1. M-commerce



tool. Global mobile commerce revenue projections show revenues up to the 88 billions for 2009 (Juniper Research, 2004).

Mobile security (M-Security) and mobile payment (M-Payment) are essential to mobile commerce and mobile world. Consumers and merchants have benefited from the virtual payments that information technology has conducted. Due to the extensive use of mobile devices nowadays, a number of payment methods have been deployed that allow the payment of services/goods from any mobile device. The success of mobile payments is contingent on the same factors that have fueled the growth of traditional non-cash payments: security,

Figure 2. A classic payment operation



interoperability, privacy, global acceptance, and ease-of-use (Mobile Payment Forum, 2002).

The challenges associated with mobile payments are perhaps better understood using the example of credit card transaction. A card transaction involves at least four parties. As illustrated in Figure 2, the user as a buyer is billed by the card issuer for the goods and services he or she receives from the seller, and the funds are transferred from the issuer to the acquirer, and finally to the merchant. First, the consumer initializes the mobile purchase, registers with the payment provider, and authorizes the payment. A content provider or merchant sells product to the customer. The provider or merchant forwards the purchase requests to a payment service provider, relays authorization requests back to the customer, and is responsible for the delivery of the content. Another party in the payment procedure is the payment service provider, who is responsible for controlling the flow of transaction between mobile consumers, content providers, and trusted third parties (TTP), as well as for enabling and routing the payment message initiated from the mobile device to be cleared by the TTP. A payment service provider could be a mobile operator, a bank, a credit card company, or an independent payment vendor.

Although with mobile payments, the payment transaction is similar to that described in Figure 1, there are some differences with regards to the transport of payment details, as this will involve a mobile network operator and will use either a browser-based protocol such as WAP or HTML or will be done via Bluetooth, WIFI, or infrared. The configuration of the payment mechanism could be achieved with the installation of either an applet or a specific application on a mobile device, and it usually takes place once. The first steps following successful installation include initialization of consumer payment (i.e., transferring payment information over a wireless network), user authentication, and payment completion, including receipt generation.

Existing mobile payment applications are categorized based on the payment settlement methods that they implement: pre-paid (using smart cards or digital wallet), instant paid (direct debiting or offline payments), and post paid (credit card or telephone bill) (Seema & Chang-Tien, 2004). Developers deploying applications using mobile payments must consider security, interoperability, and usability requirements. A secure application will allow an issuer to identify a user, authenticate a transaction, and prevent unauthorized parties from obtaining any information on the transaction. Interoperability guarantees completion of a transaction between different mobile devices or distribution of a transaction across devices. Usability ensures user-friendliness and multi-users.

M-COMMERCE SECURITY

Security is a crucial concern for anyone deploying mobile devices and applications, because personal information has to be delivered to a number of mobile workers engaged in online activities outside the secure perimeter of a corporate area. That increases the threat for unauthorized access and use of private and personal data. In order to authenticate the users accessing shared data, developers are using a number of authentication mechanisms, such as simple usernames and passwords, special single-use passwords from electronic tokens, cryptographic keys, and certificates from public key infrastructures (PKI). Additionally, developers are using authentication mechanisms to determine what data and applications the user can access (after login authorization). These mechanisms, often called policies or directories, are handled by databases that authenticate users and determine their permissions to access specific data simultaneously.

The current mobile business (M-Business) environment runs over the TCP/IPv4 protocol stack, which poses serious security level threats

with respect to user authentication, integrity, and confidentiality. In a mobile environment, it is necessary to have identification and non-repudiation and service availability, mostly a concern for Internet and or application service providers. For these purposes, carriers (telecom operators and access providers), services, application providers, and users demand end-to-end security as far as possible (Leonidou et al., 2003; Tsaoussidis & Matta, 2002).

The technologies used in order to implement m-business services and applications like iMode, Hand-held Device Mark-up Language (HDML) and Wireless Access Protocol (WAP) can secure the transport of data (encryption) between clients and servers, but they do not provide applicable security layers, especially user PIN-protected digital signatures, which are essential to secure transactions. Therefore, consumers cannot acknowledge transactions that are automatically generated by their mobile devices. Besides the characteristics of the individual mobile devices, some of the securities issues issued are dependent on the connectivity between the devices. Internet2 and IPv6 also have many security concerns, such as the authentication and authorization of binding updates sent from mobile nodes and the denial-of-service attack (Roe et al., 2002).

It is important to incorporate security controls when developing mobile applications rather than deploying the applications before and without fitting security. Fortunately, it is now becoming possible to implement security controls for mobile devices that do afford a reasonable level of protection in each of the four main problem areas: virus attacks, data storage, synchronization, and network security (Brettle, 2004).

WIRELESS MIDDLEWARE

Content delivery and transformation of applications to wireless devices without rewriting the application can be facilitated by wireless middleware. Additionally, a middleware framework can support multiple wireless device types and provide continuous access to content or services (Sofokleous et al., 2004). The main functionality of wireless middleware is the data transformation shaping a bridge from one programming language to another and, in a number of circumstances, is the manipulation of content in order to suit different device specifications. Wireless middleware components can detect and store device characteristics in a database and later optimize the wireless data output according to device attributes by using various data-compression algorithms, such as Huffman coding, Dynamic Huffman coding, Arithmetic coding, and Lempel-Ziv coding. Data compression algorithms serve to minimize the amount of data being sent over the wireless link, thus improving overall performance on a handheld device. Additionally, they ensure end-to-end security from handheld devices to application servers, and finally, they perform message storage and forwarding, should the user get disconnected from the network. They provide operation support by offering utilities and tools to allow MIS personnel to manage and troubleshoot wireless devices. Choosing the right wireless middleware is dependent on the following key factors: platform language, platform support and security, middleware integration with other products, synchronization, scalability, convergence, adaptability, and fault tolerance (Lutz,2002; Vichr & Malhotra, 2001).

MOBILE ACCESS ADAPTATION

In order to offer many different services to a growing variety of devices, providers must perform an extensive adaptation of both content (to meet the user's interests) and presentation (to meet the user device characteristics) (Gimson, 2002). The network topology and physical connections between hosts in the network must be constantly recomputed, and application software

must adapt its behavior continuously in response to this changing context (Julien et al., 2003) either when server-usage is light, or if users pay for the privilege (Ghinea & Angelides, 2004).

The developed architecture of m-commerce communications exploits user perceptual tolerance to varying QoS in order to optimize network bandwidth and data sizing. This will provide QoS impacts upon the success of m-commerce applications without doubt, as it plays a pivotal role in attracting and retaining customers. As the content adaptation and, in general, the mobile access personalization concept are budding, central role plays the utilization of the mobile client profile, which is analyzed in the next section.

MOBILE CLIENT PROFILE

The main goal of profile management is to offer content targeted to users' needs and interests, using a presentation that matches their mobile device specification. Usually, this is done by collecting all the data that can be useful for identifying the content and the presentation that best fit the user's expectations and the device capabilities. The information may be combined with the location of the user and the action context of the user at the time of the request (Agostini et al., 2003).

In order to have a complete user profile, different entities are assembled from different logical locations (i.e., the personal data is provided by the user, whereas the information about the user's current location is usually provided by the network operator). Providers should query these entities to get the required information for a user. Several problems and methods for holding back the privacy of data are raised, as mobile devices allow the control of personal identifying information (Srivastava, 2004). People are instantly concerned about location privacy generated by location tracking services.

FUTURE TRENDS

During the past decade, computing and mobile computing have changed the business and consumer perception, and there is no doubt that mobile computing has already exceeded most expectations. Architectures and protocol standards, management, services, applications, and the human factor make possible the evolution of mobility (Angelides, 2004). The major areas that will be involved are the hardware, the middleware, the operating system and the applications.

In the area of software, while likely applications are being deployed, mobile services and applications will progressively distribute a variety of higher bandwidth applications, such as multimedia messaging, online gaming, and so forth. Several applications, such as transactional applications (financial services/banking, home shopping, instant messages, stock quotes, sale details, client information, location-based services, etc.) have already showed a tremendous potential for growth. Unfortunately, applications are restricted by the available hardware and software resources. As

Figure 3. Areas of mobility evolution

a result, portable devices must be robust, reliable, user-friendly, enchanting, functional, and expandable.

Additionally, mobile computing devices will have to provide a similar level of security and interoperability as usual handsets, combined with a performance level approaching that of desktop computers. The variety of wireless connectivity solutions, the operating systems, the presentation technologies, the processors, the battery technologies, the memory options, and the user interfaces are analyzed and examined, as they will enable the growth of mobile computing. The operating system also is largely dependent on the hardware, but it should be scalable, customizable, and expandable.

Third generation mobile communication systems, such as Universal Mobile Telecommunications System (UMTS), will provide optimum wireless transmission speeds up to 2 Mbits/s, and they will have voice and video connections to the mobile devices.

The future of mobile computing is looking very promising, and as wireless computing technology is being gradually deployed, the working lifestyle may change, as well.

CONCLUSION

This article discusses the more important issues that affect m-commerce. The ability to access information on demand while mobile will be very significant. IT groups need to understand the ways mobile and wireless technology could benefit m-commerce and avoid deploying wireless on top of wired, which adds incremental costs. Mobile application frameworks create a range of new security exposures, which have to be understood and taken under consideration during the design steps of the mobile frameworks. In the general view, e-commerce is concerned with trading of goods and services over the Web and the m-commerce with business transactions conducted while on the move. However, the essential difference between e-commerce and m-commerce is neither the wire nor the wireless aspects, but the potential to explore opportunities from a different perspective.

Companies need to customize the content in order to meet the requirements imposed by bandwidth and the small display size of mobile devices. Mobile services and applications, such as location management, locations, profile-based services, and banking services are some of the applications that have great potential for expansion. The demand of m-business applications and services will grow, as new developments in mobile technology unfold. Nowadays, challenging mobile payment solutions have already established their position in the marketplace. As software systems are becoming more complex and need to extend to become wireless, in some instances, it may be useful to use a wireless middleware. What we are currently observing is mobile computing becoming increasingly pervasive among businesses and consumers.

REFERENCES

Agostini, A., Bettini, C., Cesa-Bianchi, N., Maggiorini, D., & Riboni, D. (2003). Integrated profile management for mobile computing. *Proceedings of the Workshop on Artificial Intelligence, Information Access, and Mobile Computing*, Acapulco, Mexico.

Angelides, M.C. (2004). Mobile multimedia and communications and m-commerce. *Multimedia Tools and Applications, 22*(2), 115-116.

Brettle, P. (2004). White paper on mobile security. Insight consulting. Retrieved from *http://www.insight.co.uk*

Brewster, A.S., & Cryer, P.G. (1999). Maximizing screen-space on mobile computing devices. *Pro-*

ceedings of the Conference on Human Factors in Computing Systems, Pittsburgh. New York.

Dahleberg, T., & Tuunainen, V. (2001). Mobile payments: The trust perspective. *Proceedings of the International Workshop Seamless Mobility,* Sollentuna, Spain.

Ghinea, G., & Angelides, C.M. (2004). A user perspective of quality of service in m-commerce. *Multimedia tools and applications, 22*(2), 187-206.

Gimson, R. (2002). Delivery context overview for device independence [W3C working draft]. Retrieved September 12, 2002 from *http://www.w3.org/2001/di/public/dco/dco-draft-20020912/*

Julien, C., Roman, G., & Huang, Q. (2003). Declarative and dynamic context specification supporting mobile computing in ad hoc networks [Technical Report WUCSE-03-13]. St. Louis, MO, Washington University.

Juniper Research. (2004). The big micropayment opportunity [White paper]. Retrieved September 24, 2002 from *http://industries.bnet.com/abstract.aspx?seid=2552&docid=121277*

Leonidou, C., et al. (2003). A security tunnel for conducting mobile business over the TCP protocol. *Proceedings of the 2nd International Conference on Mobile Business,* Vienna, Austria.

Lutz, E.W. (2002). Middleware for the wireless Web. Faulkner Information Services. Retrieved August 25, 2004 from *http://www.faulkner.com*

Mobile Payment Forum. (2002). Enabling secure, interoperable, and user-friendly mobile payments. Retrieved August 18, 2004 from *http://www.mobilepaymentforum.org/pdfs/mpf_whitepaper.pdf*

More Magic Software. (2000). Payment transaction platform. Retrieved July 25, 2003 from *http://www.moremagic.com/whitepapers/technical_ wp_twp021c.html*

Newcomb, E., Pashley, T., & Stasko, J. (2003). Mobile computing in the retail arena. *ACM Proceedings of the Conference on Human Factors in Computing Systems,* Florida.

Roe, M., Aura, T., & Shea, G.O. (2002). Authentication of Mobile IPv6 Binding Updates and Acknowledgements. (Internet draft). Retrieved August 10, 2004 from *http://research.micro soft.com/users/mroe/cam-v3.pdf*

Seema, N., & Chang-Tien, L. (2004). *Advances in security and payment methods for mobile commerce*. Hershey, PA: Idea Group Publishing.

Sofokleous, A., Mavromoustakos, S., Andreou, A.S., Papadopoulos, A.G., & Samaras, G. (2004). Jinius-Link: A distributed architecture for mobile services based on localization and personalization. *Proceedings of the IADIS International Conference,* Lisbon, Portugal.

Srivastava, L. (2004). Social and human consideration for a mobile world. *Proceedings of the ITU/MIC Workshop on Shaping the Future Mobile Information Society,* Seoul, Korea.

Tsaoussidis, V., & Matta, I. (2002). Open issues on TCP for mobile computing. *Journal of Wireless Communications and Mobile Computing, 2*(1).

Vichr, R., & Malhotra, V. (2001). Middleware smoothes the bumpy road to wireless integration. *IBM*. Retrieved August 11, 2004 from *http://www-106.ibm.com/developerworks/library/wi-midarch/index.html*

KEY TERMS

E-Commerce: The conduct of commerce in goods and services over the Internet.

Localization: The process to adapt content to specific users in specific locations.

M-Business: Mobile business means using any mobile device to make business practice more efficient, easier, and profitable.

M-Commerce: Mobile commerce is the transactions of goods and services through wireless handheld devices, such as cellular telephones and personal digital assistants (PDAs).

Mobile Computing: Mobile computing encompasses a number of technologies and devices, such as wireless LANs, notebook computers, cell and smart phones, tablet PCs, and PDAs, helping the organization of our life, communication with coworkers or friends, or the accomplishment of our jobs more efficiently.

Mobile Device: Mobile device is a wireless communication tool, including mobile phones, PDAs, wireless tablets, and mobile computers (Mobile Payment Forum, 2002).

Mobility: The ability to move or to be moved easily from one place to another.

M-Payment: Mobile payment is defined as the process of two parties exchanging financial value using a mobile device in return for goods or services (Seema Nambiar, paper).

M-Security: Mobile security is the technologies and methods used for securing wireless communication between the mobile device and the other point of communication, such as another mobile client or a pc.

Profile: Profile is any information that can be used to offer a better response to a request (i.e., the information that characterizes the user, the device, the infrastructure, the context, and the content involved in a service request) (Agostini et al.,2003).

Wifi: Wifi (wireless fidelity) is a technology that covers certain types of wireless local area networks (WLANs), enabling users to connect wirelessly to a system or wired local network and use specifications in the 802.11 family.

This work was previously published in Encyclopedia of Multimedia Technology and Networking, edited by M. Pagani, pp. 622-628, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.22
Mobile Commerce Applications and Adoption

Krassie Petrova
Auckland University of Technology, New Zealand

INTRODUCTION

The potential advantages of mobile commerce applications have been discussed extensively in the recent literature, with many industries offering mobile services. Examples from the financial sector include instant funds transfer (mobile banking) and share trading (mobile brokerage). Commuter services such as sending schedule change alerts or using a mobile phone to pay for parking have become widespread. Applications based on the location of the user (e.g., offering mobile coupons to customers in the vicinity of a shop or a restaurant) are also being trialled (Barnes, 2002; Siau, Lim, & Shen, 2001; Varshney, Vetter, & Kalakota, 2000).

Despite the potential benefits (for example, improved customer service) mobile commerce applications have not been widely adopted across business sectors. Mobile banking illustrates the point: initially, seen as the "killer application" of mobile commerce (Kannan, Chang, & Whinston, 2001), it has now been termed a "dead end" (Semrau & Kraiss, 2001). It has also been classified as an application which has not yet matured (Mallat, Rooi, & Tuunainen, 2004). However, innovative applications continue to emerge, for example, breaking news alerts (CNN, n.d.), and a mobile tutoring service (Butte, 2004). It has become important therefore to identify the determinants of mobile commerce adoption and the emerging adoption patterns.

A significant number of results in this area have been reported in the literature. Recent examples include studies of mobile services adoption in areas characterized by relatively high penetration of mobile devices—such as Denmark (Constantiou, Damsgaard, & Knutsen, 2004), Singapore (Samtani, Leow, Lim, & Goh, 2004), and Finland (Carlsson, Hyvonen, Repo, & Walden, 2005). The identified drivers and inhibitors of mobile commerce adoption can be broadly classified as factors related to mobile infrastructure access, and factors relating to perceived consumer value. This article proposes a mobile commerce reference model which incorporates both infrastructure

access and customer value and can be used to formulate research questions related to mobile commerce adoption.

The remainder of the article is organized as follows: first, mobile commerce is defined and compared to electronic commerce. The next section introduces a mobile commerce reference model and discusses mobile commerce adoption. The article continues with a review of future trends and a brief conclusion.

BACKGROUND

The definitions of mobile commerce (m-commerce) found in the literature such as the one suggested in Varshney et al. (2000), emphasize the use of mobile telephony and a handheld device to execute transactions with monetary value (i.e., exchange of funds for goods and services). M-commerce services are offered to subscribers only.

Turban, Lee, and Viehland (2004, p. 399) classify m-commerce as a subset of electronic commerce (e-commerce). However several features of m-commerce are either not found, or are not strongly manifested in e-commerce. These include "ubiquity"—which allows the user to interact with a mobile application anywhere, even when travelling or moving (Schneiderman, 2000, p. 1); "localization"—the ability of an application to offer a service specific to the location of the customer (Köhne, Totz, & Wehmeyer, 2005) and "personalization"—the ability to tailor an m-commerce activity according to a customer profile, and use the subscriber's account for payment (Siau et al., 2001).

The m-commerce characteristics described above (ubiquity, localization, and personalization) and the profile of the potential m-commerce user as a paid mobile network subscriber provide the grounds on which to differentiate between e-commerce and m-commerce. In this article, m-commerce is defined as a value-added service that enables mobile users to conduct reliable and secure transactions through specifically-designed mobile applications. The definition implies that a company or an organization offering a mobile service needs to develop and implement an appropriate business model which will incorporate the value proposition of the service, the revenue model, and the interactions of the company with business partners, suppliers and customers (Veijalainen, Terziyan, & Tirri, 2003).

MODELLING M-COMMERCE ADOPTION

Even the most innovative and creative mobile application or service will only be commercially successful if brought to customers through a business model that clearly focuses on the added value generated and offered by the application or service. Furthermore, the adoption of the application will depend on additional factors such as whether it is accessible from all locations, or whether it depends on the specific features of the handheld device—(e.g., WAP functionality or a small screen). General factors such as security awareness, privacy, and trust concerns might also play a role (Giaglis, 2005; Lin, 2004). To be viable, an m-commerce business model needs to:

1. Take full advantage of user mobility.
2. Offer services which would be either unavailable or prohibitively expensive if offered by means of e-commerce or brick-and-mortar commerce.
3. Offer services overcoming drawbacks caused by security and privacy related issues.

The degree to which the requirements above are met will influence the adoption of a particular m-commerce application and will act as a viability determinant of the associated business model. The investigation of the process of value creation and subsequent adoption needs to consider both

technological and social factors (Carlsson et al., 2005; Pedersen, Methlie, & Thorbjornsen, 2002) and needs to include the different players involved: network providers and operators, content contributors and aggregators, portal hosts and application developers. One of the approaches is to consider the interactions among the players and their roles in the value chain model of m-commerce (Barnes, 2002, 2003).

A REFERENCE MODEL FOR M-COMMERCE

The value chain approach breaks down m-commerce into a structured chain of entities with associated "actors" and allows the researcher to identify easily and conveniently the companies and organizations involved in creating mobility-related value (Barnes, 2002; Buellingen & Woerter, 2004; Olsson & Nilsson, 2002; Siau et al., 2001).

The reference model (Figure 1) places together the players involved in the value chain, and captures the features of technologies, applications and services related to m-commerce. It incorporates three basic layers: an infrastructure layer (devices and networks), an interface layer (mobile middleware and platforms), and a business layer (services, content, application-based business models). Direct interactions with subscribers/customers occur mostly at the infrastructure and business layers where the value chain players act as enablers and direct providers, respectively. At the interface layer, the actors perform the role of intermediaries. The layered structure complies with the m-commerce definition in the previous section and enables the systematic investigation of the processes of value-creation across the m-commerce value chain. The model can be used to develop evaluation criteria and study m-commerce applications and their adoption within an industry segment, at national or at a regional level.

Other proposed approaches towards conceptual modelling of m-commerce include the bundled value proposition (Anckar & D'Incau, 2002), the open-plane framework (Varshney & Vetter, 2002), the reference model for m-commerce applications (Stanoevska-Slabeva, 2003), and extended three-dimensional models (Chen, Lee, & Cheung, 2001; Tarasewich, Nickerson, & Varkentin, 2002). The reference model introduced above is somewhat similar to Stanoevska-Slabeva's and Varshney and Vetter's models but is more comprehensive.

M-COMMERCE ADOPTION STUDIES

Research in the area has been mostly based on prior work in the areas of adoption and diffusion and the technology acceptance model (Pedersen & Ling, 2003; Pedersen, Methlie, & Thorbjornsen, 2002). Though based on different background concepts, the research directions in adoption studies have much in common as they focus on end users and customers in different contexts.

The m-commerce reference model (Figure 1) includes two customer contexts where customers interact directly with the model constructs: "infrastructure" (the customer as a subscriber) and "business" (the customer as user of mobile application). The two interaction types correspond to the "technology user" and "consumer" perspectives defined by Pedersen et al. (2002). Factors influencing adoption related to the two perspectives have been identified as: (1) interoperability of devices and protocols, bandwidth availability, device features and functions, connectivity (technology); and (2) content personalization and localization, service ubiquity, timeliness, convenience, cost, privacy issues (consumer) (Chen, Lee, & Cheung, 2001, Petrova, 2004a, Turban et al., 2004, p. 423).

To improve the understanding of the adoption process and "move from description of the process into explaining it" (Pedersen & Ling, 2003), and

Figure 1. A reference model for m-commerce

Business layers Companies and organizations interact indirectly with customers (consumers of mobile services and end-users of mobile applications)		
7	Business Model	Companies/organizations offering an mCommerce application to customers (*direct providers*).
6	Mobile Content	Companies/organizations providing or developing content for the application (*intermediaries*).
5	Mobile Service	Companies/organizations offering the application to the customer, or providing a related service (*intermediaries*).
Interface layers Companies and organizations mostly interact indirectly with users (consumers of mobile services and applications)		
4	Application Platform	Developers of portals, integrators, and/or network providers (*enablers*).
3	Mobile Middleware	Developers of general purpose middleware or specialized consortia (include network providers and device manufacturers).
Infrastructure layers Companies and organizations interact directly or indirectly with subscribers		
2	Mobile Device	Manufacturers and vendors.
1	Mobile Network	Providers and vendors (*enablers*).
	Layers	Players in the value chain (enablers, direct providers, intermediaries)

to include the value chain perspective, the broader research questions can be formulated as:

1. Which of the factors contributing to the creation of mobility-related value are critical success factors for the viability of an m-commerce business model?
2. What other specific factors contribute to (or inhibit) business model viability, for example environmental factors (e.g., legislation) and demographic factors (e.g., age, gender)?
3. How might m-commerce actors such as "intermediaries" and "enablers" encourage or inhibit adoption processes?

4. What is the role of m-commerce payment mechanisms in the adoption processes?

Obtaining the answers to those questions might help develop appropriate applications and business models, taking into account not only customer perceptions but the roles played by all actors in the value chain (Heikkilä, Heikkilä, & Lehmonen, 2004). An example of such an innovative application is "interactive mobile TV" (Kihlström, 2005) which is driven by a partnership between the vendor (Ericsson) and content providers (media companies). Interactive mobile TV uses the handset screen to show images or information, to download multimedia content, to enable voting for TV shows, and to display advertising messages. Another developing area is mobile education (Kurbel & Hilker, 2003). Examples of mobile learning scenarios include the use of text messaging for interactive revision (Petrova, 2004b), and mobile Internet access to online seminars (Hino, Terashima, & Bunno, 2002).

FUTURE TRENDS

Using the reference model in Figure 1 as a framework, some of the current trends in m-commerce can be summarized as follows:

- **Layers 1 and 2 (Infrastructure):** To take advantage of new technology developments such as 3G networks and smart phones, developers focus on improving browsing capabilities (Lai et al., 2004) and improving the functionality of handheld devices (personal digital assistants—PDAs, smart phones).
- **Layers 3 and 4 (Interface):** Both general purpose mobile middleware and specific platforms are developed. Raatikainen, Baerbak and Nakajima (2002) state that future systems research will focus on software systems able to provide a seamless service in environments which are both dynamic and heterogeneous, specifically aiming to support large scale applications such as home entertainment. Tarumi, Matsubara and Yano (2004) envisage developing platforms such as the "virtual city", which will serve as a common infrastructure for location based services, including entertainment and advertising. The Japanese mobile portal and services provider NTT DoCoMo continues to expand its i-mode services, including multimedia and payment (NTTDoCoMo, n.d.).
- **Layers 5, 6, and 7 (Business):** Developments in the m-commerce landscape occur in different industry sectors, two of which provide interesting and innovative examples: entertainment, and tourism and travel.

Entertainment applications include streamed music, downloadable or interactive games, streamed movie shows, and chat rooms. Data indicate that the market for mobile entertainment services might be significantly fragmented based on professional background, age and culture, therefore customer preferences and requirements need to be better understood (Leavitt, 2003; Moore & Rutter, 2004; Vlachos & Vrechopoulos, 2004).

In tourism and travel, applications include providing maps or textual guides for tourists and visitors, including vehicle drivers. Such services are location based. The geographical location of the customer determines the content of the application (e.g., directions to the nearest hardware store, or instructions for reaching a particular destination). A possible impediment to the spread of these applications is the relative lack of compatibility across devices and provider networks (Brown & Chalmers, 2003; Köhne et al., 2005; Pavón, Corchado, Gómez-Sanz, & Ossa, 2004; Tarumi et al., 2004).

Other industries with potential to develop successful applications include mobile learning (Leung & Chan, 2003) and event management (Olsson & Nilsson, 2002).

Current research focuses on theory building and analysis and classification of the m-commerce landscape (Camponovo, Debetaz, & Pigneur, 2004), on general adoption models and patterns (Vrechopoulos, Constantiou, Mylonopoulos, Sideris, & Doukidis, 2002), and on services adoption in a specific industry-country context (for example, banking in Finland—Suoranta, Mattila, & Munnukka, 2005).

CONCLUSION

This article defines m-commerce and compares it to e-commerce. It identifies the important characteristics of m-commerce applications and the main research perspectives in application adoption studies. While it is recognized that technology is the primary driver behind m-commerce development, it has also become clear that the adoption processes are aligned with socioeconomic factors. The proposed layered reference model for m-commerce accommodates all actors in the m-commerce value chain and includes the m-commerce adopter as a mobile technology user and as a mobile services consumer. It can be used to derive research questions related to the adoption of emerging applications. The future trends in infrastructure development–such as improved device functionality and greater network capacity—will be able to support innovative and consumer-attractive applications across industry sectors including entertainment, travel, education and financial services.

REFERENCES

Anckar, B., & D'Incau, D. (2002). Value-added services in mobile commerce: An analytical framework and empirical findings from a national consumer survey. *Proceedings of the 35th Hawaii International Conference on System Sciences* (pp. 1087-1096).

Barnes, S. (2002). The mobile commerce value chain: Analysis and future developments. *International Journal of Information Management, 22*(2), 91-108.

Barnes, S. J. (2003). The mobile commerce value chain in consumer markets. In *M-business: The strategic implications of wireless technologies* (pp. 13-37). Burlington, Boston: Butterworth-Heinemann.

Brown, B., & Chalmers, M. (2003). Tourism and mobile technology. In K. Kuutti & E. H. Karsten (Eds.), *Proceedings of the 8th European Conference on Computer Supported Cooperative Work* (pp. 335-355). Helsinki, Finland.

Buellingen, F., & Woerter, F. (2004). Development perspectives, firm strategies, and applications in mobile commerce. *Journal of Business Research, 57*, 1402-1408.

Butte, R. (2004). Dialing up better college test scores. *California Virtual Campus.* Retrieved May 10, 2005, from http://www.cvc.edu/catalog/mnews.asp?mode=view &idx=2985.

Camponovo, G., Debetaz, S., & Pigneur, Y. (2004). A comparative analysis of published scenarios for m-business. *Proceedings of the 3rd International Conference on Mobile Business,* New York. Retrieved May 1, 2005, from http://www.hec.unil.ch/gcampono/index.php?option=content&task=view&id=12.

Carlsson, C., Hyvonen, K., Repo, P., & Walden, P. (2005). Asynchronous adoption patterns of mobile services. *Proceedings of the 38th Annual Hawaii International Conference on Systems Sciences* (pp. 189a-199a).

Chen, Z., Lee, M., & Cheung, C. (2001). A framework for mobile commerce. *Proceedings of the*

7th Americas Conference on Information Systems (pp. 443-449).

CNN (n.d.). *CNN Mobile.* Retrieved May 13, 2005, from http://edition.cnn.com/mobile/

Constantiou, I., Damsgaard, J., & Knutsen, L. (2004). Strategic planning for mobile services adoption and diffusion: Empirical evidence from the Danish market. *Proceedings of the 2004 IFIP TC8 Working Conference on Mobile Information Systems* (pp. 245-256).

Giaglis, G. M. (2005). Critical success factors and business models for mobile and wireless applications. *International Journal of Management and Decision Making, 6*(1), 1-6.

Heikkilä, J., Heikkilä, M., & Lehmonen, J. (2004). Joint development of novel business models. *Proceedings of the 4th IFIP Conference on E-Commerce, E-Business, E-Government* (pp. 433-454).

Hino, K., Terashima, K., & Bunno, T. (2002). Online seminars with the use of Internet-connectable mobile phones. *Proceedings of the International Conference on Information technology applications,* Paper 202-7.

Kannan, P., Chang, A-M., & Whinston, A. (2001). Wireless commerce: Marketing issues and possibilities. *Proceedings of the 34th Hawaii International Conference on System Sciences* (pp. 3526-3531).

Kihlström, L. M. (2005, May). Interactive mobile TV creates buzz at Milia. *Ericsson Mobility Global Newsletter.*

Köhne, F., Totz, C., & Wehnmeyer, K. (2005). Consumer preferences for location-base service attributes: A conjoint analysis. *Journal of Management and Decision Making, 6*(1), 16-32.

Kurbel, K., & Hilker, J. (2003). Requirements for mobile e-learning platform. In M. Hamza, (Ed.), *Proceedings of the IASTED Conference on Communications, Internet, and Information Technology* (pp. 467-471).

Lai, A. M., Nieh, J., Bohra, B., Nandikonda, V., Surana, A. P., & Varshneya, S. (2004). Improving Web browsing on wireless PDAs using thin-client computing. *Proceedings of the 13th World Wide Web Conference* (pp. 143-154).

Leavitt, N. (2003). Will wireless gaming be a winner? *Computer, 36*(1), 24-27.

Leung, C. H., & Chan, Y. Y. (2003). Mobile learning: A new paradigm in electronic learning. In V. Devedzic, J. Spector, D. Sampson, & Kinshuk (Eds.), *Technology enhanced learning. Proceedings of the 3rd International Conference on Advanced Learning Technologies* (pp. 76-80).

Lin, B. (2004). Mobile computing and networking. *International Journal of Electronic Business, 2*(3), 227-228.

Mallat, N., Rooi, M., & Tuunainen, V. K. (2004). Mobile banking services. *Communications of the ACM, 47*(5), 42-46.

Moore, K., & Rutter, J. (2004). Understanding consumers' understanding of mobile entertainment. In K. Moore & J. Rutter (Eds.), *Proceedings of Mobile Entertainment: User-centred Perspectives* (pp. 49-65).

NTTDoCoMo (n.d.) *Press Center.* Retrieved June 21, 2005, from http://www.nttdocomo.com/press-center/index.html

Olsson, D., & Nilsson, A. (2002). MEP: A media event platform. *Mobile Networks and Applications, 7*(3), 235-244.

Pavón, J., Corchado, J. M., Gómez-Sanz, J. J., & Ossa, L. F. C. (2004). Mobile tourist guide services with software agents. *Lecture Notes in Computer Science, 3284,* 322-330.

Pedersen, P., & Ling, R. (2003). Modifying adoption research for mobile Internet service adoption: Cross-disciplinary interactions. *Proceedings*

of the 36th Hawaii International Conference on System Sciences, CD-ROM, 10pp.

Pedersen, P., Methlie, L., & Thorbjornsen, H. (2002). Understanding mobile commerce end-user adoption: A triangulation perspective and suggestions for an exploratory service evaluation framework. *Proceedings of the 35th Hawaii International Conference on System Sciences* (pp. 1079-1086).

Petrova, K. (2004a). Mobile commerce adoption: End-user/customer views. In N. Delener & C. N. Chao (Eds.), Beyond boundaries: Navigating crisis and opportunities in global markets: Leadership, strategy and governance. *Proceedings of the 2004 International Global Business and Technology Association Conference* (pp. 604-615).

Petrova, K. (2004b). Mobile learning using SMS: A mobile business application. In S. Mann & T. Clear (Eds,), *Proceedings of the 17th Annual New Zealand National Advisory Committee on Computer Qualifications Conference* (pp. 421-426).

Raatikainen, K., Baerbak, H., & Nakajima, T. (2002). Application requirements for middleware for mobile and pervasive systems. *Mobile Computing and Communications Review, 6*(4), 16-24.

Samtani, A, Leow, T. T., Lim, H. M., & Goh, P. G. J. (2003). Overcoming barriers to the successful adoption of mobile commerce in Singapore. *International Journal of Mobile Communication, 1*(1/2), 194-231.

Schneiderman, R. (2000). *The mobile technology question and answer book: A survival guide for business managers.* New York: AMACOM.

Semrau, M., & Kraiss, A. (2001). Mobile commerce for financial services—Killer application or a dead end? *ACM SIGGROUP Bulletin, 22*(1) (abstract).

Siau, K., Lim, E. P., & Shen, Z. (2001). Mobile commerce: Promises, challenges and research agenda. *Journal of Database Management, 12*(3), 4-13.

Stanoevska-Slabeva, K. (2003). Towards a reference model for m-commercece applications. *Proceeding of the 2003 European Conference on Information Systems.* Retrieved March 1, 2004, from inforge.unil.ch/yp/Terminodes/papers/03ECISSG.pdf

Suoranta, M., Mattila, M., & Munnukka, J. (2005). Technology-based services: A study on the drivers and inhibitors of mobile banking. *International Journal of Management and Decision Making, 6*(1), 33-46.

Tarasewich, P., Nickerson, R., & Warkentin, M. (2002). Issues in mobile commerce. *Communications of the Association for Information Systems, 8*, 41-64.

Tarumi H., Matsubara, K., & Yano, M. (2004). Implementations and evaluations of location-based virtual city system for mobile phones. *Proceedings of the 47th Global Telecommunications Conference Workshops. Workshop on Network Issues in Multimedia Entertainment,* Dallas, Texas (pp. 544-547).

Turban, E., Lee, J., & Viehland, D. (2004). *Electronic commerce: A managerial perspective* (3rd ed). NJ: Prentice Hall.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications, and networking support. *Mobile Networks and Applications, 7*(3), 185-198.

Varshney, U., Vetter, R., & Kalakota, R. (2000). Mobile commerce: A new frontier. *Computer, 33*(10), 32-38.

Veijalainen, J., Terziyan, V., & Tirri, H. (2003). Transaction management for m-commerce at a mobile terminal. *Proceedings of the 36th Hawaii International Conference on System Sciences* (pp. 89-98).

Vlachos, P., & Vrechopoulos, A. (2004). Emerging customer trends toward mobile music services. *Proceedings of the 6th International Conference on Electronic Commerce* (pp. 566-574).

Vrechopoulos, A., Constantiou, I., Mylonopoulos, N., Sideris, I., & Doukidis, G. (2002). The critical role of consumer behavior research in mobile commerce. *International Journal of Mobile Communications, 1*(3), 329-340.

KEY TERMS

3G: Stands for "third generation" mobile telephony—a wireless communication technology which supports multimedia, video streaming, and video-conferencing.

i-mode: A packet-switching wireless technology, used by NTT DoCoMo (Japan). A range of commercial and financial services are offered, including browsing the Web from a mobile phone.

LBS: Stands for "Location Based Services"—applications which can obtain information about the customer location and use it to customize the service offered.

Microbrowser: Client software (Web browser), designed to operate within the constraints of a handheld mobile device: low memory, small screen, relatively low bandwidth.

MMS: Stands for multimedia message service; similar to text messaging (SMS) but allows the transmission of graphics, sound files, video clips and text. It is based on WAP and can be used to send e-mail.

Mobile Applications: A broad range of applications and services accessible through a mobile handheld device. Examples include banking, news, betting, games, travel directions.

Mobile Commerce: Broadly refers to any value-added service providing access to a mobile application.

Smart Phone: An enhanced handheld device which combines the functions of a mobile phone and a handheld computer.

SMS: Stands for short message service (also known as text messaging, or "texting").

WAP: Stands for wireless application protocol. A set of standards which enable data display for handheld devices and support Web and e-mail access.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 766-771, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.23
Mobile Computing:
An Enabler in International Financial Services

N. Raghavendra Rao
SSN School of Management & Computer Applications, India

ABSTRACT

Information and telecommunication technologies are the major stimulus for changes in trade and commerce. Recent convergence of the above technologies has become possible due to the rapid advancements made in the respective technology. This convergence is termed as information and communication technology (ICT) and considered as a new discipline. The new discipline has made cross border commerce in the present globalization scenario a reality. This chapter talks about a model for financial services sector in international market under the new discipline. The model explains the creation of knowledge based financial services system incorporating the sophisticated concepts of information technology. Further, it provides an access to the system with devices which can be used under wireless communication environment, across the globe.

INTRODUCTION

The effects of the convergence of telecommunication and information technology are being felt in the present global corporate world. This new discipline has made economics across the globe closely interconnected and integrated. Business processes are constantly changing at an exponential rate. The new discipline is also advancing by delivering exponential increase in computing power and communication capability. The result of this advancement has created a new generation of computers working on wireless technology, cell phones having the features of portable computers, and notebooks offering similar performance of desktop computers by using the same software. Portable computers and cell phones are no longer just for globetrotting executives. Innovations and radical changes are taking place in these products. The approach of the makers of these products is to provide fast and unwired connections in their

products, enabling their clients to make use of the rich resources of their organizations located across the globe.

The policy of globalization followed by many countries is changing the world's financial markets. In this context, Buckley (2003) observes that the world economy is internationalizing and, further, firms may engage in the international business by undertaking portfolio investment (p. 35).

This has led to deregulation. This is also providing opportunities to many financial institutions across the globe who are rendering investment advisory services. Accordingly every country is rapidly adapting itself to the new global changing vistas in the financial market. It is high time the investment advisory service providers take advantage of the benefits from the new discipline. A model is suggested to help investment advisors who are involved in the international financial market analyzing data and information for investment. Further it provides information to their team members who are located at various locations across the globe for providing services.

Business Process

The international financial market mainly comprises the corporate securities, Forex, metals, and commodity segments. Investment decision and advice in these segments need vast information. Information is required for corporate companies regarding the industry, natural resources such as metals, commodities, and the country level of each segment. The types of databases which can hold a high volume of data and information are required for this model. Sophisticated software tools are also needed for analyzing the data and information from these databases.

Investment financial analysts often explore an incredible amount of data about instruments, markets, and the corporate sector. They analyze the different market segments, price movements, economic forecasts, and news events. They react on the basis of market information, price trends, historical data, and their own experience. In this process, they can make many observations from the data and information available. They can try to determine the patterns from their observations.

Case Study for International Financial Services

A London-based investment consultancy organization, which has been operating in securities trading at the London Exchange market, has decided to go global. The organization decided to add other activities such as securities related to companies in different countries, Forex, metals, and commodities as their core services under its umbrella. It also changed its name to the Global Finance Services Advisory Group (GFSAG). GFSAG hired domain experts located in different countries under its business process outsourcing strategy. The group decided to follow the concept of virtual office for its operations in different countries. Domain experts and their team members can operate from anyplace of convenience. Their approach for virtual offices is to save the cost of infrastructure and to avail the benefits under the new discipline. The respective domain expert groups are expected to monitor, guide, and assist their counterparts and team members at different locations across the globe. The corporate office in London provides services for all activities to existing clients and prospective clients through the executives located at various locations across the world. In case of additional information and clarification, the executives are permitted to be in touch with the respective domain experts while they are at their clients' offices. The activities of GFSAG are summarized in Table 1. The places of operations are assumed for the purpose of case study.

These domain experts can analyze the information from the knowledge-based system for forecasting and identifying the risks associated with the operations in investments in the international

Table 1. Activities of GFSAG

Country	Location	Activities	Controlled By
USA	New York	Corporate Securities	Domain Experts
Australia	Sydney	Foreign Exchange	Domain Experts
Middle East	Bahrain	Metals	Domain Experts
Japan	Tokyo	Commodities	Domain Experts
UK	London	All the above activities	Corporate Group

financial market. On the basis of their analysis, inferences can be drawn and solutions can be suggested by them. These solutions are stored in an application database in a mobile computing server. The executives at the respective locations of their offices across the globe will be guiding their clients by having access to this server.

MODEL FOR GLOBAL FINANCIAL SERVICES ADVISORY GROUP

The business process explained in the case study will be the base for creation of knowledge-based international financial services systems under a wireless communication environment. This model will be referred as the GFSAG model. The GFSAG model has the following four stages:

- **Step 1:** Creation of Knowledge-Based System in GFSAG Model
- **Step 2:** Simulation and Forecasting for the Probable Risks in Financial Market
- **Step 3:** Mobile Computing Function in GFSAG Model
- **Step 4:** Requirements for GFSAG Model

Knowledge-Based System

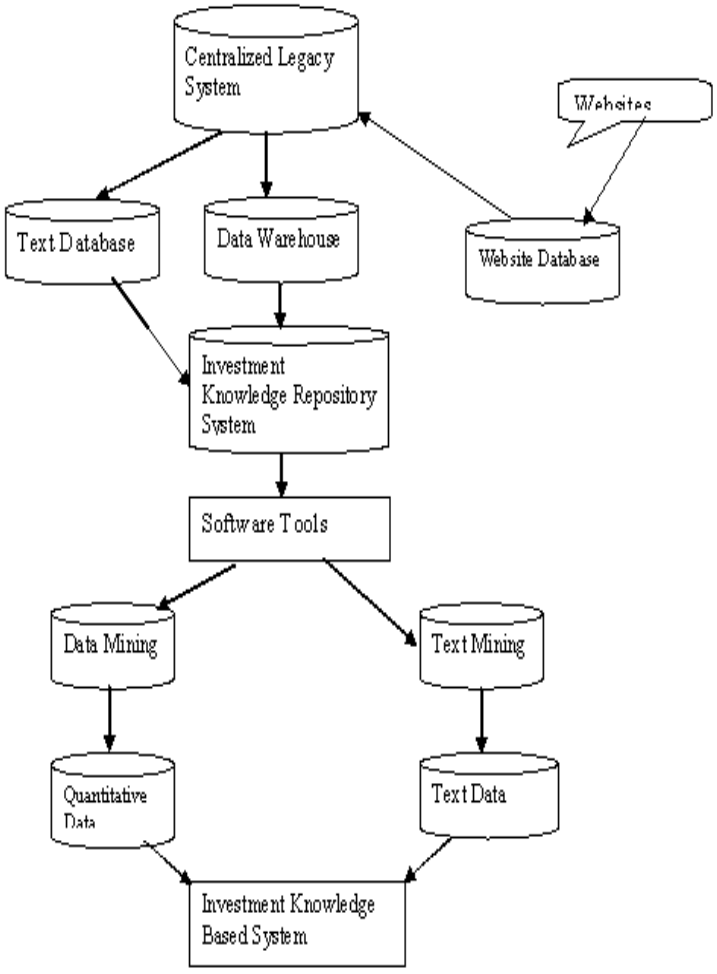
Step 1: Creation of Knowledge-Based System in GFSAG Model

A core team of GFSAG will consist of the domain experts, hardcore software professionals, and telecommunication experts. The importance of the services of hardcore software professionals and telecommunications experts cannot be underestimated because they are the backbone of the knowledge-based and core team. The macro-level design of an investment knowledge-based system is described in Figure 1.

The inputs received from the respective domain teams and other related information are stored in the centralized legacy system transferring to text database, and data warehouse depends on the type of data and information required for the creation of an investment knowledge repository system. The importance of a data warehouse in the financial sector is best described by Humphries, Hawkins, and Dy (1999):

A data warehouse contains data extracted from the many operational systems of the enterprise,

Figure 1. Knowledge base in GFSAG model



possibly supplemented by external data. For example, a typical banking data warehouse will require the integration of data drawn from the deposit systems, loan systems and the general ledger, just to name three. (p. 34)

The significance of a data warehouse is also highlighted by Adriaans and Zantinge (1999) when they say:

In order to perform any trend analysis you must have access to all the information needed to support you and this information stored in large data bases. The easier way to gain access to this data and facilitate effective decision making is to set up a data warehouse. (p. 25)

Text Database

This database will contain business practices, procedures, policies, culture, legal, taxation, accounting standards, political environment, various organizations profiles, information pertaining to natural resources of metals and commodities, and views and opinions of domain experts at each country and global level.

Data Warehouse

This will contain the quantitative data related to corporate securities, foreign exchange, metals, and commodities at each country and global level.

Web Site Database

This will contain the downloaded relevant information from the various sites in respect to the financial services sector at each country and global level.

Knowledge Repository

The data and information in the text database and data warehouse are to be grouped and stored as per the segments of the business activities of GFSAG.

Software Tools

The analysis of quantitative data stored in the investment knowledge repository system is carried through data mining. This tool helps one to know the relationship and patterns between data elements. The analysis of textual data is carried through text mining like data mining; it helps to identify relationships among the vast amount of text data. Pujari (2002) also states that text mining corresponds to the extension of the data mining approach to textual data (p. 239).

The investment knowledge-based system is created after the analysis by domain experts. This will contain how financial markets react to an event and the behavior of the market in the recent times.

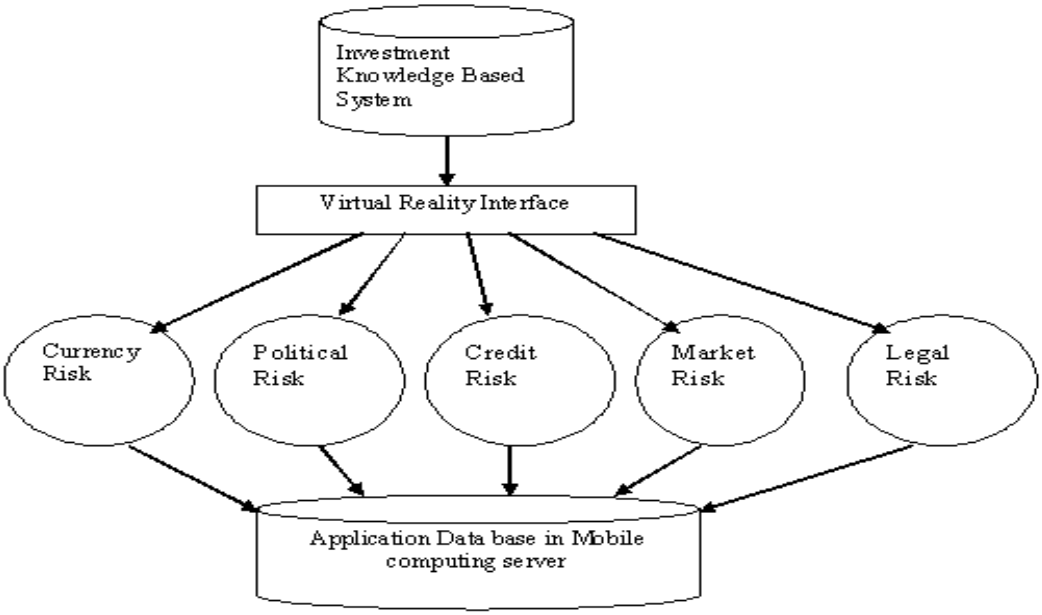
Step 2: Simulation and Forecasting for the Probable Risks in the Financial Market

Virtual reality is one of the concepts among the number of other concepts provided by information technology. Simulation is the basic element in virtual reality applications. The time dividing between simulated tasks and their real-world counterparts is very thin.

The synergy between real-world and simulated facts yields a surprising amount of effectiveness. The possible financial risks can be envisaged through simulations from the investment knowledge-based system. The simulated information can be provided through a mobile computing system. The use of features of virtual reality for simulation and forecasting for probable risks in financial markets are illustrated in Figure 2. Chorafas and Steinmann (1995) observe that each financial institution has a different way of looking at the market and business opportunity. The strategic approach must be mapped into the machine and then interactively visualized. Not surprisingly, some banks are very advanced and are leaving their competitors in the dust (pp. 174-175).

The harsh realities of risks are well known and understood by domain experts and their team members. The sophisticated tools help financial analysts form their views and opinions. It must be remembered that these tools are useful for keeping the unpleasant surprises to a minimum. The culture of the country influences the percentage of risks one takes.

Figure 2. Virtual reality concepts in GFSAG model



Wireless Environment: Concept of Mobile Computing

Mobile computing can be defined as a computing environment over a physical mobility. Schiller (2004) rightly says that GSM (Global System for Mobile communication) is the most successful digital mobile telecommunication system in the world today (p. 96).

The main features of a GSM system are indicated in Figure 3. A GSM has three subsystems: RSS (Radio SubSystem), NSS (Network and Switching Subsystems), and OSS (Operation SubSystem).

Support for Mobility

A mobile computing network becomes more useful when it supports business applications on its network. Now it is becoming possible by adding components such as file systems, databases, and security in mobile and wireless communication. The Web has been designed for conventional computers and fixed networks. Several new system architectures offer the opportunity to change the phase in telecommunication technology. Mobile communication is being influenced by merging telecommunication with computer networks. The present trend in the mobile phone market is cell phones being designed that take care of some features of computers besides voice calls. It would be apt to refer to the present mobile phone as a "mobile device" because these have additional features besides the conventional cell phones. It is interesting to note the observation of Giussani (2001) on mobile phones (pp. 227-247). He classifies it into four categories of devices: (1) dedicated devices, (2) integrated devices, (3)

Figure 3. Overview of GSM system (adapted from Schiller, 2004)

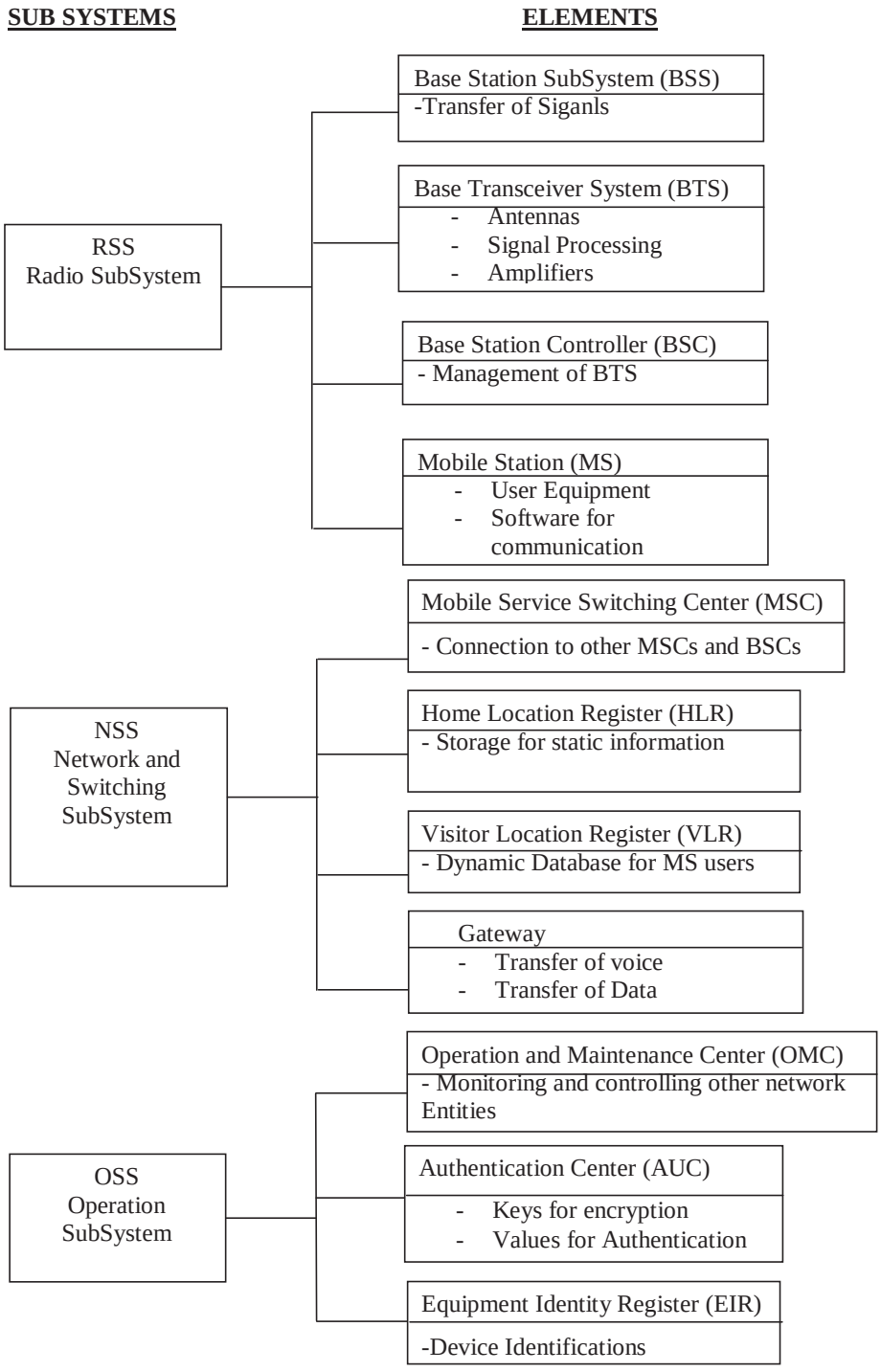


Table 2. Categories and the features of devices

Categories of Devices	Features of Devices
Dedicated	Designed for a particular functionalities
Integrated Devices	Integration of Functions of Different Devices
Modular Devices	Brining devices into one shell
Federated	Connecting Different Parts of Devices

modular devices, and (4) federated devices. The essential features of the four categories of devices are mentioned in the Table 2.

Present Scenario

Recently, handset makers have been queuing up to launch a new generation of smart phones which provide many features that have been special to pocket PC and PDAs, showing it is a buyer's market. Dood (2003) rightly points out that competition has benefited customers by triggering price decreases and wider availability of service (p. 387).

Makers of pocket PCs are enhancing many new features in their new models. Here it would be apt to quote Dornan (2001), who states:

The hype surrounding mobile data is ultimately founded on one thing: The Internet, Vendors and Operators alike use slogan such as 'Internet anywhere' and 'Internet in your pocket', promising to cut the Internet free from its PC-based roots. (p. 190)

In the case of smart phones, the additional features indicated are sending and receiving mail, Excel spreadsheets, PowerPoint presentations, and PDF files. PDAs allow the users to go online even while enjoying the usual office tools like Word, Excel, and Internet browser. A camera phone is used to capture visual information such as a phone number on a billboard instead of looking for a paper or pen to jot down the number. In the work environment some workers take pictures of finished projects to secure a visual record of completed work in case management requests such a record. On the same lines of handsets, the manufactures of notebooks and laptop computers are launching their products with wireless technology. These products have convenient mobility available with modem, integrated LAN, or wireless connections with desktop power. Now it has become a necessity to establish synergy between mobile computing and knowledge-based business systems through these sophisticated devices. Taulkder (2002) confirms this view by saying:

Mobile computing not only offers instant information to a mobile worker; to a mobile worker it is indeed a productivity tool. Further the list of possible mobile applications can never be complete. (p. 18)

Security

The international financial services knowledge-based system is more sensitive and critical. Encryption is a solution that ensures the data content is not altered during the transmission between originator and recipient in a wireless environment. This is elaborated by Minoli and Minoli (1999), who observe:

Cipher technique[s] lend themselves more readily to automated. Theses technique[s] are uses [sic] in contemporary security tools and there are three kinds of cryptographic functions, such as Hash functions, Secret key and Public key. (p. 215)

The concept of an asymmetric crypto system for encryption of data may be used in a wireless environment. An asymmetric crypto system is a system of secure key pair consisting of private key for creating a digital code and public key to verify the same. Hash function in this system means obtaining "hash result" by applying a predefined logic, control, or arithmetical process. Hash result means that every time the predefined procedure is applied, it should give the same result. The components in an asymmetric crypto system, taking the results obtained from the hash results, are explained in Figure 4.

With the advancement of information technology, the concept of cryptography used earlier by kings for secret communication is becoming popular in global commercial applications.

Step 3: Mobile Computing Function in GFSAG Model

The integration of the functions in mobile computing with a GFSAG model are illustrated in Figure 5.

Step 4: Requirements for GFSAG Model

The user groups for the GFSAG model at the respective country level and end users across the globe will be making use of it. The requirements of hardware, software, and mobile devices for the GFSAG model are mentioned in Table 3.

Figure 4. Components of asymmetric crypto system

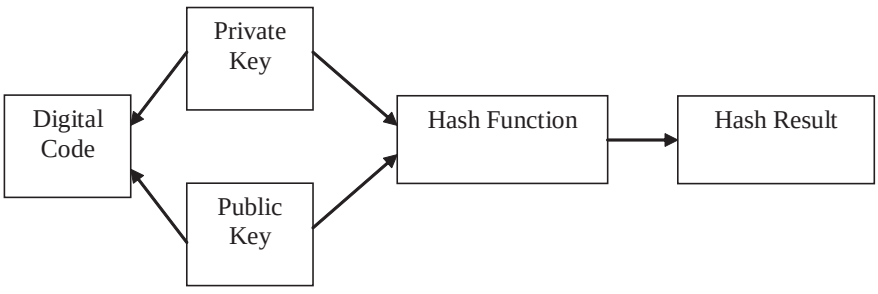
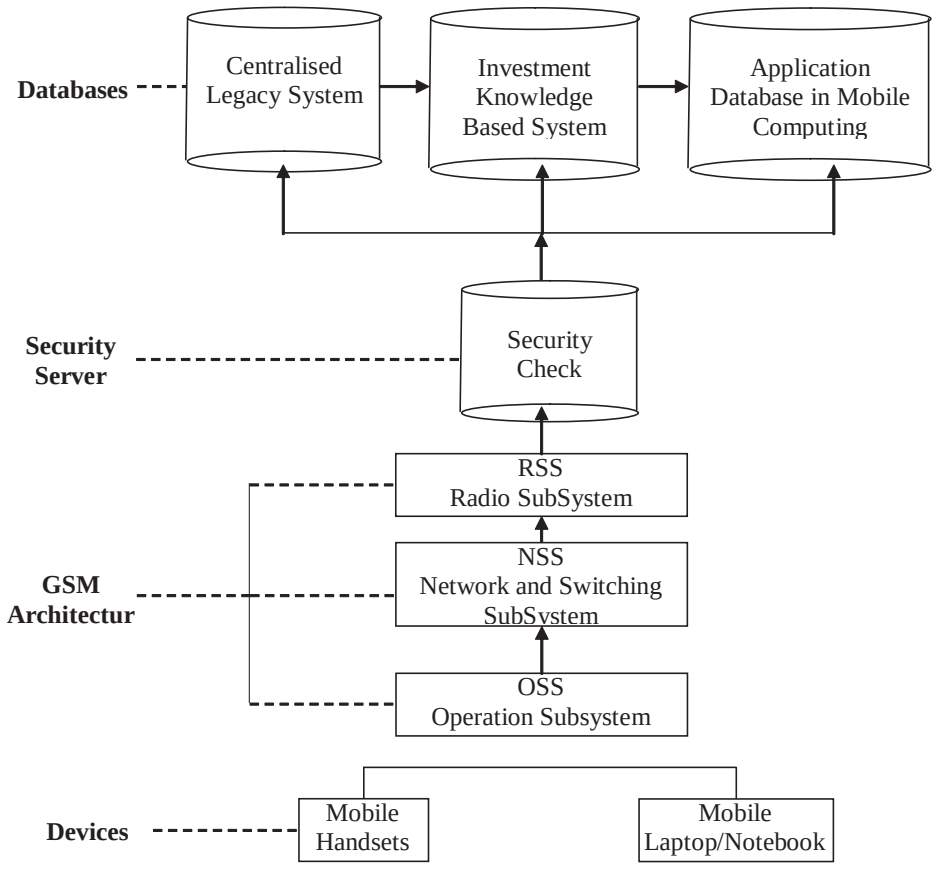


Figure 5. Mobile computing functions in GFSAG model



CONCLUSION

The disintegration of barriers in previously protected and insulated markets has created a new era of competition in the global economic environment. The present challenge for global players in international financial markets is how they should take advantage of the opportunities from the severe competition and survive in the market. The rewards for the opportunities are always are accompanied by risks. Assessing risks and incorporating the same in the final decision is an integral part of the decision making. The new discipline plays an important role for acquiring and processing information for analysis and decision making. The GFSAG model offer an idea for using the services of domain experts across the globe and for minimizing the risks in financial market. The word "international" prefixed to "financial market" will become redundant once the concepts of new discipline are taken advantage of by the corporate world.

Table 3. Requirements for GFSAG model (hardware, software, and mobile devices)

Particulars		**Purpose**
Hardware	Server	System and Application programmes
	Server	Exclusively for Encryption and Recognition of users
	Server	Storage of Data and information of various activities of financial services
	Desktop/Workstation	Development of programmes and updating and retrieving data
Software	Data warehouse	Quantitative data of various segments of the financial services market
	Text Data warehouse	Text data of various segments of the financial services market
	Data mining	Analysing data from data warehouse
	Text Mining	Analysing text data from text data warehouse
	Virtual Reality Interface	Visualizing through Simulation and applying forecasting techniques
	Other Related Software	Software for supporting the system and routine software for business purpose
Mobile Devices	Laptop/Notebook	1. Interaction with domain experts and corporate office 2. Viewing the selected information from the knowledge base system 3. Sending reports from the market place to corporate office and domain experts. 4. Downloading the analysed data for understanding
	Handsets	1. Discreetly informing some specific information from clients place 2. Capturing important data by using camera features in handsets and transmitting to corporate and domain experts offices. 3. Browsing the knowledge base system for specific purpose 4. Sending latest short news from the financial market

FUTURE STUDY

Many more financial markets are opening up and becoming integrated with global markets. Many virtual financial services organizations will make their presence felt in the global market. There will be an increase in reliable forecasting by using real-time data and financial modeling. Many more different activities related to financial services will be required to be added to GFSAG model.

RECOMMENDATION

The concept of virtual organization is the key to GRID computing. All the virtual organizations will share the common resources for computing power and accessing data across the globe. Grid and mobile computing concepts will be required to be integrated, once many financial markets are interconnected with each other in the global market under the concept of virtual organization. Referring to GRID computing in the financial sector, Joseph and Tein (2004) state that grid computing provides the financial analysis and services industry sector with advanced systems delivering all the competitive solutions in grid computing. These solutions exemplify the infrastructure and business agility necessary to meet and exceed the uniqueness that the financial analysis and services industry sector requires. This particular value statement is accomplished by the fact that many of these solutions in this industry are dependent upon providing increased access to massive amounts of data, real-time modeling, and faster execution by using the grid job scheduling and data access features (p. 14).

REFERENCES

Adriaans, P., & Zantinge, D. (1999). *Data mining and data warehouse data mining* (pp. 25-36). Harlow, UK: Addison Wesley Longman.

Buckley, A. (2003). *The internationalization process, multinational finance* (pp. 35-46). New Delhi: Prentice-Hall.

Chorafas, D. N., & Steinmann, H. (1995). Implementing virtual reality in financial institutions. In *Virtual reality: Practical applications in business and industry* (pp. 161-179). Englewood Cliffs, NJ: Prentice-Hall.

Dodd, A. Z. (2003). *Wireless services, the essential guide to telecommunications* (pp. 371-408). New Delhi: Pearson Education Asia.

Dornan, A. (2001). *Inside a mobile network, the essential guide to wireless communications applications* (pp. 175-195). New Delhi: Pearson Education Asia.

Giussani, B. (2001). *The intimate utility: Roam making sense of the wireless Internet* (pp. 227-247). London: Random House Business Books.

Humphries, M., Hawkins, M. W., & Dy, M. C. (1999). *Data warehouse concepts, data warehousing architecture and implementation* (pp. 31-48). Englewood Cliffs, NJ: Prentice-Hall.

Joseph, J., & Fallenstein, C. (2004). *Introduction, the grid computing anatomy* (pp. 12-14, 47-57). New Dehli: Grid Computing, Pearson Education.

Minoli, D., & Minoli, E. (1999). *Encryption, Web commerce technology handbook* (pp. 213-225). New Delhi: Tata McGraw-Hill.

Pujari, A. K. (2002). *Text mining, data mining techniques* (pp. 239-250). Hyderabad: Universities Press (India).

Schiller, J. (2004). *Telecommunication systems, mobile communications* (pp. 93-130). New Delhi: Pearson Education.

Talukder, A. K. (2002). Mobile computing—impact in our life. In C. R. Chakravarthy, L. M. Patnaik, T. Sabapathy, & M. L. Ravi (Eds.), *Harnessing and managing knowledge* (pp. 12-24). New Delhi: Tata McGraw-Hill.

ADDITIONAL READING

Cudworth, R. (2003). *The demand for continuous information: The source online business* (pp. 12-17). London: Kogan Page.

Haugen, R. A. (2002). *Securities and markets: Modern investment theory* (pp. 6-31). New Delhi: Prentice-Hall.

Lasserre, P. (2003). *Global financial management, global strategic management* (pp. 335-351). Hampshire, UK: Palgrave MacMillan.

Lumby, S. (1998). *Foreign exchange risk management, investment appraisal and financial decisions* (pp. 579-596). London: International Thomson Publishing.

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 828-838, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.24
E-Commerce and Mobile Commerce Applications Adoptions

Charlie Chen
Appalachian State University, USA

Samuel C. Yang
California State University, Fullerton, USA

INTRODUCTION

E-commerce applications are primarily used at home and in the workplace. Utilitarian elements, including cognitive beliefs of perceived usefulness, perceived ease of use (at the individual level), industry pressure, organizational readiness, economics, and trust (at the business level) are key determinants contributing to the usage of e-commerce applications. Mobile devices redefine the meaning of workplace. The use of mobile services could be in and outside the workplace. Hedonic elements, such as fun, culture, life style, and hype are key determinants contributing to the usage of mobile commerce applications. The purpose of our article is to discuss and clarify immediate determinants of e-commerce and mobile commerce applications based on the technology acceptance model.

BACKGROUND

A joint study by eMarketer and Forrester (2005) estimates that business-to-customer (B2C) revenues in the U.S. will reach $229.9 billion by 2008 and business-to-business (B2B) revenue will reach $8.8 trillion in 2005. According to the Computer Industry Almanac (ClickZ Stats, 2005), by 2007 the number of Internet users will grow to 1.46 billion worldwide with the U.S. market representing only about 20% of worldwide Internet users. It is clear that e-commerce (EC) is becoming a global transactional forum.

Along with the dominance of EC comes an increased demand for mobile commerce (MC). The total number of mobile telephone subscribers in the world grew to 1.34 billion in 2003 from 317 million in 1998 (International Telecommunication Union, 2003). More than half of Americans (158

million) were mobile telephone subscribers in the United States. Unlike EC, only a very limited number of MC applications are making profit (Beck & Wade, 2002). The difference in the adoption pattern of EC and MC prompts practical reasons as well as research motives to investigate what drive consumers to purchase or use a particular EC and MC application.

The goals of adoption of EC and MC applications can be grouped into two broad categories: utilitarian (productivity-oriented) and hedonic (pleasure-oriented). *Utilitarian* elements are those determinants of productivity and usefulness that should be considered by a rational user or company before deciding to adopt a particular EC or MC application. For instance, an individual uses e-banking and online job search engine to improve personal productivity. A company adopts e-marketplace or Internet EDI applications to improve operational efficiency, reduce cost, and increase customer services. *Hedonic* elements are those determinants that are associated with personal enjoyment and pleasure. (See Table 1.) A user subscribes to a gaming or dating service to meet friends who share common interests. Knowing the dichotomizing difference between utilitarian and hedonic goals can help us understand why we accept a particular EC or MC application.

E-COMMERCE AND M-COMMERCE ADOPTION

E-Commerce Adoption

Electronic commerce (EC) refers to electronic business with a broader meaning than just buying and selling on the Internet. EC is the process of transacting, transferring, or exchanging products and services over communication networks, including the Internet (Turban, King, Lee, & Viehland, 2004). Note that the underlying network may encompass different broadband (i.e., > 1 Mbps) segments such as DSL, cable modem, power line, Asynchronous Transfer Mode (ATM), and Gigabit Ethernet. Straub (2004) defined all forms of EC organizations as Net-enhanced organizations. Many EC applications are avail-

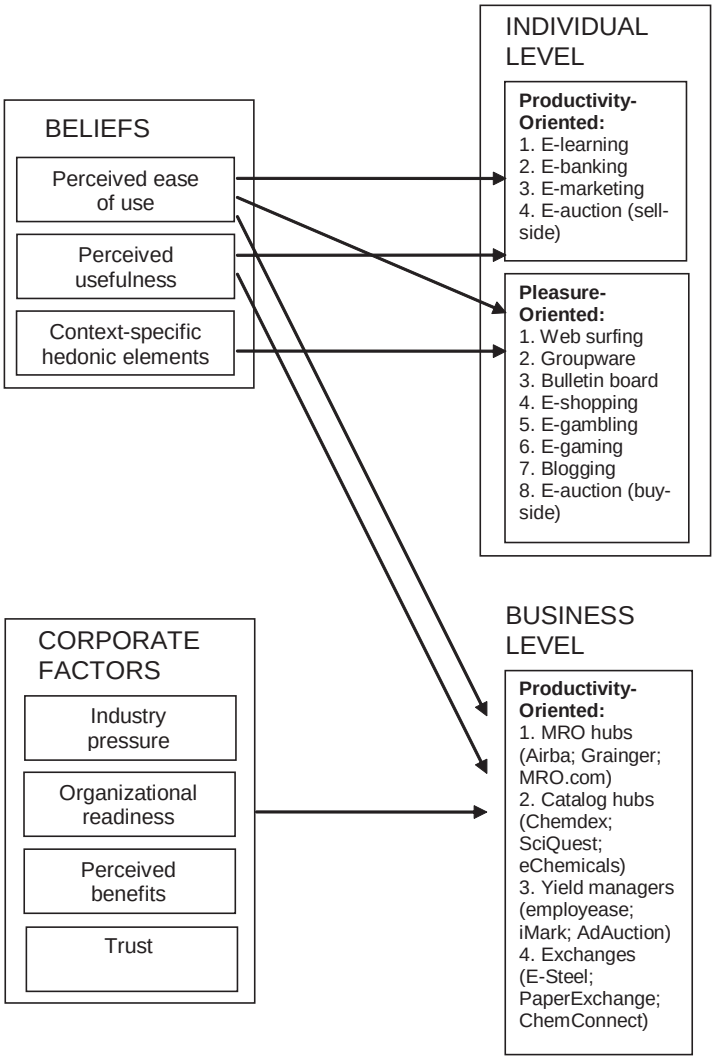
Table 1. Summary of utilitarian and hedonic factors of EC and MC adoption

	E-Commerce	M-Commerce
Utilitarian Factors (Firm)	Industry presure, organizational readiness, perceived benefits, trust	Critical mass, perceived benefits
Utilitarian Factors (Individual)	Perceived ease of use (PEOU), perceived usefulness (PU)	Perceived ease of use (PEOU), perceived usefulness (PU), cost, perceived system quality
Hedonic Factors (Individual)	Perceived playfulness, perceived enjoyment, network size, perceived user resources	Social influence, entertainment, hype, lifestyle

able to support the operation of Net-enhanced organizations. EC applications that are widely adopted at the individual level include e-tailing, Internet marketing, online travel services, online banking, e-grocery, online gaming, e-auction, etc. EC applications at the business level facilitate the inter- and intra-organizational transactions over communication networks.

There are four basic types of B2B applications: sell-side market, buy-side market, many-to-many marketplaces, and collaborative commerce (Turban & King, 2003). Sell-side applications comprise of online catalog (e.g., Cisco's Connection Online service), e-auction, and online intermediary (e.g., Boeing's Part Analysis and Requirement Tracking service). Buy-side applications are primarily

Figure 1. Immediate determinants for the adoption of EC applications



e-procurement management (e.g., Schlumberger's Oilfield Services), internal and external aggregation (e.g., Grainger.com and allbusiness.com), reverse auctions, and e-bartering. Many-to-many marketplaces allow many buyers and sellers to conduct transactions in an online pubic or private marketplace (e.g., epapertrade.com, mro.com, chemconnect.com, and employease.com). Collaborative commerce enables business partners to collaborate with each other by engaging in non-transactional activities.

Factors driving the adoption of EC applications can be examined at two levels of analysis: individual and business. At the individual level, both utilitarian and hedonic factors are responsible for customers' adoption of EC applications. At the business level, utilitarian factors (organizational productivity) are primarily responsible for adoption in for-profit and not-for-profit organizations (Figure 1).

Utilitarian and Hedonic Perspectives of Adopting EC Applications

The Technology Acceptance Model (TAM) (Davis, 1989) is recognized as the most robust and influential model that predicts an individual's adoption behavior of information technology (Davis, Bagozzi, & Warshaw, 1992; Venkatesh & Morris, 2000). At the individual level, the TAM provides a utilitarian perspective of EC adoption. This model asserts that perceived ease of use (PEOU) and perceived usefulness (PU) are important in forming customer attitude, satisfaction, and trust towards EC applications (Devaraj, Ming, & Kohli, 2002). Favorable attitude, satisfaction, and trust can lead to the adoption of EC applications. These two cognitive beliefs—PEOU and PU—have adequately explained the widespread adoption of personal productivity-oriented EC applications (i.e., most B2C applications).

However, the TAM is weak in explaining hedonic EC applications. Many studies extended the model to correct the weakness. For instance, along with PEOU and PU, perceived playfulness (Moon & Kim, 2001) and perceived enjoyment (Davis, Bagozzi, & Warshaw, 1992; Teo, Lim, & Lai, 1999) are immediate determinants for the adoption of the Internet. Perceived critical mass or network size is another complementary factor for the adoption of online groupware (Luo & Strong, 2000). Perceived user resources is the immediate determinant for the adoption of bulletin board system (Mathieson & Chin, 2001). Compatibility (Chen, Gillenson, & Sherrell, 2002) and intrinsic motivation of online users (Venkatesh & Morris, 2000) directly result in the adoption of virtual stores (Chen, Gillenson, & Sherrell, 2002). Social influence and flow experience are direct causes of the adoption of online activities that require the total involvement of online users, such as betting on sports events, gambling (Hsu & Lu, 2004), and shopping online.

Blogging is another prevalent online activity. The clustering of demographics (age, geography, ethnics, and lifestyle) and common interests (MTV, sports, gardening, traveling) into a network of virtual communities resulted in the mass adoption of blogging applications (Kumar, Novak, Raghavan, & Tomkins, 2004). When entertainment and fun are the goals of EC applications, the TAM alone is inadequate to predict and explain their adoption. Extending TAM by incorporating hedonic factors in different contexts can better explain the adoption behavior of hedonic EC applications.

EC applications at the business level differ from those at the individual level in the difficulty of measuring benefits, unselective release of confidential information to competitors (trust) and insufficient time to develop internal new skills (Teo & Ranganathan, 2004). Hence, there are additional factors leading to the adoption of EC applications beyond the utilitarian elements of PEOU and PU (Igbaria, Parasuraman, & Baroudi, 1996). Other corporate factors maybe as important as utilitarian factors, including industry pressure, organizational readiness, perceived strategic

value (Grandon & Pearson, 2004), perceived cost, and trust.

Industry pressure originates from inter- and intra-industry competition, business partners, standards organizations, regulatory bodies, and government. These environmental factors can influence the adoption of B2B EC applications. RosettaNet (EDI for computer industry), UN/EDIFACT (EDI for commerce and transportation industry), and other XML-based EDI e-marketplaces are products of industry pressure. Industry-level compliance can lower the degree of asset specificity and uncertainty (imperfect information) but increase the number of input resources (Williamson, 1981). From the comparative institutional perspective, B2B e-marketplaces can potentially achieve three cooperative and exchange gains: (1) creating awareness of gains through joint efforts, (2) discouraging parties from bargaining their own gains, and (3) enforcing agreed-upon agreements (Hennart, 1994). Industry pressure increases existing network size and increases perceived benefits of adopting B2B EC applications. Hence, industry pressure is a direct determinant for the adoption of B2B EC applications.

The organizational readiness factor is concerned with the availability of financial and technological resources and alignment of B2B EC applications with a company's vision, values, culture, and preferred work practices (Grandon & Pearson, 2004). The varying degree of organizational readiness among business partners contribute to their difference in PEOU and PU (Iacovou, Benbasat, & Dexter, 1995).This contextual variable influences the successful adoption of B2B EC applications.

Many B2B EC applications fail to help companies realize their benefits primarily because sellers and buyers misjudge their perceived scope of benefits (Pandya & Dholakia, 2005). Primary benefits of B2B EC applications for buy-side marketplaces (e.g., GE's Global Exchange Services) include lowering purchasing price, streamlining the bidding process, and reducing requisition cycle time. Benefits of B2B EC applications for sell-side marketplaces (e.g., Intel, Cisco and Dell) range from the reduction of operation costs and cycle time, to the improvement of inventory control. Adoption of a particular B2B EC application depends on whether the application can deliver specific perceived benefits for different users of the e-marketplace.

EC applications applied to many-to-many marketplaces have different adoption issues because they requires companies to collaboratively plan, transact, design, and develop products and services. Unreliable and insecure B2B EC applications can have profound impacts on a company's operation. Trust in B2B EC can be improved from (1) organizational and economic, (2) technological, and (3) behavior perspectives (Ratnasingam & Phan, 2003). Issues about the adoption of B2B EC applications implicitly incorporate the perspective of technological trust (Ratnasingam & Pavlou, 2003). Therefore, when considering whether to adopt a B2B EC application the trust of business partners in B2B EC applications must be ensured.

Mobile Commerce Adoption

Mobile commerce (MC) is the process of buying, selling, or exchanging products and services wirelessly over mobile communication networks. MC is conducted on the Internet via mobile devices that are able to connect to the Internet via wireless application protocol (WAP) or conventional hyper-text transfer protocol (HTTP). Although most of the MC transactions take place on the Internet, MC can occur over any public or private network. The most important characteristic of MC is that products and services are made available to the customer independent of the customer's location (Turban & King, 2003). B2B MC applications include different classes, such as financial applications, inventory management, service management, product locating, business process

reengineering, and data retrieval. On the other hand, B2C MC applications include classes such as financial applications, advertising, inventory and service management, product locating and shopping, auction or reverse auction, entertainment, mobile office, distance education, and data retrieval (Varshney & Vetter, 2002).

Similar to factors driving the adoption of EC, utilitarian factors heavily influence the decision to adopt EC at the business level; at the individual level, both utilitarian and hedonic factors contribute to the adoption (Coursaris & Hassanein, 2002; Haque, 2004). Note that since a wireless laptop computer essentially emulates the experience of a networked desktop computer, we will instead focus our attention on MC over handheld devices. In addition, we will place our emphasis on the consumer in our consideration of hedonic factors in adopting MC.

Utilitarian and Hedonic Perspectives of Adopting MC Applications

At the business level, one important factor is critical mass (Carroll, Howard, Peck, & Murphy, 2002), or the increased externalities of mobile devices ranging from Internet-enabled cellular phones and personal digital assistants (PDAs) to networked laptop computers. In 2002, the number of cellular phones exceeded that of fixed wireline phones on a global scale (International Telecommunication Union, 2003). In the U.S., over half of the population already carries a cellular phone. The number of network-capable mobile devices is also increasing rapidly. The growing number of MC-capable devices indicates that more service providers will be willing to develop and deploy new MC applications.

Another factor is the demonstrable benefit of adopting MC applications for businesses. With the increased diversity of data services and contents that can be delivered over wireless networks, organizations are beginning to adopt more MC applications. Many of the MC application deployments are vertical applications and aim to serve a specific company or industry need. Examples are fleet tracking, field sales force, and education. These applications typically have clearly defined goals (e.g., to streamline a specific process) and have well-defined net present value (NPV) justifications. Businesses typically develop these applications in-house or utilize content enablers and middleware.

At the individual level, both utilitarian and hedonic factors are important (Coursaris & Hassanein, 2002; Haque, 2004). Mobile voice communication is still the most successful mobile wireless application. Mobile voice is a focused application in that it concentrates a user's attention on one thing—conversation, and it is easy to use. Time and attention are more critical for mobile applications because users have a limited time span and may be distracted by their environment (Turban & King, 2003). It is clear now that PEOU is a factor that made mobile voice communication successful. PEOU is also an important factor in users' adoption of MC on modern networked handheld devices (Pagani, 2004). Another factor contributing to adoption is PU (Pagani, 2004). MC can be beneficial to the user, the benefit being that the user is able to carry out business activities, and conduct business transactions any where, any time. In addition, users are able to leverage unproductive time (e.g., during commute and travel) for productive tasks (Perry, O'Hara, Sellen, Brown, & Harper, 2001).

The third factor is the cost of using mobile devices (Haque, 2004). Accessing the Internet and checking e-mails using network-capable devices represent a lower-cost and higher-accessibility alternative to using a traditional personal computer (PC). The fourth factor is perceived system quality (Kleijnen, Wetzels, & Ruyter, 2004). Mobile devices have more limited bandwidth than their fixed counterparts, so it is important to ration the amount of information to be downloaded to the device (Treese & Stewart, 2003). In addition, network quality (e.g., unnoticed latency, jitters

control and high resilience) has to be excellent so that problems arising from weak signal strength do not frustrate the users and distract them from their tasks.

Other hedonic factors such as culture, fun, hype, and lifestyle are also important to a user's decision to adopt. The widespread use of mobile devices is becoming a social phenomenon, especially among high-school and college age students. These users have grown up in a time when mobile devices are already popular. Many users in this age group will consume more MC services once they start working and will already be comfortable with making purchases and transacting via mobile devices (Turban & King, 2003). Kleijnen, Wetzels, and Ruyter (2004) cite social influence as an antecedent in adopting mobile financial applications, and Lu, Yu, Liu and Yao (2003) also suggest social influence in their proposed theoretical model as an antecedent in adopting wireless Internet.

Recent research has begun to suggest fun and hype as emerging factors in adoption. Pagani (2004) cites enjoyment as a factor after conducting an exploratory study with focus groups in six countries (i.e., Brazil, Germany, Italy, Singapore, United Kingdom, and U.S.) and a survey study of 1,000 mobile users in Italy. Haque (2004) also cites entertainment as an antecedent as a result of his survey in Malaysia. In terms of hype, consumers are constantly bombarded with images such as "…executives reclining on sun drenched beaches, cheerfully pecking away at their laptops, or strutting, like alpha males through airports checking their stock portfolios on PDAs and mobile phones" (Sherry & Salvador, 2001, p. 108). In spending millions on advertising, companies expect to have a positive impact on their sales. In fact, Pagani (2004) cites perceived innovation as a factor in the adoption of 3G mobile multimedia services, the perception of which is also affected by various innovations.

Carroll, Howard, Peck, and Murphy (2002) identified lifestyle as an antecedent in the continued use of SMS in their survey of 16 to 22 year-old mobile device users in Australia. Lifestyle is often defined as a way of life or style of living that reflects the attitudes and values of a person. It is reasonable to treat lifestyle as an important factor leading to continued use; if device use is part of a user's way of life already (e.g., communicating with friends via SMS instead of face-to-face), its continued use is facilitated.

Overall, from a hedonic perspective, "killing time" may be the "killer application." Even for mobile users that use mobile devices for utilitarian reasons (e.g., work productivity), it is reasonable to suggest they would turn to their mobile devices for entertainment when they have a few minutes in between tasks or meetings. In addition, in those parts of the world where owning a mobile device costs less than owning a desktop PC, mobile devices may be the primary source of electronic entertainment (Coursaris & Hassanein, 2002).

FUTURE TRENDS OF E-COMMERCE AND MOBILE COMMERCE APPLICATIONS

As far as EC applications are concerned, EC applications have become diversified and more sophisticated after many years of development. In the future, however, EC applications need to overcome a variety of problems which Dekleva (2000) grouped into four themes: (1) trust, (2) legal framework, (3) information infrastructure, and (4) benefits maximization. Different issues exist in each of these themes. Regarding the trust theme, privacy, security, as well as PEOU and PU of EC applications will remain to be important issues. Regarding legal framework, laws related to taxation, intellectual property protection, and payment systems will continue to evolve. Regarding information infrastructure, reliable Internet infrastructure, effective systems integration, and common industry standards will continue to be needed to support future EC applications. Lastly,

regarding benefits maximization, organizations will continue to institute systems to measure benefits and costs of adopting EC applications. In addition, new EC applications will emerge to address issues present in each one of these four themes. For example, digital rights management (DRM) applications have already emerged to address the issues of trust and legal framework related to the transmission and consumption of entertainment-related contents (e.g., music and movies).

As far as MC applications are concerned, several trends are emerging in MC which can be examined in terms of the (1) mobile device, (2) network, and (3) application. Regarding the mobile device, there is clearly a trend toward the convergence of several functions into one physical device (Turban & King, 2003). For instance, the PDA/phone combination allows users to not only look up a phone number directly in the PDA and dial it, but also avoid carrying multiple devices while traveling. Furthermore, mobile devices now possess more hedonic features such as still-image camera, camcorder, and music player.

Regarding the network, wireless networks will continue to offer higher data-rate services. In wireless wide-area networks (WWANs), network operators have continued to deploy next-generation systems, such as General Packet Radio Service (GPRS) and Universal Mobile Telephone Service (UMTS). GPRS, which is based on a packet-switched core network, offers data rates of between 50 and 60 kbps. UMTS, which has both a packet-switched core network and an enhanced air interface, provides a higher data rate of at least 384 kbps. In wireless local-area networks (WLANs), wireless hot spots have continued to proliferate, covering hotels, airports, convention centers, and coffee shops. These networks, although nowhere near ubiquitous, offer stationary users much higher data rates ranging from 11 Mbps (peak for 802.11b) to 54 Mbps (peak for 802.11g and 802.11a). These higher-speed mobile networks can enable those applications that traditionally have been run on desktop PCs, such as e-mailing large attachments and distance learning (Varshney & Vetter, 2002).

Regarding the application, the increased data rates and more coverage areas will enable new kinds of MC applications to emerge. Many of these applications will target the hedonic elements of MC. We have already seen that mobile entertainment is one factor contributing to the success of I-Mode service in Japan. In the U.S., consumers can download full-length music titles to their phones and store them locally on the phones. In video services, mobile users in Korea routinely watch live television broadcasts streamed to their cell phones. With the advent of 3G networks, higher data rates will help improve the mobile user's gaming experience (Coursaris & Hassanein, 2002; Harmer, 2001). Furthermore, with the popularization of online blogging, many mobile users are now engaged in mobile image blogging (i.e., recording their daily activities in pictures taken by cellular phones and uploading them to a Web site).

CONCLUSION

Tim Berners-Lee, the inventor of the World Wide Web (WWW) and the director of WWW Consortium (W3C), foresees the emergence of a more open "resource description framework" (Klyne, Carroll & McBride 2004, p. 1) that can draw connections between all sorts of objects and information via wired and wireless devices. The first step towards the realization of this vision depends upon the adoption of EC and MC applications. A framework for including utilitarian and hedonic considerations to clarify reasons for the adoption of EC and MC applications by individuals and businesses is proposed in this article. EC applications at the individual level need to address both utilitarian factors of perceived usefulness

and perceived ease of use, and hedonic factors of perceived enjoyment in different contexts. Immediate utilitarian determinants to the adoption of EC applications at the business level include industry pressure, organizational readiness, perceived benefits and trust. MC applications have different adoption issues to resolve. These issues can also be grouped into utilitarian and hedonic categories. Utilitarian determinants include the perceived ease of use and perceived usefulness. Hedonic determinants range from culture, fun, and hype to lifestyle.

REFERENCES

Beck, J. C., & Wade, M. (2002). *DoCoMo—Japan's wireless Tsunami: How one mobile telecom created a new market and became a global force*. New York: American Management Association.

Carroll, J., Howard, S., Peck, J., & Murphy, J. (2002). A field study of perceptions and user of mobile telephone by 16 to 22 year olds. *Journal of Information Technology Theory and Application, 42*(2), 49-60.

Chen, L., Gillenson, M., & Sherrell, D. (2002). Enticing online consumers: An extended technology acceptance perspective. *Information and Management, 39*, 705-719.

ClickZ Stats. (2005, February 8). *Population explosion!* Computer Industry Almanac Inc. Retrieved from www.clickz.com/stats/big_picture/geographics/print.php/5911_151151

Coursaris, C., & Hassanein, K. (2002). Understanding m-commerce: A consumer centric model. *Quarterly Journal of Electronic Commerce, 3*(3), 247-272.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology, 22*(14), 1111-1132.

Dekleva, S. (2000). Electronic commerce: A half-empty glass. *Communications of the Association for Information Systems, 3*(18), 1-99.

Devaraj, S., Ming, F., & Kohli, R. (September 2002). Antecedents of B2C channel satisfaction and preference: Validating e-commerce metrics. *Information Systems Research, 13*(3), 316-333.

eMarketer, & Forrester (2005). Comparative estimates: B2C ecommerce revenues in United States, 2000-2008. *Essential Metrics.* Retrieved from www.emarketer.com/Products.aspx

Grandon, E. E., & Pearson, J. M. (December 2004). Electronic commerce adoption: An empirical study of small and medium U.S. businesses. *Information and Management, 42*(1), 197-216.

Haque, A. (2004). Mobile commerce: Customer perception and its prospect on business operation in Malaysia. *Journal of Financial Services Marketing, 8*(3), 206-217.

Harmer, J. (2001). 3G products: What will the technology enable? *BT Technology Journal, 19*(1), 24-31.

Hennart, J. F. (1994). The "Comparative Institutional" theory of the firm: Some implications for corporate strategy. *Journal of Management Studies, 31*(2), 193-207.

Hsu, C., & Lu, H. (2004). Why do people play online games? An extended TAM with social influences and flow experience. *Information and Management, 41*(7), 853-868.

Iacovou, C. L., Benbasat I., & Dexter A. S. (1995). Electronic data interchange and small organizations: Adoption and impact of technology. *MIS Quarterly, 19*(4), 465-485.

Igbaria, M., Parasuraman, S., & Baroudi, J. (1996). A motivational model of microcomputer usage. *Journal of Management Information Systems, 13*(1), 127-143.

International Telecommunication Union. (2003). *World telecommunication indicators database 2003*. Geneva: International Telecommunication Union.

Kleijnen, M., Wetzels, M., & Ruyter, K. D. (2004). Consumer acceptance of wireless finance. *Journal of Financial Services Marketing, 8*(3), 206-217.

Klyne, G., Carroll, J. J., & McBride, B. (2004). *Resource Description Framework (RDF): Concepts and abstract syntax*. World Wide Web Communications. Retrieved from http://www.w3.org/TR/rdf-concepts/

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (December 2004). Structure and evolution of blogspace. *Communications of the ACM, 47*(12), 35-39.

Lu, J., Yu, C. S., Liu, C., & Yao, J. E. (2003). Technology acceptance model for wireless Internet. *Internet Research: Electronic Networking Application and Policy, 13*(3), 206-222.

Luo, W., & Strong, D. (2000). Strong, perceived, critical mass effect on groupware acceptance. *European Journal of Information Systems, 9*(2), 91-103.

Mathieson, K., & Chin, W. (2001). Extending the technology acceptance model: The influence of perceived user resources. *The Data Base for Advances in Information Systems, 32*(3), 86-112.

Moon, J., & Kim, Y. (2001). Extending the TAM for a World Wide Web context. *Information and Management, 38*(4), 217-230.

Pagani, M. (2004). Determinants of adopting of third generation mobile multimedia services. *Journal of Interactive Marketing, 18*(3), 46-59.

Pandya, A. M., & Dholakia, N. (2005). B2C failures: Toward an innovation theory framework. *Journal of Electronic Commerce in Organizations, 3*(2), 68-81.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: Understanding access anytime, anywhere. *Transactions on Computer Human Interaction, 8*(4), 323-347.

Ratnasingam, P., & Pavlou, P. (2003). Technology trust in Internet-based interorganizational electronic commerce. *Journal of Electronic Commerce in Organizations, 1*(1), 17-41.

Ratnasingam, P., & Phan, D. D. (2003). Trading partner trust in B2B e-commerce: A case study. *Information Systems Management, 20*(3), 39-50.

Sherry, J., & Salvador, T. (Ed.). (2001). *Running and grimacing: The struggle for balance in mobile work*. London: Springer-Verlag.

Straub, D. W. (2004). *Foundations of net-enhanced organizations*. Hoboken, NJ: John Wiley & Sons, Inc.

Teo, T., Lim, V., & Lai, R. (1999). Intrinsic and extrinsic motivation in Internet usage. *OMEGA International Journal of Management Science, 27*(1), 25-37.

Teo, T. S. H., & Ranganathan, C. (2004). Adopters and non-adopters of business-to-business electronic commerce in Singapore. *Information and Management, 42*(1), 89-92.

Treese, G. W., & Stewart, L. C. (2003). *Designing systems for Internet commerce*. New York: Addison Wesley.

Turban, E., & King, D. (2003). *Introduction to e-commerce*. Upper Saddle River, NJ: Prentice Hall.

Turban, E., King, D., Lee, J., & Viehland, D. (2004). *Electronic commerce: A managerial perspective*. Upper Saddle River, NJ: Prentice Hall.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7,* 185-198.

Venkatesh, V., & Morris, M. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly, 24*(1), 115-139.

Williamson, O. E. (1981). The economics of organization: The transaction cost approach. *American Journal of Sociology, 87*(3), 548-577.

KEY TERMS

E-Commerce: The process of transacting, transferring, or exchanging products and services over communication networks, including the Internet.

Electronic Data Interchange (EDI): The transfer of data between different companies using value-added networks, including the Internet.

M-Commerce: The process of buying, selling, or exchanging products and services wirelessly over mobile communication networks.

Perceived Ease of Use (PEOU): The degree to which a person believes that using a particular system would be free of effort (Davis, 1989, p. 320).

Perceived Enjoyment: The extent to which the activity of using the computer is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated (Davis, Bagozzi, & Warshaw, 1992, p. 1113).

Perceived Usefulness (PU): The degree to which a person believes that using a particular system would enhance his or her job performance.

Short Message Service (SMS): The transmission of short text messages to and from mobile devices, including cellular phones and PDAs.

Technology Acceptance Model (TAM): A user-behavior theory which states that user acceptance of information technology can be explained by two beliefs: perceived usefulness and perceived ease of use.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 284-290, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.25
Consumer and Merchant Adoption of Mobile Payment Solutions

Niina Mallat
Helsinki School of Economics, Finland

Tomi Dahlberg
Helsinki School of Economics, Finland

ABSTRACT

As payments by mobile phones are an enabling technology, the adoption of mobile payments is believed to significantly influence the successful emergence of electronic and mobile commerce. In recent years, several mobile payment solutions have been launched, including the Mobipay in Spain, Moxmo in The Netherlands, M-pay in UK, and Pan-European SimPay. With the exception of mobile service purchases, however, consumer and merchant acceptance of these solutions has remained marginal. We discuss consumer and merchant adoption of mobile payments and suggest drivers and barriers for this adoption. We also describe potential mobile payment application areas and identify areas in which mobile payments have the highest and lowest possibility to succeed. The information is based on extensive research conducted among Finnish consumers and merchants. The relevance of the results to other markets also is considered briefly in the discussion section of this chapter.

INTRODUCTION

Proliferation of mobile telephony during the 1990s and the success of mobile content services, such as ringtones and logos, raised high expectations for mobile commerce. Mobile commerce is a form of electronic commerce, where at least one part of the transaction is conducted via a mobile device, mainly a mobile phone. The differences between mobile and electronic commerce lie in access device and network technologies, use experience, and use contexts. While e-commerce is conducted through stationary desktop and

portable laptop computers and requires basic PC and Internet literacy, the most common devices for m-commerce are mobile phones and PDAs. These mobile devices enable transactions anytime and anywhere, require limited technical understanding, and are more personal in nature, because they commonly are used by single users, who keep the devices with them most of the time (Lyytinen & Yoo, 2002; May, 2001). On the other hand, the small size of devices and slower wireless networks limit data display, input, and transfer, as compared to e-commerce environments.

To serve the needs of the developing new mobile commerce arena, banking and telecom industries, among others, have developed mobile payment solutions. The list of applications handled by current mobile payment solutions include vending, ticketing, purchase of mobile content services (e.g., ringtones, logos, news, mobile games, etc.), electronic banking, peer-to-peer fund transfers, purchases on the Internet, and purchases of goods and services in the physical world. The most common way to make a mobile purchase is to call or send an SMS to a premium-rate service number or to send a service request to a mobile Internet site. The purchase is then charged via a monthly mobile phone bill or, in the case of prepay subscribers, deducted from a call credit. Other charging alternatives are mobile credit card billing, debiting a separate mobile account, and debiting a bank account. In Japan, the latest mobile payment solutions utilize RFID and other wireless and contactless (e.g., smart card) technologies (NE Asia online, 2004).

Since both electronic and mobile commerce environments currently lack prevailing and standardized global payment solutions, especially for micro-payments, mobile payments have the possibility to become a solution for this payment problem. For widespread acceptance and value to users, however, mobile payment solutions should be adopted in physical retailing, as well, not just on Internet and mobile networks.

To better understand mobile payment adoption, we conducted an empirical research focusing on the following three research questions: (1) Are consumers and merchants aware of mobile payment solutions? (2) What factors increase or inhibit the adoption of mobile payment solutions? (3) Which applications do consumers and merchants perceive most suitable for mobile payments? Our research draws from information systems adoption and acceptance theories, such as Diffusion of Innovations (Moore & Benbasat, 1991; Rogers, 1995) and Technology Acceptance Model (Davis, 1989; Davis et al., 1989). Empirical data were collected in 2003 among Finnish consumers and merchants with a qualitative and quantitative approach. The qualitative consumer study included focus group interviews with 46 consumers and a quantitative survey with 672 valid consumer responses. The qualitative merchant study included 15 individual merchant interviews and the quantitative study a survey with 143 valid responses. The merchants contacted represent various sectors in Finnish B2C business.

KNOWLEDGE AND EXPERIENCE IN MOBILE PAYMENT SOLUTIONS

The adoption of a new innovation goes through a five-stage process: (1) knowledge, (2) attitude, (3) decision, (4) implementation, and (5) confirmation (Rogers, 1995). Awareness and knowledge are necessary preconditions for adoption. Therefore, the amount of knowledge people have of an innovation is an important predictor of the likelihood of adoption. Furthermore, research on technology diffusion has found that user experience in terms of trials or use of previous similar technologies is an important predictor of adoption (Agarwal & Prasad, 1999; Rogers, 1995; Taylor & Todd, 1995). We investigated both the awareness and the experiences of consumers and merchants about mobile payments and the level of knowledge the groups have about these new innovations.

Consumer Knowledge and Experience

Approximately 50% of the Finnish consumers surveyed were aware of ways to make purchases or pay bills with a mobile phone. Despite this general knowledge, however, most consumers do not know about mobile payment solutions in detail, because the solutions are just emerging, and consumers do not have enough information or personal experience about mobile payments other than mobile operators' mobile phone bills. Although consumers are interested in having more information about new payment solutions, the average level of consumer knowledge suggests that consumers are not yet informed enough to make decisions on mobile payment adoption.

Consumer experiences of mobile payment solutions are still marginal, with the exception of paying for mobile content, as shown by the survey results of Table 1. These results also suggest that most consumers do not consider paying for ringtones and logos as mobile payments, since over 50% of the respondents had made such purchases, but only 11% reported that they had conducted payments with a mobile phone.

We have named the 11% of consumer respondents who have experience with mobile payments and who understand the concept better than the majority of the respondents as *early mobile payment adopters.* They are relatively younger and have higher interest in mobile technology and services. Other demographic variables, such as gender or education, were not found to be statistically significant in determining who becomes an early mobile payment adopter.

In addition to the early adopters, another interesting group of consumers is *potential users.* Over 46% of the survey respondents indicated interest in using mobile payments in the future and were identified as potential users. Compared to other respondents, on average, potential users have more education and higher professional positions, and are more interested in both technology in general and mobile technology and services in particular. Together, the early adopters and the potential users form the most likely user group for mobile payments in the future.

Merchant Knowledge and Experiences

Merchants who have participated in mobile payment pilots and tested mobile payment solutions in their businesses have the most experience with the solutions. Almost all interviewees and 46% of the survey respondents reported that they knew possibilities to offer mobile payments for their customers. Merchants are most likely to take part in pilots when they need new payment solutions, when they are interested in practical learning about a new technology, or piloting does not require

Table 1. Consumer usage of mobile services and mobile payment solutions (N=672)

Purchase (have you purchased with a mobile phone)	**% of Respondents**
Ringtones, logos	50%
Other mobile content services	30%
Ticketing, vending services	10%
Payment (have you paid with a mobile phone)	11%

too much of their resources. Our results suggest that successful pilot testing is a precondition for merchant adoption. The most likely merchant user group for mobile payments is comprised of the merchants who have participated in piloting and who find the solution suitable for their businesses.

DRIVERS FOR MOBILE PAYMENT ADOPTION

Innovations need to offer superior characteristics over existing solutions to be adopted. In technology adoption research, these characteristics often have been found to relate to usefulness, ease of use, compatibility with current values, experiences and needs, and positive effect on status and image (Brancheau & Wetherbe, 1990; Davis et al., 1989; Moore & Benbasat, 1991; Rogers, 1995).

Consumer Adoption Drivers

Our research suggests that several mobile payment characteristics are perceived as beneficial by a majority of survey respondents and, thus, drive the adoption of mobile payment solutions. The identified drivers are as follows:

- **Independence of Time and Place:** The possibility of making purchases anytime and anywhere was the most valued feature of mobile payments. Although this freedom may be limited in some user contexts and during some inconvenient times, the mobility factor is still the biggest asset of mobile payments when compared with other payment technologies (i.e., physical, electronic fixed line).
- **Availability:** Consumers find mobile phones a suitable payment device, because they have mobile phones with them most of the time, and, therefore, mobile payment technology is conveniently available in different situations.
- **Bypass of Queues:** Consumers consider avoiding or bypassing queues as one potential benefit of mobile payments. One example of bypassing queues is the possibility to purchase movie tickets by ordering the tickets from and into a mobile handset in advance and, thus, avoid queuing at the box office.
- **Substitution of Cash:** The possibility of reducing cash use is another benefit of mobile payments. Consumers do not always have enough cash with them, and, thus, the mobile payment option could be useful if cash were needed urgently. Lack of correct change may happen, for example, with vending machines, other coin functioning devices, and small payments in shops and kiosks. Compared to cash, the benefits of mobile payments are that the payer always has the exact change, and there is no need to find an ATM from which to withdraw money.

A typical example of cash replacement is car parking. After parking, a person needs exact change for a car-parking meter but may lack the right kinds of coins. Most countries with high mobile phone penetration have mobile car-parking service providers such as Mint in Stockholm, Sweden (Smith, 2004). Typically, a user calls to or sends an SMS to a premium-rate service number before parking and perhaps also after parking. Sophisticated service providers debit for only the exact parking time, and the user does not need to worry about having correct change or adding coins to a car-parking meter.

Another critical situation where cash is urgently needed is when a consumer hurries to a transportation vehicle and notices that he or she does not have enough cash to pay for the fare. Helsinki city public transport in Finland, for example, has successfully offered SMS-based

tram and subway tickets since 2001, and they currently sell over 50% of all single tram tickets via mobile phones. Mobile transportation ticketing services are provided in many major Asian and European cities.

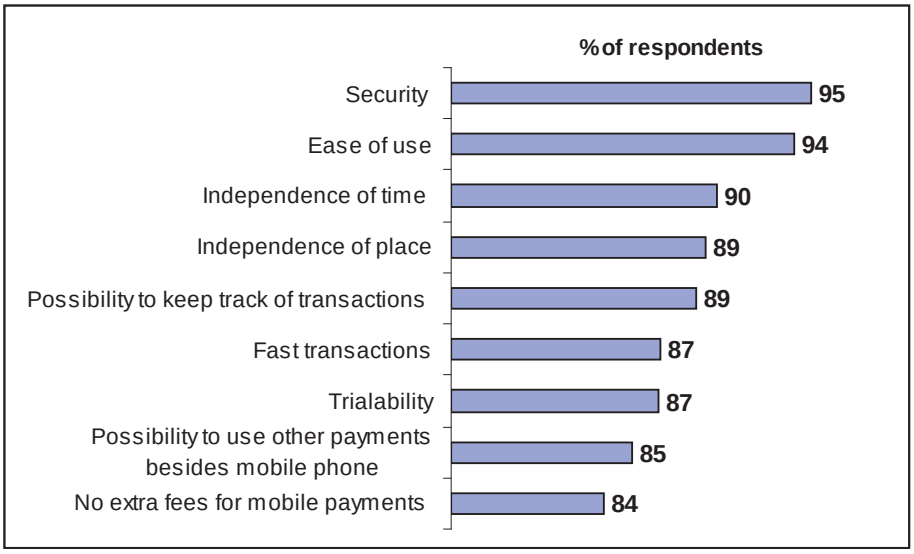
In the long run, mobile payments could replace wallets, plastic, and smart cards. Consumers favor this development, because the current multitude of cards needed for different purposes is inconvenient. However, the perceived lack of safety and security of mobile devices is a potential problem in this direction of development.

- **Economic Benefits:** Economic benefits such as price discounts, loyalty programs, and purchase bonuses are lucrative for consumers. These economic benefits could increase consumer interest in mobile payments, especially during service launch.
- **Other Benefits Consumers Require:** Currently used payment instruments (typically debit cards, credit cards, and cash) are easy and fast to use, inexpensive, widely accepted, and secure. Moreover, cash is currently the only payment method that allows anonymous payments and person-to-person money transfers, (e.g., lending money to a friend). To be able to compete with the dominant payment instruments and to fulfill consumer expectations, mobile payments should offer these benefits, too. Figure 1 presents the factors consumers consider most important in mobile payment solutions, according to our survey.

Merchant Adoption Drivers

Merchants benefit from the new mobile payment solutions, if they are able to offer new services, increase sales, or reduce costs. These benefits are dependent on the market penetration and costs of implementing and using the solution. To ensure

Figure 1. Mobile payment characteristics perceived as important by consumers (N=672)



widespread market penetration, merchants are interested in solutions that add value to consumers and are likely to be adopted. Given that the pricing of mobile payment solutions is competitive and that many consumers use these solutions, interviewed merchants identified several possibilities to offer new services and products, to enhance existing ones, to increase sales, and to reduce costs.

- **Increased Impulse Purchases:** Mobile payments enable consumers to buy more things on impulse and, therefore, have the potential to increase merchant sales. For example, there may be an advertisement in a newspaper, on TV, or on the radio, along with instructions on making a purchase with a mobile phone. The possibility of the purchase increases when consumers can make the purchase instantly after receiving an impulse.
- **New or Better Services:** Mobile payments enable merchants to offer new or better services to customers. Especially the micropayment price category includes products and services that are impossible to offer with current payment instruments. An efficient mobile payment solution could inspire merchants to create new, additional, or enhanced services that would increase their overall sales.
- **Enhanced Product and Service Availability:** Mobile payments have the capacity of enhancing the availability of products and services and of creating an additional order and delivery channel. Purchasing becomes independent of time and place and more convenient to customers. Especially digital products that can be transmitted directly to customers' mobile phones benefit from the enhanced remote purchase possibilities. Transport for London, for example, offers customers a possibility to pay for the congestion charge by mobile phone using an SMS text message. In early 2004, 20% of the payments were handled via text messages, and their usage continues to increase at the expense of retail and call-center channels (TfL, 2004).
- **New Customers:** Mobile payments may attract new customers, especially if purchases become easier and more available. Merchants estimate that mobile payments most likely attract young people who are experienced mobile phone users, teenagers who do not have debit cards, individuals who do not carry much cash with them, and technologically oriented persons who may prefer the use of mobile phones to other payment alternatives. McDonald's Slovenia, for example, recently teamed up with the local Mobitel mobile network operator to better serve those customers who prefer non-cash payments (Mobitel, 2004).
- **Improved Company Image:** For some merchants, mobile payments may improve their companies' images. Consumers regard them as innovative forerunners in their industries. Merchant opinions on the image are somewhat contradictory, however, and some perceive that the solutions are currently too immature to have a positive effect on company image.
- **Lower Commissions and Costs:** Merchants would benefit if mobile payment service providers charged lower commissions than the commissions charged on competing payment alternatives (e.g., credit cards). At least, the costs of mobile payments should equal the costs of competing alternatives.
- **Faster Settlements:** Similar to lower costs, payment settlement times should be faster than or equal to those of competing alternatives in order for mobile payments to be competitive. Only if single payment clearances are very small may merchants prefer the bundling, clearing, and netting of payments into larger sums.

- **Improved Efficiency:** Mobile payments could reduce costs, if they improve the efficiency of payment process or reduce employer work on payment processing. One example of improved efficiency would be the transformation of manned service stations to unmanned self service stations during off-peak hours.

BARRIERS TO MOBILE PAYMENT ADOPTION

In addition to adoption drivers, innovations also face adoption barriers. Barriers are complexities in understanding and using the innovation, costs of adoption and use, and perceived risks (Davis et al., 1989; Moore & Benbasat, 1991; Rogers, 1995). Recent research in electronic and mobile commerce has discovered that trust and security play a significant role in the adoption of these new forms of commerce (Gefen et al., 2003; Jarvenpaa et al., 2000; McKnight et al., 2002; Pavlou, 2003).

Consumer Adoption Barriers

Several factors currently limit consumer adoption of mobile payments. The following barriers were identified in our empirical study, starting from the most common.

- **Premium Pricing:** Many consumers abstain from paying with mobile phones because of premium pricing. Consumers are especially reluctant to pay separate fees for mobile payment services. Bundling mobile payment fees with other financial services would be a more acceptable solution. The perceived advantages of the mobile payment service will tend to increase the willingness to pay for the service.
- **Limited Possibilities to Use:** Lack of widespread merchant acceptance prevents consumer adoption. Mobile payments are currently a marginal payment technology and are not available or accepted widely enough. Mobile payments should become more common in daily purchase situations so that consumers would learn the new payment instrument and become accustomed to using it.
- **Lack of Convenient Billing Solutions:** A majority of consumers perceive current mobile payment billing solutions as difficult. Separate accounts require inconvenient preparations before use, including registration and opening the account. During use, it is difficult to keep track of the balance of the account and to transfer money to and from the account. Mobile phone or credit card billing may end in an unexpectedly large bill at the end of a month. Finally, bank account statements may be littered with small account entries, if a bank account is used to debit numerous small value mobile payments.

  While billing preferences may vary from country to country, our results suggest that mobile payment solutions should be easy to use and compatible with existing payment instruments and financial services. Lack of compatible billing procedures seems to reduce or postpone consumer adoption of mobile payment solutions.
- **Trust and Security:** Finnish consumers regard banks, credit card companies, and telecom operators as reliable mobile payment service providers. Of these enterprises, banks are perceived as slightly more reliable than others, probably because Finnish consumers have learned to trust banks due to their long customer relationships with banks. In general, the results indicate that, in mobile payment services, consumers are likely to rely on traditional financial institutions, which they also have learned to use for their other financial affairs. Tele-

com operators are perceived as reliable, because consumers expect them to have the knowledge and skills about the new mobile technology. Small and unknown payment service providers, including startups, were perceived as unreliable by consumers.

Despite the general reliability of the dominant mobile payment service providers, consumers perceived several risks in using mobile payments. The perceived risks identified in our study fall into the following six categories:

1. Unauthorized use of the payment instrument
2. Transaction errors
3. Lack of transaction record and documentation
4. Vagueness of the transaction verification
5. Privacy concerns
6. Device and mobile network reliability

To reduce perceived risks, mobile payment service providers should ensure that the payment infrastructure is secure and reliable. They also need to communicate the terms and policies of the service to consumers and provide enough feedback and documentation during the use of the service. Transaction errors, unauthorized use, and the like can be prevented with certificates such as security PIN codes (e.g., a WPKI digital signature). There is, however, a tradeoff between security and ease of use. Although certification techniques increase the security of the service, they also make it slower and more difficult to use. It is the challenging task of service providers to find the golden mean between security and ease of use.

- **Lack of Information:** Many consumers are interested in mobile payments but have little information. Perceived trust and security risks clearly indicate this. As mobile payments are not yet widely used and, therefore, not visible to consumers, better methods to spread information about the new payment services are needed.
- **Difficulty of Use:** Consumers expect that mobile payment solutions should be easy to use, whereas current experiences suggest the opposite. Difficulties lie both in the introduction phase of the new payment technology and during continuous use. The introduction phase may appear difficult, if the mobile payment service requires complex registration procedures and separate billing arrangements. During continuous use, the payment procedure should be simple (i.e., press a single key) and fast. If payments are conducted with SMS messages or mobile Internet browsing, however, the message formats often are complicated and slow to key in, and the various codes and premium service numbers or sites are difficult to remember. It is also difficult to keep track of the conducted mobile payments due to lack of documentation and to find instructions to conduct mobile transactions, not to mention charge-back situations.
- **Sufficient Existing Payment Instruments:** Current payment instruments are widely accepted and cover a majority of purchasing situations. With the exception of lack of cash and problems encountered in some special occasions, there are few situations in the physical world where new payment instruments are clearly needed. Further, as purchases on the Internet and via mobile networks are still infrequent, the lack of mobile payment instruments in networks does not appear to be a significant disadvantage for the majority of consumers. In the survey study, only 5% of the respondents indicated that they would consider mobile payments as their primary payment technology. This result suggests that existing payment methods are sufficient for the majority of the respondents most of the time.

- **Employer-Paid Mobile Phone Bills:** Consumers whose mobile phone usage and subscription fees are paid by their employers may have limited access to mobile purchases and payment services due to restrictions set by their employers.

Adoption Barriers for Early Adopters and Potential Users

Early adopters and potential users find the adoption barriers significantly lower than other consumers. Compared to consumers in general, early adopters have a better understanding of mobile payments, show more trust in mobile payment service providers and technology, perceive fewer risks in payments, and consider mobile payments easier to learn and use. Service cost, however, is a critical barrier for early adopters and potential users as well. To conclude, pricing is an especially important mobile payment adoption factor.

Merchant Adoption Barriers

Similar to consumers, merchants also experience barriers for mobile payment adoption. The following are barriers identified in our empirical study.

- **Incompatibility with Business Practices:** One significant barrier to the merchant adoption of mobile payments is the perceived incompatibility of mobile payment technology with existing business practices. For example, while mobile payments are especially suitable for digital content, their applicability in supermarket checkout, is less obvious. Merchants' businesses largely determine whether they find mobile payments compatible or not.

  Further, many merchants feel that the current mobile payment solutions have been developed to serve the payment service providers' interests, whereas the benefits to merchants and consumers have not been emphasized.
- **Lack of Users and Use:** Another essential problem with mobile payments is that too few consumers currently use this payment technology. Only a large amount of users makes it profitable for merchants to offer payment with a new instrument. For some merchants, widespread adoption is acceptable in the long run, whereas others require high usage rates before they adopt mobile payment solutions.
- **Investment and Usage Costs:** Similar to consumer adoption, the high costs of mobile payment solutions also prevent merchant adoption. In particular, the commissions of mobile payment service providers were perceived to be too high, even to the point that it becomes unprofitable for merchants to offer certain mobile services.

  Cost-effectiveness and competitive pricing are prerequisites for mobile payment adoption. Commissions should be competitive with or lower than the commissions of the current payment methods. Further, mobile payment service providers should support the early adoption of mobile payments with reduced fees or free-of-charge trials until the solution gains widespread acceptance and use. Currently, merchants see little or no cost advantages in mobile payment solutions.
- **Difficulty of Use:** Merchants are concerned about the usability of mobile payment solutions, which they see as an essential factor in consumer adoption. These solutions become commonly used, if they are simple and fast to use. Many merchants perceive current solutions as cumbersome and slow to use. Especially SMS messages and separate mobile payment accounts are seen in this light. Speed is considered important by retailers as well as by many digital service providers.
- **Lack of Standards:** Lack of standards between new mobile payment solutions is

another barrier for adoption. Merchants expressed three generic standardization requirements for mobile payments. First, mobile payment solutions need to be independent of single banks, telecom operators, and devices in order to be usable for a large variety of consumers. Second, they need to be compatible with cash register and electronic payment technology (standards) at point of sale. Third, mobile payment solutions should support mobile commerce and payment transaction roaming between service providers, because there is little room in the market for several new, competing, and incompatible payment standards. Different solutions may confuse consumers and make it too costly for merchants to provide services for all the different payment standards. The worst outcome is that both merchants and consumers become disinterested, and all new solutions remain unused. Cooperation and standardization between payment service providers is, therefore, recommended.

- **Trust and Security:** Trust in payment service providers and security of payment solutions are other prerequisites for merchant adoption of mobile payments. The surveyed merchants considered mobile payment solutions as fairly secure. Similar to consumers, Finnish merchants perceive banks and mobile operators as reliable mobile payment service providers. However, approximately one-half relied more on banks in security issues, including certification.

Precise and timely settlements are an important means of enhancing merchant trust in mobile payments. Verification and certification during payment transaction processing confirms that the payment has been successful and that the payee's account is credited. The reliability of the payment infrastructure in all situations is another critical security issue, as reliability prevents errors, fraud, and lost sales.

PREFERRED APPLICATIONS

Our results indicate that both consumers and merchants have clear ideas about suitable applications when they are asked to consider the use of mobile payments.

Applications Preferred by Consumers

Our findings suggest that consumers are most interested in using mobile payments in small value purchases, where mobile payments are used instead of cash instruments or to pay for digital content. In the survey, we asked consumers to list those purchases for which they would be most interested in paying with a mobile phone. Table 2 shows that the most often selected applications for mobile payments are mobile content and services, mobile banking and money transfer, and payments related to various types of travel. The interest in paying for the listed items with a mobile phone is significantly larger than those currently using mobile phones to pay for these services. It is possible that this finding relates more to expected removal of barriers than to drivers. Although the indicated interest does not necessarily result in future use, it is a good sign for the latent use potential among consumers.

Merchant Preferred Applications

Merchants regard mobile payments as most suitable for small value purchases of digital content and services, because they can be sent directly to a mobile phone. Ticketing is especially seen as a potential service. Also, mobile content purchases, vending and parking services, remote payments, and self-services in general were seen to be well suited for mobile payments. Table 3 shows the most suitable applications for mobile payments, as listed by the interviewed merchants.

In the merchant survey, the respondents were asked to choose products and services that are best

Table 2. Consumer interest toward mobile payment applications (N=672)

Purchase/Payment	% of Respondents
Ringtones, logos	57%
Mobile content such as weather, news, directory, route information, or games	55%
Car parking	54%
Ticketing, such as movie or concert tickets	52%
Mobile banking, such as bill payment	52%
Money transfer from person to person	51%
Taxi	44%
Car wash, refueling, and other car services	42%
Long-distance traffic ticket	41%
Vending machines for drinks, candies, passport photos, video rental	39%
Local transportation fares	38%
Hotels, accommodations	37%
Betting, lottery, gaming	36%
Lockers, storage rooms	35%
Lunch, fast food	33%
Restaurants	33%
Airline ticket purchases via the Internet	31%
Groceries	31%
Purchases in kiosks	28%
Internet purchases of books, clothing, music, games, films	27%
Consumer durables, such as electronics, CDs, videos	23%
Purchases on TV shop and digital-TV	21%
Dealing in shares and securities, related mobile services	16%

Table 3. Suitable products and services for mobile payments listed by the interviewees

Product/Service
Ticketing (e.g., events, travel, entrances, movies)
Vending machine purchases (e.g., drinks, candies, newspapers)
Digital and mobile content and services (e.g., news, weather, games, music)
Parking
Self-services, purchases from unmanned service stations (e.g., carwashes, travel insurances)
Remote payments, subscriptions, home deliveries

suited for mobile payments in the future from a list similar to that in the consumer survey. Of the 143 respondents, 19 selected Internet purchases, 17 daily consumer goods, 13 lunch and fast food, and 11 consumer durables. The relatively high figures of retail and restaurant purchases may be explained by the industries of the respondents: 28% of respondents represented restaurants and 17% stores and supermarkets.

The results indicate that a majority of the merchants have opinions similar to consumers and see mobile payments to best suit purchases of mobile content, ticketing, vending, and information networks. A small minority of retailers were willing to try mobile payments in physical point of sale, but the potential of mobile payment applications still lies in the digital goods and services and in small-value, real-world payments.

CONCLUDING REMARKS

The recently introduced new electronic payment technology—mobile payment—has come boldly to the market to conquer its place among established payment instruments such as cash, credit cards, and debit cards. Since it is still in its infancy, the adoption rates of the new payment technology have been modest, and, for the most part, mobile phones are used to pay for mobile services and content. This chapter has discussed consumer and merchant adoption of mobile payments in terms of awareness, adoption drivers and barriers, and preferred applications, based on empirical evidence provided by consumer and merchant research.

The word *curiosity* best characterizes current consumer and merchant awareness of mobile payments. While mobile payment service providers have piloted their solutions, often with low profiles, more efficient marketing is needed to bring the new services to the consciousness of intended users. Consumers need information about the use and benefits of these solutions, and merchants need information about the advantages in relation to their specific business.

Consumer adoption of mobile payment technology is driven by independence of time and place, the possibility to avoid queues, the ability to substitute cash, and envisioned economic benefits. Yet, mobile payments also should offer ease of use, widespread merchant acceptance, speed, and security matching that of the traditional payment technologies and instruments.

Merchant adoption drivers include sales increases and cost reductions. Sales increases take place through increased impulse purchases, easier or faster payments, new customers, or enhanced opportunities to sell products and services. Cost reductions may result from lower commissions or payment processing costs, more efficient payment processes, or faster settlement of payments. Some merchants and service providers highlight the importance of positive image effects, as well.

The primary barriers for consumer mobile payment adoption are premium pricing, cumbersome SMS interface, limited merchant acceptance, perceived risks, and satisfaction with existing payment instruments. Correspondingly, merchant barriers include uncertainty about the profitability and other benefits of mobile payments, low penetration rates of mobile payments among consumers, and non-standardized solutions with many payment service providers and competing technologies. Some merchants also perceive mobile payments as incompatible with their business.

There are three applications that are seen to best suit mobile payments. First, in most situations, a mobile phone is the most convenient and sometimes the only possible payment technology for mobile content and service purchases. Therefore, we expect that mobile payments will continue to be used in mobile and digital services, including content such as electronic ticketing. Second, the diminishing use of cash in the physical world makes it important to develop new compensatory payment instruments for small value purchases on

automated machines, cash desks, and self-service stations. Mobile payments have the opportunity to provide efficient payment instruments to complement cash payments. Finally, there is a need for a cost effective means to charge small value payments in remote electronic commerce. Mobile payment instruments are suitable candidates in this area, as well.

Are our findings relevant to other markets? As our empirical findings are based on a sample on a single country, we offer the following general suggestions derived from our experience, from international contacts with experts in the field, and from literature review.

Payment cultures, instruments, jurisdictions, and infrastructures differ significantly between nations. For example, the U.S. payment system is heavily characterized by check use; Japan relies on cash; and most European countries rely on bank giro or postal giro payments, all with significantly different local payment infrastructures (Hancock & Humphrey, 1998). Another example concerning regional regulation comes from the EU jurisdiction, where a limited credit institution concept was introduced recently to allow limited credited commerce for third-party services and products. This makes it possible for mobile operators to offer third-party mobile services via an operator's network and charge for the services with an operator's phone bill without acquiring a banking license, which was required previously (T2R Final Report, 2003).

Caution with the generalization of our findings is necessary, since mobile payment instruments introduced to specific markets most likely will differ and integrate regionally or globally only gradually with overall economic integration and international network roaming. If mobile operators are able to establish international mobile payment transaction roaming similar to international mobile voice roaming, local mobile payment solutions could become internationally accessible. As has happened with physical and fixed-line electronic payments, mobile credit card instruments supported by local issuers have a high probability to succeed. Consistency with local payment culture, jurisdiction, existing payment infrastructure, and instruments is a prerequisite for the adoption of mobile payments. On the other hand, we believe that within the context of cultural, jurisdictional, and infrastructure differences, the fundamental factors that make a new technology appealing to consumers and merchants are most likely similar, although some specific adoption drivers and barriers would differ. We expect this to be true, especially in the context of mobile service purchases.

Interest in the future use of mobile payments is notably higher than the current use (A.T. Kearney, 2002). Although actual future use cannot be predicted directly from indications of interest, research results suggest that intended users are willing to consider the adoption of this new payment technology. To make new payment services successful, mobile payment service providers need to ensure that the advantages of adoption drivers exceed the disadvantages of barriers and that the new payment technology is competitive with established traditional payment technologies.

In the short run, mobile payments are not likely to substitute traditional payment technologies but will be used in digital environments and in the physical world, when cash is not available, payment cards are not accepted, or it is inconvenient to pay with a payment card. In the long run, it is possible that payment cards will be integrated into mobile devices. This would result in a more profound change in payment instruments and would require years to diffuse. Before that, mobile payment solutions need to overcome their current adoption barriers and establish a solid reputation and a position as a viable payment technology.

REFERENCES

Agarwal, R., & Prasad, J. (1999). Are individual differences germane to the acceptance of new

information technologies? *Decision Sciences, 30*(2), 361-391.

A.T. Kearney & Judge Institute of Management, Cambridge University. (2002). Mobinet 4 study. Retrieved October 4, 2004, from *http://www.atkearney.com/main.taf?p=5,4,1,50*

Brancheau, J.C., & Wetherbe, J.C. (1990). The adoption of spreadsheet software: Testing innovation diffusion theory in the context of end-user computing. *Information Systems Research, 1*(2), 115-143.

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3) 319-339.

Davis, F., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1002.

Gefen, D., Karahanna, E., & Straub, D.W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly, 27*(1), 51-90.

Hancock, D., & Humphrey, D.B. (1998). Payment transactions, instruments, and systems: A survey. *Journal of Banking & Finance, 21*, 1573-1624.

Jarvenpaa, S.L., Tractinsky, N., & Vitale, M. (2000). Consumer trust in an Internet store. *Information Technology and Management. 1*(1-2), 45-71.

Lyytinen, K., & Yoo, Y. (2002). Research commentary: The next wave of nomadic computing. *Information Systems Research, 13*(4), 377-388.

May, P. (2001). Mobile commerce—Opportunities, applications, and technologies of wireless business. Cambridge University Press.

McKnight, H.D., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(13), 334-359.

Mobitel. (2004). Paying with moneta also in McDonald's. Retrieved October 4, 2004, from *http://www.mobitel.si/eng/*

Moore, G.C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

NE Asia Online. (2004). NTT DoCoMo to debut i-mode smart-card handsets capable of making electronic payments. Retrieved October 4, 2004, from *http://neasia.nikkeibp.com/wcs/leaf/CID/onair/asabt/news/314459*

Pavlou, P.A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce, 7*(3), 101-134.

Rogers, E.M. (1995). *Diffusion of innovations.* New York: Free Press.

Smith, B. (2004). Time's up for parking meters. *Wireless Week, 10*(13), 24.

T2R Final Report. (2003). Trusted transaction roaming EU project final report. Retrieved October 4, 2004, from *http://www.radicchio.cc/home/t2r.htm*

Taylor, S., & Todd, P.A. (1995). Assessing IT usage: The role of prior experience. *MIS Quarterly, 19*(4), 561-570.

Transport for London. (2004). Congestion charging central London, impacts monitoring, second annual report. Retrieved October 4, 2004, from *http://www.tfl.gov.uk*

This work was previously published in Managing Business in a Multi-Channel World: Success Factors for E-Business, edited by T. Saarinen, M. Tinnila, and A. Tseng, pp. 32-50, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.26
An Electronic Auction Service Framework Based on Mobile Software Agents

Sheng-Uei Guan
National University of Singapore, Singapore

INTRODUCTION

Electronic Auction Overview

With electronic commerce revolutionizing the traditional way of doing business, electronic auction service has been one of the many business models that were proven to be a success. The existence and development of numerous auction Web sites, such as eBay (www.ebay.com) and OnSale Inc. (www.onsale.com) have demonstrated the survivability of electronic auctions in online transactions. Considering some of the new forms of electronic auctions currently on the Internet, such as the "Get it together" network (www.accompany.com), where group bidding and negotiation is applied, it could be said that the definition of auctions is no longer restricted to that of its traditional meaning but also has been extended electronically. An auction may be an ideal way for a business to sell excess inventory and goods because it has attracted many of the common people that do not really participate in the real-world counterpart. However, current Web-based auction (e-auction) systems suffer from shortcomings in the following aspects:

- **Fairness and Friendliness:** Different conditions of Internet connections, such as varying speeds, introduce unfairness among participating bidders.
- **Security and Privacy:** The messages transmitted via the Internet are exposed to malicious attacks and may incur security problems. Also, in an auction, users may wish to be guaranteed privacy, for example, a bidder may not want to disclose his or her real identity until the auction closes and he or she is declared the winner.
- **Intelligence and Flexibility:** It is important for an e-auction service to be intelligent to cater to the needs of potential auction customers who are not into the Internet. However, current Web-based auction sys-

tems require too much user intervention. Because the process can be tedious and risky for these users, they may not want to engage in e-auction services. Thus, it would be commercially profitable if intelligent assistance is provided.

Software Agents: A Paradigm for Mobile Computing

The popularity of the Internet as the platform of electronic commerce not only brings opportunities but also challenges in organizing information and facilitating its efficient retrieval (Pham & Karmouch, 1998). Many researchers believe that the mobile software agent paradigm could propose attractive solutions to deal with such challenges and problems.

Mobile agents refer to self-contained and identifiable computer programs that can move within the network and act on behalf of the user (Pham & Karmouch, 1998). Despite the current differences in definition, the mobile agent paradigm as reported in the literature has two general goals: reduction of network traffic and asynchronous interaction. Research on agent-based e-commerce is still underway (Franklin & Reiter, 1996; Maes, Guttman, & Moukas, 1999; Poh & Guan, 2000; Subramanian, 1998; Yi, Wang, Lam, Okamoto, & Hsu, 1998). Mobile agents have demonstrated tremendous potential in conducting transactional tasks in e-commerce. The architecture we are proposing here, compared to most of the current practices on the Internet, is based on mobile agents. Specifically, the features of our system will be as follows:

- **Fairness:** The deficiency of excessive network traffic will be overcome.
- **Autonomy:** Based on the preferences of a user, agents can be fully automated to participate in the auction with little or even no intervention from the user. The bidding strategies are self-contained in the agents and can be changed dynamically. Users can still control the behavior of the agents by remote monitoring.
- **Security and Privacy:** We are introducing third-party involvement to enhance the security and privacy throughout the auction. By security, we mean that agents are protected from malicious attacks during transportation and bidding. By privacy, we mean that with the assistance of the coordinator and the encryption mechanism, the real identity of each participating bidder is protected.
- **Flexibility:** The architecture we have proposed will serve as a unified framework for various auction types, for example, English auctions, Dutch auctions, and so forth, as long as the bidding strategies and competing rules are well defined.

RELATED WORK

Background

Auctions are more complex than people can realize (Agorics, n.d.). There are different ways to classify auctions. There are open auctions as well as sealed-bid auctions. Generally, there are five major auction formats: English, Dutch, First-Price Sealed-Bid, Vickrey (uniform second-price), and Double auctions (see Table 1). One difficulty is the lack of commonality in naming conventions.

Related Research

While electronic auctions are complex, they are also equally popular and desirable. Consequently, much research has been conducted in the area of electronic auctions and in particular agent-based auction systems.

Table 1. Types of auction

Auction Types	
Type	**Rules**
English, or ascending-price Open	Seller announces reserve price or a low opening bid. Bidding increases until demand falls. Winner pays the highest valuation. Bidder can re-assess evaluation during auction.
Dutch, or descending-price Open	Seller announces a very high opening bid. Bid is lowered until demand rises to match supply.
First-price sealed-bid, known as discriminatory auction when multiple items are being auctioned	Bids submitted in written form without knowledge of bids of others. Winner pays the amount he bid.
Vickrey auction or second-price sealed-bid, known as uniform-price auction when multiple items are being auctioned	Bids submitted in written form without knowledge of the bids of others. Winner pays the second-highest bid amount.
Double auction	Sellers and buyers submit bids at the same time. Bids are matched at a middle point.

The Michigan Internet AuctionBot

The Michigan Internet AuctionBot is a project carried out at the University of Michigan, Artificial Intelligence Laboratory. It sees itself as an information service that collects the bids, determines the resulting price, and notifies the participating parties about the outcome.

The Fishmarket Project

The Fishmarket project at the Artificial Intelligence Research Institute in Barcelona evaluates a very narrow field of electronic commerce. Its main focus lies in rebuilding a commerce structure that is found in real life on downward-bidding fish markets of Spain, and it supports Dutch auction style. Mobile agents are not supported.

CASBA

The CASBA (Guttman et al., 1998) project at the Technology Management of University of Stuttgart offers flexibility and support for all common auctions types including auctioning of multiple units. It does not have sophisticated negotiation strategies and learning mechanism to improve agent performance on the market. It is not designed with mobile agent capability.

The KASBAH

The KASBAH project of the AmEC Initiative at the Massachusetts Institute of Technology introduced agents that negotiate following three time-constrained rules.

ARCHITECTURE FOR THE AGENT-BASED AUCTION SYSTEM

A complete auction service involves the following aspects: information shopping, auction process, payment, and shipping. During the auction process, the bidders compete according to the published bidding rules but may use their own bidding strategies. When the auction is closed, the auctioneer and winner will identify each other and further complete the payment and shipping matters.

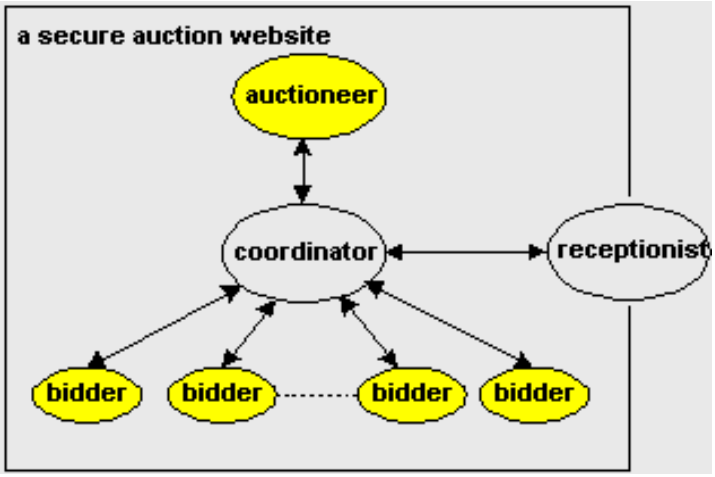
Overview

Let us first give a typical scenario of an English auction: the buyers gather together to bid for a certain product, according to the published rules and preferred strategies. In the proposed architecture, we typically have the following:

Table 2. Fuctions of the participating agents

Participating Agent	Owner	Function
Auctioneer	Seller	Decide the winner
Bidders	Customers	Bid
Coordinator	Third party	Coordinate auctions
Receptionist	Third party	Receive agents

Figure 1. A typical auction scenario



- **Participants:** They are the auctioneer agent, the bidder agents, the coordinator agent, and the receptionist agent. The functions and particulars of each agent are listed in Table 2.
- **Place:** The auction Web host is a secure auction environment provided by a third certified party widely trusted by the participants. Instead of the buyer agents roaming to the seller's host to perform the auction, we use this environment to ensure high security.

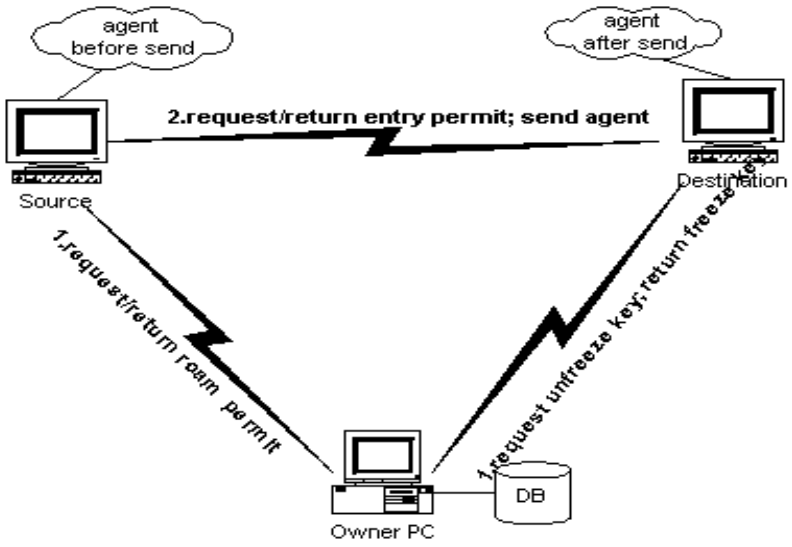
A typical auction process is divided into three stages, namely, the admission, bidding, and conclusion stage. In the admission stage, all participants will be received by the receptionist and followed up with necessary registration procedures for the auction. The bidding process is the stage in which the bidders would elect whether to submit bids or to remain silent to compete using its owner-customized strategies. In the final conclusion period, the auctioneer will decide on the final winner, and with the help of the coordinator, the auctioneer and the winner will identify each other and exchange a token of the auction result for nonrepudiation purpose.

Admission

Admission is the preparation, namely, the reception of agents and the build up of the auction relationship. The admission process will be further divided into two periods: *SAFE* transport and auction registration.

SAFER (*Secure Agent Fabrication, Evolution and Roaming for electronic commerce)* (Guan & Yang, 1999; Yang & Guan, 2000; Zhu, Guan, Yang, & Ko, 2000) has been proposed as a framework for intelligent mobile agent mediated e-commerce. SAFER agent transport protocol allows intelligent agents to roam from one host to another in a secure fashion. Our system adopts one of the three proposed transport protocols, the supervised agent transport protocol for the secure roaming of agents to prevent agents from malicious attacks

Figure 2. Supervised agent transport protocol in SAFER



during their transportation. Figure 2 illustrates the supervised agent transport protocol.

After the agents have successfully roamed to the destination—the secure auction host provided by a third trusted party—all agents are welcomed by the auction receptionist. The agents then communicate with the receptionist to complete other registration formalities

The receptionist will not close the admission process even if after the auction starts, as there will probably be some late arrivals and intermediate withdrawals.

Bidding

Once the bidding period starts at the published time, the bidder agents start submitting bids. Each bidding agent is equipped with the owner-customized bidding strategies as instructions for submitting bids.

The bidding period is divided into several rounds: in each round, bidders may elect to submit bids to compete or remain silent. Note that throughout the bidding process, an agent uses its alias to communicate with each other and the bidders are not allowed to communicate directly with the auctioneer but have to do so via the intermediate coordinator, and vice versa.

The coordinator together with the encryption mechanism used is instrumental in achieving the following goals:

- First, the auctioneer is kept blind from the bidding process, in that the auctioneer can verify the validity of each bid, but is not able to know actually who has submitted bids;
- Second, all the bidders are notified of the highest successful winning bid in each round, together with the alias of the originator. This is to facilitate the decision of the bidding strategies of intelligent agents to be adopted in the next round as some agents may wish to watch carefully the bidding situation and trace the strategies of the other bidders.

Let us further consider two exceptional cases that might occur during this stage: early withdrawal and late arrival. An early withdrawal is the case when an agent decides not to bid anymore and wishes to retreat before the auction is finally

Figure 3. Communication between agents and their receptionist

closed. Before he or she leaves, he or she needs to consult with the receptionist with his or her withdrawal and obtains permission from him or her so that the receptionist may forward the most updated bidding status to the coordinator. On the contrary, a late arrival is the case when a bidding agent fails to catch up with the starting time of the auction but wishes to participate anyway. In this case, despite the requirement that agent must follow the standard procedure before entering the auction, he or she may also need to consult with the receptionist about the latest bidding situations.

Conclusion

The final stage of the auction is conclusion, in which the auctioneer announces the result of the auction and the final winner and the auctioneer identify each other to ensure nonrepudiation with the assistance of the coordinator.

IMPLEMENTATION

A prototype has been developed and implemented to prove the feasibility of agent-based auction systems. Consequently, it has been shown that such a system can be successfully implemented. It is deployed to ensure that security, privacy, user anonymity, and fairness are attainable in agent-based e-auctions.

Overview

In this project, a typical English auction has been chosen as the prototype implementation. At the auction host, the buyers gather together to bid for a certain product according to the published rules and the preferred strategies.

Descriptions of the Prototype

At the time of implementation, Java is chosen to build the prototype. The three main components realized in this prototype are as follows:

1. The user interface consists of the agent factory panel and the auction host panel. The user must fill in the required parameters before submitting the request to the agent factory to fabricate a bidding agent. The auction panel is for sending a fabricated agent to the auction host.

Figure 4. Auction system architecture

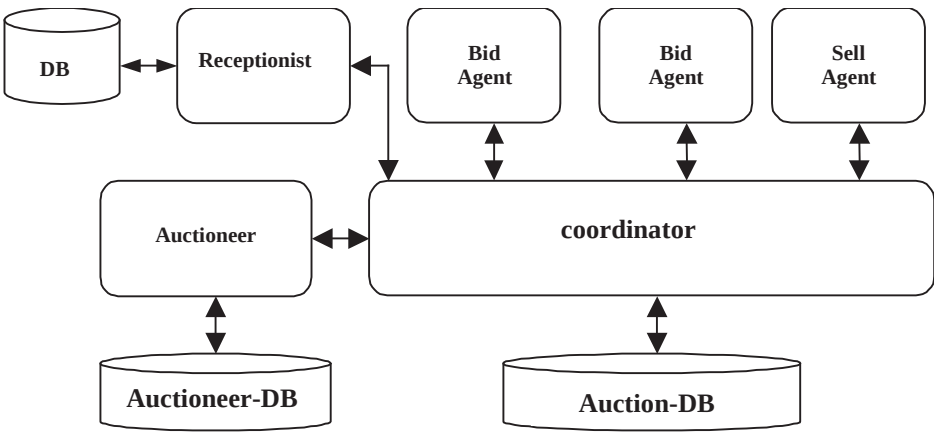


Figure 5. Implemented auction system overview

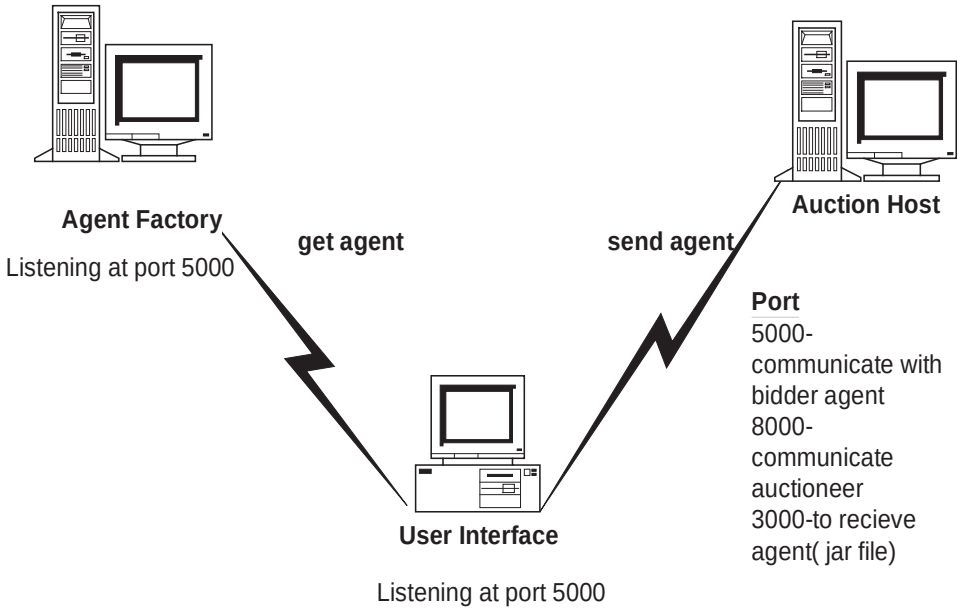
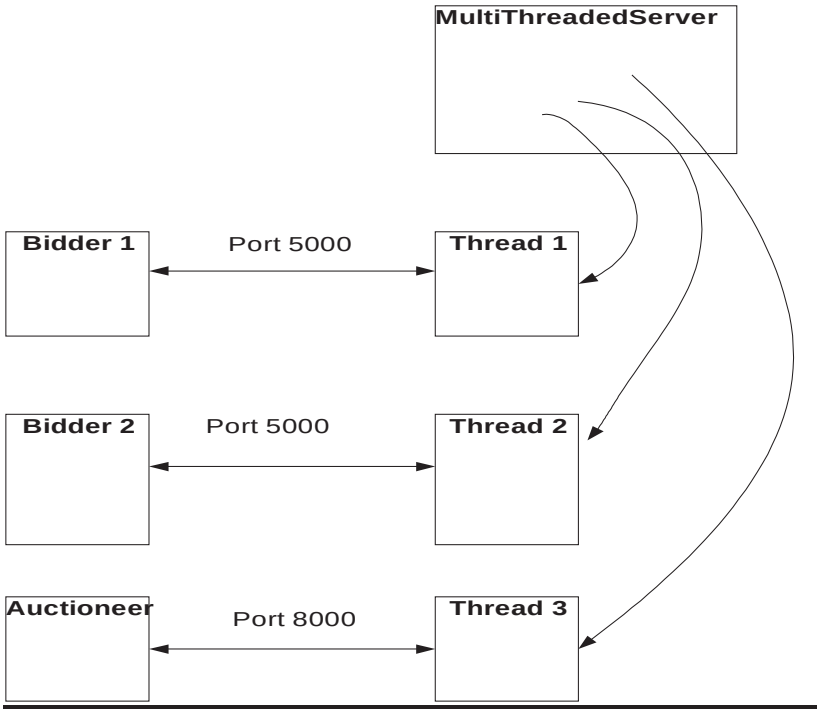


Figure 6. Multithreaded auction server



2. The agent factory, where a bidding agent is fabricated according to the user's needs.
3. The auction host system, whereby the auction is conducted.

In this prototype, a user customizes an agent by setting parameters such as user identification number, maximum and minimum bids, user IP address, port number, and the desired product he or she wants to buy. When the user clicks the submit button, a Java agent is automatically generated according to these parameters. The agent is sent to the user's machine.

The user will receive the Java agent in the form of a ".jar" file. This archive file basically contains two files, namely, the preprogrammed Java class file "Agent.class" and the "par.dat" text file. The "par.dat" file contains the user's original agent parameters, and also the agent ID assigned by the agent factory. The user has to send the Java agent to the server machine by means of a socket connection. Once on the auction host's machine, the two files are extracted and copied into the respective class directories. The "Agent.class" file is a stand-alone class file invoked to become a bidder agent. The bidder agent initializes itself by reading the "par.dat" file.

Running on the auction host is a multithreaded server "coordinator" that talks to the bidding agents and the auctioneer agent in a synchronized manner as shown in Figure 6. Once the bidding agent is invoked, it begins to establish a socket connection to its home host.

Screenshots

In this section, the screenshots of the various components are shown and explained.

- **The User Interface:** The user interface (Figure 7) consists of several parameter fields that must be filled in before submission to

Figure 7. Screenshot of user interface

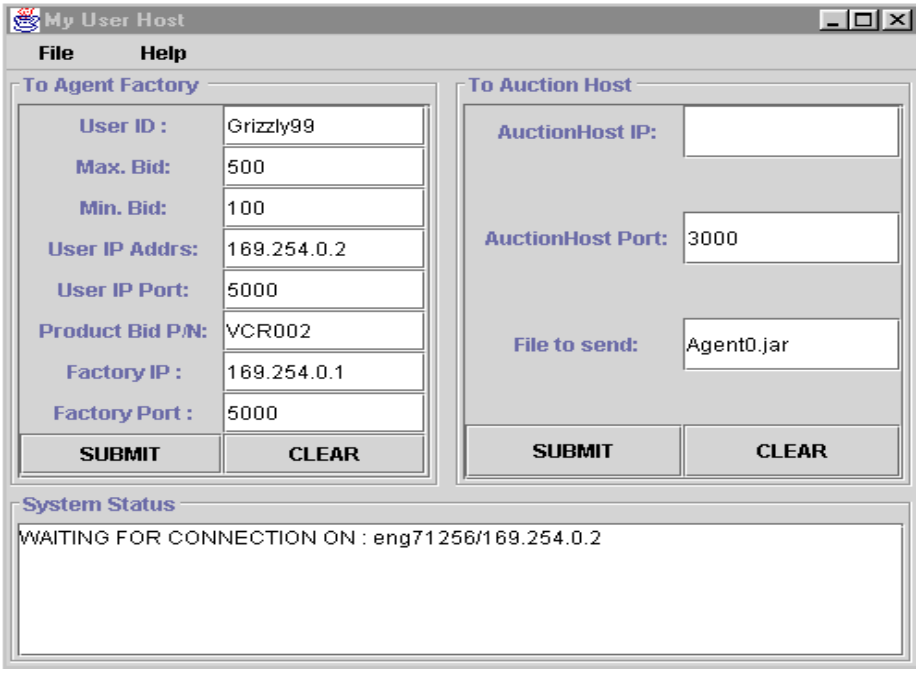


Figure 8. Screenshot of the agent factory

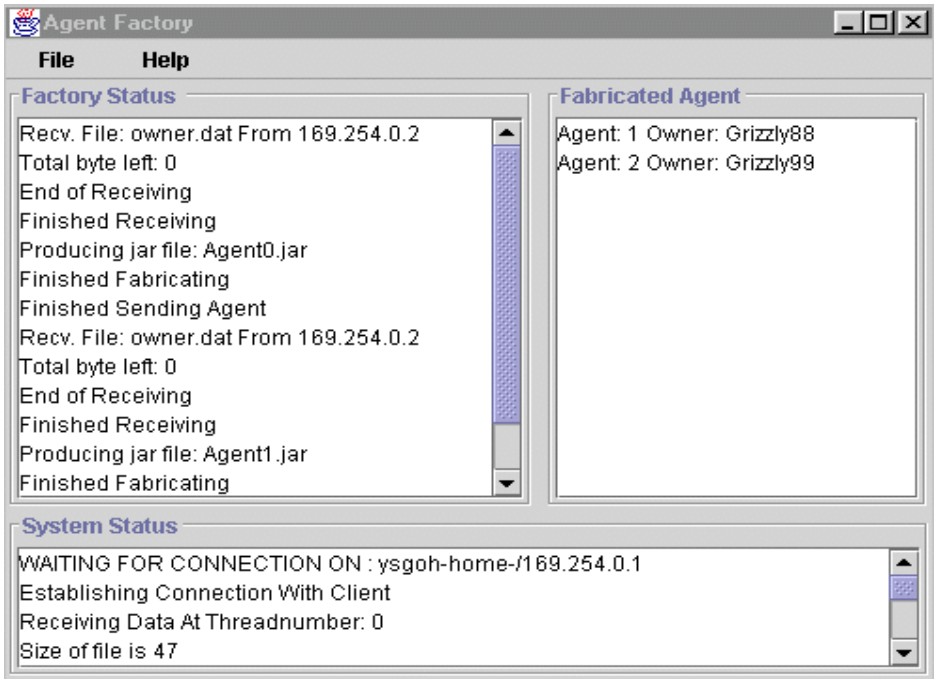
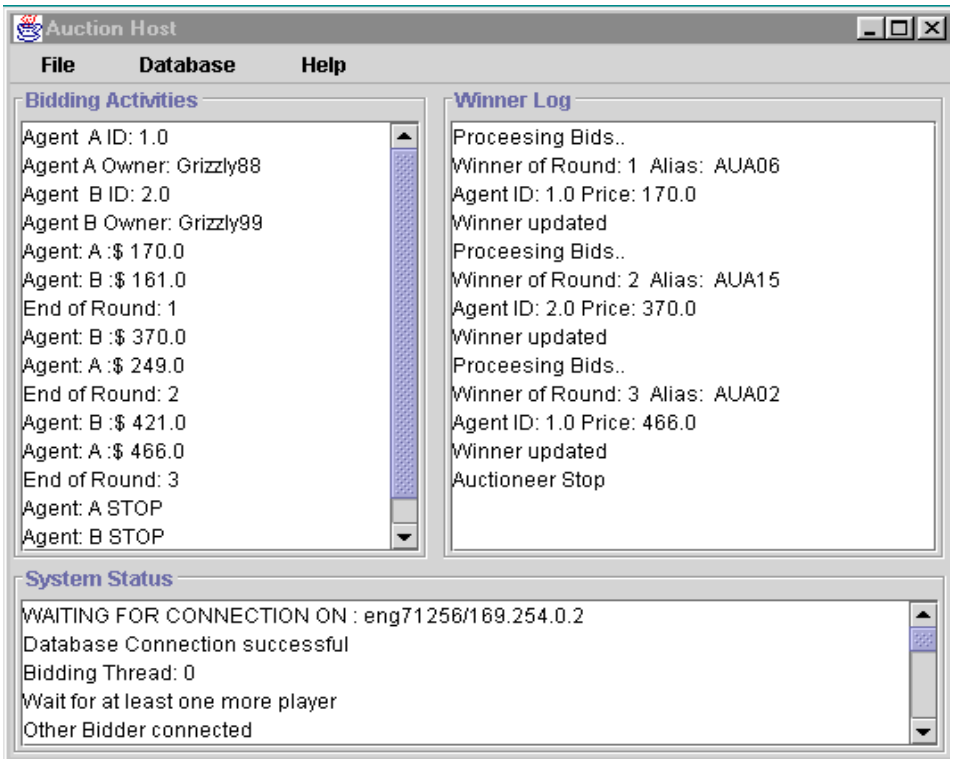


Figure 9. Screenshot of an auction host



the agent factory for the fabrication of a bidding agent. These parameters include user ID, the maximum bidding credit, minimum bidding credit, the user IP address and port, factory IP address and port, and finally the product part number.

- **The Agent Factory:** The agent factory (Figure 8) will verify and validate the user's identity and fabricate the bidding agent according to the user's needs.
- **The Auction Host System:** The auction host prototype (Figure 9) simulates the bidding process. The host waits for agents to arrive before the bidding can start. In this case "Agent0.jar" and "Agent1.jar" are received by the host. These files are extracted into the respective directories and each agent is then invoked by the host.
- **Coordinator Database:** The coordinator database (Figure 10) consists of two tables, the auction table and the winner table. The auction table stores for each agent its particulars such as agent ID, agent user, bidding price of each round, and auction alias. The

Figure 10. Screenshot of the coordinator database

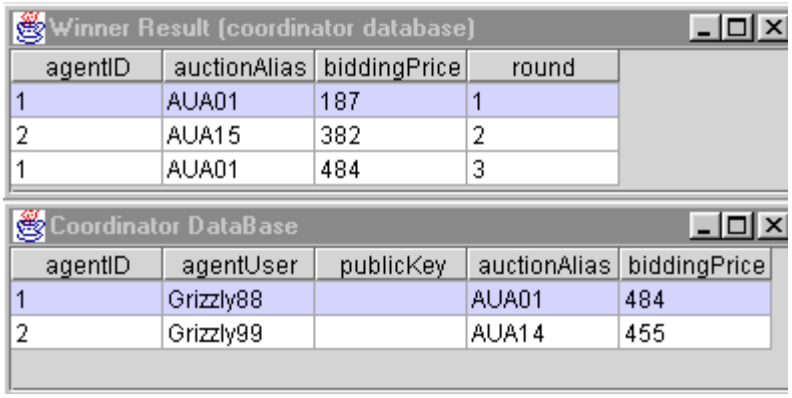
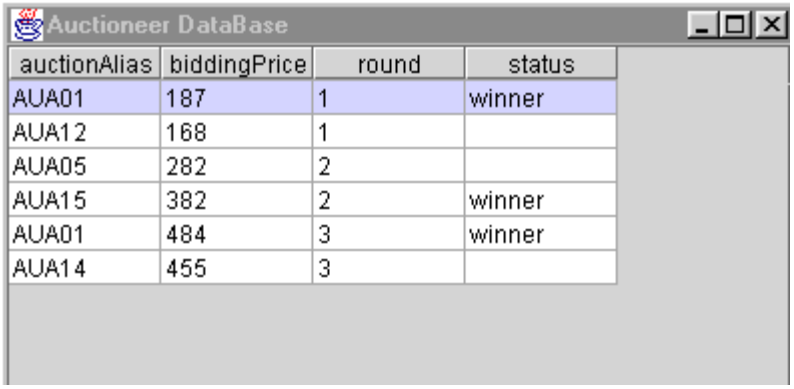


Figure 11. Screenshot of the auction database



winner table includes agent ID, alias, the bidding price, and the round count of each bidding round. Only the successful winner of each bidding round is recorded in this winner table.

- **Auctioneer:** The auctioneer database (Figure 11) consists of action alias, bidding price, round count, and status. The auctioneer records the result of each successful bid round. The winner alias and bidding price are then sent to the coordinator. The coordinator will tally the alias and price within the auction table and store the winner result to the winner table. The agents are informed of the result of each round by the coordinator.

Implementation Issues

There are two possible ways of customizing agents:

- **Local Agents and Local Customization (LALC):** Agents are software programs downloaded and possessed by owners. Agents can be customized by the owner with his or her preference in bidding strategies and other specifications.
- **Remote Agents and Local Customization (RALC):** Agents are stand-alone programs residing in the auction hosts. Users may use the browser to select the proper agent and customize it with his or her bidding strategies and other preferences by way of standard CGI forms. The user-customized agents will then be sent and automatically invoked to start the auction at the remote host. We have adopted the latter scheme in our implementation.

CONCLUSION AND FUTURE WORK

In the above sections, we have discussed in detail the architecture for an agent-based electronic auction system. Compared to existing Web-based auctions, our proposed scheme exhibits some unique features and advantages when addressing the issues of security, privacy, fairness, and flexibility. However, the auction service is not complete until the bidding strategy part is realized. The active research on agent-based e-commerce will help in reshaping the proposed scheme.

One of the attractive features of using software agents in an auction is its autonomy and intelligence. With the advent of the self-running agents representing the owners, the labor of this time-consuming job is greatly relieved. The owner can nevertheless play an active role in the process in that he or she still has good control of his or her roaming out agents. Given such a system, a good bidding strategy becomes the critical factor to win. A good strategy should be adaptive enough so as to respond rapidly and intelligently to the behaviors of the other partners.

The proposed agent-based auction system may also provide an interface for future implementation with communication to devices such as Wireless Application Protocol (WAP) phones. WAP empowers mobile users of wireless devices to access live interactive information services and applications from mobile phones.

REFERENCES

Aaron, R., Decina, M., & Skillen, R. (1999). Electronic commerce: Enablers and implications. *IEEE Communication Magazine, 37*(4), 47-52.

Agorics, Inc. (n.d.). A survey of auction types. Retrieved from *www.webcom.com/~agorics/new.html*

AmEC Initiative at the Massachusetts Institute of Technology.Retrieved from *http://ecommerce.media. mit.edu*

The Fishmarket Project. Retrived from *http://www. iiia.csic.es/Projects/fishmarket/*

Franklin, M. K., & Reiter, M. K. (1996). The design and implementation of a secure auction service. *IEEE Transactions on Software Engineering, 22*(5), 2-14.

Guan, S. U., & Yang, Y. (1999,). SAFE: "Secure-Roaming Agent for E-Commerce." *Proceedings of the 26th International Conference on Computers and Industrial Engineering,* Australia.

Guttman, R. H., Moukas, A. G., & Maes, P. (1998). Agent-mediated electronic commerce: A survey. Retrieved from *http://ecommerce.media.mit.edu/papers/ker98.pdf*

Maes, P., Guttman, R. H., & Moukas, A. G. (1999). Agents that buy and sell. *Communications of the ACM, 42*(3), 81-91.

Michigan Internet AuctionBot. *http://auction.eecs.umich.edu*

Pham, V. A., & Karmouch, A. (1998). Mobile software agents: An overview. *IEEE Communication Magazine, 36*(7), 26-37.

Poh., T. K., & Guan, S. U. (2000). Internet-enabled smart card agent environment and applications. In S.M. Rahman & M.S. Raisinghani (Ed.), *Electronic commerce: Opportunities and challenges* (pp. 246-260). Hershey, PA: Idea Group Publishing.

Subramanian, S. (1998). Design and verification of a secure electronic auction protocol. *Proceedings of the 17th IEEE Symposium on Reliable Distributed Systems.*

Webopedia. Retrieved from *http://www.webopedia.com*

Yang, Y., & Guan, S. U. (2000). Intelligent mobile agents for e-commerce: Security issues and agent transport. In S.M. Rahman & M.S. Raisinghani (Ed.), *Electronic commerce: Opportunity and challenges* (pp. 321-336). Hershey, PA: Idea Group Publishing.

Yi, X., Wang, X. F., Lam, K. Y., Okamoto, E., & Hsu, D. (1998). *A secure auction-like negotiation protocol for agent-based Internet trading. Proceedings of the 17th IEEE Symposium on Reliable Distributed Systems.*

Zhu, F. M., Guan, S. U., Yang, Y., & Ko, C. C. (2000). SAFER e-commerce: Secure agent fabrication, evolution & roaming for e-commerce. In R. Bignall & S.M. Rahman (Eds.), *Internet commerce and software agents: Cases, technologies and opportunities* (pp. 190-206). Hershey, PA: Idea Group Publishing.

KEY TERMS

E-Auctions: The process of selling online with the highest bidder winning the product.

E-Commerce: The conducting of business transactions over networks and through computers.

Bandwidth: Measure of the amount of information that may be transmitted over a channel.

Information Shopping: The gathering of auction information, including the time and place of the auction and the other related info.

Open Auctions: Bids are known to all bidders.

Sealed-Bid Auctions: Each bidder submits a bid without the knowledge of the other bids.

This work was previously published in Encyclopedia of Virtual Communities and Technologies, edited by S. Dasgupta, pp. 179-187, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.27
Mobile Advertising:
A European Perspective

Tawfik Jelassi
Ecole Nationale des Ponts et Chaussées, France

Albrecht Enders
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

ABSTRACT

This chapter is based on research conducted in cooperation with 12Snap, the leading European mobile marketing company, which has implemented large-scale mobile advertising campaigns with companies such as McDonald's, Nestlé, Microsoft, Coca-Cola, Adidas, and Sony. To set the overall stage, we first discuss the advantages and disadvantages of the mobile phone in comparison to other marketing media. Then we propose a framework of different types of advertising campaigns that can be supported through the usage of mobile devices. These campaign types include (1) mobile push campaigns, (2) mobile pull campaigns, and (3) mobile dialogue campaigns. Building on this framework, we analyze different campaigns that 12Snap implemented for different consumer goods and media companies. Drawing from these experiences we then discuss a number of key management issues that need to be considered when implementing mobile marketing campaigns. They include the following themes: (1) the choice of campaign type, (2) the design of a campaign, (3) the targeting of the youth market, and (4) the combination of different media types to create integrated campaigns.

INTRODUCTION

The market for mobile phones has expanded rapidly during the past decade and continues to grow quickly. In some European countries such as Finland, Sweden, Norway, and Italy, the mobile phone has reached almost ubiquitous penetration with levels of 80% and higher (*Economist*, 2001). In Germany, mobile phones are more widely used than fixed-line connections (Brechtel, 2002). In addition to voice communications, German users send out 2.2 billion text messages through their mobile phone every month (Brinkhaus, 2002).

The fast spread of mobile phones has created immense profit expectations in the telecommunications industry. Telecommunication companies in many countries have invested large sums of money into acquiring third-generation licenses and building the necessary infrastructure. Yet, as it turns out, it is more difficult to generate revenues than initially anticipated.

In addition to call charges, there are three main revenue sources in mobile communications: (1) **transactions**, (2) **information**, and (3) **advertising**. Transactions are of high interest, yet as of now only to a limited extent, because of the small size of the screen and the clumsy usage of the keypad. With information services (such as weather forecasts or banking services) the crucial issue is the user's willingness to pay for these types of services.

Does mobile advertising have the potential to be a significant source of revenue in the future? First studies on this new advertising medium indicate that mobile advertisement campaigns can be very successful, generating response rates as high as 40%, compared with the 3% response rate generally expected for direct mail and less than 1% for Internet banner ads (Borzo, 2002).

Because of the novelty of the technology, using mobile phones for advertising campaigns presents some challenging questions for marketing departments:

- What are the strategic advantages of the mobile phone in comparison to other advertising media?
- What campaign types can leverage these characteristics?
- What critical issues need to be considered when launching a mobile advertising campaign?

In the remainder of this chapter, we discuss these questions drawing on field research conducted in cooperation with the German mobile marketing company 12Snap.

ADVERTISING THROUGH MOBILE PHONES

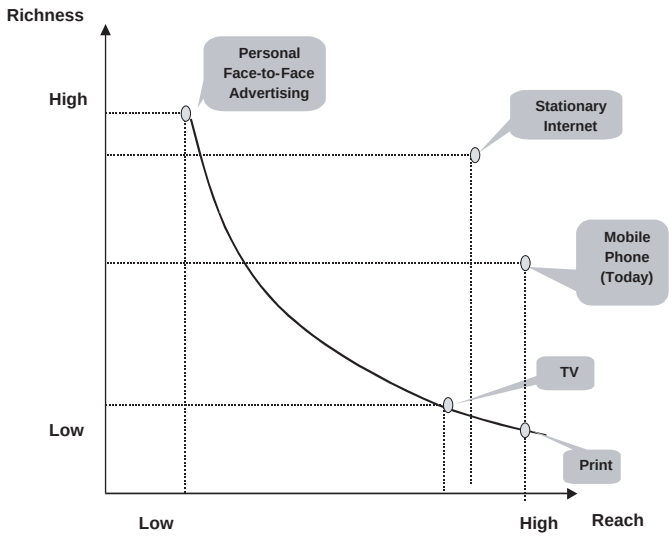
With the increasing number of media types, it has become more and more difficult for marketing managers to find appropriate strategies to target potential customers with their messages. First, while it was possible in the past to capture a large segment of society by placing advertisements with the main TV networks, the rise of private channels has led to a high degree of fragmentation, thereby complicating access to consumers. Similar fragmentation can be observed with other mass-media types such as print or radio. As a result, getting time and attention from their audience has turned into a major challenge for advertisers (Davenport & Beck, 2000).

Second, different media types require different approaches because of differences regarding their reach or richness. **Reach** is a function of how easily customers, or in this case, participants in advertising campaigns, can be contacted through a given medium. **Richness**, on the other hand, is defined by (1) bandwidth, that is, the amount of information that can be moved from sender to receiver in a given time, (2) the degree of individual customization of the information, and (3) interactivity, that is, the possibility to communicate bidirectionally (Evans & Wurster, 1997). The communication of rich marketing information, that is, information that ranks high on all three aspects, has traditionally required physical proximity to customers and/or channels specifically dedicated to transmitting the information (see Figure 1).

How does the mobile phone fare within the richness versus reach framework? It can serve as a powerful platform to get in touch with end consumers because it simultaneously provides expanded reach and a number of richness advantages vis-à-vis most other media types:

- **Ubiquitous Access:** Mobile phone users always have their phone with them and

Figure 1. The trade-off between richness and reach in advertising (adapted from Evans & Wurster, 1997)



turned on at almost all times (Balasubramanian, Peterson, & Jarvenpaa, 2002; Magura, 2003). This is especially true for teenagers and young users who use the mobile phone to stay in touch with their peers—primarily through SMS (Bughin & Lind, 2001). Ubiquitous access becomes especially important in places like buses, trains and subways, airport lounges, and so forth. The time that people spend traveling is prime time for marketing since it presents a time when people are not occupied with other activities and are thus receptive to other kinds of entertainment. A study by the Boston Consulting Group (2000) found that among private users, the categories "having fun" (71%) and "killing time" (55%) belong to the main motivators for using mobile phones—ranking only behind "keeping in touch with friends" (85%).

- **Detailed user information:** While traditional marketing campaigns only have access to very limited customer information, mobile campaigns can draw on extensive and individual information about each user (such as age, sex, usage profile, etc.). This information helps to launch highly targeted campaigns for specific products and services based on individual preferences of the user.
- **Integrated response channel:** The mobile phone presents the opportunity to interact directly with the user and elicit responses through the same medium. This has two advantages. First, it provides the opportunity for rich interaction. The interactivity and ubiquity of the mobile phone opens up the possibility to turn existing traditional media formats (such as the TV, radio, print, or packaging) interactive. For instance, companies can contact consumers via TV and then

subsequently, stay in touch with each one of them through the mobile phone. Second, the integrated response channel also allows mobile marketing companies to measure precisely the impact of their campaigns and then to adapt their strategies accordingly—something that is much more difficult to do with traditional marketing media. For instance, a customer buys a product—with a mobile phone number on the packaging—at a retailer, and as s/he exits the store, s/he completes a quick survey of the shopping experience, which is then transmitted immediately to corporate headquarters. This not only allows the consumer the satisfaction of immediate feedback if they had a positive or negative experience, but it also allows the company to measure quality control in an extremely timely and cost-effective manner (Carat Interactive, 2002).

- **Personal channel:** Unlike other advertising media such as TV, radio, or billboards, the mobile phone belongs to only one person. Therefore, it receives much more attention and, if handled properly (see risks below), can be much more powerful than other, less personal media channels. John Farmer, a cofounder of the SMS application and service provider Carbon Partners, points out that the personal character of the mobile phone is especially important to teenagers (Haig, 2001): "The mobile phone presents the teenage market with the distinct opportunity to take control of their own communications, free from the previous limitations of the home phone or computer, which were more closely monitored by parents."

Brian Levin, CEO of Mobliss, a U.S. wireless marketing firm, sums up the advantages (Stone, 2001): "When you have a little time to spare, such as in the airport or at the bus stop—then you want to be engaged or entertained. Once you are there, the proximity of this device [the mobile phone] to your face, the intimacy there, is very powerful both in terms of direct response and in terms of branding." At the same time, however, the mobile phone also presents shortcomings and risk factors:

- **Limited media format:** Mobile phones today still have to cope with a very limited set of visual and audio capabilities. In second-generation (2G) phones, screens are typically small, have only low resolution, and are typically not in color. Sound effects are also limited due to the small speakers, and text messages cannot be longer than 160 characters. The challenge is then to ensure at this stage that consumers do not expect an identical experience to what they receive through other devices such as the TV or PC (Carat Interactive, 2002).
- **Private sphere:** The fact that mobile phones belong to only one person does not only present an opportunity but also a challenge for mobile advertisers. Unlike the TV or Internet, the mobile phone is a very personal device to which only family, friends, coworkers, and a selected few others will gain access. Thus, "spamming" is considered much more intrusive than in other media formats (Carat Interactive, 2002).

DEVELOPING EFFECTIVE MOBILE ADVERTISING CAMPAIGNS

One of the main challenges and opportunities for mobile advertising companies is the personal nature of mobile phones. Advertising campaigns over mobile phones are very sensitive and companies that engage in this type of marketing need to be careful not to offend users. Will Harris, global marketing director for Genie, British Telecom's mobile Internet service, emphasizes (Pesola,

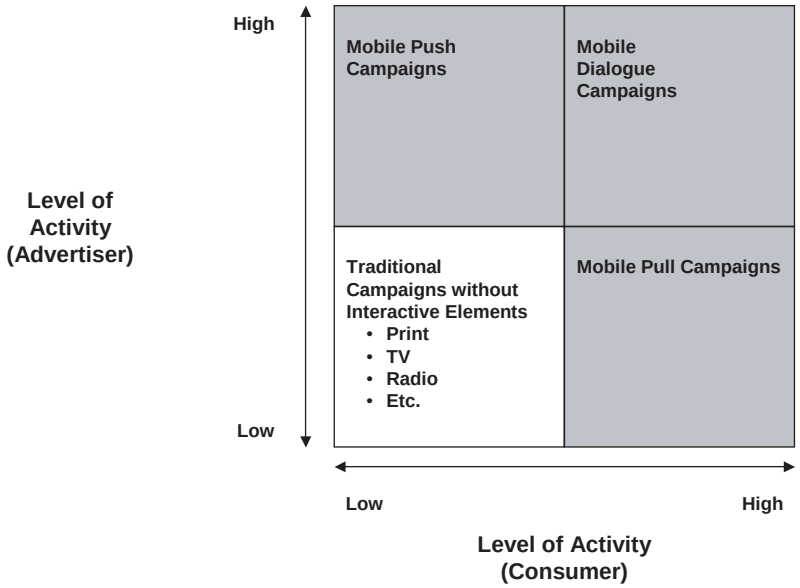
2001): "Sending unsolicited messages is tantamount to brand suicide. Our business is entirely dependent on the goodwill of our customers."

The mobile advertising industry is trying to protect mobile phone users by establishing guidelines for responsible advertising. The main feature of these guidelines is consent, that is, consumers agree or opt-in to receive the advertisements. In addition, they must have a clear understanding of what their personal information is being used for, and if they wish, be removed from the advertiser's databases.

As a result, mobile advertisers have to find ways to entice customers to opt into their campaigns. Cyriac Roeding, 12Snap's marketing director, explains why many companies have difficulties attracting mobile phone users (Pesola, 2001): "A lot of companies make the mistake of coming to this from a technological angle, rather than thinking about what the consumer wants. If advertising is entertaining, if it engages the emotions, it will be accepted."

Although mobile advertising is a relatively recent phenomenon, a number of large corporations, including McDonald's, 20th Century Fox, and Sony, are using this medium in their marketing mix, especially to target young customers. These campaigns differ according to the degree of active involvement of advertiser and recipient (see Figure 2). Level of activity refers here to the involvement both advertiser and consumer show throughout the course of an advertising campaign. Traditional campaigns, which still present the most prominent advertisement type, display low levels of activity on both the advertiser's and the consumer's side since they consist of noninteractive, one-way advertisements in the form of TV spots, radio or print ads, or posters. Mobile campaigns, on the other hand, show high levels of activity either on the side of the advertiser, the consumer, or both.

Figure 2. Mobile advertising framework



High level of activity on the side of the advertising company implies that the consumer is approached proactively, whereas a high level of activity on the side of the consumer implies that s/he reacts actively to an advertisement or a newspaper ad, for instance, by soliciting further information via the mobile phone.

Through the built-in response channel, mobile phones are suitable both for push and pull campaigns. According to the mobile advertising framework, mobile campaigns can be categorized as follows:

- **Mobile Push Campaigns:** Push advertising is categorized as messages that are proactively sent out to wireless users. Companies use databases with existing customer profiles, which can be proprietary or purchased externally, to address their target groups (Carat Interactive, 2002). Because of the sensitivity of the mobile phone, it is important to ensure that all members of the database have agreed beforehand (i.e., given their opt-in) to receive mobile advertising. In addition, for the success of a campaign, it is essential to ensure that the selected target group from the database is interested in the specific advertising, which requires extensive profiling of the database (Pearse, 2002). Doing so avoids the alienation of uninterested users, while at the same time maximizing the impact of the advertising budget on those targeted.
- **Mobile Pull Campaigns:** Applying a pull approach, advertisers use their traditional marketing media mix, such as TV, radio, print, or packaging, to promote an interactive mobile campaign. For instance, a telephone number on a French fries box might invite a customer to participate in a raffle for free

Figure 3. Taxonomy of mobile advertising campaigns

	Push Campaigns	Pull Campaigns	Dialogue Campaigns
Set-up	Targeted SMS to user in existing database • Company-owned database • Database from external provider	Advertisements for mobile campaigns through other media types, e.g. • Flyer or "on-pack" ad • TV ad	Continuous interaction between advertiser and user
Opt-In	Need to have explicit "opt-in" prior to sending out SMS	Users "opt-in" by calling the phone number	Generation of "Opt-in" either through pull or push
Scope	Single theme • Game • Raffle • Etc.	Single theme • Game • Raffle • Etc.	Multiple themes during the course of a campaign • Different games • Greetings
Length	Short duration (2-4 weeks)	Short duration (2-4 weeks)	Extended duration (several months)
Implementation	• Wella Hair Fashion • Sony • 20th Century Fox • Adidas	• Nestlé KitKat Chunky • Toyota	• McDonald's

food through his/her mobile phone. By calling, the consumer gives the "opt-in"—an explicit consent to the campaign—and can then participate.

- **Mobile Dialogue Campaigns:** Dialogue campaigns differ from the above-mentioned campaign types in their duration and the intensity of interaction between advertiser and customer. While simple push and pull campaigns tend to last only 2 to 4 weeks and center around one single theme such as a raffle or a game, dialogue campaigns last several months and include various different themes that build on one another. Their goal is to establish a long-lasting relationship with consumers so as to generate extensive insights into consumers' preferences. A mobile horoscope service, for instance, allows the advertiser to capture the birthday of the consumer, which can then be used for sending out personal birthday greetings later on. The in-depth consumer information serves then to distribute mobile coupons—for instance, a free candy bar as a birthday present—to introduce new products or to do market research in a very targeted fashion.

Building on this framework, it is now possible to categorize the actual implementations of mobile advertising campaigns (see Figure 3).

Push Campaigns

Wella, a German manufacturer of hair-care products, developed a push campaign, which featured a "mobile kiss." Wella sent an SMS to members of an externally acquired database offering them to send a kiss message to their friends, who received a voice file with a kiss sound. This was followed by an SMS revealing who had sent the kiss and also providing details on how to return the kiss or send it to someone else. The maximum number of kisses sent by one person was 160. Other components of this mobile campaign included an SMS quiz and a free kissing-lips logo for the mobile phone.

Sony launched a push campaign in the UK which integrated e-mail, Internet, and traditional print media to promote a new PC-compatible MiniDisc Hi-Fi system. The campaign, which was based on MiniDisc Island—an online, interactive adventure playground—had the objective of driving large numbers of users to the Web site. Through an initial SMS 100,000 participants from a database of 14 million permission-based, profiled users were selected and invited to enter a competition to win a stereo system and the PC Link product. Interested participants replied via SMS and were mailed a winning number, with which they could then go to the Web site to see if they had won. Throughout the course of this campaign, 18% of those originally contacted responded to the initial SMS. When contacted with the winning number, over 9% logged on to explore the online adventure game and see if their numbers had come up.

The movie studio **20th Century Fox** launched a push campaign in partnership with the mobile phone operator Vodafone to advertise the UK release of *Planet of the Apes*—a post- apocalyptic movie where apes rule over humans who struggle to survive. The campaign, which started 2 weeks prior to the UK release of the movie, targeted the 2 million 16- to 24-year-old Vodafone customers. They received messages, which invited them to survive a variety of challenging interactive voice response and text games—with names such Ape S-cape and Ape@tak, where callers are asked questions relating to the movie and have to shoot down apes using the keypad when they hear a roar of an ape.

The sports article manufacturer **Adidas** used interactive betting game for the Soccer World Cup 2002 to promote its products in Germany. Users bet on games and received immediate notification after the game about how they did, how they ranked overall within the betting competition, and

if they had won a prize such as a shirt from the soccer idol Zinedine Zidane or a personal meeting with the German national soccer team.

Pull Campaigns

Nestlé used a mobile pull campaign to promote the KitKatChunky chocolate bar in Germany. The campaign, which lasted 2 weeks, complemented the overall marketing presence consisting of TV and radio spots and the Web site www.chunky.de. The campaign worked as follows: an SMS offered community members the opportunity to win a 1-year supply of KitKatChunky if they called a specified number. Then, callers were shown the face of a taxi driver who also appears in a KitKatChunky TV commercial, and two other new characters on their screen who were presenting riddles to them. An automated voice then explained: "Each of the three protagonists names a number which makes him shut up. Once you have discovered the number, push the appropriate button on your mobile and a KitKatChunky is stuck in his mouth and you go on to the next round." In the first round, 400,000 users were identified to receive a kick-off SMS at the beginning of the campaign. In the following rounds, only those players who had actively opted-in in the previous round received an SMS. In order to maximize the number of responses, users received alert messages the day before the ad's TV premiere and again 30 minutes before the TV show in which the questions were sent. This illustrates to what extent different media types—here the TV and the mobile phone—can be interlinked, using the respective strengths of each medium, to generate a seamless and entertaining marketing experience for the end user.

The car manufacturer **Toyota** also launched an interactive TV pull campaign during the Soccer World Cup 2002 that displayed a quiz question at the bottom of the TV screen. The question asked viewers to find the license plate number of the Toyota shown in the TV ad and to send this number in via SMS. Within seconds they received notification whether they had made it to the final drawing. In addition to having the chance to win a prize, all callers also received a Toyota ring tone for their mobile phone.

Dialogue Campaigns

Extensive dialogue campaigns are still a rarity because of the novelty of the mobile phone as an advertising medium. In Germany, **McDonald's** launched a mobile dialogue campaign with a focus on interactive mobile games and an evaluation of McDonald's products. The campaign targeted mobile phone users who were informed through in-store flyers placed in McDonald's restaurants. By activating the service (active "opting-in"), participants received automatic messages when music CDs or vouchers were raffled off. In addition, they also received SMS promotions of McDonald's products. The goal of the campaign was (1) to increase in-store traffic, (2) to build a McDonald's customer database of mobile phone numbers, and (3) to increase overall brand awareness. Following this initial pull activity, registered users continued to receive other services such as horoscopes, which in turn allowed McDonald's to capture users' birthdays and to send them personalized birthday greeting subsequently. An additional part of the campaign was a viral activity: McDonald's sent the Christmas greeting "Rockin' Rudi" to users who could then forward it to their friends. The recipient then listened to a taped version of the "Rockin' Rudi" song in combination with a short message from the sender and from McDonald's.

OUTLOOK AND MANAGEMENT ISSUES

There are plenty of opportunities in mobile advertising for companies that thoroughly understand how consumers can benefit from these types of

services. The new technology will not be very useful, however, if companies simply use their existing advertising approaches and translate them to the mobile world without addressing the specific characteristics of this new medium (Nohria & Leestma, 2001).

The different innovative types of mobile advertising campaigns mentioned above offer many useful benefits—for instance, highly targeted advertising and interactivity—to those companies that want to add a mobile component to their advertising approach. They also illustrate the difficulties and challenges that are associated with this new approach. Therefore, before embarking on mobile advertising campaigns, managers need to carefully address the following questions.

Which Campaign Type Should We Employ?

For starting a mobile advertising campaign, there are two basic options: push or pull. A push campaign requires an extensive database of customers. Some companies such as telcos or retailers have built up these types of databases in the past through CRM efforts and can now tap into them. However, they always need to keep in mind the personal nature of the mobile phone when doing so. "Spamming" existing customers with unwanted SMS is a sure way to alienate them. Another option is to buy existing profiles from other companies. MTV, for instance, markets its permission-based database through an external mobile advertising company to other companies that want to target the attractive youth market. These companies benefit since they can tap into an extensively profiled, permission-based database of their target group while MTV generates additional revenues.

Setting up a pull campaign is not as sensitive regarding the opt-in, since consumers themselves decide whether they want to participate when they see the advertisement printed on a poster or watch it on TV. Here, the challenge is much more to create compelling advertisements that have the desired pull effect to entice consumers to call in and participate.

How Should We Design Attractive Mobile Advertising Campaigns?

The challenge for any mobile marketing company is to create enough interest within the target group to justify the required investment. Based on the campaigns we have analyzed, four key success factors need to be considered when launching a mobile advertising campaign (Brand & Bonjer, 2001):

- **Interactivity:** just like the Internet, the mobile phone allows advertisers to solicit immediate feedback when contacting recipients. Since the mobile phone is usually always turned on, the inherent interactivity of mobile phones should be integrated in mobile marketing campaigns where possible. The interaction can have many different facets: the number pad can be used to answer riddles or mental agility can be tested through reaction tests. A mobile marketing campaign that does not integrate interactivity would be the equivalent to the broadcasting of a slide show on TV. It would leave a main asset of the medium untapped.
- **Entertainment:** interaction is only fun for users if they find the advertisement exciting. Therefore, mobile campaigns need to combine advertising and entertainment in such a way that users are willing to lend their time to an advertisement. In this respect, the creation of mobile campaigns is similar to more traditional campaigns on TV, for instance. TV viewers watch advertisements mainly because they are entertaining. Ideally, they do not just watch them but they also talk about them to friends thereby creating a viral effect in which the message is passed on by people other than the original

sender—as was the case in the Wella and McDonald's campaigns. Therefore, the inclusion of entertaining elements such as a game or a story ought to present an integral part of a mobile marketing campaign.

- **Emotion:** the inclusion of emotional elements—such as visual sequences or music clips in TV ads that aim beneath the conscious understanding of the viewer—has long presented a valuable marketing tool to subliminally reinforce the intended message with consumers. In mobile marketing, however, text, especially if shown on a small mobile phone display, can hardly carry this emotional dimension. Here, just like with TV advertising, it is necessary to leverage the admittedly limited resources of the mobile phone to create "emotion." This can be achieved through the combination of voice and sound. For instance, music jingles such as a short sequence of the soundtrack of the movie *Titanic* can be used as the opening for a partner test or an activity aimed at single people. Again, it is not primarily the technology that drives the quality of any given campaign but instead the creative combination of different effects that ultimately determines its success.
- **Incentive:** the offering of incentives such as product samples increases the willingness of consumers to participate in interactive mobile games. The prospect of winning a prize is especially important due to the above-mentioned opt-in nature of mobile marketing campaigns, as it provides the potential participants with a direct and tangible incentive to participate in a mobile marketing campaign. However, although instant-win competitions are effective in driving volume, they are less suitable to generate a long-term relationship with consumers, since they do not offer incentive to return (Cowlett, 2002).

The overall goal of combining these four factors is to create a game, an image or a jingle that, despite the limitations of the small screen and tiny ring tone of the mobile phone, is so compelling that it is no longer seen as an ad, but takes on a value of its own.

How Should We Target the Difficult-to-Reach Youth Market?

Addressing the lucrative youth market gives marketers a perennial headache, since they do not only vary in their habits, interests, and attitudes and are swayed by rapidly changing fashion trends. They are also hard to pin down, since they do not primarily watch three or four TV stations anymore as was in the past. Instead, their media usage is fragmented between hundreds of TV stations, radio, magazines, newspapers, and the Internet. One thing is generally guaranteed, though—they almost certainly carry a mobile phone and consider SMS an intrinsic part of their lifestyle since it allows them to stay in touch with their peers in a cost-effective and entertaining way (Cowlett, 2002).

Mobile campaigns can effectively leverage the characteristics of this new youth market. Viral effects, used in the McDonald's and Wella campaigns, fulfill the desire to communicate with peers in a fun way. Teens enjoy quizzes or greeting cards they can pass on to friends, because this type of promotion focuses on using the mobile phone for what it was made to do—communicate with other people (Centaur Communications, 2002). In addition, viral elements help to expand significantly the group of recipients beyond the database of the company conducting the campaign and it increases the impact since marketing messages sent from a friend are, because of their personal nature, much more effective than those sent directly from the company itself (Haig, 2001; Kenny & Marshall, 2000). The communication from consumer to consumer helps to generate "buzz"—explosive

self-generated demand—where people share their experiences with a product or a service amongst one another (Dye, 2000). At the same time, this approach helps to lower costs since users themselves target new consumers.

How Should We Combine the Mobile Phone with Other Media Types to Create Integrated Campaigns?

Because of its limitations regarding screen size, sound, and handling, the mobile phone is not suitable for stand-alone campaigns. Instead, it should be used to extend the presence of a company into an additional channel (Carat Interactive, 2002). Doing so, the mobile phone plays the role of the natural glue between other media types because of its ubiquitous nature: it is handy and turned on when watching TV, looking at a billboard on the subway, buying groceries at the supermarket, or listening to the radio. All the campaigns mentioned above make extensive usage of this cross-linking of different media types, leveraging the unique strengths of each. It is not only other media types that benefit from the integration of the mobile in multichannel advertising campaigns: tangible support mechanism from other media types that have been around for years—such as a flyer or an in-store promotion—give mobile campaigns higher legitimacy because they have a physical component (Enders & Jelassi, 2000).

From a market research perspective, the inclusion of mobile components in advertising campaigns has the added benefit that it allows to measure directly the effect of different advertising approaches. Take, for instance, a TV advertisement that is aired on different channels and broadcasting times, or a billboard advertisement placed in different locations that asks viewers to participate in an SMS contest. Based on the measurement of actual response rates in different channels or locations, it becomes possible to steer placement more effectively than via traditional indirect measurements.

REFERENCES

Balasubramanian, S., Peterson, R., & Jarvenpaa, S. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Sciences, 30*(4), 348–361.

Borzo, J. (2002). Advertisers begin dialing for dollars. *Asian Wall Street Journal*, February 18.

Boston Consulting Group. (2000). Mobile commerce—winning the on-air consumer. November.

Brand, A., & Bonjer, M. (2001, November). *12Snap: Mobiles Marketing im Kommunikations-Mix innovativer Kampagnenplanung* (White paper). Munich: 12Snap AG.

Brechtel, D. (2002). Bei Anruf Werbung. *Horizont, September 12*, 80–81.

Brinkhaus, G.B. (2002). Keine Massenmailings: Wie Mobile Marketing Funktioniert. *FAZ-online, September 29.* Retrieved from *www.faz.net*

Bughin, J., Lind, F. et. al. (2001). Mobile portals mobilize for scale. *McKinsey Quarterly, March*, 118–125.

Carat Interactive. (2002). *The future of wireless marketing* (White paper).

Centaur Communications. (2002). Good text guide. *In-Store Marketing, October 7*, 23–27.

Cowlett, M. (2002). Mobile marketing—"text messaging to build youth loyalty." *Marketing, October 31*, 29–34.

Davenport, T., & Beck, J. (2000). Getting the attention you need. *Harvard Business Review, September*, 118–125.

Dye, R. (2000). The buzz on buzz. *Harvard Business Review, November*, 139–144.

The Economist. (2001). The Internet untethered. *The Economist, October 13*, pp. 3–26.

Enders, A., & Jelassi, T. (2000). The converging business models of Internet and bricks-and-mortar retailers. *European Management Journal, 18*(5), 542–550.

Evans, P., & Wurster, W. (1997). Strategy and the new economics of information. *Harvard Business Review, September–October*, 71–82.

Haig, M. (2001). KIDS—talking to the teen generation. *Brand Strategy, December.*

Jelassi, T., & Enders, A. (2005). *Strategies for e-business: Creating value through electronic and mobile commerce*. Essex, UK: Financial Times/Prentice Hall.

Kenny, D., & Marshall, J. (2000). Contextual marketing: The real business of the Internet. *Harvard Business Review, November*, 119–124.

Magura, B. (2003). What hooks m-commerce customers? *MIT Sloan Management Review, Spring*, 9.

Nohria, N., & Leestma, M. (2001). A moving target: The mobile-commerce customer. *Sloan Management Review, 42*, 104.

Pearse, J. (2002). NMA wireless—mobile conversations. *New Media Age, October 31*, pp. 37–43.

Pesola, M. (2001). The novelty could quickly wear off. *Financial Times.com*, July 17.

Stone, A. (2001, January 3). Mobile marketing strategies Q & A. Retrieved from *www.mcommercetimes.com/Marketing/200*

ENDNOTE

- This chapter is based on a teaching case study that features the German wireless advertising company 12Snap (see Jelassi & Enders, 2005).

This work was previously published in Unwired Business: Cases in Mobile Business, edited by S. Barnes, and E. Scornavacca, pp. 82-95, copyright 2006 by IRM Press (an imprint of IGI Global).

Chapter 4.28
China:
M–Commerce in World's Largest Mobile Market

Nir Kshetri
University of North Carolina at Greensboro, USA

Nicholas Williamson
University of North Carolina at Greensboro, USA

David L. Bourgoin
University of Hawaii at Manoa, USA

ABSTRACT

China is emerging as a global capital of m-commerce applications. China is the world's biggest mobile market in terms of subscriber base and the fastest growing in the history of telecommunications. Although China currently lacks advanced mobile applications compared to Europe, North America, Japan and Korea, a number of mobile players are rapidly launching sophisticated mobile applications. Unique institutions and the nature of mobile market conditions in China, however, superimpose in a complex interaction that harbors a paradoxical nature. The Chinese m-commerce market is thus drastically different from that of the Western world. This chapter examines the Chinese m-commerce landscape and analyzes its drivers. We also examine the Chinese market from the CLIP perspective.

INTRODUCTION

Consider the following observations on the development of mobile technology and the potential of m-commerce in China:

For the tech industry, it's China — not Europe, or Japan, or other Asian countries — that will soon be its [USA's] main rival. The implications are profound. No longer content to cheaply make other people's products, a task it has clearly mastered, China wants to be a global standards setter. … One place to watch the flexing of power is in

mobile phones[1]*. The mobile Internet has really saved China's Internet industry*[2].

As the above statement indicates, China has emerged as a capital of the global mobile market. The growth rate achieved by the Chinese mobile network, the biggest in the world, is the fastest in the history of telecommunications. Rapid development in the Chinese mobile market has driven increasingly China-centric activities of major players in the global mobile market (Kshetri, 2004a, 2004b).

Unique institutions and the nature of mobile market conditions in China, however, superimpose in a complex interaction that harbors a paradoxical nature (Kshetri, 2005). This chapter provides a brief survey of the paradoxical Chinese m-commerce market and analyzes its driving force. The chapter structure is as follows: The next section provides a brief overview of the Chinese mobile market. The paper then analyzes some major forces behind China's rapid growth in this market. Next, we examine the Chinese market from the CLIP perspective. Finally, the paper presents its conclusions.

A BRIEF SURVEY OF THE CHINESE MOBILE MARKET

The Chinese mobile market became the largest in the Asia Pacific in 2000 and the world's largest in 2002 (Stout, 2001). By the end of 2004, there were over 300 million mobile subscribers in China and an additional 75 million "Little Smart"[3] users (BBC News, 2004). During 2004, mobile phone users in China grew by 27 per cent (Reuters, 2004). An estimate by the telecom analyst firm, EMC, suggests that China will have 36 million 3GSM[4] (W-CDMA) subscribers by 2009 (globalsources.com, 2004).

In 2002, 120 million handsets, or 27 percent of the world total, were produced in China. The proportion increased to 33 percent in 2003, 35.1 percent (233.5 million) in 2004 (*SinoCast, China Business Daily News,* 2005a) and is estimated to reach 50 percent by 2008 (Symbianphone.com, 2003). China's handset exports increased from 22.75 million in 2000 (*GIS News*, 2001) to 55 million in 2002. During the first half of 2003, China exported 37 million handsets.

China's Time Division — Synchronous Code Division Multiple Access (TD-SCDMA) — developed by Datang is currently accepted as a global third generation (3G) standard. Among 16 proposals submitted for IMT-2000[5] standards, TD-SCDMA was one of the three 3G mobile standards selected by the International Telecommunications Union (ITU) in May 2000. The other two standards are the U.S. based CDMA2000 system and Europe's WCDMA. The Third Generation Partnership Project (3GPP) accepted the TD-SCDMA in March 2001.

TD-SCDMA is scheduled for launch in the second half of 2005. Estimates are that following its launch, the TD-SCDMA will capture 30 percent of the Chinese market and 10 percent outside China (Einhorn, 2003). It is also interesting to note that some innovative m-commerce applications were first developed and employed in China. To take one example, the world's first electronic stock trading over a wireless network, took place on byair.com in Shanghai, China, in 1998[6] (see Box 1).

By 2008, China's wireless mobile market is estimated to exceed US$200 billion[7]. An estimate by Lehman Brothers Inc. suggests that revenues of mobile portals generated by sending news updates, games and online dating amounted to $200 million in 2001, and was estimated at $3 billion in 2004. China Mobile, the world's largest mobile operator, was the most profitable telecom operator in the Asia-Pacific region in 2002, with a profit of $3.5 billion on revenues of $12.2 billion. A 28 percent return on revenue is an excellent indication of where the Chinese market is developing.

China introduced Wireless Fidelity (Wi-Fi) technology in 2002 and is diffusing rapidly (Clark

Box 1. GWcom's mobile portal in China[13]

GWCom is a mobile wireless ASP in China. It launched its wireless portal byair.com[1] *in 1998 to provide timely information and e-commerce capabilities such as stock trading and banking to users with mobile phone or wireless palmtop devices in the U.S. and Greater China. The company provides its networks and handheld device (netset) to individual investors. By 2002, GWCom had partnered with over 30 Internet content providers and e-commerce portals in the U.S. and Greater China and connected with more than 20 securities trading firms.*

By March 2000, byair.com had over 6,000 subscribers with the number of stocks traded as high as 3,500 daily and number of page views 250,000. By the early 2002, it delivered services to over 250,000 mobile users and more users on Information on Demand (IOD) and messaging services. GWCom users mostly use the two-way paging capability for trading stock electronically and such transaction-type services have turned out to be the 'killer application' (TDAP, 2002).

The company's pricing structure made stock investment on its paging network more attractive than on the fixed network. Because of low PC penetration and relatively higher Internet access fees, the only way to trade stock for a large proportion of Chinese is to read newspapers or magazines and then pick up a phone[2]*. These factors have made GWcom's web portal more attractive (Ebusinessforum.com 2000). GWCom describes its network product, PLANET, as a "high-capacity and low cost cellular packet data network that is optimized for serving wireless palm computers and PDAs"*[3]*. The users pay a monthly service charge of only about US$5-10. With the increasing demand, GWCom has decided to specialize in the mobile wireless data network infrastructure and outsource the equipment manufacturing to Ericsson and some Chinese vendors. This is likely to result further reduction in the price.*

China's stock market is growing very fast[4] *and the stock exchange companies are located in Shanghai and Shenzhen. GW Trade selected these two cities for the initial trial. Wireless users have been using GWCom's application platforms to conduct online trading since 1998 in Shanghai and since 1999 in Shenzhen. In March 2000, 3,000 investors in Shanghai, and 100 in Shenzhen, were trading stocks over the paging networks managed by GWCom. The average daily volume of 3,000 Shanghai users in early-2000 was $3.6 million, about 30 times as much as the average trading volume on stockstar.com, the largest and most popular Web-based stock trading company.*

In developing countries like China, non-voice technologies (such as paging) have potential to offer a cheap and reliable way to transmit data that will be a viable alternative to the mobile phone. In other parts of the world, big players are not following such paging route (Holland, 2000). The GWCom case also provides some evidence of leapfrogging potential of mobile technologies. For instance, the world's first electronic stock trading over the wireless network took place on the GWCom network in 1998 in Shanghai.

& Harwit, 2004). By mid 2003, 80 percent of China's five-star hotels, airports and high-grade office buildings in its four largest cities were connected to China Netcom's Wi-Fi network[8]. At the end of 2003, China Telecom, China Netcom and China Mobile had about 10,000 hotspots deployed or planned for rollout (Clark & Harwit, 2004).

One estimate suggested that the Chinese Wi-Fi market was at $24 million in 2003 (compared to the worldwide market of $600 million) (Koprowski, 2004). Another estimate suggests that China's Wi-Fi market will reach $250 million by 2005[9]. This growth from 4 percent to 30 percent of the market is astonishing. Venture capital companies such as Intel Capital are capitalizing on the huge Wi-Fi potential in China by funding the deployment of Wi-Fi technology (Clark & Harwit, 2004).

The Chinese mobile market, however, is characterized by a high degree of bias towards urban areas. For instance, in 1999, 78 percent of the population owned mobile phones in the three wealthy cities — Beijing, Shanghai and Guangzhou (Tsuchiyama, 1999) — compared with the national average of 3.42 percent that year (UNDP, 2001). This disparity can indicate a huge potential market for future investment and development, or, conversely, show the sophistication of the Chinese wealth sector, with its high technology development and implementation that is close to the cutting edge.

CHINA'S RAPID MOBILE DIFFUSION[10]

Starting the mid-1980s, China invested heavily in the telecom sector. The heavy investment was supplemented by a series of programs designed to accelerate telecom development, including extensive re-engineering of and intense competition in the mobile sector. China Unicom, formed in 1994, competes with the former monopoly China Telecom and is licensed for mobile, paging, data, Internet and long-distance (James, 2001).

Fierce competition in the Chinese mobile sector led to low connection fees as well as lower subscription fees for mobile services. In the late 1990s, for instance, monthly subscription rates as well as connection charges for mobile services in China were lower than the average in lower-middle income countries or in general across the world (Table 1). Fixed line connection, on the other hand, was more expensive than both of these comparative averages. Competition and technological development have steeply reduced the costs of mobile phones. When mobile handsets were first introduced in China in 1994, the price was US$850, which decreased to about US$200 in 1999 (Tsuchiyama, 1999). Similarly, the connection fee declined from US$600 in 1994 (Tsuchiyama, 1999) to US$60 in 1999 (Table 1). By mid-2001, China Mobile eliminated its mobile connection fees in many cities.

To penetrate the market, mobile subscribers have reduced subscription fees and introduced other promotional measures. Examples of such measures include China Mobile's heavy discount plans in Beijing, Shanghai and Guangzhou (*Wall Street Journal*, 2003), and China Unicom's plan allowing users to set five local phone numbers at 1.2 cents per minute (one-sixth of normal rate) and up to 60 percent discounts for heavy users of short messages. The planned rollout of Xiaolingtong (Little Smart)[11] and the ongoing heavy price competition among mobile suppliers have intensified the promotion.

The Chinese government expects that a richer and more technology-orientated economy might help increase respect for the nation. The government also has the ambition of providing every household with a telephone. To achieve these objectives, top priority was given to building R&D capacity in mobile telephony in the late 1990s (Niitamo, 2000). The government is also promoting mobile phones as the "people's phone" and is actively encouraging Chinese consumers in cities and the countryside to buy mobile phones (Kshetri & Cheung, 2002).

Innovations in mobile pricing, such as the introduction of mobile prepaid cards, have been the major driving force for the rapid diffusion of mobile phones across the world. Mobile prepaid cards were introduced in China at the end of 1999 and are contributing to the high mobile growth rates.

Table 1. A comparison of fixed and mobile charges in China and the world

	China	Lower-middle income countries average	World average
Mobile network (1999)			
Connection ($)	*60*	90	86
Monthly subscription ($)	*6.04*	20.99	21.40
Tariff per minute (peak) ($)	*0.05*	0.25	0.27
Tariff per minute (off-peak) ($)	*0.05*	0.18	0.18
100 minute basket ($)	*10.87*	39.69	38.15
Fixed network (1998)			
Residential connection ($)	*226*	133	109
Business connection ($)	*226*	212	155
Residential monthly subscription ($)	*1.9*	4.8	6.9
Business monthly subscription ($)	*2.9*	8.8	11.5
Cost of 3 minute local call ($)	*0.01*	0.05	0.09
Subscription as a % of GDP per capita	*3.1*	3.8	7.5

Source: ITU (1999), Adapted from Kshetri and Cheung (2002)

CHINESE MOBILE MARKET FROM THE CLIP ANGLE

China differs widely from the developed world in terms of the four CLIP dimensions: communications (C), information (I) exchange, payments (P) and "locatability" (L). Compared to developed countries, voice communications account for a significantly higher proportion of the Chinese mobile market.

In data communications, technologies that are a decade old and outdated in most developed countries are among the most profitable in China. For instance, SMS, a standard feature on almost every wireless phone, which sends text messages at 80 percent less cost than voice transmission is very popular in China[12]. The popularity can be attributed to its low cost and the fact that even basic wireless handsets can perform it.

Nevertheless, more advanced m-commerce applications are rapidly emerging in China. For instance, through revenue-sharing deals with China Mobile and China Unicom, China-based Web portals such as Sina, Sohu and NetEase have launched new business models that are tailor-made for the Chinese market. These companies charge the users for news updates, games and online dating information on mobile phones. The services have thus evolved beyond simple text messages. Sohu, for instance, sends out color greeting cards accompanied with voice messages from basketball star Yao Ming (Einhorn, 2004). By April 2003,

Sohu provided 150 fee-based wireless products (such as the Japanese game "Kung-Fu Boy") to its over one million SMS services subscribers (*World IT Report,* 2003).

Virtual games played on mobile devices are also growing rapidly in popularity. By the end of 2004, there were over 10 million mobile game players in China generating about $100 million in revenue (*SinoCast China Business Daily News,* 2005c). The Chinese mobile game market is expected to grow by 80 percent in 2005 (*SinoCast China Business Daily News,* 2005b). IDC suggests that the Chinese online game market will reach $809 million by 2007. Virtual games offered on mobile devices are becoming increasingly popular. For instance, in a game offered by Mtone Wireless Corp in late 2003, 500,000 people signed up in three months (Einhorn, 2004). Chinese and foreign mobile players are planning to launch a wide array of business models for the Chinese mobile market. In mid 2005, for instance:

a. Global cellular NetVillage was at a final stage to distribute its pinball game to Java-enabled handsets in China and aimed to develop a variety of business models (Tsukioka, 2005).
b. Tom Online and Warner Bros. Online were planning to launch a Chinese-language Web site featuring cartoon characters such as Bugs Bunny and Scooby-Doo for mobile phones (*Wall Street Journal*, 2005).

Mobile payment is less attractive in China compared to Europe and North America. A major hindrance is China's cash-based economy — over

Box 2. M-payment in China

Although m-payment is in a nascent stage in China, it is growing exponentially. One estimate suggests that wireless payment will be 15 percent of e-commerce payments in 2006 (Rashtchy, 2004). Mobile companies with innovative business models are capitalizing on this rapid growth rate. Smartpay, a multi-province mobile payment system, which allows subscribers to pay phone bills simply by sending an SMS message (DMAsia.com, 2004) had over 100,000 users as of December 2004 (Ortolani, 2005). The company also has partnerships with seven banks, including China Construction Bank and Agricultural Bank of China (Ortolani, 2005).

The Chinese mobile market, however, lacks more sophisticated m-payment services. M-payment in China is hindered by a number of factors including a lack of secure network with an efficient authentication system in banks; a lack of retail network that accepting codes that link to customers' bank accounts and perceived fraud in transactions (chinanex.com 2004). Basic services such as handset banking in which a cell phone user can check his/her account online and transfer funds within the same bank have been available for some time (chinanex.com, 2004). China Mobile and China Unicom launched one-way payment services in 2003. China Unicom's one-way payment card launched by in 2003 attracted over 20,000 new users in three days (SinoCast China Business Daily News, 2005d). Nonetheless, major players such as China Mobile are aggressively expanding m-payments (Rashtchy, 2004). In May 2005, for instance, China Mobile and China UnionPay announced their cooperation with ten banks to launch m-payment services in Beijing in 2005 (pacificepoch.com, 2005).

90 percent of business transactions takes place on a cash basis. Nonetheless, companies are launching a variety of innovative business models to facilitate m-payment (see Box 2). In the payment realm of business, in 2002 Sumit Mobile Systems Ltd., working with mobile network providers, banks and utility companies in Shanghai, designed a system that allows users to pay their bills by cell phone. When a payment is due, a message displays on the screen that shows the amount due and where the user can then authorize the payment from a bank account by typing a secret code. The service attracted 90,000 users in Shanghai in nine months (Manuel, 2003). Similarly, in mid 2005, The Music Engine (TME), a UK-based technology and online marketing solutions provider, was planning to provide m-payment services in China (Salz, 2005).

The Chinese mobile landscape is also developing rapidly on the locatability dimension. In the mid-2004, Cambridge Positioning Systems (CPS) partnered with Wisemax, a Beijing-based supplier of multimedia messaging and SMS,to provide the Chinese wireless market with location-based services supported by its Matrix software solution (Geospatial Solutions, 2004). Similarly, in November 2004, Sichuan Yingda, a China-based mobile value-added service provider, launched location-based services based on the Matrix location system of CPS to track its vehicles, security personnel as well as other assets of the company (*Wireless News*, 2004).

CONCLUDING REMARKS

The discussion in this paper makes clear that m-commerce in China is expanding quickly while unusual paradoxes coexist. For instance, the Chinese mobile market is the biggest in the world and penetration rates in some of the wealthiest Chinese cities are much higher than the averages of many developed countries. Yet mobile phones are virtually non-existent in many Chinese villages. Similarly, some of the most modern m-commerce applications are emerging from China. At the same time, decade-old mobile technologies that are outdated in the developed countries are widely used in China and are among the most profitable applications. A technology marketer's success in the Chinese m-commerce market, thus, is a function of its capability to go beyond superficial indicators and understand how the company can capitalize on the paradoxes.

QUESTIONS FOR DISCUSSION

1. What are the most important factors that are driving the diffusion of mobile technology in China?
2. How does the Chinese mobile market differ from European and the U.S. mobile markets? What do you think are the opportunities created by the Chinese mobile market that do not exist in European and U.S. markets?
3. Basing your examination from the CLIP perspective, how does the Chinese mobile market differ from mobile markets in Western Europe and North America?

REFERENCES

BBC News (2004). Virgin plans China mobile service. Retrieved December 7, 2004 from, http://news.bbc.co.uk/1/hi/business/4074933.stm

Chinanex.com (2004). Mobile payment market. Retrieved July 15, 2004, from http://www.chinanex.com/insight/mopay0704.htm

Clark, D., & Harwit, E. (2004). *Wi-Fi in China.* Paper presented at the Pacific Telecommunications Council, Honolulu, HI.

Dholakia, N., & Kshetri, N. (2003) Mobile commerce as a solution to the global digital divide: Selected cases of e-development. In S. Krishna &

S. Madon (Eds.), *The digital challenge: Information technology in the development context* (pp. 237-250). Aldershot: Ashgate Publications.

DMAsia.com (2004). Smartpay launches payment services in Anhui. Retrieved October 28, 2005 from, http://www.digitalmediaasia.com/default.asp?ArticleID=4074

Ebusinessforum.com (2000). *GW trade: Serving a high-tech niche in China*. Retrieved 2005 from, http://www.ebusinessforum.com

Einhor, B. (2003, April). Master of innovation? China aims to close its technology gap with Korea and Japan. *Business Week Online* (pp. 22-28). Retrieved July 19, 2005, from http://www.businessweek.com/magazine/content/03_15/b3828010.htm

Einhor, B. (2004, March 15). China.Net; China will soon be No. 1 in Web users. That will unleash a world of opportunity. *Business Week*.

Geospatial Solutions. (2004, May). Location-based services. *Geospatial Soutions, 14*(5), 51.

GIS News (2001). China outstrips Korea in IT industry growth: Samsung. Retrieved September 11, 2005, from http://www.gisdevelopment.net/news/2001/oct/news111001.htm

globalsources.com (2004). Asia to lead worldwide 3GSM subscribers. Retrieved November 26, 2004, from http://www.globalsources.com/gsol/I/GSM-phone/a/9000000058313.htm

Holland, L. (2000, Feb 24). Turning a new pager. *Far Eastern Economic Review, 163*(8), 44.

ITU (1999). *World telecommunications development report.* Geneva: International Telecommunications Union.

James, D. (2001). China's sizzling circuits. *Upside, 13*(1), 60-67.

Kahn, G. (2003, Sept 22). World business; what's old is new: A Chinese Internet company has thrived by focusing on a seemingly obsolete technology. *Wall Street Journal*, R.4.

Koprowski, G. J. (2004). Wireless world: China's WiFi revolution. United Press International. Retrieved July 5, 2004, from http://chineseculture.about.com/gi/dynamic/offsite.htm?site=http://www.upi.com/view.cfm %3FStoryID=20040506%2D101652%2D8297r

Kshetri, N. (2004a, March 11-12). China's emergence as an epicenter of the global mobile market. *Proceedings of the Austin Mobility Roundtable, Center for Business, Technology and Law*, McCombs Business School at the University of Texas at Austin.

Kshetri, N. (2004b, July 10-13). Internationalization of the Chinese cellular industry: The inward-outward connection. *The 2004 Annual Meeting of the Academy of International Business (AIB)*, Stockholm, Sweden.

Kshetri, N. (2005). *Six paradoxes of China imperative for technology companies.* Working Paper, Department of Business Administration, University of North Carolina at Greensboro.

Kshetri, N., & Cheung, M. K. (2002). What factors are driving China's mobile diffusion? *Electronic Markets, 12*(1), 22-26.

Manuel, G. (2003, Oct 20). E-commerce; dialing for dollars: Will people use their cell phones to buy lots of stuff? Phone companies are determined they will. *Wall Street Journal*, 3.

Niitamo, V. (2000). Making information accessible and affordable for all. *Wider Angel*. Retrieved July 20, 2001, from http://www.wider.unu.edu/newsletter/angle2000-1.pdf

Ortolani, A. (2005, Feb 2). Chinese begin paying by cellphone. *Wall Street Journal*, p. 1.

Pacificepoch.com (2005). *China Mobile joins China UnionPay for mobile payment service.*

Retrieved March 25, 2005, from http://www.pacificepoch.com/newsstories/25207_0_5_0_M/

Rashtchy, S. (2004). *The China analyst.* Retrieved September 2, 2005, from http://www.piperjaffray.com/760

Reuters. (2004). *China's mobile subscribers rise 27% in 2004.* Retrieved January 20, 2005, from http://147.208.132.198/news/181_1206466,0003.htm

Salz, P. A. (2005). Power to the people: Do it yourself content distribution. *EContent, 28*(6), 36-41.

SinoCast. (2005a, July18). Shenzhen produced 10% world's cellphones in 2004. *China Business Daily News,* p. 1.

SinoCast. (2005b, July 12). Mobile game market to grow 80% this year. *China Business Daily News*, p. 1.

SinoCast. (2005c, July 5). China mobile game market to reshuffle in 2005. *China Business Daily News*, p. 1.

SinoCast. (2005c, May 31). China unicom Chongqing branch launches one-way payment service. *China Business Daily News*, 1.

Stout, K. L. (2001). China mobile has eyes only for 2.5G. *CNN.Com*. Retrieved June 5, 2005, from http://archives.cnn.com/2001/BUSINESS/asia/06/05/hk.chinamobile2.5g/

Symbianphone.com. (2003). Every third mobile phone made in China. Retrieved April 1, 2005, from http://www.symbianphone.com/?l=apr03.htm

TDAP. (2002). China's new regulatory environment spurs UNICOM's subscriber growth. Interview with Wang Jianzhou, Executive Vice President of China Unicom, *Telecommunications Development, Asia-Pacific*. Retrieved July 10, 2003, from http://www.tdap.co.uk/uk/archive/interviews/inter(unicom_0012).html

Tsuchiyama, R. (1999). The cellular industry in China: Politics, rewards, and risks. *Pacific Telecommunications Review,* 2nd quarter, 149-160.

Tsukioka, A. (2005, July 15). NetVillage, KA-ZeNet form alliance in mobile game business in China. *JCNN News Summaries - Japan Corporate News Network,* p. 1.

UNDP. (2001). *Human development report 2000.* New York: United Nations Development Program. Retrieved July 11, 2001 from, http://www.undp.org/hdr2001/completenew.pdf

Vogelstein, F., Boyle, M., Lewis, P., & Kirkpatrick, D. (2004, Feb 23). 10 tech trends to bet on. *Fortune,* 75-80.

Wall Street Journal. (2003). China's mobile operators bring discounts to big cities. Retrieved April 21, 2005, from http://www.bdachina.com/content/about/pressquotes/P1051109819/en

Wall Street Journal. (2005, June 21). Tom online signs deal with Warner Bros. to distribute content. p 1.

Wireless News. (2004, Nov 26). China's Sichuan Yingda turns to CPS software for location-based solution. p. 1.

World IT Report. (2003, April 9). Sohu launches interactive cellular game in China. p. 1.

ENDNOTES

1. See Vogelstein et al. (2004).
2. Victor Wang, CEO of Mtone Wireless Corp, quoted in *Einhorn* (2004).
3. "Little Smart" is a low-cost, limited-roaming service provided by several fixed-line operators.
4. The 3GSM standard builds on GSM to allow the integration of IP (Internet Protocol) technology and new features such as video calling, faster Internet, WAP access and

downloadable content (http://www.wave-telecom.com/content/shared_content/mobile/downloads/Wave_FAQs.doc).

5 International Mobile Telecommunications (IMT- 2000) is a general term for technologies planned for inclusion in ITU world standards for 3G mobile communications.

6 See http://www.mobic.com/news/2000/01/gwcom_receives_capital_investmen.htm.

7 See http://mobile2004.com/ (4G/B3G-OWA Will Reshape China's Future Mobile Communications Research).

8 WiFi Makes Internet Plus Coffee Possible in Beijing and Tianjin , SinoCast, *China Business Daily News.* July 25, 2003, p. 1

9 See http://www.itfacts.biz/index.php?id=P448

10 This section draws from Kshetri and Cheung (2002).

11 Xiaolingtong is provided by fixed-line phone companies and is much cheaper than mobile services provided by China Mobile and China Unicom.

12 See Kahn (2003).

13 This case draws from Dholakia and Kshetri (2003).

14 GWcom restructured the corporation in April 2002, dividing the business into two companies. The short messaging service (SMS) business has been renamed to byair Corporation, which encompasses the mobile media services. The network business is GWcomPlanet Corporation.

15 See http://www.gwcom.com.cn/gwcom_news-m17.htm

16 See http://www.chinatelecomconference.com/china-dc/bio/bio13.html

17 See http://www.gwcom.com.cn/gwcom_news-m17.htm

This work was previously published in M-Commerce: Global Experiences and Perspectives, edited by N. Dholakia, M. Rask, and R. Dholakia, pp. 34-45, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.29
Canada:
Mobile Commerce Under Construction

Detlev Zwick
Schulich School of Business, York University, Canada

ABSTRACT

This chapter sketches the Canadian mobile commerce space. Relative to other parts of the world, Canadians have been slower in their embrace of the brave new wireless world. The Canadian mobile commerce market is characterized by relatively high costs for consumers combined with inflexible pricing strategies and the rather limited content available for wireless communication and commerce. In addition, limited network speed and robustness to handle more advanced services remain obstacles for wider application rollout and adoption. As a result, the Canadian market for mobile communication and commerce will remain dominated for some time to come by consumers' demand for reliable and low-cost voice services.

INTRODUCTION: MOBILE COMMUNICATIONS IN CANADA[1]

Canada's telecommunications market is characterized by a relatively strong growth in wireless services, high levels of penetration of broadband Internet services and moderate competition for local and long-distance services (Budde, 2005). A recent slowdown in revenue growth for the industry as a whole is due to increasing competition that puts pressure on prices and fuels investment in product and service innovation. Together with wireline services, such as cable and direct-to-home, and multi-point distribution systems (DTH/MDS), the wireless segment has driven growth in the communications service industries. Together, these three market segments account for more than a fifth (up from 14 percent in 1998) of the entire communications service market in Canada (*Statistics Canada*, 2003). Revenue in the resellers,

satellite and other telecommunications segments has remained relatively stable since 2001.

Canada's mobile industry has been expanding significantly from its inception and, according to Informa Telecoms and Media Group (2005), demand is expected to remain strong for the foreseeable future. This assessment is likely to hold true as currently only just above 50 percent of Canadians (about 15 million) have access to a wireless device and subscribe to wireless products and services.[2] Hence, with regard to wireless adoption, Canada ranks quite low internationally, leaving room for growth.

One reason for such slow and low adoption of wireless services is the inflexible and comparatively high cost structure in the Canadian cell phone market, protected by what could be called an oligarchic market structure. Within such a market, lowering price points or pushing different payment models, such as the prepaid phone card, to attract higher volume of what are typically lower margin customers, is not essential. In addition, wireless operators have invested a rather modest C$12 billion in infrastructure and services since 1990 (Informa Telecoms Media Group, 2005). Yet, in 2003 alone the Wireless Service Providers segment, which in Canada includes Bell and its partners (Bell Mobility, Aliant Mobility, MTS Mobility and SaskTel Mobility), TELUS Mobility, Rogers AT&T Wireless, Microcell Telecommunications Wireless services as well as paging companies and other radio communication carriers, generated C$8.2 billion in revenues (*Statistics Canada*, 2003). Relative to the total economy, the telecommunications service industry's *total* share of the economy's capital investment was 2.9 percent in 2003, its lowest level over the past seven years (*Statistics Canada*, 2004).

The Canadian market is currently split between the CDMA and the GSM/GPRS network standards. Among the dominant carriers, Bell Mobility and Telus use CDMA (or updated versions such as CDMA 2000 1xRTT or CDMA 1xEV/DO) while Rogers AT&T and Microcell (acquired by Rogers in 2004) runs on GSM/GPRS (while phasing out its TDMA network). Rogers Wireless has initiated efforts with its United States partner, AT&T, to enhance the Canadian and United States networks by introducing EDGE (enhanced data rates for GSM evolution) capability across the network. Rogers AT&T claims this to be a 3G network enhancement, but according to the ITU (International Telecommunications Union) the data rates for 3G have to be 384 Kbps and above, which EDGE does not provide. So, although EDGE is a 3G technology, the Rogers network is a true 2.5G enhancement. The other carriers follow the United States model and are moving towards CDMA 1X standards with comparable transmission speeds.

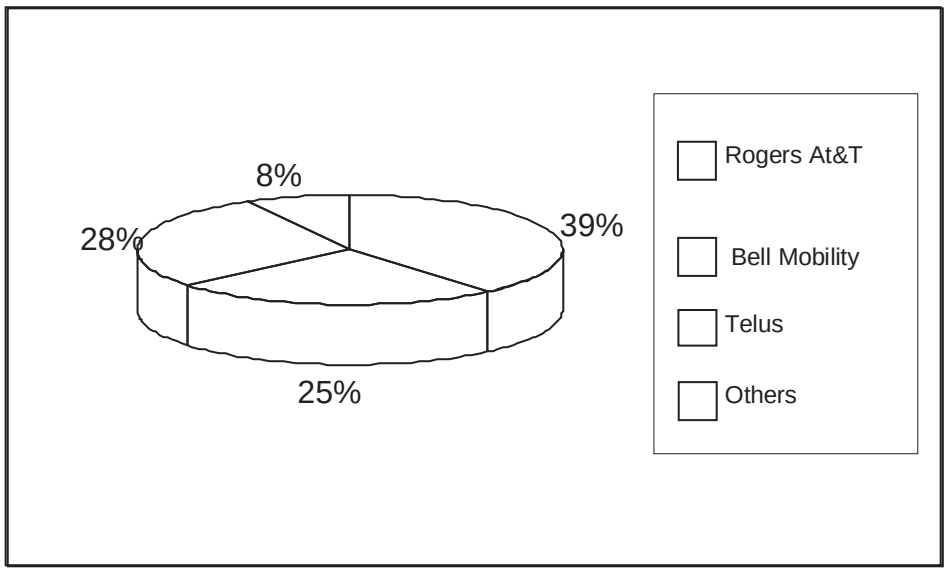
In the following sections we first provide a brief historical context for the emergence of Canada's mobile commerce industry. Then we discuss current success stories in the Canadian mobile service business and subsequently place our observations in the CLIP framework. Finally, we offer our views on the future direction of the industry and respective managerial opportunities.

MOBILE COMMERCE IN THE CANADIAN CONTEXT

Mobile commerce is comprised of many kinds of wireless services, driven by different technologies that operate within various frequency bands. Four major wireless service providers — Rogers Wireless, TELUS Mobility, Microcell and Bell Mobility (see Figure 1) — offer such services in Canada.

In 2004, however, Rogers Communications Inc. acquired Microcell, taking over its subscriber base and the Fido brand. Since their inception in 1988, all of the incumbent telecommunication service providers have increased their wireless subscriber bases. Since 1999, the increase in the number of subscribers has been particularly robust with annual growth rates for the national

Figure 1. Wireless market share by wireless company, 2004



Source: Statistics Canada and companies' annual reports

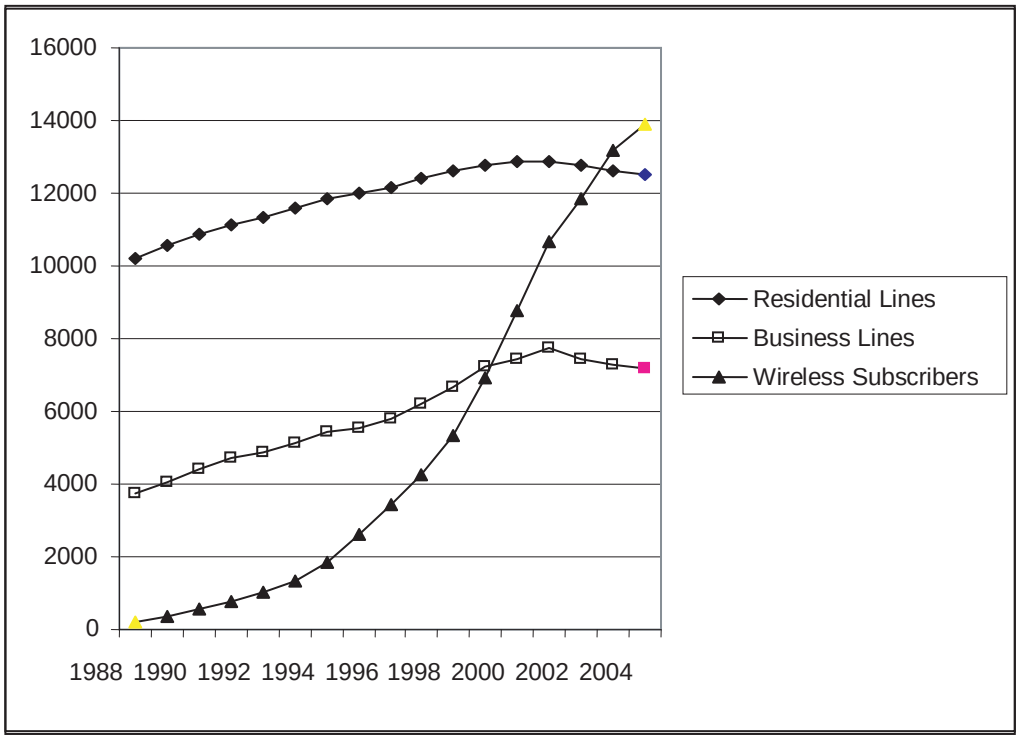
wireless service providers of more than 20 percent (see Figure 2).

The steady growth of the wireless subscriber base has been accompanied by a steady decrease in prices, thus decreasing the average contribution per user to overall revenues from almost C$80 in 1994 to below C$28 in 2004 (Miezejeski, 2004). As a rough approximation of a customer's wireless bill, declining average revenue per customer means not only falling prices for existing services but also underscores the difficulty of wireless service providers to market new value-added, high-margin services to new and existing customers. However, in the past two years average revenues per customer have stabilized as companies now focus on retaining higher margin customers rather than acquiring new ones. A strategic emphasis on revenue growth to the detriment of aggressively adding new customers may indicate that the market is maturing, even though the teledensity indicator of wireless subscribers is still just above 50 subscribers per 100 inhabitants (Miezejeski, 2004).

The wireless segment had experienced negative operating profits due, in part, to the significant start-up costs associated with the introduction of the digital Personal Communications Services (PCS) network. Operating profits significantly improved in 2002 and have since surpassed the operating profit margins of the wireline segment. This is largely due to a significant drop in capital expenditures per customer from previous years.

At the end of 2004, the wireless service provider market presents itself as transparent and is largely divided among three remaining big players in Canada– Rogers, Telus and Bell Canada.

Figure 2. Growth in wireless subscriber base relative to residential and business wirelines, 1988-2003

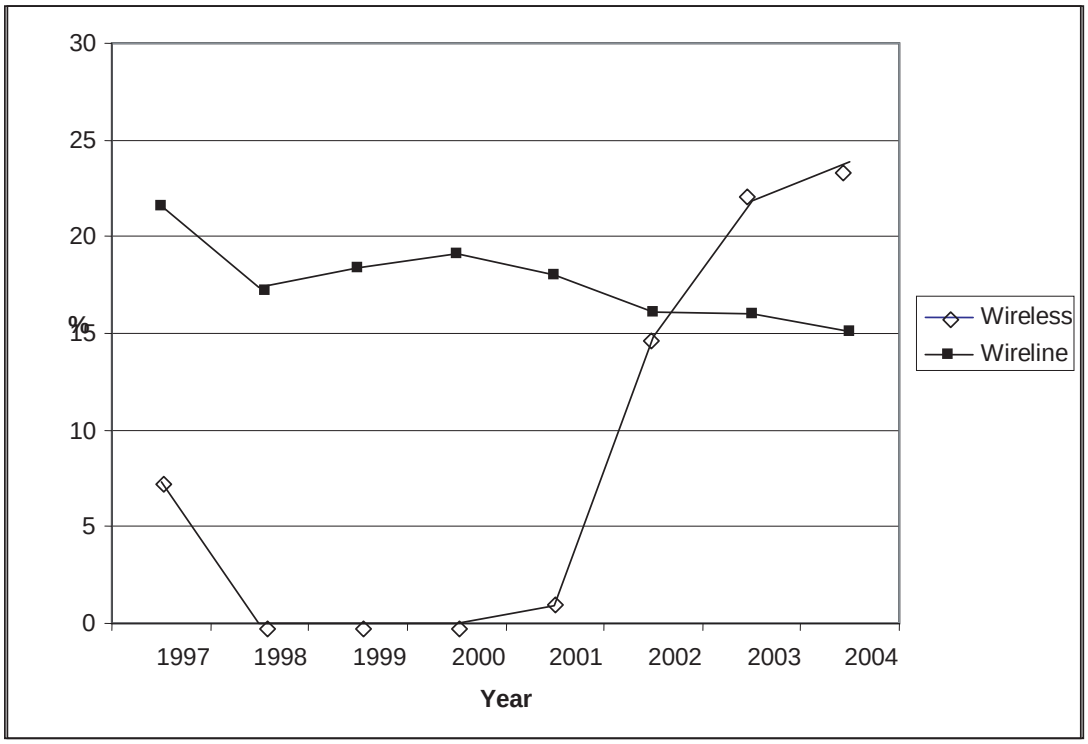


Source: Industry Canada, Survey of telecommunications service providers (April 2004), and Buttle (2005). Residential and Business line numbers for 2003/04 are based on estimates by Industry Canada and NBI/ Michael Sone Associates (2004).

Competitive pressure in the wireless market has been moderate, and with the loss of Microcell and their flagship brand Fido, it is unlikely to increase. As one commentator puts it (Englehart, 2004), "[L]ocal telephone competition in Canada has been a flop. The incumbent telcos have nearly all of the market after seven years of competition. Almost all of the competitive local exchange carrier (CLECs) have become insolvent and many have disappeared." With respect to mobile commerce and the wireless market, the combination of sustained profitability, uncertainty about what wireless services, phone features, and applications consumers really want, a strong broadband-based Internet presence and absence of high competitive pressures to innovate provide for a less dynamic market environment in Canada than perhaps in other parts of the world.

One of the unique features of the Canadian market in general, and the telecommunications marketing in particular, is the regulatory constraints put on foreign companies interested in competing. Foreign ownership is limited to 25 percent, which means that the existing oligarchic

Figure 3. Telecommunications services operating profit margin 1997-2004



Source: Industry Canada, Survey of telecommunications service providers (April 2004), and companies' annual reports.

market structure is relatively safe from foreign threats. However, an interesting development will be the arrival of Virgin Mobile in the guise of a mobile virtual network operator (MVNO). In a deal with Bell Mobility, Virgin Mobile rents wireless capacity rather than build its own network. It is a model that has been known in Europe but took longer to develop inroads in Canada because none of the big carriers wanted to open up the market. It can be expected that Virgin Mobile will shake up the market and increase competition, especially for the lucrative youth market.

SUCCESSFUL SERVICES IN CANADA'S MOBILE SECTOR

As Goggin (2004, p. 2) observes, "[C]ontemporary soundscapes now feature not only voice calls in previously quiet public spaces such as buses or restaurants but also the aural irruptions of customized polyphonic ringtones identifying whose phone is ringing by the tune downloaded." In Canada, cell phone use, while prevalent, does not yet have the same public pervasiveness it has in other parts of the world. It is far from absent, of

course, but compared to Berlin, Rome or Tokyo, where busses and subways, waiting lines, town squares, cafes, university dining halls and high schools are considered ideal space/time configurations for eclectic cell phone uses, Toronto seems conspicuously empty with wireless chatter, frantic multimedia- and text-messaging and absorbed game-playing.

SMS, for example, one of the most successful applications in Europe, has been a slow starter in Canada. A recent Ericsson survey (Ericsson Consumer Lab,2004) reported that only four percent of Canadian wireless phone users report sending or receiving text messages once a day or more. Only about half of all cell phone users send SMS messages at least once a week. On the other hand, North Americans claim to spend about 49 minutes a day talking on their cell phones, which is almost twice the global survey's average talk-time of 27 minutes. Cost remains the main reason for this different usage pattern in North America. Voice service is relatively inexpensive, with flat fee plans that provide users with a set amount of weekday minutes (typically between 100-1,000, depending on the price of the plan) and unlimited evening and weekend time. Such a price/usage structure replicates in essence the fixed-line service plans in Canada, which charge consumers a flat fee for unlimited local calls and a per-minute fee for long-distance calls; SMS, in comparison, is relatively expensive. Hence, unlike other parts of the world, in Canada the economic incentive to switch modes of communication from voice to the somewhat cumbersome and slow text messaging does not exist to the same degree.

Furthermore, the slow adoption rate of innovative services such as SMS has to do with the fact that WSPs showed little interest in compatibility between services, hoping that theirs might become the standard. Thus, interoperability between systems (enabling consumers of different carriers to exchange short messages) was not achieved until 2002, when carriers looked to Europe and noticed that almost 15 percent of WSPs' profits came from the use of the short messaging service. As a result, wireless carriers pushed for interoperability between systems to enable short message exchange that was independent of the digital wireless technology of the sender or receiver. In April of 2002, Bell Mobility (CDMA), Telus (CDMA and iDen), Rogers (TDMA, GSM/GPRS) and Microcell (GSM/GPRS) set up the CMG Interoperability gateway, which enabled short messages exchange between any cell phone on their combined systems (Crowe, 2002).[3]

While interoperability was established for SMS, the existence of two network standards has also hampered the development and functionality of other complex interactive network applications (i.e., applications that do not run from the handset), such as Multiplayer games or mobile banking services. As a result of this, and a cost structure that favors voice service, the market for data transfer is comparatively small. Yet even as domestic demand for mobile communication services, productivity applications and entertainment content may remain limited for the foreseeable future, Canada is home to a number of companies like AirG, one of the most successful mobile entertainment content supplier companies,headquartered in Vancouver, British Columbia. While AirG provides games and multimedia content to Canadian carriers, with some exclusively for Bell Mobility, the company sources and sells globally.

Mini Case Studies: From the Canadian Mobile Desert

As Dholakia, Rask and Dholakia state in the opening chapter of this book, m-commerce refers to monetary transactions conducted via a mobile telecommunications network by employing devices such as mobile phones or palmtop units. For m-commerce to happen, at the minimum, the device and the network should be configured to enable communications (C), information (I) exchange and payments (P). In the Canadian

context, such configurations are still somewhat limited. Yet, some m-commerce applications are beginning to emerge. We preset case studies that are representative of the emerging models in the Canadian market. The study of Paymint exemplifies the use of the phone as wallet. Another case study relates to Research in Motion's (RIM) BlackBerry device and the versatile communication capabilities it provides. The case of Swordfish presents the small but trendy market for location-based entertainment applications, while the Instant Talk case features a voice-based, multiple-user communication service by Telus.

Paymint

Monday afternoons are a hectic time for Angela Smith, a freelance sportswriter for the *Toronto Star* and mother of two teenage daughters. At 1 p.m. she gets in her Chrysler *Sabre* to drive from her home in one of Toronto's sprawling suburbs to the weekly editorial meeting at the newspaper's downtown headquarters. Luckily, she finds a parking spot close to the *Star*'s facilities in one of the city's public parking facilities scattered all over downtown and prominently marked with a large green P-sign. As she enters the parking garage, she takes out her cell phone and speed-dials the phone number displayed at the entrance. An automated answering service recognizes her phone and prompts her for a password. She enters a 5-digit code and immediately receives a conformation that she has been registered in the garage. The meeting with the newspaper's sport editor is short this time, and just 60 minutes later Angela returns to her parked car. As she starts it up and slowly makes her way to the exit, she again speed-dials the parking lot's phone number to "check out." The system on the other end confirms the transaction, provides Angela with the exact duration of her stay in the garage and the amount that will be charged to the credit card associated with her phone number and password. If she is lucky, she might make it in time to watch her daughters' ice skating lessons.

Paymint is a parking service provided by Mint Inc., a software company headquartered in Toronto, Ontario that allows for a simple and convenient way of paying for parking services through the use of a mobile phone. The customer is the Toronto Parking Authority (TPA), the largest municipal parking operator in North America, which manages more than 200 parking facilities and thousands of automated street-side parking meters. In Toronto, there are two typical methods of paying for parking. The first involves paying an electronic parking meter (cash or credit card) for street-side parking; the second involves paying a human parking attendant or automated teller upon entering or leaving the parking garage. With the Paymint System, customers avoid these hassles by first parking their vehicles, then dialing a specified telephone number and entering their lot number. When a customer is ready to leave, he or she dials the number again to end the parking session. The session is then charged to a pre-authorized credit card that users specified during their registration phase. Payments can then be viewed on a consolidated on-line statement for easy tracking. The city pays Mint Inc. a commission for handling payment security, billing and collection of fees.

The adoption of this technology has been promising, especially in downtown Vancouver, British Columbia where the Paymint System has been incorporated into 80 percent of the 7,500 EasyPark lots run by the city. In Toronto, the TPA has implemented Paymint in at least 30 of the 200 Green-P downtown lots. It represents one of the first e-wallet type applications in Canada.

BlackBerry (Research in Motion)

Peter doesn't know the meaning of the word "relax." As the creative director of one of the biggest ad agencies in the country, he is constantly on the

road to meet with clients, check up on designer teams and oversee implementation. Just as Peter was leaving his office in a trendy neighborhood in the east of the city to get some Sushi for lunch, he notices a beeping sound coming from his suit's breast pocket. It was his BlackBerry reminding him that he had a 12:30 p.m. appointment with the account manager and the client of his biggest current job. He had forgotten about the appointment which was scheduled a month ago, but like many other appointments over the past few days (since the largest Canadian beer company announced that it was looking for a new agency to conduct a major re-branding campaign), got lost in a deluge of new activities. He quickly ran back to his office, took the files he needed for the meeting and called a cab. In the car, he used his BlackBerry to send a message to his two team members reminding them of the meeting and what he expected them to talk about. Peter knew that unlike their cell phones, his colleagues never switched off the BlackBerry device, so they would receive his message instantly. Within minutes, he received confirmation that they were on their way to the meeting with their gameplan in hand. Peter sinks into the back seat and tries to relax as much as possible during these busy times. He still had not had lunch and decided to e-mail his client's secretary to order some sandwiches. Two minutes later his beeping BlackBerry displays a reassuring "no problem."

The BlackBerry is a hand-held mobile device created by Research In Motion (RIM), a Canadian firm based in Waterloo, Ontario. The BlackBerry is ideal for busy professionals who need to stay connected while away from their desks. The device allows individuals on the road or away from their offices to receive their e-mails instantly (always on cellular connectivity), organize their contacts and make phone calls. It also provides Internet accessibility via a fully functional web browser. On traditional PDA devices, users have to log into an e-mail program and manually retrieve their e-mails. The RIM BlackBerry eliminates the need to manually connect to such programs through what they call "push" architecture. The BlackBerry's "push" technology enables messages to be automatically and immediately routed to the hand-held device. The BlackBerry has a mini-keyboard, which makes it relatively easy to compose e-mails and text messages. The latest BlackBerry models (7100 series in North America) have a quite large (high-resolution 240x260 pixel screen) color screen, operate on a 850/900/1800/1900Mhz GSM/GPRS network and have an integrated phone with speakerphone along with Bluetooth connectivity. The operating systems provide support for MIDP 1.0 and WAP 1.2. The BlackBerry devices and subscription plans can be purchased directly from partnered network service providers such as Rogers Cable in Canada. Shipments of the BlackBerry increased 289 percent in 2004 from the previous year, boosting the company's market share to 18.6 percent from 5.3 percent at the end of 2003..

The BlackBerry has become indispensable among the Canadian corporate elite. Its cellular connectivity and always-on feature deliver reliable, anytime/anywhere communication and information services on a decent size screen. However, the device and usage rates are expensive, currently preventing private consumers from adopting BlackBerry devices in any significant numbers.

Swordfish

At 3 p.m. Peter and Claudia can't wait to get out of their midtown high school classroom and, armed with their cell phones, hit the streets of Calgary. They aim for Kensington and 10th Street, a trendy hangout for students from local colleges and universities. The district's alternative/granola/urban tech feel makes it easier for Peter and Claudia to inconspicuously disappear into their cell phone screens and launch Swordfish, North America's first location-based mobile multiplayer game. They will need to separate, though, because as

of the moment they connect to the network, they are competitors in this challenging skills game, which pits the player against a virtual school of swordfish and against other players trying to get to the fish first. Peter and Claudia arrange to meet up later for dinner at their favorite organic sandwich bar a few blocks north, where they will compare the results of their afternoon's worth of fishing. As Peter says good-bye to Claudia, his mind is already racing, scheming how to make the best use of roughly three hours of playing time before they meet up again. He loads Swordfish and immediately his location is verified and the fish become visible on his screen. As he scans the screen, he sees a few smaller fish up north, probably 300 feet. They will be easy to catch, he thinks, but Claudia went in that direction when they parted and, while he contemplates his chances of beating her to them, a message appears telling him that these fish are being approached and caught by Superfly (Claudia's screen name). The good news is two much bigger fish are swimming just about 200 feet east of him and, since Claudia will need some time to reel in her catch, he should be able to get there first, unless, of course, someone else interferes; currently the area looks clear. As he gets closer to the fish, one of them, it turns out, is a Swordfish. The Big Catch! Hard to find, and even harder to catch. It will take patience but, even if that is all he will get out of this afternoon — just one Swordfish — his high score will get a major boost, destined to move him close to the top spot in the Calgary Swordfish rankings — and finally past Claudia — at least for a few days.

Blister Entertainment Inc. of Calgary, Canada, has launched North America's first assisted global positioning satellite (GPS) game with the introduction of Swordfish, a high-tech contest that pits cellular phone users against virtual fish and other users. To play Swordfish, users need to download the game onto their java-enabled, GPS-equipped handsets first (Audiovox 8450 or 8455, a Samsung SPH-A600 or Sanyo 8100 phone), and they need to be in a Bell Mobility CDMA 1X network. The game system is comprised of three main components: the client software that resides on the phone, KnowledgeWhere's Location Application Platform (LAP) and the mobile providers' location-based system (LBS). Swordfish client software provides the gaming interface to the end-user. The first version of the game was launched in the summer of 2004. Version 2.0 was released December 2004.

Carriers charge a flat fee for the purchase and an additional usage fee of around 15¢ per scan. In addition, mobile browser airtime charges may apply. When the game is launched, the user's actual position is determined via assisted GPS and rendered on screen in relation to the nearest school of virtual fish. In order to go fishing, the user needs to physically move within range of the object, cast the lure and try to catch the fish. The game uses artificial intelligence for the creation of virtual fish and schools of fish. Hence, the objects behave relatively intelligently, depending on their type. Some will try to move away from the fisher to avoid being caught, while others struggle and fight for freedom when actually caught and skills are required to hold on to big fish.

Users can always check their scores on the phone and online and see their local and national ranking against other players. In sum, mobile multiplayer network games take advantage of high-speed networks and services that mobile service providers are eager to sell. The problem with the current game ecology is that the payment structure favors handset games rather than networked multiplayer games. In addition, the value of a network game depends on the presence of other users. Yet, presenting a typical high-technology marketing dilemma, users are unlikely to sign up if the existing user network is perceived as weak (Frels, Shervani, & Srivastava, 2003; Moore, 1991). Carriers are partly responsible for the low adoption rates of these games at the current time.[4]

Nevertheless, with network connectivity growing faster and more robust, interactive net-

work- and location-based games are likely to become important applications, spearheading the development and introduction of immersive communications environments that operate and persist across phones and networks. And if these systems make for smooth micropayments, the early multiplayer, multiplatform game systems could usher in a new era of mobile commerce and content (Hall, 2003).

Instant Talk

Sitting in one of Montreal's bohemian cafés savoring the short-lived silence before the evening crowd files in for concentrated espressos and urbane tête-à-têtes, John glances over the quarterly sales figures on his laptop. Selling high-end chocolate in Canada was a challenging business! After a large amount of coffee, John thinks up a way to improve distribution but he needs to check with a number of folks to verify whether his idea has legs! John reaches into his pocket and pulls out his cell phone with the Instant Talk feature available to his company's employees. With the push of a single button, John is able to access his "partner list" and send out "message alerts" to everyone he needs to talk to. Within seconds he is connected simultaneously with Paul at the head office in Utah, Peter, main supplier of milk in Calgary and Wendy, sales director of the western region in Vancouver. They are holding a virtual meeting. With everyone there, throwing around ideas and making suggestions, John refines his idea and, once disconnected, begins drafting the memo of his distribution scheme! As the café is filling up with the after-dinner crowd, John transmits the memo and heads home.

Instant Talk is a new feature offered by Telus Mobility, a Canadian wireless telecommunications provider. The Instant Talk feature allows users to use their cell phones like traditional walkie-talkies. By pressing a single button, clients can communicate immediately with one or many contacts simultaneously. This feature only works on selected handsets, therefore, for a party to be included in an Instant Talk session, an Instant Talk-enabled phone is needed. The service is targeted at businesses that want to enable employees to have instant meetings without having to travel.

To activate the service, the client simply pushes the button on the side of the phone, inputs the receiver's number or selects a client list then presses the button again. This sends an alert message to that person(s) and once received, they can begin their conversation. A unique feature of Instant Talk is the ability to create online client lists, which enable the client to talk to multiple people across the country simultaneously with the click of a button. Registering client lists is done online via Telus mobility's online account management system. In the scenario above, if John wants to get in contact with Paul in Utah, Peter in Calgary and Wendy in Vancouver, all he needs to do is to select the contact list that contains these contacts and press the call button, which sends an alert message to all three parties instantaneously. Instant Talk uses Telus's national CDMA 2000 1X network, which provides decent coverage across Canada. The carrier offers this service as an addition to its existing PCS phone plans and charges between C$10 and C$20 for it.

VIEWING CANADA'S MOBILE SECTOR IN THE CLIP FRAMEWORK

At this point in time, the CLIP framework captures many possibilities for mobile applications and services that will take many years to materialize in Canada. The Canadian mobile ecology is currently largely focused on and fueled by integrating communication capabilities and providing entertainment-based information exchange. Network coverage for mobile phones is generally good across the country and excellent in large urban centers. Communication integration is improving as new handsets can now be used

to send and receive e-mail. Customers of Rogers AT&T, for example, can log into their Yahoo! accounts to check and respond to e-mail. With RIM's BlackBerry leading the way and setting standards with regard to integrating all possible communication formats, adoption of handsets will be driven by their ability to bring multiple communication formats to the consumer across platforms.

The second driver of mobile commerce and wireless business in Canada is entertainment-based information exchange. Short and multimedia messaging has seen slower adoption and growth than in other parts of the world but both information formats will keep growing and fuel the data market. Handsets are improving, adding new features such as camera capabilities and Java support, running more sophisticated games and playing diverse ringtones. Here, the arrival of Virgin Mobile could have an accelerating effect because of the company's focus on aggressively marketing mobile usage as fun and entertainment to the youth market, historically the early adopters of new services. This may include pushing more handset- and network-based games, even as network standard incompatibility issues between major carries set a barrier.

Payment services, contrary to what the Paymint e-wallet mini case above may indicate, are the exception and location-based services are almost entirely absent from mainstream consumer markets. In the payment sector of the CLIP m-portal framework (see Figure 1 in keynote chapter), banks offer mobile banking services including account balance, money transfer and bill payment. However, none of these services have seen encouraging adoption rates. In fact, some banks have already terminated development and support for mobile stock trading after recognizing that there is not sufficient interest among existing and new consumers for this service. The 724 Solutions wireless banking service, launched first in Europe and later offered by the Bank of Montreal in collaboration with Nokia's network division, flopped while costing the bank a lot of money. Many problems driving up cost while not generating any measurable return had to do with the complexities of making such services available across different network standards. At a time when there were, in effect, three coexisting main network standards, banks had to make sure their interactive, network-based banking communication service was operational with all of them.

Similarly, even though information about strategies, investment and results of mobile initiatives is difficult to obtain, some of the MBA students at a leading business school who are directly involved with their employers' mobile initiatives confided to the author that so far consumer demand for any mobile service has been disappointing and that any future initiatives will be tested very conservatively for their profit potential. The conventional wisdom among these IT marketing professionals is that consumers simply will not tolerate the mobile device's small screen and slow connection speed while they have high-speed connectivity and intimate familiarity with the environment on their home or work computers. In addition, it should be noted that *electronic banking* in Canada is also much less prevalent and common than in Europe. It could be theorized that because of low profile electronic banking, an important socializing force is lacking that would support the adoption of mobile banking.

Location-based services (LBS) are increasingly available and are only very slowly becoming more popular, mostly among the youth segment interested in location-based multiplayer games such as Swordfish presented above. The lackluster LBS market does not suffer from a supply side problem. There is no dearth of choice for consumers interested in transforming their handsets into mobile game machines. However, closer scrutiny of actual adoption rates of some of the multiplayer games reveals very low numbers. For a good gaming experience, handsets need to be java-enabled and GPS-equipped and networks need to have at least 2.5G data transfer rates. Such requirements

slow down the rate of adoption of new technologies. In addition, unlike handset games that are played locally and individually, network multiplayer games suffer from the network effect, which states that a critical mass of players is needed to make the experience valuable for everyone (Lee & Colarelli O'Conner, 2003).

More promising at this point are less complex and more functionally oriented LBSs. For between C$3 and C$12 per month, Rogers sells a user plan called "navigate mobile internet," which allows users to browse the Web, download ringtones and video clips and locate the closest taxicabs. In addition to the youth market, safety needs drive location-based traffic. Bell Mobility's Roadside Assistance service has been enhanced to include location-based technology, called the e9-1-1 in North America, which allows Bell to locate customers in case of an emergency, if they consent. In 2003, the Roadside Assistance service was the first commercial application of Assisted GPS roadside assistance technology offered by a wireless carrier in North America. Bell also has a service called MapMe™, which uses Global Positioning System technology (GPS) and Cell Tower technology to locate the user and display a live, real-time map. The user can zoom in and out of the map, pan in all directions, display street names and get step-by-step directions from his current location to restaurants, movie theatres, ATMs, gas stations and hotels.

In sum, within the Canadian context, WSPs build the m-portal. Since charges occur on a pay-per-use basis, WSPs try to increase content of the wireless Internet, such as ringtones, e-mail accessibility (e.g., Yahoo! and Hotmail), chats or games. Promotion of these services is moderate and user acceptance comparably low, chiefly because of the perceived high costs of using them. Key limitations to more aggressive development and adoption of payment- and location-based services are found on the supply and demand sides. In the payment arena, meaningful offers have been rare and carriers, financial players and third parties appear unenthusiastic about building up a mobile commerce product line. On the location-based service side, product offers are increasing more rapidly but technology issues dampen their adoption. Upgrading to more advanced, GPS and java-enabled handsets is costly for consumers and 2.5G network rollouts (CDMA and GSM/GPRS) have been slower than expected; coverage will remain spotty for some time to come.

CONCLUDING REMARKS: STANDARD BARRIER AND THE FUTURE

Exchange students from Europe and Asia frequently comment on what they perceive as a backward mobile ecology in Canada. Undoubtedly, Canadian students are much less reliant on, and possibly more selective with, cell phones than their foreign counterparts to satisfy their communications, entertainment and personal organization needs. Mobile commerce in Canada largely consists of voice, some moderate short- and increasingly multimedia-messaging use, entertainment (single-user games) and some customization-related services (such as downloading ringtones and printing skins for one's phone[5]). With broadband connectivity at home among the highest in the world and the number of wireless networks in cafes, hotels, airports and campuses growing very quickly, PCs currently represent the preferred means of satisfying mobile communication, entertainment and information needs for Canadians. In addition, in the eyes of consumers the flat fee pricing strategy favors voice services over data transfer. Without a price differential between voice and data, text messaging, often perceived slow and tedious by the users, remains relatively unattractive.

Furthermore, in many parts of the world the mobile phone increasingly plays an important role in contemporary visual and material culture as a fashion item and status symbol (Goggin, 2004,

Jan 12; Harney, 2004). In Canada, however, cell phones for the moment maintain a distinctively utilitarian appearance, while their effect on the cultural and aesthetic fabric is still relatively confined to subsegments of the population, such as younger consumers and, to a lesser degree, businesspeople.

As mentioned above, Canadians have been slower in their embrace of the new wireless services that have swept Europe and Asia, including global blockbusters like SMS and, increasingly, MMS. Our analysis of the Canadian wireless context suggests that the uniquely Canadian "mobile consumer behavior" is the result of the comparably high cost of the medium and carriers' rather inflexible pricing strategies, the still somewhat limited content available for wireless communication and commerce and the lag of network speed and robustness to handle more advanced and sophisticated services. However, attempts are being made to promote what *is* available, such as SMS, at least in some market segments. During the 2004 presidential elections in Canada, Nokia's Youth Text 2004 program sponsored an unprecedented text message initiative designed to engage young people in the issues of the federal election campaign. The program staged a number of successful live text messaging chats between presidential candidates of all parties and young Canadians across the country (Schick, 2004).

As in other parts of the world, WSPs and content providers face a lot of uncertainty about consumer preferences beyond a desire for reliable and clear voice connections. Hardware features such as wireless headsets and color screens have been successful, as have personalization features such as downloadable ringtones and printable skins. Nevertheless, surveys by Forrester show that cameras, videocams, and other doodads each right now generate interest in only smaller groups of American users (LaGesse, 2004).

It appears that unlike in the United States, where competition has squeezed profit margins on voice calls, Canadian WSPs are less pressured to introduce new services to increase the data traffic sent over their networks. Still, even for Canadian carriers, increasing the data market must have high priority to remain profitable in the long run when voice will become commoditized even here. The addition of Virgin Mobile to the carrier ecology is likely to speed up this move towards data services. Hence, features and applications will be launched in the hope that something will catch on. In the final analysis, it is not a question of whether additional m-commerce functionality will be offered to Canadians, but rather when and how.

Judging from our analysis, information- and entertainment-based services prompted by the availability of new phone features (device convergence) and faster networks are most likely to be added to the current products because the infrastructure to deliver these products is already in place. LBSs are a promising, yet wholly unpredictable market, and one thing carriers have going for them is that they have been very cautions in managing consumer expectations for these services to forestall disappointment. Particular functional services such as roadside assistance and taxicab location will catch on sooner because these services are clearly defined in the mind of the consumer and they do not require the same technological level of sophistication that, for example, network games do.

Payment services using handheld devices will take longer. Banks, while initially relatively eager to develop such capabilities, have since learned that consumers are less interested than anticipated. In addition, banking in Canada is still relatively paper-based (checks are still a very common and often preferred method for paying rent, phone and utility bills, or government services). Electronic wire transfers have only very recently been introduced to the consumer market and are still considered a novelty(!) and, compared to their European counterparts, many years behind in the process of integrating information technology in business and consumer banking processes. Banks also face high cost hurdles and disadvantageous

economies because they have to make services work across handsets and networks. In addition, the banking sector is characterized by a handful of powerful incumbents, which does not make for an overly competitive and innovative environment.

In sum, growth and evolution of mobile commerce in Canada is hampered by the co-existence of the CDMA and the GSM network standards. Once the 2.5G networks have been rolled out and users upgrade to more powerful handsets, a more viable environment for developing and delivering third-party services, especially in the payment and locatability application field, will exist. However, whether the mobile Internet i-Mode style (based on the delivery of entertainment and "fun") will take hold in Canada is another question that will have more to do with the right marketing approach than merely good technology (Moon, 2004).

QUESTIONS FOR DISCUSSION

1. What are the main factors that have held back the development of large-scale m-commerce applications and services in the Canadian mobile telecom markets?
2. What actions would you recommend to expand the market for location-based services (LBS) in Canada?
3. What can Canadian companies do to leverage the global success of Research in Motion (RIM) with its BlackBerry handheld devices?

REFERENCES

Budde, P. (2005). *Broadband and consumer e-commerce in Canada Dec. 2004 review.* Retrieved January 30, 2005, from http://www.internetworldstats.com/am/ca.htm

Crowe, D. (2002). *SMS interoperability: Canada leads the way.* Retrieved November 21, 2004, from http://www.cnp-.wireless.com/ArticleArchive/Wireless%20Telecom/2002Q3-SMSInterworking.htm

Englehart, K. (2004). *No new rules: Incumbents must remain regulated if voice competition is to grow.* Retrieved January 12, 2005, from http://www.cablecastermagazine.com/issues/ISarticle.asp?id=151779&story_id=17602111609&issue=06012004&PC=

Frels, J. K., Shervani, T., & Srivastava, R. K. (2003). The integrated networks model: Explaining resource allocations in network markets. *Journal of Marketing, 67*(1), 29-45.

Goggin, G. (2004). Mobile text. *M/C: A Journal of Media and Culture,* Jan 12, *7*(1).

Hall, J. (2003). *Multiplayer—the only mobile game*. Retrieved February 12, 2005, from http://www.3nw.com/pda/gaming/mobilegaming.htm.

Harney, A. (2004). It's the catwalk calling. *National Post,* Oct 23, p. SP4.

Informa Telecoms Media Group, (2005). *2003/2004 telecoms in Canada: Industry report.* Retrieved January 30, 2005, from http://www.telecoms. com/marlin/20001000461/MARKT_EFFORT/marketingid036?proceed=true&MarEntityId=1106931287073 &entHash=1001dc55761.

Lab, E. C. (2004). *Ericsson study.* Retrieved December 3, 2004, from http://www.ericsson.com/ca/en/press/2004_11_23.shtml

LaGesse, D. (2004). The spell of the cell. *U.S. News & World Report.* Retrieved February 12, 2005, from http://www.usnews.com/usnews/culture/articles/041129/29cell.div.htm

Lee, Y., & Colarelli O'Connor, G. (2003). New product launch strategy for network effects products. *Journal of the Academy of Marketing Science, 31*(3), 241-255.

Miezejeski, T. (2004). *Wireless in Canada*. Retrieved January 16, 2005, from http://www.intelecard.com/features/03features.asp?A_ID=360

Moon, Y. (2004). Don't just do something, stand there! *Harvard Business Review, 82*(3), 16-17.

Moore, G. A. (1991). *Crossing the chasm: Marketing and selling technology products to mainstream customers*. New York: Harper Business.

Schick, S. (2004). *How political groups could improve their online approach*. Retrieved October 12, 2004, from http://www.hillwatch.com/Media/ITBusiness.aspx

Statistics Canada (2004). *Telecommunications service in Canada: An industry overview*. Retrieved January 28, 2005, from http://strategis.ic.gc.ca/epic/internet/insmt-gst.nsf/en/sf06084e.html

ENDNOTES

1 This chapter has benefited from the excellent support of my research assistant Anthony Rotondo, for which I am very grateful.

2 Different surveys generate different estimates, ranging anywhere from below 50 percent (*Statistics Canada*) to just above 60 percent (Ericsson).

3 SMS interworking is not an issue in Europe, where GSM originated and is still the dominant wireless technology. For the Canadian carriers, there is only GSM and, consequently, only one SMS format. Interoperability is assured at every protocol layer.

4 A scan of the company's Website suggests that only about 40 or 50 users currently play Swordfish, most of them located in Calgary, where Blister Entertainment Inc. is headquartered.

5 The emergence of Top 10 ringtone charts demonstrates the popularity of this option.

This work was previously published in M-Commerce: Global Experiences and Perspectives, edited by N. Dholakia, M. Rask, and R. Dholakia, pp. 15-33, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.30
Mobile Commerce in South Africa

Anesh Maniraj Singh
University of KwaZulu-Natal, South Africa

INTRODUCTION

The last decade has seen a rush among businesses to get onto the Internet. Since its introduction, e-commerce has grown in leaps and bounds. The frenzy to get online and be a part of the "new economy" was spurred on by media hype describing the Internet as the greatest technology this century. Organisations embarked on initiatives to change their business models, looking for e-strategies as a means of revolutionising their business. By mid 2000, many of the dot.coms were "dot.gones." The primary reason for this sudden death was that businesses forgot the basic rule of business: creating economic value. Economic value as defined by Porter (1985) is the gap between price and cost—the larger the gap, the greater the economic value. According to Porter (2001), gaining a competitive advantage does not require a radical approach to business; it requires building on the principles of effective strategy. Businesses that went online should not have looked for e-strategies, but should have improved on their existing strategy to include an e-strategy.

GPRS, wireless Web, handhelds, m-commerce, 2nd coming of the Internet, m-management, killer apps, 2G or 3G, always-on, have been the buzzwords in the media. Is this new hype really worth the fuss? M-commerce has failed in the United States and has made a brief appearance in South Africa. Therefore, this article asks the question "Is there potential to revive m-commerce in South Africa?" In attempting to answer this question, this article will examine issues such as uses of m-commerce, the benefits and challenges of m-commerce, trends in the wireless industry, and the technology underlying m-commerce. This article will also attempt to provide suggestions for harnessing the power of the wireless Web. Most of the discussions are based on universal experience supported with what the current situation is in South Africa; therefore, this article will not be separated into a universal section with a smaller subset focussing on South Africa.

BACKGROUND

What is Mobile Commerce?

Organisations have just begun to get comfortable with e-commerce in terms of what it can do for them and what are its limitations. Some are still coming to grips with e-commerce and have now been hit with the new wave of m-commerce. According to Rainer (2000), m-commerce refers to the use of wireless communications technology to access network-based information and applications using mobile devices. Laudon and Laudon (2004) described m-commerce as the use of wireless technologies for conducting business-to-business and business-to-consumer transactions over the Internet, hence, m-commerce can be described as the mobile Internet (Herron, 2000). Cleenwerck (2002) describes m-commerce as the wireless Web. It is evident from these definitions that the main characteristics of m-commerce are mobility, wireless, mobile devices, and the Internet.

It is evident that m-commerce is merely an extension of the Internet to wireless handheld devices, thus bringing e-commerce into the palms of users beyond the physical boundaries of bricks and mortar. If m-commerce is e-commerce on the move, why all the hype?

Uses of Mobile Commerce

The proposed uses of wireless technology seem like something out of a James Bond movie. However, users should clear the image of driving a BMW with a Nokia cell phone from their minds. Improved communication is probably the most important use of the wireless Web. People have access to text-based data such as short message services (SMS), e-mails, news broadcasts, and file transfers. Advanced functions include booking of tickets for movies and shows and making restaurant reservations. Very advanced features would involve transactions such as purchasing airtime, ordering products online, and secure banking. Some of the other uses include the following:

- **Navigation Systems:** Global Positioning Satellite (GPS) services integrates the wireless Web with satellite and Geographic Information Systems (GIS) to locate people in space. These systems will be able to assist people who are lost to find their way. GPS will also be able to calculate the shortest route between two points, saving time and money.
- **Electronic Wallets:** According to Posthumus (2001), wallets built into cell-phone technology is highly appealing and could herald a new era in business and financial systems whereby users could make payments to vending machines for the purchase of items and effect funds transfers at in-store point-of-sales (POS) systems.
- **Multipurpose Remote Controls:** Handheld devices will soon be linked to all the electronic devices in a home, allowing one to control gates, burglar alarms, televisions, sound systems, and just about anything that is electronic. Currently, some models of the Ipaq® have a multidevice interface. The use of these devices as remote controls is limited by appliance manufacturers' developing devices compatible with PDAs. Other uses of m-commerce include stock trading, weather forecasts, vehicle tracking, and instant messaging, among others.

The uses of m-commerce are limited only to the extent of one's imagination. However, Reedy, Schullo, and Zimmerman (2000) warned that certain products such as perishables and small items were not suited for sale on the Internet. Similarly, La Fontaine (cited in Brewin, 2000) noted that not all business opportunities can be translated onto the wireless Web. The uses of m-commerce are many, but do they bring with them any benefits?

M-COMMERCE IN SOUTH AFRICA

Benefits of Mobile Commerce

According to Navision (2002), the benefits of m-commerce are threefold (i.e., it provides immediate access to information where it is needed, it helps employees respond immediately to business needs, and it allows organisations to provide better field service). Wireless makes communication possible in areas of uneven terrain, such as mountains, where it is difficult to install cable. According to Haag, Cummings, and Dawkins (2000), serving customers goes beyond the provision of products and services. Businesses need to provide perfect service at the customer's moment of value; one of the dimensions of which is *place*. M-commerce makes it possible to deliver service where the customer wants it, such as at his or her work place, at home, and even at the beach. Armed with a cell phone or a PDA, a sales consultant can provide near-perfect information to assist the customer in his or her buying decision, immaterial of the customer's location.

M-management, an offshoot of mobile commerce, makes it possible to keep managers apprised of all events at the workplace, wherever in the world the manager may be. M-productivity, also an offshoot of mobile commerce, makes it possible to improve worker productivity. Employees can access their contact information, review their calendars, and respond to e-mail, which ordinarily would have to be done at a desk. Other benefits of m-commerce include the following:

- **Immediate Access:** Due to the nature of the technology, cell phones and PDAs are instantly on, which reduces overhead time, which is the time taken to get started. PCs take an extremely long time to get started, initialise all the peripherals, and then to establish a dial-up connection with the service provider.
- **Use of Niche Time:** Time that is unavailable or wasted whilst waiting for services or sitting idle in public transport can be leveraged for work (Rainer, 2000). GNER, a United Kingdom train operator, has installed WiFi on its trains. In first class, access is free, and as a result, standard-class commuters are upgrading to first class (WiFi Growth on UK Trains, 2004). WiFi hotspots are making it possible to access the Internet in airports, restaurants, and other public places in South Africa.
- **Generate New Income:** In order to generate new income, firms need to advertise extensively in traditional media such as television, radios, and billboards to stimulate impulse buying. The advertising must contain the message that these products can be ordered from ones handheld device.
- **Reduce Costs:** Cell phones, pagers, and PDAs are much cheaper than computers and laptops; this reduces organisational cost and makes m-commerce accessible to larger markets.
- **New Marketing Medium:** Short-range broadcast systems can be used within small areas or buildings such as malls, where advertisements and special offers could be sent to their mobile devices (Gaede, as cited in Mobile Services: Less Talk, More Profit, 2005).

It is evident that the benefits of m-commerce are immense. But how does one get around those tiny keypads?

Challenges Facing Mobile Commerce

Mobile phones have extremely small screens that are capable of delivering, at maximum, approximately eight lines of text (Herron, 2000). Furthermore, the absence of a QWERTY keyboard

makes typing a painful experience. According to Ewalt (2000b), even when one gets the technology working, purchasing online using a cell phone or PDA is extremely difficult due to the low resolution of the screens, unreliable network support (dropped calls), and poor security.

In order for m-commerce to work properly, it will have to run seamlessly among different carriers, networks and handheld devices. In short, there must be interoperability (Mc Guire, cited in Ewalt, 2000b). Just like HTML offered a one-size-fits-all platform for the Internet, similarly, a common platform has to be developed for the wireless Web.

According to King (cited in Ewalt, 2000a), consumers are reluctant to use their phones for anything other than SMS and voice transmissions. However, user reluctance is not all. The number of users with data-enabled phones is low in the United States, which has only 1 million data-enabled phones (Brewin, 2000). Similarly in South Africa, only 3% of Vodacom subscribers had GPRS phones, and less than 0.5% of users had multimedia phones (Vodacom SA Launches GPRS, 2002).

The biggest challenge for m-commerce is developing and providing the technology infrastructure that underlies m-commerce.

Technology Required to Enable Mobile Commerce

Network Infrastructure

Current cell phone/wireless technology was not intended for m-commerce, which requires the transmission of data. Current networks were intended for transmission of voice and simple text messages. The evolution of cellular technology makes it easier to understand the technology underlying m-commerce. South Africa, a known follower in terms of technology development and technology implementation, has a more advanced cellular platform as a result of technology leapfrogging. Instead of investing in outdated technology and then upgrading systems, cellular providers in South Africa adopted the latest technology.

Debate still rages whether or not to implement 3G networks. However, the return on investment does not justify the cost of the infrastructure, especially because there are very few applications and services available that require the infrastructure.

South Africa has recently seen the emergence of GPRS with both Vodacom and MTN, making the service available on their networks. The business benefits of GPRS are immense. There is no need for dial-up modems, and users are always on. For private users, however, GPRS costs more than land-line downloads. MTN SA charges U.S.$ 8.33 for downloading one megabyte of data. The same amount of data downloaded on landlines take approximately 15 minutes, which at U.S. 1 cent per minute costs a mere U.S. 15 cents. Private users who opt for GPRS will have to pay the premium price for the convenience offered by the service. According to Socikwa (2004), broadband would enable a connection to each customer allowing for simultaneous combined voice and data services, this would reduce online time, which in turn will reduce cost. In a bid to increase broadband usage, South African mobile networks have discounted the price to a mere U.S. 33 cents (Masango, 2005). It is evident that that the enabling technology has to be in place to conduct m-commerce. However, focussing on technology and price will not drive the market; it is applications and services that will (Treguertha, 2005).

Applications

Users are beginning to realise that their phones are minicomputers and they are not satisfied with the limited choices of content available on their cell phones (King, 2005). Wireless Application Portal (WAP) was supposed to be the forerunner of bringing the Web into the hands of the user. However, according to Nadler-Nir (cited in Herron,

2000) and Gani (2002), WAP in South Africa has been a dismal failure. The failure of WAP has been attributed first to the boring services offered, and second to the wrong assumption that users want to browse the Web on their phones (Herron, 2000). Two of South Africa's leading banks, ABSA and Standard, have developed phone banking that allows users to check balances, pay accounts, and transfer funds. First National Bank has developed an alert system that alerts customers of the amount of money withdrawn when they use their debit or credit cards. If the card is used fraudulently, the customer can put an immediate stop on further purchases.

Handheld Devices

In order to get the full benefit of m-commerce, one requires a PDA, or a cell phone that has data-handling capabilities. GPRS phones were introduced in South Africa in mid 2002. These phones were only available on business contracts that cost more than an average contract, however, making the technology inaccessible to nonbusiness and prepaid users. The high cost of PDAs has made this technology equally inaccessible to nonbusiness users. Furthermore, using PDAs is cumbersome due to the absence of a mouse and keyboard. It is evident that the network technology and hardware for m-commerce are readily available, but is this any indicator of the future of m-commerce?

Trends in M-Commerce

In the United States, several large companies are shutting down their wireless services mainly due to the fact that people use their cell phones for voice calls rather than SMS and transacting unlike elsewhere in the World, Americans use personal computers to shop online (Wolverton 2002). The hype surrounding m-commerce promised the consumer the wireless Web. However, what was delivered did not meet these expectations (Rainer 2000). For business, the promise of increased sales and exposure to new customers never materialised. According to Stahl (2002), m-commerce is still more myth than reality. M-commerce has suffered failure every bit as dramatic as the dot. gones (Ewalt, 2002b). In a survey conducted by *Information Week,* 75% of small companies and 58% of large companies were sceptical of the revenue generation potential of m-commerce (Ewalt 2002b). Is mobile commerce truly dead?

Growth Potential of Mobile Commerce

In South Africa, there are 18.7 million active cell-phone users, which increases by more than 9,000 users per day and has the potential to reach 21 million users in 2006 (Statistics of Cellular in South Africa, 2004). Between MTN and Vodacom, 71% of the geographic area is covered by cellular masts, which means that people almost anywhere in South Africa can engage in m-commerce. According to Tregurtha (2005), South Africa has the highest teledensity of mobile to fixed phones in the world. Based on the sheer numbers of cell phone users, it suggests that there is growth potential for m-commerce. However, it must be noted that South Africa has an illiteracy level of 33% (Aitchison 1998). M-Commerce which is text based, would require that users are literate in order to use the service effectively. There are no statistics to prove what numbers of users are illiterate or semi-literate. However, it will impact on the m-commerce market size. Furthermore, as mentioned previously, GPRS phones are inaccessible to a large number of cell phone users, which further diminishes the potential m-commerce market. It is evident as one dissects the statistics that the potential m-commerce market is not as large as one would expect. Does this then sound the death knell of business-to-consumer (B2C) mobile commerce?

Evidently not. Cointel, a trailblazer in m-commerce in South Africa, has reported revenues of

up to R50 million per month. Rather than being distracted by new technologies such as GPRS, Cointel uses existing GSM standards to connect cell-phone users with their banks to purchase airtime. Users no longer have to go to a shop to purchase a recharge voucher—recharging can be done anywhere, 24 hours day, 7 days a week. Cointel is developing the technology to include the purchase of flowers, tickets, and limited shopping (*Comparex News,* 2001). Although m-commerce does not look very healthy, it is not entirely dead. M-Commerce needs to be given an injection in order to revive it.

What Needs to be Done to Revive Mobile Commerce?

All the stakeholders need to make an effort to revive m-commerce. Handset manufacturers, network service providers, businesses, and consumers need to work together to arrive at a solution that will harness the latent power underlying mobile commerce. Some of the issues that need to be addressed include the following:

- **Role of Network Service Providers:** Increase bandwidth to handle voice and data traffic at faster speeds. Increase capacity to hold extra traffic. Improve Security standards. Keep pricing low.
- **Role of Handset Manufacturers:** Handset manufacturers, in conjunction with networks, need to develop a transmission standard that is safe, cheap, fast, and convenient to use. Manufacture cheaper, user friendly handsets with built-in firewalls.
- **Businesses:** Businesses need to use traditional media such as newspapers, radio, television, and billboards to encourage consumers to transact through their phones. Businesses need to train their customers in the use of their m-commerce sites. Furthermore, businesses should not wait for B2C to take off; instead, they should use mobile commerce extensively to benefit from business to business transactions. Ferguson and Pike (2001) suggested that organisations choose one application and use it extensively to determine how it is adopted and how its works rather than trying to make a profit from it from the time it is implemented.
- **Content Providers:** The role of content providers is growing in stature in the revival of m-commerce. Customers are still looking for the "killer app" that will justify the move to transacting by phone. According to Cowper (cited in Mobile Services: Less Talk, More Profit, 2005), the killer app is always going to be personalisation. He stated further that Telecoms need to target the right content and services to the right customers.

It is evident that m-commerce requires a joint effort to become the new wave of electronic commerce. Stakeholders including competitors should collaborate initially to develop a set of standards. Thereafter, they can work independently at perfecting what they do best.

FUTURE TRENDS

Mobile commerce in South Africa saw a resurgence in 2004. The three networks—Vodacom, MTN, and Cell-C—are focusing on their core business that is to provide efficient connectivity. As a result, they are not engaging in the development of mobile applications. Instead, they have invited independent content providers to develop applications for the mobile Web. WAP and SMS (short message service) have come to the fore. Companies such as Xactmobile have emerged. Users can purchase ringtones, backgrounds, games, and screensavers by simply sending an SMS request to Xactmobile. The cost of the item being purchased is debited to the users cell-phone bill, which the networks then pay to the vendor. Like its predecessor e-commerce, mobile commerce

has been increasingly used to download sex and pornography. For a fee, users can purchase sex stories and pornographic pictures, which are sent by WAP onto their cell phones.

After the December 2004 Tsunami disaster in Indonesia, local television stations advertised SMS numbers that users could contact to make financial contributions to the disaster relief fund. If they sent the SMS to number *x*, their cell phone account would be debited for US$ 2.50, and if they sent it to number *y*, their account would be debited for US$ 5.00.

Motor-vehicle dealerships could send courtesy SMSs to customers after they have had their vehicles serviced or repaired in order to gain valuable feedback regarding customer satisfaction. The University of Cape Town developed a content delivery system at the end of 2004, wherein lecture halls are wireless enabled, allowing for students to link their PDAs to the lecturers system. They can download lecture slides, make notes on the slides, and even answer simple multiple-choice questions. This system has great potential for e-learning. However, until the prices of PDAs drop, this will just be another great innovation that goes nowhere.

According to Ankeney (2001), research indicates that a 6-second increase in transaction speed can boost a fast-food franchise's revenues by 1%; the ability to order fast foods by phone before reaching the outlet could benefit both the consumer and the supplier. Other uses could include short-range broadcast systems where restaurants, cinemas, and other stores could detect all the cell phone users within a particular radius and broadcast messages inviting them to place their orders or make a booking in advance of arrival, thereby saving time. The biggest challenge for the mobile industry in South Africa is disposable income. Eighty-four percent of South African users are prepaid users (Statistics of Cellular in South Africa, 2004), many of whom are willing to forgo other luxuries to enjoy the status attached to owning a cell phone (Posthumus, 2001). If disposable incomes are limited, affordability of connectivity and purchasing online services will be limited as well.

Privacy and even security may be threatened due to eminent legislation. According to Perlman (2004), a draft proposal is being finalised that will oblige all mobile and fixed operators to provide real-time monitoring of all voice and data communications services including SMS, WAP, MMS, and GPRS messages. As a result, encryption techniques need to be disabled, which would expose communication and transactions to cyber-criminals on a relatively unsecure platform. This has the potential to threaten M-commerce even further.

CONCLUSION

This article has provided a balanced perspective on mobile commerce by answering the question, "Is there potential to revive m-commerce in South Africa?" Based on recent historic performance, it would seem that m-commerce was dying in South Africa. However, there are a number of issues that were outlined that need to be addressed that could see the turnaround of mobile commerce. Furthermore, current practice has shown that creative vendors are developing applications and products that cell phone and other mobile device users are willing to purchase using their mobile devices. With the large number of cell phone users in South Africa m-commerce has the potential to be the future, perhaps all it needs to rebound, is an affordable killer app and time.

REFERENCES

Aitchison, J. J. W. (1998). *A review of adult basic education and training in South Africa*. Retrieved December 2004 from www.fsu.edu /~/v-adca/english/adeas.html

Ankeny, J. (2001). *When will m-commerce whet America's appetite?* Retrieved January 2005, from www.wireless review.com/ar/wire less_mcommerce_whet_ americas/

Booming mobile commerce revenues top R50m per month. (2001). Retrieved January 2005, from www. comparexafrica/page s/news.asp?news _id=107

Brewin, B. (2000). *Mobile commerce*. Retrieved January 2005, from www.computerworld.com/ printthis/2000/ 0,4814,52653,00.html

Ewalt, D. M. (2002a). *Palm addresses pitfalls of mobile commerce*. Retrieved January 2005, from www. informationweek.com/story/ iwk20020613S0012

Ewalt, D. M. (2002b). *Wireless e-commerce bombed. Is there any life in this strategy?* Retrieved December 2004, from www.infor mationweek. com/story/iwk20020613S0013

Ferguson, G. T., & Pike, T. H. (2001). *Mobile commerce: Cutting loose*. Retrieved January 2005, from www. accenture.com/xd/xd.asp?it= enweb&xd=ideas/outlook /1.2001/cutting.xml

Glascock, S. (2002). *Is there hope for m-commerce?* Retrieved January 2005, from www. techweb.com/tech/mobile/20020624_mobile

Haag, S., Cummings, M., & Dawkins, J. (2000). *Management information systems for the information age*. Boston: McGraw Hill.

Herron, A. (2000). *Why is SMS popular?* Retrieved January 2005, from www.useit.com/alertbox/20000709 comments.html

It's not too soon to be WAP happy. (2000). Retrieved December 2004, from www.itouch.co.za/news/archives2000/news_ 28_jul_2000_02.html

King, S. (2005). *Companies that make cell-phone content face resistance from carriers*. Retrieved February 2005, from www.menafn.com/qn_print.asp?storyid=cqgb eueicq1bulunftexdt05uru5u

Laudon, K. C., & Laudon, J. P. (2004). *Management information systems: Managing the digital firm* (8th ed.). NJ: Pearson Education.

Masango, G. (2005, May). *Cell C joins Internet fray* (Business Rep. 3). Durban, South Africa: Independent Newspapers.

Mobile services: Less talk, more profit. (2005). Retrieved January 2005, from www.sun.com/br/comms_821/feature_mobile.html

Mobile solutions for navision South Africa. (2002). Retrieved December 2004 from www.navision. c o.za/za/view.asp?categoryid= 359&documentid=435,2

MTN launches GPRS in South Africa called MTNdataLIVE. (2002). Retrieved January 2005 from www.cellular.co.za/africa/south-africa/mtndata live.htm

Perlman, L. (2004). *SA interception regulations set for implementation*. Retrieved December 2004, from www.cellular.co.za/news_2004/oct/101904sa _interception_ regulations_set.htm

Porter, M. E. (1985). *Competitive advantage*. New York: MacMillan.

Porter, M. E. (2001). Strategy and the Internet. *Harvard Business Review, 3*(79), 63-78.

Posthumus, C. (2001). *Mobile Africa: Leapfrogging the digital divide*. Retrieved December 2004, from www.thefeature/article ?articleid-12036

Rainer, R. K. (2000). *An update on wireless communications and mobile commerce*. Retrieved December 2004, from www.auburn. edu/~rainerk/mobile_commerce.html

Reedy, J., Schullo, S., & Zimmerman, K. (2000). *Electronic marketing: Integrating electronic resources into the marketing process*. Philadelphia: Harcourt College.

Socikwa, K. (2004). *Africa: Mobile renaissance?* Retrieved December 2004, from www.africafocus.org/docs04/han0405 .php

Stahl, S. (2002). *M-commerce is still more myth than reality.* Retrieved from www.informationweek.com/story/IWK20020616S0002

Statistics of cellular in South Africa. (2002). Retrieved December 2003 from www.cellular.co.za/stats/statistics_ south_africa.htm

Tregurtha, G. (2005). *2G or not 2G ? That's the hot question for S.A. mobile works.* Retrieved January 2005, from www.mediatoolbox.co.za/pebble.asp?p=63& relid=2799

Vodacom S.A. launches GPRS. (2002). Retrieved January 2005, from www.cellular.co.za/africa/south-africa/vodacom_gprs

WiFi growth on UK Trains. (2004). Retrieved January 2005, from www.cellular.co.za/news_2004/oct/100304wifi_ growth_on_uk_trains.htm

Wolverton, T. (2002). *Wireless commerce is wavering.* Retrieved January 2005, from http://news.com.com/2102-1017_3956969.htm l?tag=st.util.print

KEY TERMS

Bandwidth: Determines the rate at which information can be sent through a channel.

Broadband: A transmission facility having a bandwidth sufficient to carry multiple voice, video, or data channels simultaneously.

Internet: The Internet is a global connection of individual and networked computers to other individuals and networks.

Killer App: Technical jargon for the eternal search for the next big idea.

Mobile Commerce: Refers to the use of wireless communications technology to access the Internet to conduct purchases and sales.

Network Service Provider (NSP): Also known as network carriers, NSPs provide the infrastructure that enables mobile communication. Cell phone users pay for using the infrastructure.

Teledensity: The number of telephones per 100 people in a region.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 772-778, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.31
Mobile Payment Issues and Policy Implications:
The Case of Korea

Youngsun Kwon
Information and Communications University, Republic of Korea

Changi Nam
Information and Communications University, Republic of Korea

ABSTRACT

This chapter introduces three mobile payment plans that have been launched in Korea: mobile banking service, mobile prepaid electronic cash service and mobile phone bill service. Based on the recent experiences of the Korean economy, this chapter discusses the regulatory and monetary policy issues associated with mobile payments. Mobile payments are superior to existing means of payments because of their efficiency and convenience and mobile network operators (MNOs) are on the verge of turning into non-bank financial institutions in their nature. The government needs to facilitate the crossbreed between banks and MNOs to accelerate the development of efficient payment instruments rather than hindering innovation in banking industry.

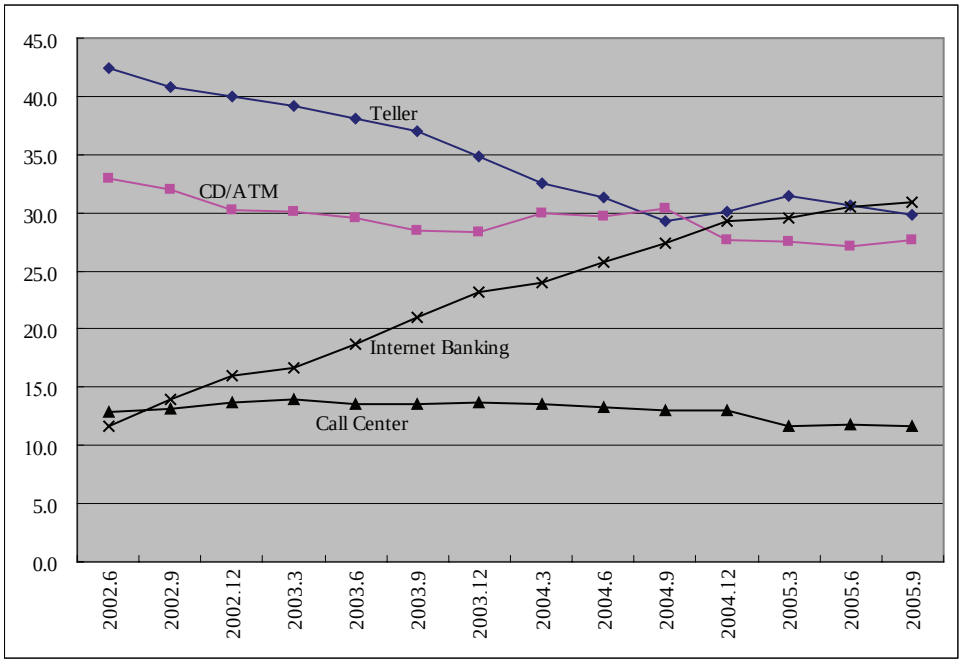
INTRODUCTION

All business transactions entail at least one method of payment. Traditionally, fiat money has been the most popular method of payment for retail business transactions. But as computer and telecommunication technologies have developed, electronic payment methods, including credit cards, debit cards and electronic payments via online banking, have appeared and are now widely used in most economies. By the latter part of the 1990s, many advanced economies, including Korea, had experienced three important changes: an electronic commerce boom, wider Internet penetration and the spread of mobile communication. As electronic commerce sales expanded, firms and customers needed new electronic payment systems that were more convenient and efficient.

As electronic commerce has grown, Internet banking in Korea also has increased rapidly since its introduction in 1999. In mid-2002, teller service was still the dominant delivery channel of banking services in Korea, with face-to-face transactions accounting for 42.4 percent of total retail banking transactions. By contrast, only 11.7 percent of retail banking transactions was delivered through Internet banking. However, the gap between traditional teller banking and Internet banking has been fast shrinking. Indeed, the number of Internet banking transactions has outgrown teller banking transactions more recently, as shown in Figure 1. The fact that the proportion of transactions done through banking call centers has been relatively stable over time suggests that Internet banking is the main substitute channel for teller service and cash dispensers (CDs)/automated teller machines (ATMs). As a result, the number of Internet banking customers has grown from 1.2 million at the end of June 2000 to 25.4 million at the end of the third quarter 2005. In addition, daily fund transfers made through Internet banking over the same period have increased about 22 times, from 608 billion Won to 13.5 trillion Won.

However, one limitation is that Internet banking cannot fully satisfy the human desire for mobility, because it cannot provide users with a ubiquitous connection to the communication system. As a result, Internet banking has a critical drawback as a means of payment for retail business transactions because it does not accommodate human beings' mobility. In order to use an Internet banking service, customers often need to use a specific computer, usually their own, which has an authentication file. Although customers can carry their authentication file on a diskette, they have to find a computer hooked up to a network

Figure 1. Share of Korean banking service channels, 2002-2005 (Source: Bank of Korea [2005])



to use the Internet banking service. Therefore, Internet banking cannot satisfy that part of human nature that makes us prefer to move about, rather than staying in one place all the time. By contrast, mobile payments, which we define as the transfer of a currency-denominated value from buyers to sellers using mobile devices such as mobile phones and personal digital assistants (PDAs), provide better mobility for customers.

Although mobile banking was introduced in Korea in late 1999, the number of users of mobile devices as a payment tool began to grow significantly after the integrated circuit (IC) chip-embedded mobile phone appeared in the third quarter of 2003. In addition to mobile banking, new mobile payment solutions such as mobile prepaid electronic cash service and mobile phone bill service have been recently provided to customers.

Based on the recent experiences of the Korean economy, this chapter discusses the regulatory and monetary policy issues associated with mobile payments. Mobile payment systems are evolving, so our policy discussions are confined to the current state of the Korean mobile payment systems, which we believe are relatively more advanced than those in other countries. Unfortunately, a rigorous empirical analysis of the effects of mobile payment systems on financial intermediation is almost impossible due to the lack of data. Therefore, this chapter focuses on identifying and discussing the policy issues related to mobile payment systems.

The second section introduces the concept, characteristics and types of mobile banking in Korea and describes the current state of mobile payment systems in Korea. According to Mishkin (1997), three basic forces determine the evolution (or innovation) of payment systems: efficiency (low transaction costs), convenience and security. This section addresses how the superiority of mobile payments stems from their efficiency and convenience. In addition, a brief literature survey on mobile payments is provided. Section 3 discusses the main regulatory and monetary problems caused by the new mobile payment systems in Korea. In particular, Section 3 reports on the competition and resultant lack of cooperation between banks and network operators. Section 4 addresses government reaction to the problems and discusses what the government can do to accelerate the innovation of payment systems. Finally, the fifth section concludes the chapter and suggests future research topics.

MOBILE PAYMENTS: CONCEPT, CHARACTERISTICS, AND TYPES

Concept of Mobile Payments

There are two viewpoints on mobile payments. First, a "process perspective" views mobile payments as just one method of transferring monetary value among people. Kuttner and McAndrews (2001) defined a payment as "a transfer of monetary value from one person to another" (p. 37). Following this, a mobile payment can be defined as a monetary value transferred among economic agents using mobile devices.[1] The alternative "business perspective" views mobile payments as the product of the collaboration between financial institutions, including banks and credit card companies, and mobile network operators (MNOs). In other words, a mobile payment is a new crossbreed service enabled by the collaboration of firms belonging to two different industries—the financial industry and the mobile network industry.

Mobile payments are a new payment method, which can either substitute for or complement existing payment instruments, such as cash, checks, credit cards and Internet banking. They substitute for existing payment instruments if people prefer to use mobile payments for convenience in circumstances where multiple payment instruments are available, whereas they complement existing payment instruments if mobile payments are the only available payment instrument.

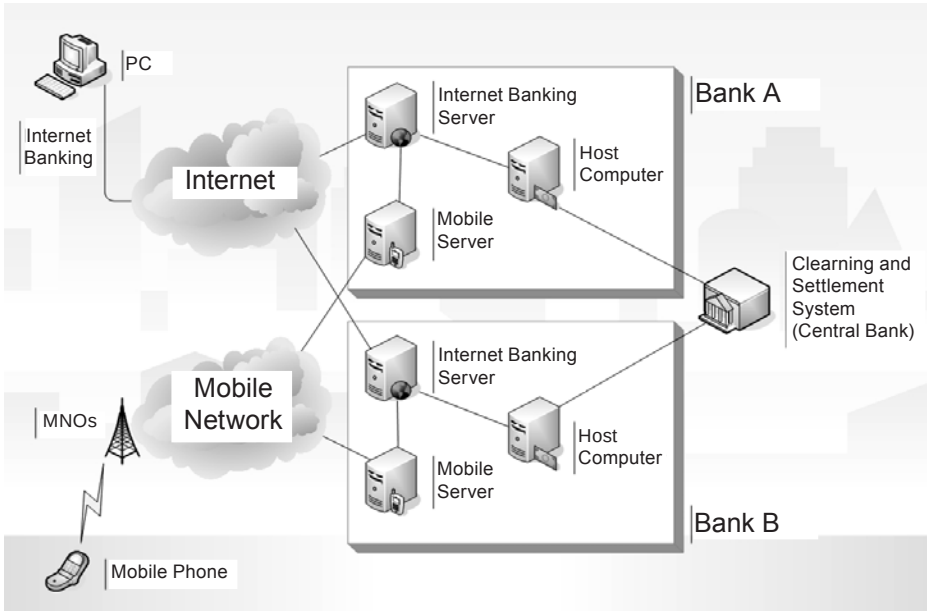
But no matter what role mobile payments play, their usefulness grows as electronic commerce expands. This is because, as electronic commerce transactions increase—for example, downloading music, avatar and virtual game tools—the need for instantaneous completion of small payment amounts also grows.

Characteristics of Mobile Payments

Mobile payments have unique characteristics that distinguish them from similar electronic payments. First, they provide users with mobility, which Internet payments cannot.[2] If we focus on payments based on bank accounts, the path of money transfer in mobile banking is exactly the same as that of Internet banking. As illustrated in Figure 2, a person with an account in Bank A has access to his/her account using a mobile handset and can transfer some monetary value to a person with an account in Bank B. MNOs provide network connections between mobile handsets and the mobile server located in a bank, which in turn is connected to an Internet banking server. Contrary to Internet banking, users carry a payment instrument, the mobile handset, with which they can complete bank transactions anytime, anywhere. In short, mobile banking is much more convenient to use than Internet banking.

Second, mobile payments provide users with greater ubiquity of service than other electronic payment instruments, such as credit and debit cards, which can be used only at stores that have card readers. By contrast, mobile payments can be used anywhere. As well as being more convenient, online payments, including mobile payments, are usually less expensive payment methods than credit and debit cards, as noted by Kuttner and McAndrews (2001).[3] Third, although cash provides people with high mobility and ubiquity, the major drawbacks are weak security and the inconvenience involved in carrying large

Figure 2. Mobile payment system



amounts. By contrast, even though there is some chance of losing the mobile handset, the chance of losing money using mobile payments is almost zero. In addition, mobile payment instruments substitute for paper currency and coins, because carrying mobile handsets enables users to pay virtually any amount of money for retail transactions. Finally, although checks are close to mobile payments in terms of mobility, they are inferior because many small stores do not accept them and their settlement takes a longer time, whereas settlement is virtually instantaneous for mobile payments. In addition, mobile payments do not require face-to-face contact. This can be contrasted with checks, which hinder instantaneous or flexible payment.

As summarized in Table 1, mobile payments are the most convenient, secure and efficient payment method available among the competing retail payment instruments. They are more convenient than other conventional payment instruments because people can transfer funds anytime and anywhere as long as the mobile network works effectively. Mobile handsets are also as easy to carry as a credit card or cash. For instance, people cannot transfer funds to those who live in remote areas at midnight with conventional payment instruments, but they can with a mobile payment device. Even cash (not to mention credit cards and checks) is inferior to mobile payments because it is not appropriate for electronic payment. Cash is the lowest payment instrument in terms of security, especially when value is large. In conclusion, it is mobility and ubiquity that make mobile payments unique.

Types and Current Status of Mobile Payments in Korea

In Korea, there are presently three types of mobile payments: mobile banking (usage of which has grown significantly recently), mobile prepaid electronic cash and mobile phone billing. First, mobile banking involves transferring monetary value between the bank accounts of buyers and sellers using mobile handsets. Figure 3 illustrates funds transfer between banking accounts using a mobile banking service. To purchase a product or a service, a payer (a buyer) sends funds to a payee's account by accessing his/her bank account using a mobile device.[4] After confirming that the money is transferred, the payee (seller) sends the product or service to the payer. This type of e-commerce has tended to be limited to retail transactions of less than $100 mainly because there is some risk that the seller may send a defective product or one different from that ordered. Hence, sales have usually been completed directly between the buyer and seller.

Table 1. Comparison of conventional payment instruments with mobile payments

	Mobility	Ubiquity	Security	Settlement*
Mobile payment	Very High	High	High	Instantaneous
Internet banking	Low	Low	High	Instantaneous
Credit card	High	Medium	High	–
Cash	High	Medium	Low	Instantaneous
Checks	High	Medium	High	1 day

** Settlement indicates how quickly a person can withdraw money that is transferred by another person.*

Mobile banking service was introduced in Korea at the end of 1999. Initially, adoption of the service was not great, as shown in Table 2. However, with the launch of a new enhanced mobile banking service based on the IC-chip method in the third quarter of 2003, mobile banking service use began to skyrocket.[5] Prior to this, Korea's mobile banking service was based on the Web browser method, which was virtually the same as the Internet banking method. It was clearly inconvenient for users to both surf the Web and type in account information on a tiny display using miniature buttons. Hence, use of the mobile banking service market was not widespread until the MNOs invented new technologies to enhance the convenience and speed of use and reduced the data input process. IC-chip mobile banking reduced the data input process to between five

Figure 3. The mobile payment process based on mobile banking

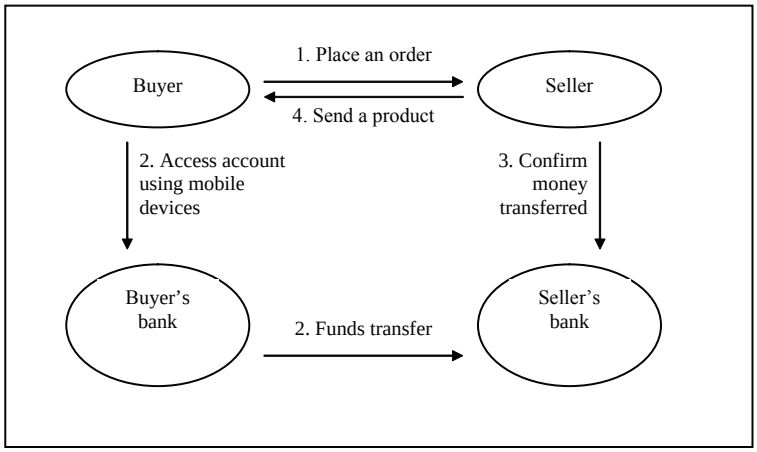


Table 2. Mobile banking use in Korea, 2000-2004 (Note: Transaction numbers are in thousands and the figures in parentheses are the percentage growth rates)

	Dec. 2000	Dec. 2001	Dec. 2002	Dec. 2003	Dec. 2004
Balance checking	200 (–)	692 (246.0)	1,081 (56.2)	2,173 (101.0)	5,013 (130.7)
Funds transfer	2 (–)	18 (800.0)	14 (–22.2)	387 (2,664.3)	1,269 (227.9)
Total	202	710	1,095	2,560	6,281
Ratio*	0.6	0.6	0.6	1.0	1.6

** As a ratio to Internet banking*

and six steps, down from the nine to 16 steps required under Web browsers. In addition, the need for mobile banking in Korea increased by 2003, as electronic commerce expanded continuously after 2000. The new technology and the growing need for electronic transactions brought about a dramatic increase in the use of mobile banking from 2003, as shown in Table 2.

The second type of mobile payment is mobile prepaid electronic cash. The mobile prepaid electronic cash service that Korean MNOs currently offer is a digital substitute for bank notes and coins, with which people can transfer funds and pay for small transactions. In order to use the mobile prepaid electronic cash service, a mobile phone user transfers a certain amount of money from his/her bank account to his/her virtual account opened at an MNO, which is linked to his/her mobile phone number. After storing a certain amount of money in his/her virtual account, users can initiate transactions using mobile devices. Simply by using a mobile phone number, a user can send electronic cash to others, without being required to provide any further information. However, transferring money from a bank account to a virtual account then means users' purchasing electronic cash by paying cash into their bank accounts and MNOs' issuing electronic cash.[6] The service is not popular in Korea because of a fraud that occurred in April 2004.[7]

The third type of mobile payment is a mobile phone bill service. The mobile phone bill service was developed to enable those with mobile handsets to pay for small retail transactions in Internet shopping malls. Online shoppers who chose to pay using the phone bill service are asked to type in their mobile phone number and a certification key when they place an order on the Web. The shoppers type their mobile phone numbers on a Web page and wait a few seconds to receive a certification key sent through a short message service (SMS) by the membership MNOs. The transaction process finishes when online shoppers type in the certification key. They usually receive the commodity after a few days, or, in the case of a digital commodity or online service like virus checking/fixing, receive it immediately. The transaction record is sent to the MNOs, and the price of the service or product is included in the monthly mobile phone bill. After online shoppers pay for the service or product, the MNOs remit the collected funds to the Internet shopping mall. In other words, MNOs simply play the role of rate collectors on behalf of the online shopping malls. SK Telecom, the largest MNO in Korea in terms of its customer base, reported that about seven million users were using the mobile phone bill service in 2003, and that value of transactions completed through the mobile phone bill service amounted to some $US330 million. One merit of using the mobile phone bill service is that online shoppers do not need to provide personal information to shopping malls (Financial Supervisory Service, 2002).

A Brief Literature Survey

Research papers on mobile payments and banking are scant, even though the number of papers on mobile payments and banking are increasing. Major concerns of existing papers on mobile payments have been mobile payment systems, the effects of mobile payments on the banking business and industry, mobile banking consumer behavior, security aspects of mobile payments and financial regulation.

Mobile payments are new services enabled by mobile communications technology, so it is not difficult to find previous literature on mobile payment systems. Examples are Lin, et al. (2006), Kreyer, Pousttchi, and Turowski (2003), Herzberg (2003), Mobile Payment Forum (2002) and Kuttner and McAndrews (2001). These papers summarize and introduce the mobile payment processes and services succinctly.

A group of papers explore the impacts of mobile communications technology on banking business and industry. Kumar and van Hillegersberg (2004)

discuss briefly how deregulation, globalization and new communication technology change the shape of traditional financial markets. Warwick (2004) and Orr (2006) describe the evolution of our society toward a cashless society. Mallat, Rossi, and Tuunainen (2004) and Kumar and van Hillegersberg (2004) introduce various mobile banking services. Shin and Lee (2005) introduce an electronic payment platform, MONETA, developed by SK telecom in order to provide mobile payment services.

The papers on mobile banking consumer behavior investigate the determinants of mobile banking usage. Luarn and Lin (2005) draw on an extended technology acceptance model (TAM) to understand mobile banking users' behavior. Wang, Lin, and Luarn (2006) apply the extended TAM model to a set of mobile services. Laukkanen and Lauronen (2005) study the factors affecting consumer value creation in mobile banking services by utilizing a qualitative interview method. Suoranta and Mattila (2004) study the diffusion pattern of mobile banking services in Finland. These papers intend to help financial institutions develop marketing strategies.

Claessens et al. (2002) discusses the security issues of electronic banking systems in terms of engineering perspective. Herzberg (2003) argues that electronic payments with mobile devices are securer than other wired electronic payments, and Mallat, Rossi, and Tuunainen (2004) also point that security is one of characteristics that make people prefer mobile banking to other electronic banking services.

There is not much literature dealing with regulation issues caused by mobile payment service. Penny (2001) discusses the opportunities and threats for UK financial services providers and payments disintermediation issues. Warwick (2004) suggests the adoption of a secure government-operated electronic cash that can replace traditional fiat money. This chapter intends to explore and discuss new policy issues caused by mobile payments in the Korean financial market.

THE MAIN ISSUES CAUSED BY MOBILE PAYMENT SYSTEMS

As discussed, mobile payments are more convenient than other payment instruments, especially in terms of mobility and ubiquity of use, and they are as secure as existing electronic payments, including credit cards and Internet banking. If people choose to use mobile payments when other payment instruments are available, this means that mobile payments are more efficient. In fact, even if people use mobile payments because they are the only available means of payment, again this indicates that mobile payments are more efficient than other forms of payment.[8] However, despite their obvious convenience and efficiency, use of mobile prepaid electronic cash and mobile phone bill services has not spread rapidly in Korea, except for mobile banking. This section attempts to shed some light on the lack of cooperation in the mobile payment industry supply chain, and the regulatory and monetary problems caused by mobile payment systems, all of which have contributed to the lukewarm reception of mobile prepaid electronic cash and mobile phone bill services in Korea.

Lack of Cooperation in the Mobile Payment Industry Supply Chain

Mobile payment services came into existence as a result of the collaboration between banks and MNOs. In the case of mobile banking and mobile prepaid electronic cash, banks are the service providers and MNOs take the role of providing a ubiquitous wireless service delivery channel. As shown in Figure 2, the major role of MNOs is to set up the wireless connection between the mobile server at banks and the mobile handsets. If we focus on this complementary relationship between banks and MNOs, there is no apparent reason for the two parties to compete with each other. However, in the process of launching mobile payment services in the form of mobile banking

and mobile prepaid electronic cash in Korea, banks and MNOs have engaged in competitive, rather than cooperative, behavior. Banks view mobile payments as an encroachment by MNOs on the conventional financial market, although they recognize that the devel opment of more efficient and convenient service channels is inevitable. In addition, banks often overstate the possibility that mobile payments can weaken the stability of the payment and settlement system and gradually reduce central bank authority.

This ambivalent attitude of the banks resulted in an inefficient mobile banking service and the near death of prepaid electronic cash. As shown in Table 2, the use of the mobile banking service increased exponentially after the mobile banking service evolved from the Web browser method to the IC-chip method. Therefore, mobile banking in Korea appears to be a success. However, there is still much room for improvement. Currently, one serious problem is that each bank individually issues an IC-chip for mobile banking. Therefore, if users have two accounts at two banks, they have to carry two chips and exchange one with another to access a different account. This is a very inefficient outcome. Technically, MNOs can mount multiple account information on an IC-chip, which would allow users to access multiple accounts one at a time without replacing their chip. However, despite the feasibility of such a solution, it has not been implemented in Korea because banks have the power to issue chips. This awkward mobile banking system was the result of a tug-of-war between banks and MNOs without government intervention.To date, banks in Korea have succeeded in maintaining MNOs as simple network service providers.

Potential Issues Related to the Overall Financial System

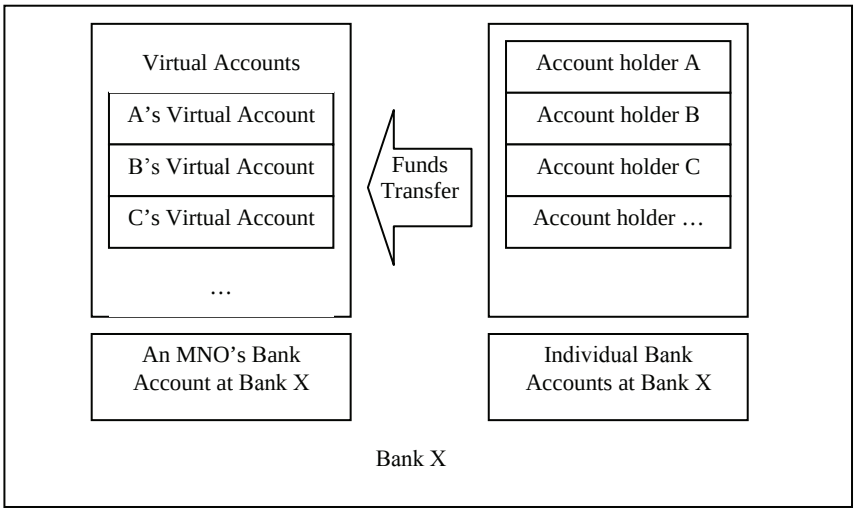
There is potential risk involved in the fact that MNOs may undertake some financial institution functions without direct central bank surveillance, which in turn may erode the stability of the financial system. This subsection addresses the financial issues related to mobile prepaid electronic cash and mobile phone bill services that are likely to alter the nature of MNOs.

To understand clearly the impact of the mobile prepaid electronic cash service on the financial system, it is necessary to appreciate in detail how the service works. In order to use a mobile prepaid electronic cash service, users first need to obtain membership of the service. Upon joining, they have a virtual account to which they can transfer funds from their conventional bank accounts. The virtual account is created in the MNO firm's banking account and identified by the user's phone number. In other words, as illustrated in Figure 4, virtual accounts are individual users' accounts within an MNO's own bank account, with phone numbers as identifiers. Upon transfer from a bank account, the money stored in the virtual account becomes electronic cash that users can, by simply using the phone number, remit to friends or purchase products or services.

The virtual accounts created in an MNO's bank account can be understood as depository accounts leased from the bank because the virtual account is, in fact, identical to a conventional bank account. Transferring money from bank accounts to virtual accounts is comparable to depositing money in virtual accounts. In other words, the electronic cash stored in virtual accounts is not money earned by MNOs from their business, but money deposited by users. The MNOs are then de facto depository institutions without the necessary government license for financial business. However, as long as the MNOs do not create credits based on the balance of virtual accounts, and the balance is treated separately in the accounting process, prepaid electronic cash service should not disrupt financial market stability.

After a certain amount of cash is stored in the virtual account, it takes time for account holders to drain their electronic balances. Therefore, on average, at least some balance will remain

Figure 4. The relationship between virtual and conventional bank accounts



in the MNOs' bank accounts during the year. A MNO that has a large number of mobile prepaid electronic cash service users will have a large balance upon which it earns financial gains, at least at prevailing market interest rates, without paying interest to depositors (the users). If MNOs make loans using the balance of virtual accounts, they can earn higher financial gains than banks because they are not required to hold some share of balances as noninterest-bearing reserves. Thus, they have an incentive to make loans using the virtual accounts. If they cannot make loans, they can instead use the funds for investment in their original business without cost (interest). Therefore, the mobile prepaid electronic cash service can disturb the long-standing order of the financial market and make it more difficult for a central bank to estimate the amount of money in the economy. Only in the case where MNOs do nothing with the balance—that is, they keep the cash in the virtual accounts intact—will the mobile prepaid electronic cash service not upset the financial market. However, this is implausible without government intervention.

In addition, the mobile phone bill service can potentially work as a disturbing factor in financial markets. If MNOs simply collect rates for products or services that people purchase and then deliver the collected rates to online shopping malls, they function as rate collectors on behalf of the online shopping malls. In this case, the mobile phone bill service does not create any monetary issues. However, MNOs can provide loans to online shopping malls based on transaction records before the rates for products or services are collected and they can charge fees and interest on these loans. MNOs also have their customers' transaction records so they can evaluate default risk. Therefore, MNOs have sufficient motivation to extend themselves to nonbank financial business.

GOVERNMENT ACTIONS: EVALUATION AND SUGGESTION

To date, the Korean government has taken a hands-off approach to mobile payment services. Mobile payment services are on the verge of burgeoning and the technology for their use is still developing, rather than becoming more stable and standardized. In addition, the size of mobile payment transactions is still very small compared to total financial transactions. It therefore seems appropriate for the government simply to monitor mobile payment market developments, without any regulation.[9]

However, there is much room for the government to facilitate the development of efficient payment instruments by acting as a constructive rule maker in the financial payment market. First, the government needs to monitor and become involved as a market participant in the evolution of the mobile payment system in an attempt to maximize the public interest. This is because new industries, such as online gaming businesses, cannot prosper without an efficient, convenient and secure electronic payment system. An example of inefficiency is the Korean mobile banking service. As shown in Table 2, the demand for the mobile banking service has increased exponentially, but the system is still inefficient because users with multiple accounts at different banks have to change IC-chips. The inefficient mobile banking system is the result of market competition between banks and MNOs. If MNOs had won, a more efficient mobile banking system would have prevailed in Korea, by allowing users access to multiple accounts in different banks with the one chip. With hindsight, the Korean government should have become more actively involved in the standard-setting process for mobile payment systems, to facilitate the evolution of the mobile banking service and the development of the mobile payment market.

Second, the mobile prepaid electronic cash service that utilizes the virtual account system transforms the MNOs into de facto nonbank financial institutions. As the MNOs are not regulated financial institutions, they can compete with banks and other financial intermediaries from a more advantageous position. Therefore, as long as the MNOs act as financial institutions, the government should ensure they compete with conventional financial institutions on a level playing field by imposing the same regulations on MNOs as applied to other financial institutions. The regulations governing MNOs should then include setting a level of non interest-bearing reserves, and imposing rules as to whether the deposit insurance system should apply to the balances in the virtual accounts, protection of customer privacy, and so on.

The government can require MNOs to hold the total balances in virtual accounts as non interest-bearing reserves because MNOs do not pay interest for the funds saved in virtual accounts. The government does then not need to worry about the liquidity and credit risk of the MNOs and extend a deposit insurance system to the balance of virtual accounts. Clearly, however, this inactive policy will result in the inefficient use of financial resources because the balance of virtual accounts will be submerged in accounts unavailable for productive purposes. In addition, such a policy will hinder the advent of new efficient payment media, and in turn, reduce the competition in the banking industry, and deter crossbreeding between banks and MNOs. If the government wants to facilitate crossbreeding between banks and MNOs to accelerate innovation in payment services, it should foster competition in the electronic payment service market by allowing MNOs to provide banking services, and therefore regulate MNOs as banks.

Third, the government should adopt an accounting separation rule for the MNOs because they engage in businesses that belong to different industries. Otherwise, the MNOs can enhance their position in certain markets by cross-subsidizing one business with the profits from another. In

addition, a dominant market player in the mobile phone industry can extend its market power over the finance industry by utilizing its customer base in the mobile phone market. This possibility is especially problematic in Korea because it has been a long-standing tradition that industrial capitalists are prohibited from extending their businesses into banking. The MNO providing the mobile prepaid electronic cash service can be considered as either a bank or a nonbank financial institution, depending on how the bank is defined. If the bank is defined as a financial intermediary taking deposits and lending loans, the MNO is not a bank as long as it does not make loans by employing virtual account balances. However, if the ban is defined as a financial institution simply taking deposits, the MNO can be considered a bank.[10] If the government adopts the latter interpretation, the MNO cannot offer the mobile prepaid electronic cash service, which means that the government stifle the evolution of the payment system with outdated restrictive regulation. At the center of the ongoing payment system evolution are computer and telecommunication technologies, with which banks usually are not very familiar. Therefore, the government should allow the MNOs to lead the evolution process of payment systems and abolish or lessen the outdated regulation that separates financial capital and industrial capital. and industrial capital.

Fourth, even though it is too early to state the effects of mobile payments on monetary policy, we can see hat the definition of money and the central bank's monetary control policy are to be redefined in the future. This can be seen from the fact that mobile payments can affect the money creation process outside the central bank's surveillance. For example, MNOs making loans by capitalizing on virtual accounts will eventually undermine the effectiveness of conventional monetary aggregates. While for the time being, the mobile prepaid electronic cash will not have a discernible impact on monetary aggregates and monetary policy because of its small size, the central bank should still monitor the evolution process of electronic cash market and stand ready to remove conventional regulations that block the development of more efficient and convenient payment instruments.

CONCLUSION

Historically, as new technology has developed, new services have appeared in markets. When a new service takes root firmly and successfully prevails in markets, it can develop into a new industry or displace an existing industry through absorption or transformation. In the last decade, computer and telecommunication technologies have created the mobile communication industry and have recently begun to dismantle other Industries, including the banking industry. The MNOs, by providing wireless ubiquitous communication service, can merge and collaborate with conventional firms in existing industries to create better products or services. Some MNOs will definitely attempt to evolve into non bank financial institutions by refurbishing existing services with their superior technology. By nature, innovation threatens the conventional order of market, so incumbent firms and regulatory authorities are likely to be h stile to such change. However, history has repeatedly demonstrated that innovation cannot and should not be annulled by self-protecting incumbents. The central bank needs to facilitate the development of efficient payment instruments by acting constructively, rather than conservatively, as a rule maker in the financial payment market.

The convergence of the telecommunication and financial sectors creates a dynamic confluence, with many potential research topics being evident. Some future research topics are as follows. First, one can simply trace the evolutionary process of the industry in terms of an historical perspective because mobile payments are an emerging industry. Second, the impact of mobile payments

on conventional financial orders are a critical concern for central banks, so research on the effect of mobile payments on monetary systems will be of great importance. Third, the adoption of mobile payments is accelerating because they are more efficient, convenient, and secure than traditional payments. Mobile payments help reduce both customers' and banks' transaction costs, and improve customers' convenience. This is a logical presumption, not confirmed by existing empirical studies. Research on the impact of mobile telecommunication technology on the transaction costs of banking industry would therefore be highly valuable. However, it may not be possible to perform rigorous empirical study in the near future because of the scarcity of relevant data.

ACKNOWLEDGMENTS

This research was financially supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program of the IITA (Institute of Information Technology Assessment).

REFERENCES

Bank of Korea. (2005). *Press release on the use of Internet banking in Korea.* (Quarterly press releases from 2002-2005).

Bank of Korea. (2004a). Payment and Settlement System of Korea. BOK, Seoul: Bank of Korea.

Bank of Korea. (2004b). *Press release on the use of mobile banking in Korea.* (Payment Information 2004-6).

Bank of Korea. (2004c). *A comprehensive survey on electronic banking.* Seoul: Bank of Korea.

Claessens, J., Dem, V., de Cock, D., Preneel, B., & Vandewalle, J. (2002). On the security of today's online electronic banking systems. *Computers & Security, 21*(3), 253-265.

Financial Supervisory Service. (2002). *Supervisory information on electronic finance.* Seoul: Bank of Korea.

Herzberg, A. (2003). Payments and banking with mobile personal devices. *Communications of the ACM, 46*(5), 53-58.

Kreyer, N., Pousttchi, K., & Turowski, K. (2003). Mobile payment procedures. *e-Service Journal, 2*(3), 7-22.

Kumar, K. & van Hillegersberg, J. (2004). New architectures for financial services. *Communications of the ACM, 47*(5), 27-30.

Kuttner, K. N., & McAndrews, J. J. (2001). Personal on-line payments. Federal Reserve Board of New York Economic Policy Review, 7, 35-50.

Laukkanen, T., & Lauronen, J. (2005). Consumer value creation in mobile banking services. *International Journal of Mobile Communications, 3*(4), 1-11.

Lin, P., Lin, Y. B., Gan, C. H., & Jeng, J. Y. (2006). Credit allocation for UMTS prepaid service. *IEEE Transactions on Vehicular Technology, 55*(1), 306-316.

Luarn, P., & Lin, H. H. (2005). Toward an understanding of the behavioral intention to use mobile banking. *Computers in Human Behavior, 21*(6), 873-891.

Mallat, N., Rossi, M., & Tuunainen, V. K. (2004). Mobile banking services. *Communications of the ACM, 47*(5), 42-46.

Mishkin, F. S. (1997). *The economics of money, banking, and financial markets* (5th ed.). Reading, MA: Addison-Wesley.

Mobile Payment Forum (2002). *Mobile payment forum white paper: enabling secure, interoperable, and user-friendly mobile payments.* Retrieved November 21, 2005, from http://www.mobilepaymentforum.org/pdfs/mpf_whitepaper.pdf

Orr, B. (2006). Cashless society, ahoy! *ABA Banking Journal, 98*(3), 44-45.

Penny, J. (2001). The payments revolution: the growth of person-to-person and 'Generation Y' payments services. *Journal of Financial Services Marketing, 6*(2), 190-201.

Shin, B., & Lee, H. G. (2005). Ubiquitous computing-driven business models: A case of SK Telecom's financial services. *Electronic Markets, 15*(1), 4-12.

Suoranta, M., & Mattila, M. (2004). Mobile banking and consumer behavior: New insights into the diffusion pattern. *Journal of Financial Services Marketing, 8*(4), 354-366.

Wang, Y. S., Lin, H. H., & Luarn, K. N. (2006). Predicting consumer intention to use mobile service. *Information Systems Journal, 16*(2), 157-179.

Warwick, D. R. (2004). *Toward a cashless society. Futurist, 38*(4), 38-42.

ENDNOTES

1 The Mobile Payment Forum (MPF) defines a mobile payment "as the process of two parties exchanging financial value using a mobile device in return for goods or services" (MPF, 2002, p. 10).

2 Internet payments refer to monetary value transfers by people using the Internet. A good example is Internet banking.

3 In Korea, credit card companies typically charge a fee of 2.25% of the transaction value for credit and debit cards (Bank of Korea (BOK), 2004c).

4 Korea's central bank, BOK, launched an E-Commerce Payment Gateway System in 2000, which is the settlement network for interbank obligations arising from electronic business-to-consumer transactions. This system has been utilized for electronic commerce over the Internet.

5 An IC-chip contains an account number, ID number, and a security program (BOK, 2004b).

6 SK Telecom, a dominant mobile phone service provider in Korea, offers a mobile prepaid electronic cash service called MONETA cash. The maximum amount of money a user can store in his/her virtual account per day is about $US500.

7 In April 2004, 36 million Won (about $US35,000) was fraudulently transferred from bank accounts to virtual accounts and then withdrawn.

8 If people want to purchase a product but do not have the means to pay, they have to spend time finding a store that accepts the payment instrument they have or to obtain a payment instrument circulating in the market. This is quite similar to the situation where a person is traveling in a country without the currency circulated in that country.

9 As at December 2004, some 29.3% of total banking transactions were conducted through Internet banking, as against 1.6% of mobile banking transactions (BOK, 2004a).

10 Kuttner and McAndrews (2001) argue that 'bank' can be defined in two ways (p. 42).

This work was previously published in Strategies and Policies in Digital Convergence, edited by S. Park, pp. 88-101, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.32
Payment Mechanism of Mobile Agent-Based Restaurant Ordering System

Jon T. S. Quah
Nanyang Technological University, Singapore

Winnie C. H. Leow
Singapore Polytechnic, Singapore

Chee Chye Ong
Nanyang Technological University, Singapore

INTRODUCTION

The Internet, especially the World Wide Web, is moving from a free, academic domain to a profitable commercial world. This underscores the importance of a digitally secure means of electronic payment for an electronic commerce application. The payment is usually an important part of an electronic commerce transaction, and it deals with the transfer of trust, either as cryptographically signed promises, or as digital cash, between the customer, the merchant, and the payment service provider.

Due to the explosive growth of e-commerce transactions, many electronic modes of payment are devised to address a diverse set of Internet user requirements (Guida, Stahl, Bunt et al., 2004; Tsiakis & Sthephanides, 2005; Garfinkel, 2003; Usher, 2003; Polk, Hastings, & Malpani, 2003; Evans & Yen, 2005; Marchesini, Smith, & Zhao, 2005; Lancaster, Yen, & Huang, 2003; Lekkas, 2003; Medvinsky & Neuman, 1995; Schoenmakers, 1997; Levi & Koc, 2001; Mahony, Peirce, & Tewari, 2001; DigiCash Press, 1994; Neuman & Tso, 1994; Vivtek, 2000).

The background of this article is that we have developed a mobile agent-based restaurant reservation and ordering system whereby users are able to search for restaurants that fulfill a list of user-entered parameters (e.g., type of cuisines, ambiance, specialties such as steaks, etc.) (Quah & Leow, 2003). The system is built on the IBM

Aglet mobile agent platform. (A mobile agent is a small executable code/program that can migrate itself to remote hosts and execute predefined instructions—e.g., information retrieval, and return the processed information to its originating host system) (Lanage & Oshima, 1998). Due to the uniqueness of our system, we find the existing e-commerce payment methods inadequate to fit our system's need. As such, we studied several existing methods and adapted one into our system operation structure. The use of mobile agent to implement the payment system adds robustness and scalability to the system.

DESCRIPTION OF THE ELECTRONIC PAYMENT SYSTEM

To support electronic commerce, various Internet payment protocols have been proposed and adopted by a variety of organizations. In fact, the existence of different payment mechanisms are justified because there are different needs to be satisfied in terms of:

- Cryptographic needs (strong, symmetric, exportable, importable, etc.)
- Latency of the transaction (micropayment must be very fast)
- Minimal and maximal amount for the transaction itself
- Minimal and maximal amount for the cost of the transaction
- Repudiation, notarization needs
- Involvement of financial institution (i.e., online vs. off-line)

Some of the above requirements may call for contradictory system requirements, and as such, trade-offs have to be made. In a nutshell, an electronic payment system should meet the following requirements:

1. Sufficient security means based on the amount of money transferred in a transaction.
2. Similar running scenario as the traditional business whenever possible to ease the doubts of the public and encourage them to participate.
3. Minimum changes on the current financial system to avoid tremendous costs when electronic commerce is introduced.

The participants of an electronic commerce transaction must be able to exchange trade and payment information over a network. The implementation addresses the problem of online payment by credit card in which anyone with knowledge of the customer's credit card number can create an order for payment. It also tries to eliminate the requirement of a Certificate Authority (CA), and consequently a CA-based Public Key Infrastructure (PKI), in order to verify a public key-based digital signature.

Characteristics of the Mobile Agent-Based Restaurant Order Payment System

Secure Socket Layer- (SSL) (Rainbow Technologies, 2001; Freier, Karlton, & Kocher, 1996; Albrecht, 1998) based protocols used in credit card payment are convenient but have some authentication and non-repudiation problems. Secure Electronic Transaction standard (SET) (MasterCard, 1997) and other payment-card-based protocols, which require either intermediary agents or CA-based PKI, are secure, but not so convenient, particularly for financial institutions (FIs). The mechanism of our implementation tries to find a middle ground in the "security vs. convenience" trade-off.

In our payment mechanism implementation, both the customer and the merchant need to be registered off-line with a network payment service

(or trusted party) with their credit card data and given a unique persona. This persona acts as a mapping between an identified user and that user's public key and credit card information stored in the trusted third-party system. The trusted third party then acts as an intermediary in collaborating with the customer's issuing bank and the merchant's acquiring bank in the settlement of the credit card bill.

The basic idea behind the mobile agent-based restaurant order payment system is to avoid the necessity of consumer certificates. It also provides a remedy to the inability of the traditional credit card payment system in authenticating the customer's identity. The trusted third party makes use of the stored public key to authenticate the identity of the customer and merchant. Only signed payment request from the customer and an order endorsed by the merchant can effect the payment. The merchant verifies the digital signature of the consumer in most of the electronic payment protocols. In our payment scheme, the trusted third party is the authority who verifies the consumer's digital signature.

The payment system serves like a credit card system without the online authorization with the issuing bank. This payment method does not require changing the existing credit card settlement infrastructure tremendously to adopt this scheme. Another important characteristic of the system is that messages transmitted among the consumer, merchant, and trusted third party are not encrypted. Justifications for this challenging characteristic are as follows:

1. The persona is not valid unless there is an accompanying digital signature issued by its owner. Any third party cannot take advantage of knowing the persona, since it cannot produce a digital signature. Thus, personas need not be encrypted.
2. The strong public private key authentication is sufficient to prevent the majority of consumer and merchant frauds. Using the persona concept and strong authentication make encryption a luxury in this payment system.

Mobile Agent-Based Payment Protocol

A payment protocol based on mobile agents is devised to structure the interactions and information exchange among agents to complete a payment transaction. The payment structure integrates the use of public key cryptography into the mobile agent framework for enhancing security in electronic payment.

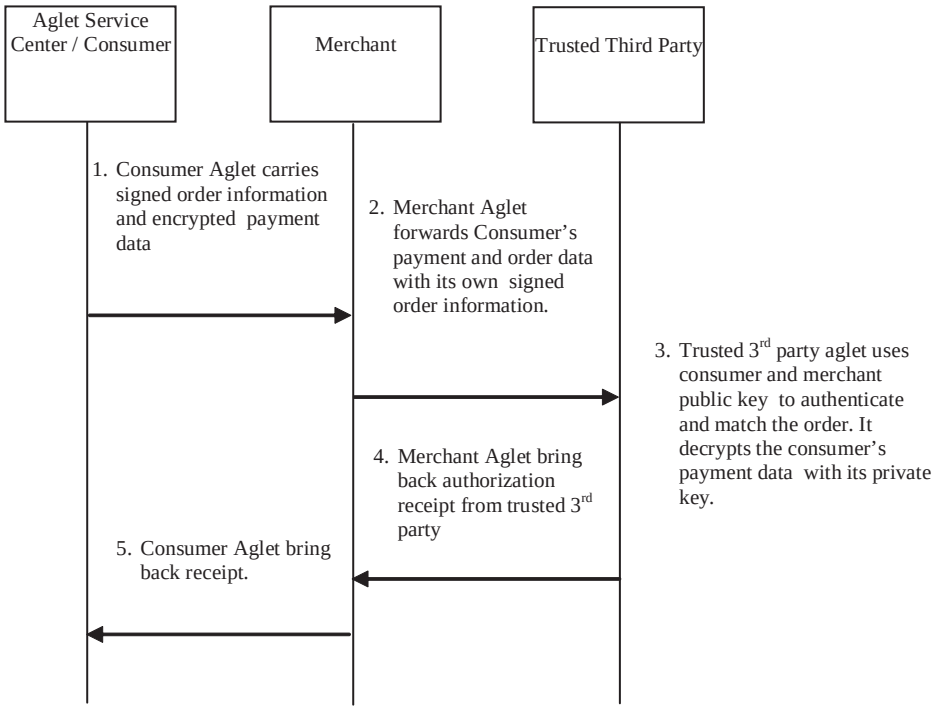
The payment protocol consists of the following steps, as shown in Figure 1:

- CC—Consumer
- TTP—Trusted third party
- Mer—Merchant
- SigCC—Consumer's signature of payment data and order information
- SigMer—Merchant's signature of order information

1. Having decided the food to order, the consumer clicks on the "Pay" button at the Aglet Service center. The ASC launches the consumer aglet that carries the payment request which most importantly contains the merchant's persona and the payment amount.
2. At the merchant site, the consumer aglet passes the payment request to the merchant aglet. The merchant aglet carries the consumer's payment request with its own signature of the order and dispatches to the TTP site.

   Consumer payment request:
   Consumer persona, [payment data,
   H(Order)], SigCC

Figure 1. Payment protocol

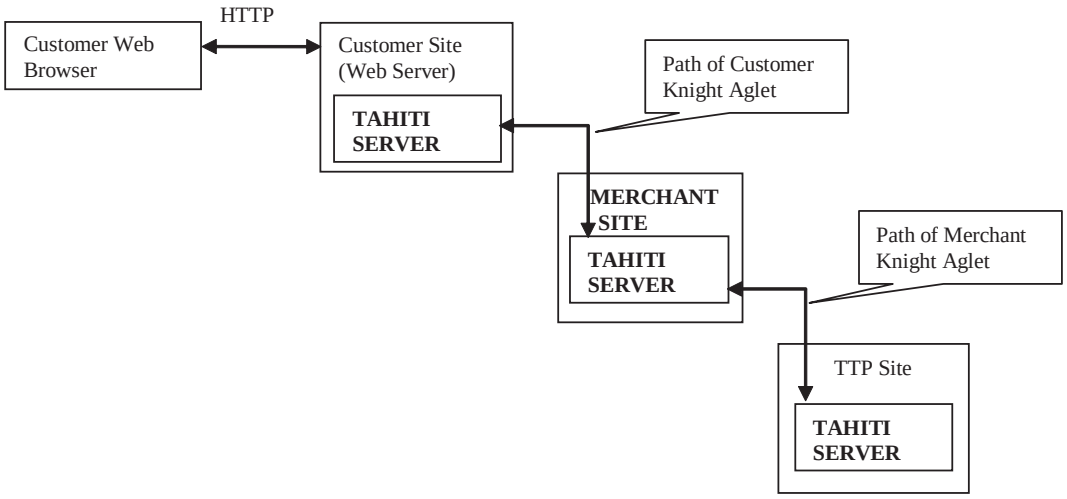


Merchant signature of the order:
Merchant persona, [H(Order)], SigMer

3. At the TTP site, the merchant aglet passes the consumer's payment request and merchant's signature of the order to the TTP aglet. The TTP aglet verifies the consumer's signature on the payment request and the merchant's signature on the order. The TTP aglet verifies that both the customer and merchant agree on the order and authorizes the payment. The TTP aglet returns a receipt to the merchant aglet.
4. The merchant aglet returns to the merchant site with the receipt.
5. The merchant aglet forwards the receipt to the consumer aglet waiting at the merchant site. The consumer aglet returns to the consumer site with the receipt.

Figure 2 shows the setup of the payment system prototype in the aglet environment. Our system implements a master-slave mechanism of agents. The master agent is a stationary agent residing in the Tahiti Server. It spawns off one or more slave agents to accomplish a task. The slave agents are mobile agents, as they need to execute on a remote host. The mobile slave agent keeps information about its origin, the destination site on which it is going to run on, and the aglet information of the master

Figure 2. System setup



who created it, as they need to inform their master on their return.

IMPACT AND USEFULNESS OF THE PAYMENT MECHANISM

Features of the payment system that is used in our mobile agent-based restaurant system include:

1. Anonymity, meaning the merchant would not know the customer's identity. Only the trusted third party knows the identities to process the payment. But the trusted third party would not know the order detail. He can only verify that both the merchant and customer agreed with the order.
2. No credit card number is required to be sent over the network by using the trusted third-party approach. Instead, the customer's persona, sent with the signed payment request, acts as an identifier that is only recognized by the specific trusted third party. This unique persona will map to the credit card information preregistered by the customer with the trusted third party.
3. Non-repudiation using public private key authentication. In the proposed payment system, it can be later proven that the customer had agreed to pay for the food and the merchant had endorsed the order, as the trusted third party kept record of the signed payment request from the customer and signature of the order from the merchant. This addresses the problem of the credit card and token-based payment system where there is no proof that the cus4. Off-line settlement via third-party avoid transaction latency resulted from online authorization via the banking infrastructure as in a typical credit card transaction. The transaction and its associated authorization are stored in a "batch,"

along with the rest of the transactions for the day. The trusted third party will submit the batch request to the processing network to process the transactions for which authorizations have been recorded.

5. Auditability, meaning the trusted third party has records of transaction signed by the customer and the merchant.
6. Widespread use and acceptance of credit card payment is a proposed scheme which can leverage on the existing credit card payment infrastructure. The credit card billing procedure that current customers are accustomed to can be utilized. As for merchants, there is no overhead since the trusted third party acts as a bridge for payment between the acquirers and merchants.
7. Achieve interoperability and reduce message exchange across network using Java mobile agent technology.

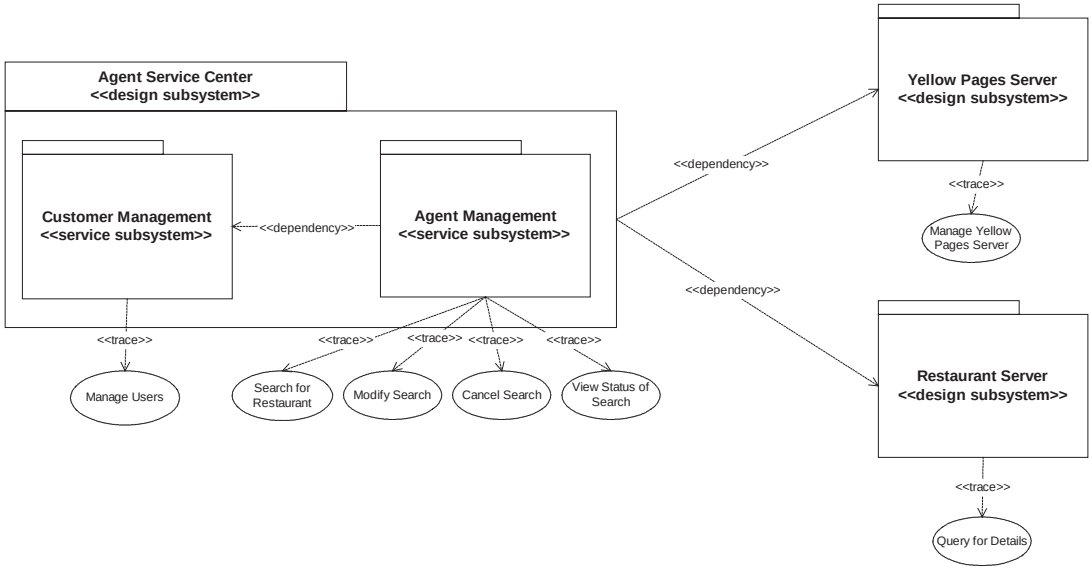
A potential weakness of the payment mechanism is the bottleneck at the trusted third party during peak traffic volume because all payment requests are sent to the trusted third party for signature verification. This, however, is a common problem for most online payment systems. One possible solution is to have a cluster of high-power servers at the verification site.

RESULTS

The mobile agent framework consists predominantly of Client, Agent Service Center (ASC), and the Restaurant Server Platform (see Figure 3).

The hotel guest will be able to access the online restaurant recommendation system through a Web browser in his hotel room. The search request(s) made by the guest will be registered with the ASC, which will then process each request and generate a list of online restaurants sites that will likely provide the pertinent information (i.e., food dishes) requested by the guest. The Yellow Page (YP) server that provides a database of such online sites helps to facilitate the compilation and gen-

Figure 3. Architecture framework for online restaurant system



eration of a list of such relevant online sites. This list of online sites will constitute the itinerary list that the mobile agent (MA) will have to visit. As mentioned, another main functionality of the ASC server would also be to generate a mobile agent that will begin traveling to the online restaurant sites to gather data on behalf of the user.

The mobile agent will abide by the generated itinerary list as it travels from one online restaurant server to another in order to complete its search for food and restaurant information. Upon arrival at each online site, the mobile agent will enquire the restaurant server to search for the food based on the user's search requirements. In addition, the restaurant server can retrieve its promotional information and push it to the ASC server, which in turn displays the promotional information to the users.

When the returned information is consolidated, the user (the customer) will be able to browse through the information and decide on his purchases.

CONCLUSION

A mobile agent system has been developed to perform restaurant food searches for customers from restaurants with Web presence. Details on promotional offers are also 'pushed' to the user. The security aspect of the system and integrity of the data are ensured by means of cryptography and digital signature schemes. The system provides a user-friendly environment for easy usage.

There are areas that we are currently exploring to further improve the system. Firstly, we are extending the aglets server functions to handle advance reservation requests from the user; this will require further mobile agent activity such that the mobile agent will interact with the restaurant reservation system to place a booking. This will also involve expanding payment options that must be provided for the user to pay for his meals. To push the technology further, we are exploring the possibility of allowing autonomous negotiation by the mobile agent. Basically, mobile agents representing the users and the restaurant servers will meet at some cyberspace negotiation room to transact their requests for their respective hosts that they are representing.

REFERENCES

Albrecht, C. (1998). An analysis of SSL and SET for electronic commerce. *Capstone Proceedings.*

DigiCash. (1994). *World's first electronic cash payment over computer networks.* Press Release. Retrieved March 15, 2003, from http://ntrg.cs.tcd.ie/mepeirce/Project/Press/ec_press.html

Evans, D. M., & Yen, D. C. (2005). Private key infrastructure: Balancing computer transmission privacy with changing technology and security demands. *Computer Standards and Interfaces, 27*(4), 423-437.

Freier, A. O., Karlton, P., & Kocher, P. C. (1996). *The SSL protocol version 3.* Retrieved from http://home.netscape.com/eng/ssl3

Garfinkel, S. L. (2003). Email-based identification and authentication: An alternative to PKI? *IEEE Security and Privacy, 1*(6), 20-26.

Guida, R., Stahl, R., Bunt, T., Serest, G., & Mooroones, J. (2004). Deploying and using public key technology: Lessons learned in real life. *IEEE Security and Privacy, 2*(4), 67-71.

Lancaster, S., Yen, D. C., & Huang, S. M. (2003). Public key infrastructure: A micro and macro analysis. *Computer Standards and Interfaces, 25*(5), 437-446.

Lange, D. B., & Oshima, M. (1998). *Programming and deploying Java mobile agents with aglets.* Boston: Addison-Wesley.

Lekkas, D. (2003). Establishing and managing trust within the public key infrastructure. *Computer Communications, 26*(16), 1815-1825.

Levi, A., & Koc, C.K. (2001). CONSEPP: CONvenient and Secure Electronic Payment Protocol based on X9.59. *Proceedings of the 17th Annual Computer Security Applications Conference,* (pp. 286-295).

Marchesini, J., Smith, S. W., & Zhao, M. (2005). Keyjacking: The surprising insecurity of client-side SSL. *Computer and Security, 24*(2), 109-123.

MasterCard. (1997). *SET Secure Electronic Transaction specification* (Book 1: Business Description). Purchase, NY: MasterCard Inc.

Medvinsky, G., & Neuman, B. C. (1995). Requirements for network payment: The NetCheque perspective. *Proceedings of IEEE Compcon'95,* (pp. 32-36).

Neuman, B. C., & Tso, T. (1994). Kerberos: An authentication service for computer networks. *IEEE Communications Magazine, 32*(9), 33-38.

O'Mahony, D., Peirce, M., & Tewari, H. (2001). *Electronic payment systems for e-commerce.* Norwood, MA: Artech House.

Polk, W. T., Hastings, N. E., & Malpani, A. (2003). Public key infrastructures that satisfy security goals. *IEEE Internet Computing, 7*(4), 60-67.

Quah, J. T. S., & Leow, C.H. (2003). Mobile agent technology in e-businesses—An implementation example. In S. K. Sharma & J. N. D. Gupta (Eds.), *Managing e-business in the 21st century* (chap. 5, pp. 71-89). Heidelberg, Australia: Heidelberg Press.

Rainbow Technologies. (2001). *The Secure Sockets Layer protocols—Enabling secure Web transaction.* Retrieved from http://www.rainbow.com/library/library.asp

Schoenmakers, B. (1997). *Basic security of the ecashTM payment system.* Retrieved from http://www.win.tue.nl/~berry/papers/cosic.pdf

Tsiakis, T., & Sthephanides, G. (2005). The concept of security and trust in electronic payments. *Computers and Security, 24*(1), 10-15.

Usher, M. (2003). Certificate policies and certification practice statements—A practical approach. *Information Security Technical Report, 8*(3), 14-22.

Vivtek. (2000). *CyberCash.* Retrieved from http://www.vivtek.com/cybercash.html

KEY TERMS

Aglets: Platforms for mobile agents to operate on and on which to perform transactions.

Electronic Commerce: Business transactions over the Internet.

Encryption: Data coding schemes to protect information privacy.

Mobile Agent: Software code that can migrate host to host autonomously to perform operations.

Payment System: A mechanism for enabling payment for Internet transactions.

Trusted Third Party: A trust center that serves as an intermediary for Internet transactions.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 908-913, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 4.33
Structural Effects of Platform Certification on a Complementary Product Market:
The Case of Mobile Applications

Ankur Tarnacha
Pennsylvania State University, USA

Carleen Maitland
Pennsylvania State University, USA

ABSTRACT

This article examines the structural effects of platform certification on the supply of complementary products. Drawing on the exploratory case of mobile application markets, the article highlights the broader market effects of competing platforms and their certifications on a platform-based complementary product market. The case suggests that platform certifications influence market intermediation, entry barriers, and deployment fragmentation. We present these market effects in a conceptual model that can be applied to understand similar complementary product markets. As such, the article contributes to the literature on compatibility standards by emphasizing some of the complementary product market effects of employing certification in enhancing compatibility.

INTRODUCTION

One of the challenges resulting from complex, specialized information technology systems has been maintaining vertical compatibility, typically through compliance to a standard, while ensuring the development of a vibrant market of complementary goods. One tool used to foster compatibility is that of platforms, which are

technology architectures composed of subsystems and interfaces between those subsystems and the external environment (Greenstein, 1998; Meyer & Seliger, 1998; West & Dedrick, 2000). The interfaces provide access to platform subsystem functionality that can be used to design complementary products. Standardized platform interfaces in that sense facilitate vertical compatibility between the product implementing the platform and the complementary product (Schmidt & Werle, 1998). However, without consistent interface implementation, compatibility can suffer (Egyedi & Dahanayake, 2003; Egyedi & Hudson, 2005). As vertical incompatibilities reduce complementary network externalities for the platform (Gandal, 1995; West & Dedrick, 2000), it is in the best interest of platform promulgators (firms or industry alliances that develop, promote, or support a particular platform) to ensure its "correct" implementation and use.

Extant literature has discussed various strategies to ensure compatibility, including standards certification, wherein products are assessed for conformance to a developed standard (see, for instance, Egyedi and van Wendel de Joode, 2003). Research has primarily investigated the strategic implications of compatibility certification for the standard promulgators and organizations directly involved in the implementation (Egyedi, 2001b; Rada, 1996), with little emphasis on how certification influences the market of *complementary products.* As complementary products contribute to the value of the platform and are often more closely aligned with end users, understanding these effects will have broad implications not only for suppliers of technology but for end users as well.

In this article, we explore the structural effects of platform certification (the conformance of a platform's complementary product to the platform-sponsor-defined best practices in platform interface usage). In particular, we seek to answer the question: What are the structural effects of platform certification on a platform's complementary product market? We examine these market effects using an exploratory case (Yin, 1994) of the mobile applications market, where multiple mobile computing platforms competing for dominance offer mobile application certification programs. The data for the case were gathered from 18 open-ended interviews with the top-management of mobile application vendors in the United States in the fall of 2005. These data were supplemented by several on-site interviews with the top-management team and certification program managers at a leading global certification intermediary in the summer of 2006. Further, the case is supported by secondary sources such as trade press articles and news released by various platform promulgators. Our analysis is exploratory in nature, and as such, the goal is to identify relationships in the case of mobile applications market, which can be more systematically investigated in the future.

The article is structured as follows. We first review the extant literature on compatibility standards, platforms, and certification. Subsequently, we present the case of mobile applications markets, wherein we first provide an overview of the market highlighting the prevalence of computing platforms, their implementations, and certifications. We then present some of the observed market effects of platform certification on the mobile applications market. The case is followed by the discussion of these effects, grounding them in the standards and information technology literature. Finally, we conclude with recommendations for future research.

LITERATURE REVIEW

Compatibility Standards and Platforms

While standards exist in various forms (David & Greenstein, 1990; David & Steinmueller,

1994; Tassey, 2000), one of the more common classification schemes distinguishes reference, minimum quality, and "interface" or compatibility standards (David & Greenstein, 1990). Reference and minimum quality standards provide benchmark characteristics that define the quality, performance, or other desirable attributes of the standardized product. Compatibility standards, on the other hand, are identified as a set of technical specifications that provide an interface to develop compatible complementary products. Compatibility standards in that sense provide "vertical compatibility" across two complementary components such as a printer and cartridge or a computer hardware and software (Schmidt & Werle, 1998). Compatibility standards assure that a complementary product can be successfully integrated into a larger system built from subsystems conforming to the same compatibility standard and serve as a functional subsystem.

From an economic perspective, compatibility decisions in networked industries are influenced by network externalities, which can be direct or indirect (Katz & Shapiro, 1985, 1986). While direct network externalities result from the increased utility a consumer derives from a product with increased users (e.g., a telephone network), indirect network externalities result from the increased availability of complementary products (e.g., a telephone). Research on network externalities has been elucidated by the "complementary product" or "hardware/software" paradigm (Church & Gandal, 1992; Katz & Shapiro, 1994), which emphasized the importance of indirect or complementary network externalities in influencing technology adoption (Church & Gandal, 1993), market structure (Economides & Salop, 1992), and strategic behaviors of firms in vertically networked industries (Church & Gandal, 1992; Matutes & Regibeau, 1987). Further, various empirical studies have documented the influence of complementary network externalities on the economics (Brynjolfsson & Kemerer, 1996; Gandal, 1995) and structural evolution of ICT industries (Bresnahan & Greenstein, 1999; Gallagher & Park, 2002; West & Dedrick, 2000).

The complementary product paradigm has also been explored in the compatibility literature under the term "platform," which is technology architectures that provide a framework for developing complementary products (Bresnahan & Greenstein, 1999; West & Dedrick, 2000). Computing platforms have been viewed as a cluster of technically standardized components called Application Programming Interfaces (APIs) that define the interaction of complementary applications designed for the platform (Greenstein, 1998). These APIs provide the necessary information for developers to design applications compatible with the given platform. Hence, API clusters or computing platforms mediate software compatibility (West & Dedrick, 2000) and in that sense can be referred to as a "compatibility standard" (David & Greenstein, 1990).

Platforms, as compatibility standards, have had significant effects on the evolution of the computing industry in general (see Greenstein, 1998; Bresnahan & Greenstein, 1999), as well as in various ICT industry segments in particular (Gandal, Greenstein & Salant, 1999; Iversen & Tee, 2006; Karvonen & Warsta, 2004; Tilson & Lyytinen, 2006; West & Dedrick, 2000). For instance, in the mobile computing industry segment, Karvonen and Warsta (2004) highlight the importance of development platforms in providing access to various technological layers of mobile computing. Iversen and Tee (2006) further document the case of the Symbian platform in shaping the evolving industry structure in the mobile telecom sector. Studies have also documented the strategic implications of developing and managing platforms (Garud & Kumaraswamy, 1993; Methlie & Gressgård, 2006; Windrum, 2004) as well as their adoption by organizations (West & Dedrick, 2006).

Platforms, Compatibility, and Certification

Development and specification of compatibility standards is an important facet of ensuring compatibility. However, without consistent implementation, compatibility can suffer (Egyedi & Dahanayake, 2003; Egyedi & Hudson, 2005). This is particularly true for platforms, as they are complex technological systems that encapsulate multiple component interfaces and their relationships. Although the specifications for component interfaces are documented to explicitly provide information about their usage, their complexity can leave room for interpretation and, hence, inconsistent implementations (Egyedi & Dahanayake, 2003). Varying implementations are a central issue for compatibility standards, particularly complex computing platforms, as inconsistent implementations can create incompatible products and reduce complementary network externalities, thereby defeating the very purpose of a platform. Hence, in many cases, it is in the best interest of platform promulgators (i.e., firms or industry alliances that develop, promote, or support a particular platform) to ensure "correct" implementation and use.

Various studies have examined the issue of managing compatibility and implementations. Egyedi and van Wendel de Joode (2004), for instance, identify coordination mechanisms that can be used to manage compatibility in open source software. These coordination mechanisms include the use of regulation (through contracts, licenses, and member agreements), operations (through reference implementations, tools, and training), and authority (through gatekeepers and hierarchies) that can converge the strategic behavior of various stakeholders toward compatibility (Egyedi & van Wendel de Joode, 2004). Further, Egyedi (2001), discusses various strategies employed by Sun Microsystems to foster compatibility of the Java technology. She highlights compatibility fostering "input controls" such as providing training and software development kits (SDKs) during the early stages of specification development, as well as "output controls" such as providing reference implementation and compatibility certification and logos during the later stages of specification implementation (Egyedi, 2001a, 2001b). Output controls are particularly relevant for platform standards, as they can be employed to control implementation variations of proprietary as well as open platforms (Egyedi, 2001a). Compatibility certifications wherein products are assessed for conformance to a developed standard is one such output control, which can be employed by platform promulgators to ensure vertical compatibility.

Despite the use of compatibility certification by industry consortia such as CTIA[1] and firms such as Sun and Microsoft[2], not to mention the long history of reference and minimum quality certification employed by various governmental agencies (e.g., in the United States, FCC-EA[3] and in the EU, CE markings[4]) as well as international quasigovernmental standards development and compliance organizations such as IEC[5] and UL[6], there has been little academic research on the topic of certification (studies on Java and other open source software compatibility strategies by Egyedi are notable exceptions).

While fostering compatibility is an important goal, the mechanisms used to drive compatibility can have implications for the suppliers of complementary goods that in turn add value to the standardized technology. Although research on compatibility recognizes certification as a mechanism to foster compatibility, it does not directly address the resulting competitive landscape of the complementary product market. Thus, research on the effects of compatibility certifications for the complementary product market is needed and, to the best of our knowledge, has not been performed.

In this article, we address this gap by reporting on our exploration of the market effects of multiple compatibility standards certifications in the emerging mobile computing industry, where

complementary mobile applications are certified for compatibility on various evolving computing platforms.

THE CASE OF MOBILE APPLICATIONS MARKET

The mobile industry is comprised of multiple interrelated segments, while a complete depiction of the mobile industry might best be achieved using a so-called value network or ecosystem model; here, for the sake of simplicity, we employ a value chain metaphor (Porter, 1985). Examining the value chain from the mobile application perspective, researchers have identified five core segments in the mobile industry (Barnes, 2002; Karvonen & Warsta, 2004). These segments include (1) the mobile content providers that create, aggregate, and distribute mobile content; (2) mobile application developers/vendors that develop and distribute software providing computational functions such as e-mail/chat clients, word/spreadsheet processing, and mobile games on mobile devices; (3) mobile platform providers that provide the necessary implementation tools for deploying mobile applications; (4) mobile device manufacturers that provide information processing capable mobile devices; and (5) mobile network operators that deploy and manage mobile network infrastructure to provide mobile access to end users.

In this article, we are primarily interested in mobile application developers as they develop complementary applications for mobile computing platforms. These application developers are essentially software development agencies that specialize in developing applications for mobile devices. They are the core technical facilitators of the emerging medium, with a high degree of interdependence with the various segments of the mobile value chain.

Computing Platforms and Certification

As mobile applications execute on mobile devices, they typically interact with various technological layers such as the mobile operating system, device hardware, and the network infrastructure (Karvonen & Warsta, 2004). Depending on its functionality, an application might interact with many interfaces across these layers. This interaction is achieved by APIs, which provide access to the layer-specific features required by the application. For instance, a location-based multimedia application might require and use the mobile network's location APIs for mapping; the mobile device's camera-controlling APIs for collecting visual imagery, and the operating system's file-system APIs for data storage. While APIs enhance interoperability, they also limit applications to devices and networks that have those specific APIs.

In order to circumvent the complex dependency of mobile applications across multiple layers, various computing platforms have been developed. These platforms essentially package available APIs at various layers and provide a standardized mechanism to access various layer-specific features. In the mobile environment, operating systems such as Symbian-OS, Palm-OS, and Windows Mobile have been extended to provide access to layer-specific features. Additionally, middleware platforms such as Sun's Java ME and Qualcomm's BREW provide APIs that indirectly provide access to various layer-specific features. While these middleware platforms have the same basic aims, they differ in their development and management. Java ME, for instance, is a semi-open standardization initiative that has evolved from its success in the desktop and server markets[7]. Similar to Java, it is a platform that runs on top of a mobile operating system, allowing Java ME applications to execute on Java ME-implementing

devices. BREW, on the other hand, is a proprietary standard developed by Qualcomm that is specifically designed for network operators. Similar to Java ME, it is a development platform that runs on top of a mobile operating system, allowing BREW applications to execute on various BREW-implementing mobile devices[8].

While the purpose of a platform is to provide integrated access to various components through APIs, due to rapid device and network innovations, both operating systems and middleware platforms are often not able to provide full device- and network-specific functionalities. Device manufacturers, realizing this dependency, provide their own flavors of integrated device-specific platforms. Nokia, for instance, provides platforms such as Nokia Series 60 and 80 that are based on the integration of Symbian and Nokia device APIs. RIM's Blackberry is another example, where additional Blackberry-specific APIs are integrated with the Java ME platform. As the complex integration of multiple APIs takes on various forms, we use the term "computing platform" to refer to the mobile operating systems (e.g., Symbian-OS, Palm-OS, and Windows Mobile), middleware platforms (e.g., Java ME and BREW), as well as device-specific platforms (e.g., Nokia Series 60 and 80, and Blackberry Java).

Contributing to the complexity of various computing platforms are the certification programs sponsored by operating system and middleware platform promulgators such as Symbian, Microsoft, Sun, and Qualcomm. Further, various device manufacturers sponsor certification programs that partially or completely adopt certification criteria from platform promulgators. Motorola, for instance, sponsors separate certification programs for testing Java ME applications and Windows Mobile applications (Mahmoud, 2002; Motorola, 2006)[9]. Furthermore, various network operators also sponsor customized certification programs to ensure proper application behavior on their networks[10].

These certification programs are essentially designed to test and verify the adherence of mobile applications to a platform sponsor's defined best practices in platform usage to allow appropriate application execution. The certified applications are digitally signed and can use proprietary logos for deployment on various mobile operator networks for eventual end-user consumption.

From the application developer's perspective, these certification programs are critical, as most network operators mandate that applications be certified prior to deployment[11]. The prevalence of computing platforms and their certification programs provides an opportunity to explore the effects of certification on a complementary product market. In the following section, we examine these effects in the mobile applications market.

Market Effects of Platform Certification

Clearly, the attempt to mitigate the complex nature of mobile technologies through the creation of application development platforms has, at least temporarily, been thwarted by the emergence of multiple competing platforms that enjoy a varying degree of support from actors across the mobile value chain. While the competitive dynamics surrounding the emergence, adoption, and support of various platforms in the mobile value chain is bound to influence the structural landscape of the entire value chain, in this article we focus on some of the core issues raised by the prevalence of platforms and their certification programs upstream in the value chain.

It was expected that the multiplicity of platforms coupled with high prevalence of certification would have structural and strategic implications for the complementary mobile applications market. In order to investigate these possible effects, we conducted 18 open-ended interviews of the top management of mobile application vendors that develop and distribute mobile applications

across multiple platforms, during the fall of 2005 in the United States. In addition, several on-site discussions and interviews with the top management team and certification program managers at National Software Testing Labs (NSTL), a leading global certification intermediary providing certification for multiple certification programs, were also conducted in the summer of 2006[12]. Based on these interviews and secondary sources such as trade press articles and publicly available documentation of certification programs, we outline some of the compatibility-driving service niches and their consequent effects on the mobile applications market.

Compatibility Services and Certification Intermediation

The existence of multiple platforms creates vertical application compatibility issues, wherein the applications based on one platform standard are not able to execute on devices and networks supporting another. Given this situation, we sought to find how developers deal with this issue. Accordingly, they reported that the vertical incompatibilities have given rise to various types of services that assist application developers in reaching consumers using mobile devices implementing various platforms.

An important category of such services is *compatibility and interoperability testing.* Such testing services test applications for compatibility with various mobile devices implementing the platform on which the application was developed. The service provides compatibility test reports based on test plans, which are either designed by the compatibility service providers or negotiated between the compatibility service provider and the developer. In addition to testing services, application migration services known as *cross-platform application porting* are also abundant, wherein applications are reprogrammed and migrated to execute on a competing platform. Although various technological tools are available to developers that assist in developing applications for multiple platforms simultaneously,[13] such tools often do not achieve complete compatibility. All the developers interviewed reported that such cross-platform application porting services are quite commonly used to systematically redesign and reprogram an application to execute on competing platforms.

Although such services are common in markets where multiple standards compete for dominance, the mobile applications market diverges from other markets with multiple competing standards due to the prevalence of platform certification. In addition to various compatibility, interoperability, and cross-platform porting services, the use of *certification services* has become a necessity in the business of mobile application development and distribution. In addition to application testing on various certification test criteria, certification services also manage the applications through various certification processes of application submission, testing, and signing. All the application developers we interviewed underscored the importance of certification services in the mobile applications market.

Existence of these service niches (i.e., compatibility testing, application porting, and certification services) has had some key structural effects on the mobile applications market. One such effect comes in the form of certification intermediation. Platform certification sponsors typically outsource certification testing to specialists often referred to as Authorized Testing Labs (ATLs). Some of the major application platform certification programs and their authorized certification labs are shown in Table 1. Although the ATLs are agents of certification sponsors in certifying applications, they are in a unique position to serve multiple certification programs simultaneously and address the certification service niche comprehensively. In that sense, they act as intermediaries in the market by serving both the platform certification sponsors and the mobile application developers. We refer to this phenomenon as *certification intermediation*.

Table 1. Key mobile platform certification programs, sponsors, and authorized certification labs

Platform	**Certification Program**	**CertificationSponsor**	**Authorized Certification Labs***
BREW	True BREW[14]	Qualcomm	NSTL
Java ME	Java Verified[15]	Sun Microsystems	Babel Media, Capgemini, NSTL, and RelQ
Symbian-OS	Symbian Signed[16]	Symbian	Capgemini, Mphasis, and NSTL
WindowsMobile	Designed for Windows Mobile with Mobile2Market[17]	Microsoft	QualityLogic, NSTL, and Veritest

**Source: Certification program Web sites, accessed May 2007*

In the mobile applications market, certification intermediaries aggregate technical resources, such as an inventory of various mobile devices supporting various platforms, to perform certification testing. Certification intermediaries procure most, if not all, devices available in the market. The procurement sometimes also involves obtaining precommercial mobile devices, which are typically supplied by certification sponsors. Additionally, these intermediaries also buy access to multiple network operator voice and data service plans to test applications on live networks.

In addition to aggregating technical resources, certification intermediaries face risks typically related to certification programs. The ability to achieve and maintain the designation of an authorized test lab for a certification program by generating and managing a steady volume of application certifications is one such risk. In the longer term, certification intermediaries face risks from the possible dissolution of the certification program altogether[18]. Competition also exists with other application testing labs, although the number is fairly limited (approximately 10 worldwide).

As aggregators of various technical resources facing risks associated with the certification programs, certification intermediaries have incentives to diversify across service niches such as compatibility testing, porting, and other quality assurance services such as functionality, usability, and performance testing. Certification intermediaries are able to leverage their technical expertise and access to resources (often exclusive such as access to precommercial mobile devices and platform upgrades) to provide such additional services. By positioning themselves as a one-stop-shop for getting applications to the market, most certification intermediaries have evolved as a critical resource for mobile application developers.

In their activities, certification intermediaries support both the platform certification sponsors and the application developers. As a technical enterprise requiring both technical competency and detailed information about evolving platforms, certification intermediaries are critically positioned to influence the emerging mobile applications market. In the following section, we present some of the implications of certification intermediation on the supply of mobile applications.

Supply-Side Market Effects

As the core technical facilitators of mobile content in the mobile value chain, application developers essentially represent the supply side that delivers mobile applications and services to consumers. Our interviews with the application developers pointed to two essential supply-side effects of prevalent platforms and their certification programs on the mobile applications market; namely, the degree of *deployment fragmentation* and extent of *market entry barriers*. We describe each in turn.

Application developers attempt to address various specialized niches, which are characterized by specific mobile application functionalities such as navigation, monitoring and tracking applications for fleet management, games, ringtones, and screen-saver applications for entertainment. Irrespective of the addressed niche, the interviewed application developers indicated that all developers must first contend with multiple platforms and their varying implementations on various mobile devices. It essentially necessitates the development of multiple application versions to target a particular customer segment. For instance, in order to develop a specialized application for a segment of customers (e.g., truck drivers or Hispanic immigrants), the application developer has to provide multiple versions of the same functional application so it can execute on various platforms and their implementations on mobile devices used by the intended customer segment. From the developer's perspective, this fragments the market along multiple operating systems and middleware platforms, as well as along their various implementations on a plethora of mobile device models. We refer to this fragmentation as *deployment fragmentation*, which restricts application deployment across multiple platforms ("between-platform" deployment fragmentation) as well as their varied implementations on mobile devices ("within-platform" deployment fragmentation).

Although certification does not directly influence the between-platform deployment fragmentation, it was expected that certification can, at least partially, address the within-platform deployment fragmentation. However, the developers we interviewed noted that the certification testing criteria are typically very basic, and applications are only tested on limited mobile devices. Additional testing on mobile devices implementing the same platform requires additional certification fees. Further, the developers also experienced different levels of within-platform deployment fragmentation across platform standardization approaches. Open platforms such as Java ME, in spite of platform certification, seem to exhibit greater within-platform deployment fragmentation compared to proprietary platforms such as BREW[19]. Platform certification in that sense does not comprehensively address the within-platform deployment fragmentation, even though it provides a framework to manage variability in platform implementations.

Another market effect of multiple platforms, their implementations, and platform certification is an increase in entry barriers for application developers. As discussed earlier, multiple platforms and their implementations necessitate creation of multiple versions of an application in order to sell it to its intended customer segment. This creates additional costs for developing an application, reducing the return on investment in developing new applications and thereby reducing the incentive to enter a new market segment. For established application developers, such costs can be spread over multiple applications they develop and sell. However, for new market entrants, such costs act as entry barriers. The interviewed developers suggested various reasons for the same. First, mobile operators typically do not deploy uncertified applications on their networks for eventual sale to their subscribers. A developer entering the market has to certify the application to make a sale, thereby increasing application development and deployment costs relative to an environment

without certification. Second, ATLs are given the flexibility by the certification sponsors to provide volume discounts for certification. Volume discounting assists incumbent application developers that develop and certify multiple applications. However, it provides a competitive disincentive to market entrants who compete with incumbents for new applications, as market entrants typically cannot offer multiple application volume. Finally, the certification costs are based on testing an application on one handset model. As the number of targeted handsets to reach the intended consumer segment increases, the certification costs skyrocket, increasing the upfront application deployment costs. To understand the magnitude of such upfront costs, the range and average pricing for the major platform certification *per handset* are presented in Table 2 . According to the interviewed developers, the number of targeted handset models typically range from tens to hundreds, creating sunken certification investments of thousands of Euros, which are often higher compared to the application development cost itself. In that sense, certification raises the entry barriers for the supply of mobile applications and potentially hampers innovation as well[20].

Certification intermediation, together with mobile application deployment fragmentation and entry barriers, constitute the market effects of multiple platforms, their implementations, and the prevalence of platform certification on the supply of mobile applications. In the following section, we discuss the implications of these findings for similar complementary product markets that are characterized by compatibility standards or platform certification.

DISCUSSION

The case of platform certifications in the mobile application market provides the basis for investigating the overarching research question: What are the structural effects of platform certification on a platform's complementary product market? In this section, we analyze this question, drawing on the literature on compatibility standards and industrial organization. We conceptualize how platform certification can alter the market effects of competing platforms. For analytical clarity, we first discuss the issue of vertical compatibility and identify the market mechanisms that can potentially address it, and then discuss the implications of these mechanisms on the complementary product market.

Table 2. Platform certification program pricing

Platform Certification Program	**Application Certification Price** *per application per handset*	
	Range	**Average Price**
True BREW	€ 700	€ 700
Java Verified	€ 150 - € 500	€ 240
Symbian Signed	€ 185 - € 560	€ 332
Designed for Windows Mobile with Mobile2Market	€ 196 - € 314	€ 275

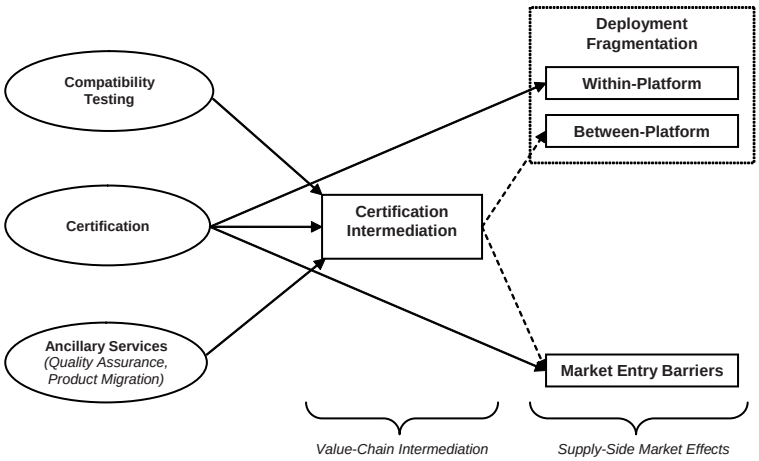
Source: Program sponsor websites and own calculations, accessed May 2007

As has been observed, vertical compatibility cannot be ensured by standardization and documentation of platform specifications because incompatibilities can result from inconsistent or selective implementation and use of a standardized platform (Egyedi & Dahanayake, 2003; Egyedi & Hudson, 2005). However, in some cases, these incompatibilities can be resolved by the market ex-post through provision of various tools such as cross-standard product converters, adapters, and gateways (Farrell & Saloner, 1992). On the other hand, in cases of rapid technological change and where multiple platforms and their varied implementations exist, such incompatibility-resolving-tools can become technically and economically unfeasible. In such a situation, the service niche of compatibility and interoperability testing assume greater significance. Further, demand for "ancillary services" such as cross-platform product migration and quality assurance services such as performance, functionality, and usability testing, can also be enhanced. Although such ancillary services do not resolve vertical incompatibilities, their demand increases as the functioning of the complementary product becomes intricately coupled with the implementation of the underlying platform.

Platform promulgators can also proactively address vertical incompatibilities through certification of complementary products and use of proprietary compatibility logos (Egyedi & van Wendel de Joode, 2003). Platform certification in the mobile applications market can be viewed as such a mechanism, where platform promulgators exercise authority and act as "gatekeepers" (Egyedi & van Wendel de Joode, 2004) over platform implementations and use by authorizing certification labs. In that sense, the authoritative mechanism of platform certification attempts to reduce implementation variation and increase complementary product compatibility through validation by experts (platform promulgators and authorized testing labs).

This raises two questions: Do these authoritative mechanisms achieve their goals, and what other effects do they generate? Based on the findings from the mobile applications market, we outline the structural implications of these mechanisms on the complementary product market as depicted in Figure 1. Starting on the lefthand side, the model presents the *primary market effects*

Figure 1. Market effects of platform certification on complementary product market



(bold lines) of certification, compatibility testing, and ancillary services on three outcomes; namely, value chain intermediation, the level of deployment fragmentation, and the extent of entry barriers in the complementary product market. In turn, value chain intermediation, in the form of certification intermediation, produces *secondary market effects* (dotted lines) by further influencing market fragmentation and entry barriers of the complementary product market. We consider each in turn.

One of the primary effects of platform certification is on the degree of fragmentation in the complementary product market. In general, certification aims to ensure consistent use of the platform in designing complementary products, thereby addressing within-standard vertical incompatibilities of complementary products. In theory, the initially disjointed submarkets of various product versions for varying platform implementations are merged by certification, thereby reducing within-platform deployment fragmentation of the complementary product market. However, the case of mobile applications market suggests that this effect depends on the certification test criteria and the standardization approaches of the platform.

In addition, platform certification can also influence the entry barriers for the complementary product market. In particular, the cost of certification is incurred by complementary product developers. These costs can be especially daunting for market entrants, which typically do not offer multiple products during the startup phase and therefore are unable to take advantage of potential volume certification discounts. Hence, compared to a market without certification, *ceteris paribus*, a market with certification increases market entry barriers in the complementary product market.

Furthermore, certification also creates an opportunity for intermediation. Of course, the existence of certification intermediation largely depends on the platform promulgator outsourcing certification testing. However, the decision to outsource certification testing can be motivated by the need to obtain third-party objectivity as well as the availability of expertise in performing related services (Biglaiser, 1993; Choi, 1998). As discussed earlier, in markets with multiple platforms and their varied implementations, compatibility testing and related ancillary services are in greater demand. Therefore, both third-party objectivity and expertise can be acquired through outsourcing certification testing to firms providing such related services. Hence, it follows that the structural outcome of certification intermediation is more likely in markets where compatibility testing and related ancillary services are in greater demand.

The existence of certification intermediation can, in turn, have secondary market effects on the complementary product market. Given that multiple platforms compete in the market, certification intermediaries can potentially build on the expertise gained through services such as compatibility testing to diversify across multiple platforms. Certification intermediaries can increase both economies of scope and scale by aggregating multiple certifications and related services. Such aggregations are especially common in digital industries (Resnick, Zeckhauser & Avery, 1994; Sarkar, Butler & Steinfield, 1995; Whinston, Stahl & Choi, 1997).

Such economies (given competing intermediaries) may be passed on to the complementary product developers, thereby potentially reducing entry costs and barriers as compared to the case with certification alone. Additionally, certification intermediaries, with their increased expertise on certifying multiple platforms, can provide aggregate economies of scale across all platforms, thereby reducing complementary product development costs across multiple platforms. In that sense, they can reduce the impact of between-platform deployment fragmentation in the complementary product market.

CONCLUSION

In the rapid evolution of information technologies, platform certification represents an important mechanism for facilitating vertical compatibility. However, in markets with competing platforms and their varied implementations, the requirement to certify across a variety of platforms may stifle the development of a rich variety of complementary products. To date, there has been little research on the complementary product market effects of platform certification. Drawing on evidence from a case study of the mobile application market, the research presented here provides insights into the effects of platform certification for a complementary product market.

Our exploratory research finds that in the mobile industry, platform certification has three effects; namely, intermediation within and between platform deployment fragmentation and entry barriers. If one assumes, however, that platform certification and competition are inevitable, attention should focus on the effects of intermediation. As compared to an unintermediated market, our research suggests that certification intermediaries may potentially mitigate some of the negative effects of a platform competition through their beneficial impact on between-platform deployment fragmentation and a potential reduction in entry barriers. Interestingly, in the case that an intermediary is able to certify for more than one program, it is likely to possess valuable information about the supply of complementary products to the various platforms and to some extent the true state of the platform competition. Clearly, the value of this information is dependent on the structure of the intermediation market, with a monopoly certification provider holding fairly complete information as compared to intermediaries in a competitive market.

Thus, through identification of the market effects of platform certification as a means for managing compatibility, this work contributes to both the broader literature on compatibility standards as well as standards competition. However, our analysis is exploratory in nature, and hence, the conceptual model requires more systematic validation before its generalizability to other similar product markets can be assessed. Particular characteristics of the mobile industry that may have influenced our findings include that the suppliers of the complementary goods (mobile applications) are numerous, generally small, and often require only relatively low levels of investment, which encourages entry. This, in addition to the global nature of application supply, may have influenced the certification outsourcing decision, which in turn generated certification intermediation. Assessing the extent to which our findings are valid in, for example, more concentrated complementary product markets requires further research. Future research might also evaluate the effects of certification intermediation on extending indirect platform network externalities. Models might also be developed to improve understanding of the role of certification intermediaries across a range of software development activities to assess the generalizability of their benefits. In addition to systematic assessment of the model presented here in both similar and dissimilar contexts, future research might attempt to model the influence of compatibility certification on the standards competition to assess whether certification serves to extend or shorten the life of such competitions.

REFERENCES

Barnes, S.J. (2002). The mobile commerce value chain: Analysis and future developments. *International Journal of Information Management, 22*(2), 91–108.

Biglaiser, G. (1993). Middlemen as experts. *RAND Journal of Economics, 24*(2), 212–223.

Bresnahan, T.F., & Greenstein, S. (1999). Technological competition and the structure of the

computer industry. *The Journal of Industrial Economics, 47*(1), 1–40.

Brynjolfsson, E., & Kemerer, C.F. (1996). Network externalities in microcomputer software: An econometric analysis of the spreadsheet market. *Management Science, 42*(12), 1627–1647.

Choi, S. (1998). Market lessons for gatekeepers. *Northwestern University Law Review, 92*(3), 916–966.

Church, J., & Gandal, N. (1992). Network effects, software provision, and standardization. *The Journal of Industrial Economics, 40*(1), 85–103.

Church, J., & Gandal, N. (1993). Complementary network externalities and technological adoption. *International Journal of Industrial Organization, 11*(2), 239–260.

David, P.A., & Greenstein, S. (1990). The economics of compatibility standards: An introduction to recent research. *Economics of Innovation and New Technology, 1*, 3–41.

David, P.A., & Steinmueller, W.E. (1994). Economics of compatibility standards and competition in telecommunication networks. *Information Economics and Policy, 6*, 217–241.

Economides, N., & Salop, S.C. (1992). Competition and integration among complements, and network market structure. *The Journal of Industrial Economics, 40*(1), 105–123.

Egyedi, T.M. (2001a). Strategies for de facto compatibility: Standardization, proprietary and open source approaches to Java. *Knowledge, Technology, and Policy, 14*(2), 113–128.

Egyedi, T.M. (2001b). Why JavaTM was -not standardized twice. *Computer Standards & Interfaces, 23*(4), 253–265.

Egyedi, T.M., & Dahanayake, A. (2003). *Difficulties implementing standards.* Proceedings of the 3rd IEEE Conference on Standardization and Innovation in Information Technology, Delft, The Netherlands.

Egyedi, T.M., & Hudson, J. (2005). A standard's integrity: Can it be safeguarded. *IEEE Communications Magazine, 43,* 151–155.

Egyedi, T.M., & van Wendel de Joode, R. (2003). *Standards and coordination in open source software.* Proceedings of the Standardization and Innovation in Information Technology (SIIT 2003), Delft, The Netherlands.

Egyedi, T.M., & van Wendel de Joode, R. (2004). Standardization and other coordination mechanisms in open source software. *International Journal of IT Standards & Standardization Research, 2*(2), 1–17.

Farrell, J., & Saloner, G. (1992). Converters, compatibility, and the control of interfaces. *The Journal of Industrial Economics, 40*(1), 9–35.

Gallagher, S., & Park, S.H. (2002). Innovation and competition in standard-based industries: A historical analysis of the U.S. home video game market. *IEEE Transactions on Engineering Management, 49*(1), 67–82.

Gandal, N. (1995). Competing compatibility standards and network externalities in the PC software market. *The Review of Economics and Statistics, 77*(4), 599–608.

Gandal, N., Greenstein, S., & Salant, D. (1999). Adoptions and orphans in the early microcomputer market. *The Journal of Industrial Economics, 47*(1), 87–105.

Garud, R., & Kumaraswamy, A. (1993). Changing competitive dynamics in network industries: An exploration of Sun Microsystems' open systems strategy. *Strategic Management Journal, 14*(5), 351–369.

Greenstein, S. (1998). Industrial economics and strategy: Computing platforms. *IEEE Micro, 18*(3), 43–53.

Iversen, E.J., & Tee, R. (2006). Standards dynamics and industrial organization in the mobile telecom sector. *Info: The Journal of Policy, Regulation and Strategy for Telecommunications, Information and Media, 8*(4), 33–48.

Karvonen, J., & Warsta, J. (2004). *Mobile multimedia services development: Value chain perspective*. Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia MUM '04, College Park, Maryland.

Katz, M.L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *The American Economic Review, 75*(3), 424–440.

Katz, M.L., & Shapiro, C. (1986). Technology adoption in the presence of network externalities. *The Journal of Political Economy, 94*(4), 822–841.

Katz, M.L., & Shapiro, C. (1994). Systems competition and network effects. *The Journal of Economic Perspectives, 8*(2), 93–115.

Mahmoud, Q.H. (2002). *Testing wireless Java applications.* Retrieved August 1, 2007, from http://developers.sun.com/techtopics/mobility/midp/articles/test/

Matutes, C., & Regibeau, P. (1987). Standardization in multi-component industries. In H.L. Gabel (Ed.), *Product standardization and competitive strategy.* Amsterdam, The Netherlands: Elsevier Science.

Methlie, L.B., & Gressgård, L.J. (2006). Exploring the relationships between structural market conditions and business conduct in mobile data services markets. *Journal of Electronic Commerce Research, 7*(1), 14.

Meyer, M.H., & Seliger, R. (1998). Product platforms in software development. *Sloan Management Review, 40*(1), 61.

Motorola. (2006). *Motorola Q Developer Guide*. Retrieved August 1, 2007, from http://developer.motorola.com/docstools/developerguides/

Porter, M.E. (1985). *Competitive advantage: Creating and sustaining superior performance*. New York: The Free Press.

Rada, R. (1996). Who will test conformance? *Communications of the ACM, 39*(1), 19–22.

Resnick, P., Zeckhauser, R., & Avery, C. (1994). *Roles for electronic brokers.* Proceedings of the Telecommunications Policy Research Conference, Arlington, Virginia.

Sarkar, M.B., Butler, B., & Steinfield, C. (1995). Intermediaries and cybermediaries: A continuing role for mediating players in the electronic marketplace. *Journal of Computer Mediated Communication, 1*(3).

Schmidt, S.K., & Werle, R. (1998). *Co-ordinating technology: Studies in the international standardization of telecommunications.* Cambridge, MA: MIT Press.

Tarnacha, A., & Maitland, C.F. (2006). *Entrepreneurship in mobile application development.* Proceedings of the Eighth International Conference on Electronic Commerce, Fredericton, New Brunswick, Canada.

Tarnacha, A., & Maitland, C.F. (Forthcoming). The effects of standards competition on market entry: The case of the mobile application markets. In *The standards edge: Unifier or divider?* The Bolin Group.

Tassey, G. (2000). Standardization in technology-based markets. *Research Policy, 29*(4-5), 587–602.

Tilson, D., & Lyytinen, K. (2006). The 3G transition: Changes in the US wireless industry. *Telecommunications Policy, 30*(10/11), 569–586.

West, J., & Dedrick, J. (2000). Innovation and control in standards architectures: The rise and fall of Japan's PC-98. *Information Systems Research, 11*(2), 197–216.

West, J., & Dedrick, J. (2006). Scope and timing of deployment: Moderators of organizational adoption of the Linux server platform. *International Journal of IT Standards & Standardization Research, 4*(2), 1–23.

Whinston, A.B., Stahl, D.O., & Choi, S.-Y. (1997). *The economics of electronic commerce*. Indianapolis, IN: MacMillan Publishing Company.

Windrum, P. (2004). Leveraging technological externalities in complex technologies: Microsoft's exploitation of standards in the browser wars. *Research Policy, 33*(3), 385–394.

Yin, R.K. (1994). *Case study research: Design and methods* (2nd ed.). Thousand Oaks, CA: Sage Publications Inc.

ENDNOTES

1 Various CTIA-sponsored certifications can be found at http://www.ctia.org/business_resources/certification/

2 See, for instance, the Windows Logo Program available at http://www.microsoft.com/whdc/winlogo/default.mspx

3 See, for instance, the Equipment Authorization (EA) certification available at http://www.fcc.gov/oet/ea/

4 The CE health and safety certification details can be found at http://www.cemarking.net/

5 Various IEC-sponsored certifications can be found at http://www.iec.ch/helpline/sitetree/conformity/

6 Various UL conformity assessment marks can be found at http://www.ul.com/

7 The Java ME standard was established and is managed by an expert group of leading mobile device manufacturers, wireless carriers, and software vendors using the traditional Java Community Process (JCP) of developing publicly available specifications.

8 In addition to being an application platform, BREW also incorporates an application distribution system that allows end users to shop, purchase, download, and install software over the operator's network. This combination provides network operators with a vertically integrated distribution control system that streamlines the development, deployment, and billing of applications for both developers and the network operator. For details, see BREW Developer home hosted at http://brew.qualcomm.com/brew/en/developer/overview.html

9 Nokia as a device platform promulgator also used to provide a certification program called NokiaOK, which they later merged with Symbian. See Nokia OK press release at http://press.nokia.com/PR/200203/853384_5.html

10 Examples of such programs include Cingular Certified Solution (http://developer.cingular.com/developer/testing/index.jsp?itemId=400025), Sprint Application Testing (http://developer.sprint.com/site/global/develop/p_testing/p_virtual_dev_lab/p_virtual_dev_lab1.jsp), and Virgin Mobile Certification (www.nstl.com/about_nstl/press_docs/virginmobile.pdf), to name a few.

11 As examples, see VerizonWireless' requirement for True BREW Certification at http://www.vzwdevelopers.com/aims/public/BrewLanding.jsp, and Orange requirements for Industry Standard Testing at http://www.orangepartner.com/site/enuk/

develop/v_devcentre/tools/p_compatibility_test.jsp#5

[12] As NSTL has provided certification services for various PC and mobile technologies for over a decade, they are in a unique position to provide insights into this somewhat ignored service niche of certification.

[13] For example, software tools like AppForge Crossfire and Tira Jump Product Suite enable developers to create different platform versions of an application.

[14] The program details can be accessed from http://brew.qualcomm.com/brew/en/developer/overview.html

[15] The program details can be accessed from http://javaverified.com/

[16] The program details can be accessed from https://www.symbiansigned.com/

[17] The program details can be accessed from http://msdn2.microsoft.com/en-us/windowsmobile/bb250547.aspx

[18] The dissolution of NokiaOK certification and its merger with the Symbian Signed is an example of such a risk.

[19] For a detailed comparison of the effects of open and proprietary standards on market fragmentation in the mobile applications market, see Tarnacha and Maitland (Forthcoming).

[20] For a discussion on the effects of standards on entry barriers and entrepreneurial innovation in the mobile applications market, see Tarnacha and Maitland, 2006).

This work was previously published in the International Journal of IT Standards and Standardization Research, edited by K. Jakobs, Volume 6, Issue 2, pp. 48-65, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 4.34
Buongiorno! MyAlert:
Creating a Market to Develop a Mobile Business

Guillermo de Haro
Instituto De Empresa, Spain

José María García
Instituto De Empresa, Spain

ABSTRACT

In 1999 Jorge Mata, vice president of Broadvision and former expert in interactive solutions for Banco Santander and McKinsey, decided to leave everything to create MyAlert. The company was born on the basis of offering the same Internet services on the new and growing mobile devices. With a strong financial capitalization after raising more than 50 million euros during the bubble burst, in 4 years the company figures were in the black, and the journey had led to the creation of the European sector of mobile data services market and the European leader in that sector. As Charles Darwin emphasized, if a being wants to survive in a shifting environment, it must evolve at least as fast as the medium itself: Buorngiorno! MyAlert ruled the change.

INTRODUCTION

This case helps to understand, through the history of Buongiorno! MyAlert (García & de Haro, 2003), how the mobile data services market works, who are the main players, and what are their business models, explaining clearly the *initial development of the mobile data services market* in Europe.

MyAlert was a pioneer in the sector. It was the first company to have the idea and to define a business model. It was the first to appear in the mobile services market, leveraging on its own technology platform. Also, it was the first in raising enough money to make this market grow on the basis of products and services development. Once the market was launched, the firm consolidated the business model by merging with Buongiorno.

Continuous design and development of products, applications, and business models was a key. As Nomura pointed out, "MyAlert is positioned to move into several areas of the value chain at low cost." The flexible strategy of the company made it possible to lead a changing market, influenced by many major players. Value chain analysis and strategic diagnosis on the environment and its evolution is provided. Also product development and business models are explained.

Another key was the market-driven evolvement of the offering, maintaining a close look at customers, competitors, and environment. Firm changed on a market basis, not on a technological basis, for example, focusing on SMS instead of WAP when it was "flavour of the month."

Beginning from a "sweet spot" of the value chain with technology as their competitive advantage, by 2004 the company had evolved looking for new business, consolidating and looking for new markets.

THE COMPANY EVOLUTION

MyAlert

The idea was simple: to take advantage of the "alerts"—data messages sent through the mobile phone network—to create services like those being launched in the Internet. MyAlert started as a "portal of alerts": users defined via the Web which information alerts they were interested in receiving via a text message delivered to their mobile phone. In the midst of the Internet boom in Spain and Europe, Jorge pioneered the extension into the mobile phone's mailbox.

Jorge Mata, telecommunications engineer, MBA from New York University, 4 years at McKinsey & Co, VP at Banco Santander developing Internet and GSM banking solutions, and VP at Broadvision, had the idea. Time to market was critical: he gave up his job and also an important stock-option plan and explained his vision to Pehong Chen, Broadvision's founder. Chen decided to invest in the project. MyAlert was created in March 1999 with 500,000 euros capital.

Jorge recalled later, "That could seem a lot of money, but when you make strong expenditures in technology and staff, it runs away quickly. I got top engineers, we got an office and we tried to make the money last as much as we could . . . but in September 1999, we had nothing." In October 1999 a new capital increase of 4 million euros was subscribed mainly by Banco Santander and BBVA. Jorge attracted qualified top talent, but the bulk of the recruits happened in the technology area, reaching 35 employees in 1999, most of them engineers devoted to developing the platform.

Market Launch

The first versions of MyAlert's Internet portal were operational in July 1999 with services such as top headlines of the day or soccer match scores, based on data feeds provided by Europa Press, a leading Spanish news agency which also took a stake in the company. Due to agreements with other content and service providers the number of services increased: information about travel and services as innovative as an alert to remind you to quit smoking.

The company became the leader within that market segment and began to identify potential business services for the companies it was partnering with, starting to develop customized services for corporate customers. For instance, the recruitment consultant Adecco could contact candidates in record time by delivering them employment offer alerts in real time. Traffic in the portal increased hand in hand with product development, surpassing in less than a year 200,000 registered users without a substantial advertising expense.

MyAlert went for all European major markets replicating what was already working in Spain:

launching advertising supported free services in a way that allowed growing and developing a local presence and then starting offering business services based on its proprietary technological platform. France was first in October 1999, and by July 2000 Italy, Germany, and the United Kingdom were covered through organic growth. For R&D a development center was based in Madrid, and investments were made in companies in Bulgaria (low costs and high productivity) and Finland (control stake in Future121, specialized in WAP and 3G developments).

The team reached 80 members and consolidated internationalization but the market demanded further growth. In May 2000 MyAlert successfully closed a new financing round, bringing in an additional 48 million euros from top-tier investors such as Nomura, Brokat, 3i, or the original shareholders. Investors found highly valuable the proprietary technology platform. MyAlert's market value went from 14.2 million euros by December 1999 to 163.8 million euros a year later.

But the NASDAQ Index fell dramatically in 2000. Markets went deeply pessimistic, the economy fell into recession, and new technologies took the worst part of the crisis. Suddenly MyAlert was forced to confront its business model with the new scenario.

Buongiorno! MyAlert

After the financing round two challenges were yet to be solved. First, how to sustain the speed required for growth, and second, how to move from now unacceptable "new economy" standards to a positive P&L as required by the "old economy"?

Some subjects became important. First, competition escalated and a substantial chunk of the company's expected revenues came from advertising and mobile commerce, both sectors seriously hurt by the crisis. The company had reached a 5-million-euro turnover, one of the highest within this sector, but it was presenting sizable losses. Financials were acceptable due to the investment community supporting the strategic need for short-term strong losses to buy market share, but the financial markets were now requiring the revenue base to cover the cost base in every venture.

Second, the targets of growth and leadership required reaching in all major markets a leading position, something impossible to achieve through organic growth: it was the hardest option from a practical standpoint but could also seriously put into risk the company's bottom line. Another option was a sell out to a strategic partner, following the Finnish IOBox example, bought out by Telefónica for 230 million euros––even when its yearly turnover was 60,000 euros.

MyAlert decided to seek for a "twin soul," a company sharing the same objectives and ambition. By September 2001 the company agreed to merge with BuonGiorno!, Italian leader in personalized e-mail alerts and newsletters, creating a 260-employee company, with nearly 30-million-euro yearly revenues and 20 million subscribers (a yearly total of more than 3,000 million messages via Internet or mobile phone). Curiously both companies had exhibited an impressive story in raising funds: together more than 85 million euros.

Complementarity in their core skills was remarkable: MyAlert's strengths were in the mobile data services market; Buongiorno! was the leader in e-mail marketing services, having reached an 80% penetration over the total PC base in Italy.

Restructuring to reap synergies, decrease costs, and speed up the path to profitability was simplified by the complementarities in the strengths: Buongiorno! has advertising, marketing know-how, and Italian leadership, while MyAlert has ASP services technological excellence and reference in Spain.

Mauro del Río, founder of Buongiorno!, was president and Jorge Mata was vice president. Andrea Casalini, former CEO of EDS for Italy, was the new CEO. One year later 34 million subscribers were reached and the first positive EBITDA achieved.

THE MOBILE DATA SERVICES MARKET IN EUROPE

MyAlert and Buongiorno! grew in the midst of the two main business driving forces at the turn of the 20th Century: the development of mobile telephony as a new mass communication channel and the Internet as a universal information network.

The mobile telephony market developed in an accelerated way during the 1990s. Initial marketing was targeted to corporate users, soon becoming a mass market. In 2004 the total number of mobile telephone handsets worldwide was estimated at no less than 1,000 million, with more than 450 million users in Europe (EITO, 2004), surpassing the number of fixed telephony handsets in several countries.

A key factor for this development in Europe was the adoption of a shared digital standard, GSM, since 1992 (Huidobro, 1996). This standard helped a faster market penetration and a substantial acceleration in its innovation curve creating a pan-European market. Development of new functionalities or handset designs, and aggressive marketing offers from the telecom operators, such as subsidizing the cost of the handsets speeded up the market growth, with rates exceeding a yearly 60%. In Spain figures changed from less than a million users in 1995 to 7 million in 1998 or 15 million in 1999 after the entry of a third telecom operator (Amena). By 2004 more than 36 million users had a mobile phone.

The Internet was another force allowing the launching of new businesses. Its interactivity and universal access provoked an accelerated development of the Internet user base. By August 2004, worldwide total users reached almost 800 million (according to the Internet World Stats).

The development of both technologies revolutionized the economy and changed global society, altering social usages or creating new sectors and corporations.

Mobile Data and Value-Added Services Markets

"Mobile data services" refers to the delivery of information messages—as opposed to voice messages—through the mobile telecom networks.

Two distinctive businesses may be included: the delivery of messages from one user to another ("peer-to-peer" or "P2P") and the value-added services provided by a third party ("value-added services" or "VAS"). Both use the same technology platforms and communication networks, but they are different businesses in terms of industry players and value split. P2P services are mainly a simple communication service provided by the mobile telecom operator; VAS allow the users access to every sort of digital alerts supplied by content and service providers, similar to Internet browsing.

The mobile data services business started taking advantage of the Short Messaging Service (SMS) standard technology. Its enthusiastic usage by the youngest mobile customers—which stand for 45% of the European mobile users—made them a new revenue model. According to analysis, the European market in 2002 implied more than 10,000 million monthly text messages, and an average of 35 text messages a month per user.

Most of the revenue was coming from P2P services, although VAS were increasing. New significant submarkets appeared, such as mobile gaming (around 500 million euros in Europe; Frost & Sullivan, or ring tones' downloading (1,500 million euros in Europe; Strand Consult. Expected growth trend favors a much higher relative increase of VAS, which will exceed the P2P

messages share in 2005 according to Strand. In that year, data mobile services will account for 33% of total ARPU, and 17% will come from VAS.

Another aspect of this market is its growth rate. Telefónica Móviles, already exceeding the 15% share in 2002, expected text messages to become 30% of its total turnover by 2005. These are startling figures considering that by 2000 the mobile data revenues were only 4% of the total mobile telephony market in Spain.

Evolution of the Mobile Data Services Market

How to enable such a spectacular increase in sales? The main agents focused on data services after the SMS boom. Nevertheless and because of the SMS limitations, new technology platforms with a broader range of features were developed to guarantee future growth (DBK, 2003).

By 2000 the mobile telecom industry started to create the so-called "Mobile Internet" (Lautenschlänger & Schmidtke, 2000). The new model would be similar to the Web navigation: mobile portals translating browsing experience to the mobile phone screen interface. They failed as WAP protocol was unable to provide an appealing enough user experience (slow, no attractive contents, etc.).

Growth was fuelled by new services such as the adoption of Japanese i-Mode, an alternative technology allowing basic services (weather forecast and horoscopes) and advanced messaging features (e-mail and image delivery via color screens). So it seemed necessary to develop new technologies with faster and higher data transmission capabilities to enable a substantial increase in the number of mobile services provided (Lamont, 2001).

In Europe the industry decided to launch a new standard: UMTS (Universal Mobile Telecommunications System). The development process became much slower than expected (Gómez & González Martínez, 2001). In the meantime the industry kept on milking GSM or using other transitional technologies (2.5G), such as HSCSD (High Speed Circuit Switched Data), which works at 56Kbps and only requires software updates, GPRS (General Packet Radio Service), which works up to 115Kbps but requires new hardware routers, or EDGE (Enhanced Data rates for GSM Evolution), which is closer to 3G and able to reach 384Kbps. Furthermore, network equipment and handset manufacturers faced difficulties in developing in time the new hardware to put in place the UMTS mobile networks infrastructure.

At this very moment in 2002 MMS (Multimedia Messaging Service) messages were launched, which improved P2P applications (enhanced e-mail, image delivery, audio and video delivery, etc.), browsing and downloading applications (news, horoscopes, adult entertainment, games, etc.), and massive participation applications (voting, quizzes, etc.).

Industry expected a new increase in the usage level of its customers but this new message format required new handsets. Camera phones, phones that allow music downloading, personal organizer phones, or handsets geared towards gaming were also launched (Funk, 2004).

This phenomenon reflects the absolute need for the industry to keep its high growth rate on the basis of handset renewal and additional features. NTT DoCoMo managed to bring to the Japanese market 3 million camera phones in 6 months.

Main Players

To make a data mobile service work, participation of different players is necessary. Buongiorno! MyAlert is a specialist or "pure player" in the production of VAS. The other agents may be grouped in four areas: mobile telephony operators, technological infrastructure providers (hardware, software, and services), media, and interactive media.

VAS providers were the smaller in size. They identified a growing market niche and launched their services, directly to end users (download-

ing of logos), or developing tailored solutions for the other type of players involved such as telecom operators and infrastructure providers (developing information services customized to their customers), mass media (voting services to interact in a TV program), or interactive media (mobile versions of their offering).

Pure players are heterogeneous in terms of business lines (some specialize in ring tones, logos, and TV voting; others are mobile marketing specialists; and some others cover the whole spectrum of services) and in their background and business approach (some have strong technological foundations, other come from the advertising field).

The number of companies in this scene was high. Easy access to the necessary technology was possible once the market exploded, no longer being a barrier to compete. However, top players were concentrated, and in 2002 80% of the total SMS traffic in Spain was split among five or six companies, when there were more than 40 companies competing. Despite its pan-European development, market remains highly local. Top positions are usually taken by companies tied to the top local telecom operators and media or by "early movers" within that country.

In Europe relevant companies may be divided into four groups:

1. International companies, relevant position in several countries, covering a broad range of services, sound technological foundations: Buongiorno! MyAlert or iTouch.
2. International companies with presence in several countries, single product line: KiWee for consumer services or 12Snap for business services.
3. Relevant companies for a single market: MoviListo in Spain.
4. Companies tied to a main player: Vizzavi (Vodafone) or Terra Mobile (Telefónica).

Their business depended on share captured from mobile operators' total revenue and marketing budget that their customers channeled through them. Telcos kept a share of revenues while the rest was split among the agents involved in the production of the service (content providers, associated TV programs, etc.). In Spain and Italy

Table 1. Players in the mobile data services market

PLAYER	CHARACTERISTICS	SAMPLE PLAYERS
Mobile telephony operators	Own the network, provide voice and data telecommunication services	Vodafone, Telefónica Móviles, TIM...
Infrastructure providers	Produce and provide the necessary technological infrastructure and services (hardware, software, services) to enable telecom services and network intelligence leading the way to the provision of VAS.	Hardware: Nokia, Ericsson, HP... Software: Nokia, Microsoft, Oracle, Symbian... Services: IBM, Accenture...
Media groups	Produce content and entertainment which may be enhanced by mobile telephony enabled services	Vivendi, Bertelsmann, Endemol, PRISA…
Interactive media	Produce interactive content and entertainment which may be enhanced by mobile telephony enabled services –mostly internet portals	Terra, Wanadoo, Yahoo!, eBay, Amazon

it was usually 50%, in Japan DoCoMo was only charging 9%.

VAS providers seemed to have placed themselves in a high-growth hot spot. The Economist Intelligence Unit emphasized, "Will survive those who exploit today's technologies and successfully drive the transition to deliver other applications that consumers will pay for." And the report added that Buongiorno! MyAlert was at the forefront of the international wireless sector.

BUONGIORNO! MYALERT BUSINESS MODEL

After the merger product portfolio was conformed by technology and mobile marketing services provided by MyAlert and the interactive marketing services provided by Buongiorno. The business model was based on three revenue lines: advertising, business services, and consumer services. But when revenues were generated by a corporate customer it was called a "business service," and "consumer services" if they came from an end user.

Business Services

Services sold to corporate customers like advertising or interactive marketing. Also custom made applications to allow customer companies perform these services.

In the beginning mainly e-mail newsletters and mobile alerts' sponsorship, on the basis of its database with more than 34 million subscribers and its CRM tools which allowed segmentation and campaign hit rates higher to traditional media. A business based on volume: marketing impacts through new channels.

Lately appeared marketing services geared towards brand building or loyalty: interactive games, direct promotions and even interactive market surveys. The company developed a new business model: "Digital Marketing Project," consulting-like projects with higher margins and project size.

Head & Shoulders® launched a campaign in association with the movie *Men in Black*, giving the chance to win a Smart car prize to those sending messages to the 5556.

Another business services was the provision of some infrastructure and technology services to corporate customers. Thus CRM, e-mail marketing and SMS delivery tools were available through ASP (Application Service Provider) agreements, with relevant revenues.

Consumer Services

They are paid by the end user, who is billed by the mobile telco. Interactive games, voting, downloading of ring tones and logos, and P2P communication services such as SMS-based chats. The company was continuously investing in the development of new applications such as MMS-supported games or group messages.

Model based on the revenue split of the price paid by the end user. Telcos charged each user a previously set price (in Spain usually in the 0.3–0.9 euros range). This revenue is broken down between the telco, the content provider, and the service aggregator or VAS provider. Telcos retain 50% of revenues. If product was fully developed and managed by Buongiorno! MyAlert, it could keep the other 50%, otherwise it depended on the agreement among partners involved in the service development. KPMG estimated an average of 12% for the content provider and 38% for the VAS provider.

ORGANIZATION STRUCTURE AND OPERATIONS

After the merger a new organizational structure was defined and conformed by geographical markets rather than product units. International structure was defined along the already-exist-

ing country units: Italy, Spain, France, United Kingdom, and Germany/Austria. Each branch was run by a Country Director leading a sales managers team. Only Italy was structured per product line.

Country business units were supported by centralized staff units at the new Italian headquarters. There were two types of staff: purely corporate departments (Administration or Finance) and operational departments (IT, Content Development, and Customer Relationship Management). If a new game was needed the project manager would make a request to IT and this unit would make it work within the technological platform. This area was also in charge of the management and administration of B!3A. Content Development was accountable for negotiating and managing third-party content licensing terms, and the development of Buongiorno! MyAlert's own proprietary content. The CRM unit helped in the personalization of advertising and marketing campaigns.

The launching of a new service was centralized in the sales managers, whose role evolved into that of "product managers." In consumer services, the launching decision was made by each sales manager. With the new service defined the IT team started developing or customizing the necessary support systems. Meanwhile, the Content Development team acquired or licensed the appropriate contents required. In a parallel process, the service is added to the mobile operators' network infrastructure. The sales manager also decides on the appropriate marketing and is accountable for the profitability and final results of the service. In business services a similar process is followed with a consulting-like business approach.

This process differs from MyAlert's pre-merger product development methodology. Formerly the decisions about new products were taken by top management, based on the ideas and products developed by the more than 100 engineers in charge of MAGO (ideas such as m-auction services). In the next step these technological solutions were enhanced by features and content ideas developed by the Marketing department (for instance, defining specific alerts, such as the nonsmoking example). Finally only those new products successfully marketed by the sales managers were incorporated to the technological platform.

BUONGIORNO! MYALERT'S TECHNOLOGY

Andrea Casalini stated that "Buongiorno! MyAlert has two distinguished technological platforms. MAGO is like a jumbo jet: highly advanced technology, but very costly maintenance. B!3A resembles a fighter aircraft: fast and cheaper to maintain, but with more limited features." At merger time, these platforms were able to process more than 250 million monthly e-mails and 2,000 SMS per second, respectively. B!3A was created to manage very large Internet end-user communities, and MAGO was geared toward the management of wireless users communities.

MAGO

MAGO had allowed creating and sending every type of highly personalized alerts to end users, through different channels (SMS, e-mail). A system supported a personalization engine, several system management tools, client and end user interfaces, and connectivity devices with the different networks. From the start it was conceived with a prospective vision to serve as the basis for future development. A highly scalable architecture, it is robust and flexible, based on the compliance of industry standards which would allow future growth and multiple-user acceptance. Its architecture was structured around CORBA, an open object communications standard. Its software was coded in object-oriented languages (J2EE,

C++), JavaScript, and XML/XSL. The operating system was HP/UNIX and the primary database is Oracle Parallel Server RDBMS.

Other additional services were security (PKIs), commerce gateways, or event managers (www.buongiorno.com). Part of these applications and services were developed internally within MyAlert's development team, while others were based on adapting third-party developments.

The system included management tools and interfaces developed to allow customer companies access to the system: the foundations of the ASP business model. Easy to use and configure applications supported in a Web interface and application performance metrics are incorporated to increase the value for customers. The end-user interface was also developed in Web format easily customizable to customers.

The delivery engine integrated communication channels with end users through own- developed gateways, managing mobile networks (SMS, WAP, etc.), Internet, even a UMS (Unified Messaging System) integrating fax, voice and e-mail, among others. Because of its international ambition it was designed to support GSM, GPRS, TDMA, and CDMA standards. The company developed its own virtual network to link mobile operators' message centers. This network was a competitive advantage and barrier to entry for any competitor.

By 2002 due to the fine welcoming of its technology the company started "packaging" its technological platform into a software license agreement for its usage by third parties.

Buongiorno! and the Merger

B!3A was created only to enable the massive delivery of e-mail messages: the Buongiorno! branded daily bulletins. Afterwards delivery through platform was offered to corporate customers via ASP services. It was claimed that wherever there was a PC in Italy, there was a BuonGiorno! newsletter.

B!3A was a simpler platform, based on the Linux open-source operating system and developed on Java, built with the purpose of achieving the required performance at minimal cost. For this reason B!3A had its own application server, and could not easily work with other standards, which made impossible its license to third parties.

Performance was the main driver, as it should manage a high number of subscribers and e-mail messages to be delivered (in late 2002 the company servers delivered up to 400 million messages monthly by sending the company's different newsletters to its 34 million subscribers). This implied requirements in terms of scalability, to cope with continuous increases in traffic volumes, and flexibility, because of the seasonality and peak operation highs. The must was having a technology that could easily operate in an extremely short time, rather than a safer and sounder solution implying costly development and maintenance.

Also MAGO user interfaces were based on a Web format, accessible with an identifier and a password, without requiring software installation. The management tool allowed multichannel campaign management (wireless and e-mail) based on a single database, with CRM capabilities and campaign reporting generation. More than 500 corporate customers used this tool.

After the merger, both platforms continued delivering their services and carrying on "independent lives." Even for the same service, the technology platform used could be different depending on the country.

Buongiorno! MyAlert simultaneously marketed both platforms to its corporate customers with license agreement, ASP business model, or even customized developments to generate their own corporate platforms. This turned into a new separate business unit which focused on software licensing and technology consulting.

B!Digital Technologies

By the end of 2002 B!Digital Technologies was a separate P&L-focused business unit in break even. It brought together some of the Group's business services, such as software licensing and technology consulting, and R&D efforts.

Born with a technical team of 40 engineers (half based in Bulgaria), sales headquarters in London, and operating center in Madrid, and with an interesting customer portfolio (Hutchison 3G and the Chinese Government), was based in the software license of the MobileCast Messaging technological platform, the new commercial name for the latest evolution of MAGO. For licensing purposes an API was added allowing its modification by their clients for new developments and interfaces with their own back-end systems (billing, CRM, provisioning, etc.).

Core customer was Hutchison 3G, which selected the platform to launch its 3G alert and messaging services in seven countries. Another customer was Hellas Online which licensed also Infotainment applications, enabling it to offer in its home market services comparable to those currently offered in Spain and Italy by Buongiorno! MyAlert.

A second line of business was technology consulting professional services, for example, the creation of an intelligent traffic system for the Chinese Government, enabling alerts to the authorities about traffic signs and about congestion and routes.

B!Digital Technologies has continued development efforts towards MobileCast.

CONCLUSIONS

The Business Model

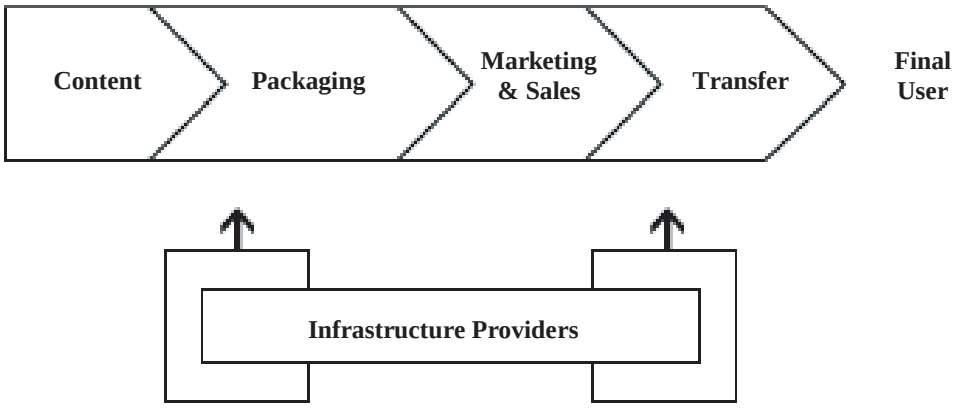
SMS alerts potential to create information and interaction services was huge. It was feasible the adoption by end users (Montero, 2000). The mobile market was under an explosive growth so the product was an open door to new services. As a pioneer the leadership was easy to achieve due to lack of competence. The matching of the entrepreneur and the project was perfect: Jorge had the experience (technology, business development), knowledge (in the new interactive channels), and ambition (refused his stock options). But there was a black spot: the lack of a clear and profitable business model from the beginning.

One of the main factors for the success and leadership of the company has been flexibility, the ability to change with the environment. To analyze this we will focus in the value chain and its evolution. The value chain of any industry is a conceptual tool that defines the activities needed for the creation and deliverance of a product or service. The methodology is to "chain" graphically the different functions that get to the final product, so it is easy to understand all the steps needed to make that product or service available.

The original value chain was as follows:

- Content: Development of the content of the alerts (the soccer info with the results of a match in real time)
- Packaging (or VAS): Integration of the content with the technical and operational requirements to send it (the soccer result in the text of an SMS)
- Marketing and Sales: To achieve new users (TV campaigns for the soccer alert)
- Transfer: Net providers who send the alert through the mobile networks to the end user (Vodafone sending the soccer message to a Vodafone end user)
- Infrastructure Providers of VAS: Development of the technology (software, hardware, services, etc.) to design and package the alerts (a tool to customize the soccer clubs' alerts to each end user)

Figure 1. Value chain



- Infrastructure Providers of Transfer: Development of the technology to deliver the alerts (the telecommunication equipment created by Ericsson and sold to Vodafone)

The value chain helps us to understand the business of the company (links where it is operating) and the opportunities available. The attractiveness of one part can differ dramatically from others. When the company was created in 1999 the Transfer link was an opportunity for telcos like Vodafone: launching new services with a marginal cost they could amortize the investment made in the network. New entrants had no network to amortize, so they focused in other links. MyAlert developed two business lines that covered different links.

One was a Portal of Alerts to Mobiles. A user registered to an alert, MyAlert developed it with its technology (proprietary) and the content needed (provided by a third party or developed by MyAlert), and sent it via the telco that operates in the user's mobile phone (negotiating with all telco providers). This business implied the VAS, and Marketing and Sales links, and sometimes the Content: also the Infrastructure Provision, but for internal operation of MyAlert only.

The other was as a service provider, developing customized solutions for companies that wanted to have their own alert services. MyAlert designed and packaged the alert and was responsible for sending it, leaving to the client company the Content, and Marketing and Sales, and acting only in the VAS link. Here the Infrastructure Provision is commercialized occasionally as a service, not as a product.

An important fact is that MyAlert acted as an Infrastructure Provider not being its business aim, but because no one in the market had developed the infrastructure needed to send alerts. Thus the company put a stake on the technology as a competitive advantage to confront competitors that want to get into the market.

As we pointed out before, the profitability of the business model depended on the announcers who paid MyAlert. The costs of sending an alert are the same, regardless of whether the ad was sold or not, and the telco has to be paid for each

SMS (alert). If the number of users increases the problem grows because advertisement income cannot cover costs.

MyAlert survived because of flexibility, correcting the business model to increase income (Gual & Ricart, 2001). Thus it became a provider of technology consultancy services and a development services provider to others in an ASP model.

The Merger

This was another good example of the strategic flexibility as a key success factor. By 2000, after the crash, MyAlert noticed that its position was strong but vulnerable, so consolidation was a good way to confront the future.

From the options valuated a merger among equals was the better option if the companies where complementary. Organic growth could not keep up with the growing rates needed for this environment, less having problems to generate cash flow. Selling the company was another possibility, but the only main had been to "make cash" and leave the project.

Buongiorno! was a good mate because of its high complementarities:

- Both were strong, MyAlert technologically, Buongiorno! in marketing services
- MyAlert income came from business services (mainly technological), Buongiorno! from advertisement
- Both were pioneers in developing direct marketing services, MyAlert for mobile, Buongiorno! for e-mail
- MyAlert was a Spanish leader, Buongiorno! was an Italian leader, and both had international presence
- MyAlert had the technology to access the mobile market, and B! the marketing and sales capabilities. The main indicator of the success: black numbers by 2002.

New Business Model

The new commercial focus forced continuous development of new products and services, possible due to the new structure based upon Product Managers. Formerly product development was developed in the R&D department with less flexibility.

In the first business plan m-commerce was expected to be 30% of the income by 2002, technology 15%, and advertising 55%. The real figures in 2002 where totally unexpected: 50% of advertising, 25% of technology, and 25% of consumer services, without any income from m-commerce.

Those deviations are normal. If prediction of new products' success is difficult, it is much more difficult if it is related to the adoption of new trends by end users. But the company was capable of reformulating its original plans to refocus towards more profitable areas. Most of the 2002 income was not expected in the original plan. Nomura Equity Research pointed out in 2000 that "MyAlert is positioned to move into several areas of the value chain at low cost."

Competitive Advantage Evolution

A competitive advantage helps a company be successful in a market on the basis of its superior conditions to those of its competitors. Those conditions come from the following:

1. Offer Advantages, due to processes or systems of the company in the design, development, and operation of products and services (lower costs due to scale economies).
2. Demand Advantages, due to the particular characteristics of the demand of products or services (relationships with clients or net economies that makes it difficult for them to change to another provider in the future).

3. Advantages related to the control of assets, so the competitive advantage in one business is related to the leadership in another business or in the influence power over the providers of other business (Microsoft leadership in operating systems market helped success in the text processing market).
4. Advantages related to innovation, when a real innovation in business term is developed and launched successfully.

At the beginning the competitive advantage was based upon the innovative product and the leadership, but it was vulnerable from the other perspectives because existing advantages were replicable by competitors. Thus to maintain the first-mover advantage acquired by being innovative in a new niche market, it had to keep on innovating quicker and better than its competitors. In the mean time, it needs to also develop offer and demand advantages to reinforce the original advantage and make it sustainable.

The merger favored this, first consolidating the leadership position. The merger implied lower costs (scale economies), developing offer advantages, and increased the number of clients: demand advantages. Not so clear and influential seem the associated assets (the extension of products from MyAlert to Buongiorno!) or the innovation (the product portfolio was not more innovative when combined).

When Buongiorno! MyAlert reached profitability the competitive original advantage had been consolidated and its position reinforced due to growth and consolidation. Its size was important in comparison with other pure players, but mean in comparison with other players like Vodafone.

When markets evolve strategic assets should change: the firm created new competitive advantages. When the market was born, business key factors were flexibility in products, and consequently good access to technology. While market is expanding keys turn into low costs and scale.

Technology

Is the property of technology important? The company played in the technology business and also used technology for the business. It changed from a technology-based company to a consolidated business model, but by 2003 it had presence in a totally technological business (B!Digital Technologies), and another two (Business Services and Consumer Services) requiring technology but not implying its development.

In terms of strategy and resources assignment, should the technological business line be prior? What is the importance of the technology development business for the rest of the business of the company? We will analyze the business lines attractiveness to define the strategic priorities.

Business lines are defined answering three questions: Who is the client (who pays, B2C or B2B?); What is the nature of the services sold (marketing/advertising or technological); and What is the business model (standard services—paid per unit; standard products—selling license; services ad hoc).

ASP and Software Licenses provided a quite similar service, but differences were in the way they were provided. A license implies a software package "as is," according to the industry standards. ASP implies a periodical payment for the operation of a technological platform.

To evaluate the attractiveness business lines can be reduce to three: business services (advertising, digital marketing projects, and ASP), consumer services, and software and technology consultancy. We decided to include new possibilities that the mobile market evolution could provide (e.g., 3G). Evaluation was made aligning capabilities of the company with the following criteria: market growth, competitors, and profitability.

First consideration is profitability. Consumer services is a margin business based upon the infrastructure already available with profitability was under 10%. Business services is a line with bigger margins but also bigger fixed cost, similar to consultancy. Software and technology consultancy is about 10% of income. Some aspects were difficult to valuate due to data availability, but according to available data, business services had fewer competitors and high margins in some of the products (e.g., Digital Marketing Projects), and also more volume in the future with consumer

Table 2. Business lines

Nature of Service	**Business Model**	**B2B**	**B2C**
Advertising / Contents	Standard Services	Advertising Impacts Selling	Consumer Services
	Ad hoc services	Digital Marketing Projects	
Technology	Standard Services (paid per unit)	ASP	
	Standard Products (selling license)	Software Licenses	
	Ad Hoc Services	Technological Consultancy	

Table 3. Business priorities for the future

Business Line	**Market Attractiveness**	**Capabilities**
Business Services	High Profitability Attractive Growth Moderated Competence	• Marketing • Projects
Consumer Services	Medium Profitability High Growth High Competence	• Marketing • Projects
Software and Technological Consultancy	Requires to Invest in R&D Attractive Growth Moderated Competence	• Proprietary Technology • Product
Future Products (3G)	Uncertain and high expectations Risk of getting out of business	• Needs?

services. To that extent, the software and technology consultancy business will be in a second priority level.

Second consideration is the influence of technology development in the business development of the other lines. By 2003 the market had changed and was looking for profitability, once demonstrated it was possible, and there was more technological availability. Basic technologies offerings increase and the development rate slows down. To add up the company did not integrate technologies after the merger, evidencing that the less sophisticated B!3A can compete with the advanced MAGO once some of the functionalities are available (mainly gateways to mobile networks). Need for technology was a must at the beginning due to a lack of market products necessary to develop the services. After the merger R&D efforts in new platforms were abandoned, focusing in developing new products and services as a response to market changes.

We consider that technological development was not a priority for the company, not even conditioned by the needs of the other business lines. Nevertheless further analysis could be interesting because of cases like Amazon, the world leader in electronic commerce also considered as a leader in developing software for creation and operation of Internet shops.

FINAL CONCLUSIONS

The company was able to become a leader in a newly created market thanks to technological development and strategic flexibility. Now less focused on technology, changing with the environment as the market consolidated. Facing new challenges like development of new products and services, to increase size of the business, and looking for new markets all over the world to complete Jorge Mata's vision: "*MyAlert aims to become the world leader in mobile commerce.*"

2002–2004

Industry presented major changes. In Spain the five main companies—Movilisto, Netsize, Terra Mobile, Buongiorno! MyAlert, and Gsmbox—represented a market share of 57.6% (DBK, 2003) by 2002. By 2004 Buongiorno! MyAlert acquired Gsmbox and I-Touch acquired Movilisto.

Telcos began developing content to position their business in another link of the value chain. Telefonica launched the E-mocion service, via WAP and i-mode, and Vodafone Life was launched all across Europe.

UMTS has come with the first services in Europe, video download and videoconferences. Market was growing at a 30% rate, with a progressive evolution towards multimedia terminals, the basis for new services, like Java Games or video chatting. Market size in Europe was almost 1,300 million euros, and expected to double in 3 years.

Nowadays Buongiorno! (B!) is the leading provider with nearly 400 employees, 56 million euros in revenue in 2003, and traded on the Milan stock exchange since July 2003, and is the only European company with presence in every European country.

REFERENCES

DBK, Informes Especiales. (2003, October). *Contenidos y Servicios para Telefonía Móvil.* Madrid.

European Information Technologies Observatory (EITO). (2004). European Union.

Funk, J.L. (2004). *Mobile disruption.* Hoboken, NJ: John Wiley & Sons.

García, J.M., & de Haro, G. (2003). *Buongiorno! MyAlert Case.* Madrid: Instituto de Empresa.

Gómez, F., & González Martínez, A. (2001). *Telefonía Móvil Digital.* Madrid: Anaya Multimedia.

Gual, J., & Ricart, J.E. (2001). *Estrategias empresariales en Telecomunicaciones e Internet.* Madrid: Fundación Retevisión-Auna.

Huidobro, J.M. (1996). *Telefonía Fija y Móvil.* Madrid: Thompson Paraninfo.

Lamont, D. (2001). *Conquering the wireless world: The age of m-commerce*. UK. Capstone Publishing.

Lautenschlänger, G., & Schmidtke, B. (2000). *Móviles: SMS, WAP y compañía (guía rápida).* Madrid: Data Ibérica de Software S.L.

Montero Pascual, J.J. (2000). *Competencia de las comunicaciones móviles: de la telefonía a Internet.* Valencia, Spain: Tirant lo Blanc.

www.buongiorno.com

This work was previously published in Unwired Business: Cases in Mobile Business, edited by S. Barnes and E. Scornavacca, pp. 29-47, copyright 2006 by IRM Press (an imprint of IGI Global).

Chapter 4.35
Location-Based Services in the Mobile Communications Industry

Christopher Ververidis
Athens University of Economics and Business, Greece

George C. Polyzos
Athens University of Economics and Business, Greece

INTRODUCTION

Advances in wireless communications and information technology have made the mobile Web a reality. The mobile Web is the response to the need for anytime, anywhere access to information and services. Many wireless applications have already been deployed and are available to customers via their mobile phones and wirelessly connected PDAs (personal digital assistants). However, developing the "killer" wireless application is still a goal for the industry rather than a reality. One direction for developing such applications points to location-based services (LBSs). LBSs are services that are enhanced with and depend on information about a mobile station's position. Location information by itself is not the ultimate service, but if location information is combined with content, useful services may be developed. These services offer the capability to users and machines to locate persons, vehicles, machines, and resources, as well as the possibility for users to track their own locations (GSM Association, 2003). The focus of this article is the analysis of the most critical success factors and challenges for LBS.

BACKGROUND

In order to show the domains on which LBS may have an impact, a list with the LBS categories, as defined by the Third-Generation Partnership Project (3GPPP, 2004), is presented in Table 1. Also, based on the information-delivery method, we identify three types of LBS: pull, push, and tracking services (GSM Association, 2003). In the case of a pull service, the user issues a request in order to be automatically positioned and to ac-

cess the LBS he or she wants. A use-case scenario demonstrating a pull service used broadly in the LBS literature (Poslad, Laamanen, Malaka, Nick, Buckle, & Zipf, 2001; Zipf, 2002) is the following. A tourist roams in a foreign city and wants to receive information about the nearest restaurants to his or her current location. Using a mobile device, the tourist issues an appropriate request (e.g., via SMS [short messaging service] or WAP [wireless application protocol]), and the network locates his or her current position and responds with a list of restaurants located near it. On the contrary, in the case of a push service, the request is issued by the service provider and not the user. A representative example of push services is location-based advertising, which informs users about products of their

Table 1. Standardized LBS types and corresponding application domains

Application Domain	Standardized LBS Types
Public-Safety Services	Emergency Services Emergency Alert Services
Tracking Services	Person Tracking Fleet Management. Asset Management
Traffic Monitoring	Traffic-Congestion Reporting
Enhanced Call Routing	Roadside Assistance Routing to Nearest Commercial Enterprise
Location-Based Information Services	Traffic and Public Transportation Information City Sightseeing Localized Advertising Mobile Yellow Pages Weather Asset and Service Finding
Entertainment and Community Services	Gaming Find Your Friend Dating Chatting Route Finding Where am I?
Location-Sensitive Charging	
Service-Provider-Specific Services	

interest located at nearby stores. In this service, users submit their shopping-preference profiles to the service provider and allow the provider to locate and contact them with advertisements, discounts, and/or e-coupons for products of interest at nearby stores. So, in this case, the service provider is the one who pushes information to the user. Finally, in a tracking service, the basic idea is that someone (user or service) issues a request to locate other mobile stations (users, vehicles, fleets, etc.).

From a technological point of view, LBSs are split into two major categories depending on the positioning approach they use to locate mobile stations. There is the handset-based approach and the network-based approach. The former approach requires the mobile device to actively participate in the determination of its position, while the latter relies solely on the positioning capabilities of elements belonging to the mobile network. For both of these approaches, several positioning techniques have been developed or are under development. What distinguish them from one another are the accuracy they provide and the cost of their implementation. The most popular network-based positioning techniques are cell-global-identity (CGI) methods, timing advance (TA), uplink time of arrival (TOA), and angle of arrival (AOA), while the most popular handset-based positioning techniques are observed time difference of arrival (OTDOA), enhanced observed time difference (E-OTD), and assisted Global Positioning System (A-GPS; Drane, Macnaughtan, & Scott, 1998; Swedberg, 1999). The accuracy provided by some of these techniques in different coverage areas of the mobile network is presented in Table 2.

In order to understand the emergence of LBS, one has to identify the major forces that brought to the surface the need for this kind of services. There exist four major forces, namely, market forces, competition forces, technology forces, and regulatory forces. Each of them is briefly discussed in the following paragraphs.

Market Forces

Market research around the globe has documented the willingness of mobile subscribers to pay for LBS. The LBS subscriber base is forecast to reach

Table 2. Positioning accuracies

	CGI	**E-OTD**
Rural Area	1 km–35 km	100 m–300 m
Suburban Area	1 km–10 km	50 m–150 m
Urban Area	100 m–1 km	50 m–150 m
Dense Urban Area	100 m–1 km	50 m–150 m
	CGI-TA	**A-GPS**
Rural Area	550 m	50 m–100 m
Suburban Area	550 m	30 m–100 m
Urban Area	100 m–550 m	10 m–20 m
Dense Urban Area	100 m–550 m	10 m–20 m
	E-CGI	**TOA**
Rural Area	250 m–8 km	85 m–100 m
Suburban Area	250 m–2.5 km	30 m–75 m
Urban Area	50 m–550 m	25 m–70 m
Indoor Urban Area	50 m–550 m	25 m–70 m

680 million customers globally by 2006. Predictions are that LBS will generate over $32 billion in Europe only by 2005. Numerous firms have already emerged to tap into this growing opportunity (Rao & Minakakis, 2003).

Competition Forces

Having established large customer bases, cellular-service providers will seek new ways to ensure customer loyalty by offering new types of services. Location-based services are the most promising type of these services (called value-added services). Some of the advantages for the cellular-service provider who offers location-based services are the following.

- Innovative service provision attracts new customers and enhances existing customers' loyalty to the provider.
- Revenues increase due to the traffic generated by the use of such services.
- There is the capability to introduce new revenue streams through deals with third-party companies (that specialize in LBS implementation and/or provision) in order to sell to these companies user location information.

Technology Forces

The first location-based services are expected to be offered or are already offered to mobile-phone users via WAP, SMS, or MMS (multimedia messaging service). Every mobile phone supports the SMS feature and most of them also support WAP and MMS. The cost for such a phone is negligible nowadays. This means that many customers can instantly make use of the location services provided. In addition, the evolution from GSM (global system for mobile commications) to general packet radio service (GPRS), which means a significant increase in the available bandwidth for data communication over mobile phones (from 9.6 Kbps to over 115 Kbps), also assists the provision of LBS, which in many cases can be bandwidth demanding (not to mention the introduction of UMTS [universal mobile telecommunications system] networks in many countries). Finally, new types of phones such as media phones and "communicators" have already entered the market, giving greater capabilities for displaying information (e.g., user interfaces enhanced with photos, buttons, etc., not only text based).

Regulatory Forces

In the USA, the Federal Communications Commission has issued a directive requiring the identification of the geographical origin of an emergency call made by a mobile-phone user. According to this directive, operators should be able to provide location information for every mobile subscriber who makes an emergency call with an accuracy of 100 m 67% of the time (GSM Association Services Expert Rapporteur Group [ASERG], 2000). A similar directive has been released for the European Union.

SUCCESS FACTORS AND RESEARCH CHALLENGES IN LBS

Despite the appealing idea of using user location information to provide highly personalized and intelligent services, there are certain challenges that should be addressed in order for LBS to succeed. We can divide these challenges into three categories, namely, technological challenges, ethical challenges, and business challenges.

The main technological challenge for LBS is the capability to create easy-to-use and satisfying services. There is much talk concerning what would be the most suitable user interface and type of service (pull or push) in terms of user satisfaction. For example, in the case of push-based services, a user is not required to manually issue queries in order to get the information he or she seeks. The system automatically informs him or her based

on the current location and a list of preferences listed in the user's profile. The problem is that in this way, user intent cannot be perfectly captured and the user may be frequently disturbed by out-of-context information. So, despite the easiness of usage (no or minimal interface), user satisfaction is not assured. On the other hand, in pull-based LBS, in which clients have to poll the server for updates, the users may experience difficulties in using these services because cell phones, PDAs, and wearable computers are less suitable for browsing and query-based information retrieval due to their limited input-device capabilities (Burcea & Jacobsen, 2003). All these restrictions along with the unpredictability in mobile environments (disconnections, frequent context differentiations, etc.) have to be taken very carefully into account when designing LBSs. Some of the implied requirements, as identified in Tsalgatidou, Veijalainen, Markkula, Katasonov, and Hadjiefthymiades (2003), are the following:

- A less intensive use of the mobile network and a minimal volume of transmitted data;
- The possibility of off-line operation;
- Simple and user-friendly interfaces, and limited and well-specified amounts of presented information content.

Therefore, it becomes apparent that LBS will not succeed in attracting users without implementing sophisticated techniques based on carefully designed interfaces and/or detailed knowledge of customer profiles, needs, and preferences. So, given existing technical limitations such as device capabilities, access speeds, and so forth combined with human limitations such as reduced consideration sets and the need for speed and convenience, in order for LBSs to succeed, they will need to deliver relevant, targeted, and timely information to consumers at the time and place of their choice (Rao & Minakakis, 2004).

Also, from a database perspective, LBSs raise critical challenges such as spatial and temporal query processing because the continuous movement of users or objects leads to the need for fast and frequent or continuous updates to the databases. Some of the most important database research challenges brought to the surface by LBS, as identified by Jensen, Friis-Christensen, Pedersen, Pfoser, Saltenis, and Tryfona (2001) and Saltenis and Jensen (2002), are the following.

- **Support for Nonstandard-Dimension Hierarchies:** In LBS, the geographical area may be divided into multidimensional regions following the pattern of network coverage. Until now, geographical-area representation models used by data warehouses were in the form of completely balanced trees (strict hierarchies), which cannot capture irregularities like those that frequently occur in mobile networks (e.g., the same region covered by more than one base station).
- **Support for Imprecision and Varying Precision:** Varying precision means that the location of the same user may be pinpointed with different accuracies depending on the positioning technology used while he or she is roaming from network to network. Imprecision means that the location data for the trace of a specific user may be incomplete (e.g., a user may have gone out of the network coverage or may have switched off the device for some time). So, varying precision and imprecision should be carefully handled by employing intelligent query-processing techniques, especially for queries on complete user traces.
- **Support for Movement Constraints and Transportation Networks:** Most of the time, users move on certain routes as defined by transportation networks (e.g., railways, roads, etc.) and their movement is blocked depending on the morphology of the land (e.g., mountains). The incorporation of such constraints in query resolution may offer increased positioning accuracy to LBS de-

spite the potentially low-accuracy positioning technology used.

- **Support for Spatial Data Mining on Vehicle Movement**
- **Support for Continuous Location Change in Query-Processing Techniques**

From an ethical point of view, a critical challenge is to protect user privacy. LBS can potentially intrude on customer privacy. The adoption of LBS is highly dependent on the successful confrontation of digital frauds, attempts of intrusion in customer databases with sensitive data and profiles, and the threat of unauthorized or uncontrolled resale of location information. As underlined in Rao and Minakakis (2003, p. 63), "LBS providers must alleviate consumer privacy fears by implementing secure network and encryption technologies to curb illegal activity and by developing clear communication strategies to interact with customers and allay their fears." It has also been shown that a privacy-intruding service (for example, an always-on tracking service), despite its usability, is not desirable by users since it does not allow them to switch it off whenever they want (Barkhuus & Dey, 2003). So when designing an LBS and in order for the service to be adopted, the provider should take into account very seriously the user's concerns on privacy.

From the point of view of the regulator of the telecommunications market, new laws have to be implemented. In order to protect user privacy, there are certain laws in the United States (Wireless Communications and Public Safety Act of 1999) and the European Union (Personal Data Processing and the Protection of Privacy in the Telecommunications Sector, 97/66/EU Directive) with direct references to the way location data should be handled. However, these laws have certain deficiencies and shortcomings, and there are ongoing efforts to achieve full legislative coverage of the LBS sector.

Finally, capitalizing on the promise of LBS requires developing sustainable and viable business models for offering such services. Unfortunately, until today there has been little effort on developing a framework with which to identify the most appropriate business models for the large variety of LBSs. The major obstacle for this arises from the fact that there is a multitude of players participating in the provision of such services forming a complex value network. The main categories under which these players are grouped are the following:

- Application developers and content providers;
- Service providers and network providers;
- Hardware manufacturers.

The roles of all these different actors or players are many times conflicting if not competitive, and fairness in revenue sharing is viewed differently buy each actor. In this context, it is difficult to determine which activities should be performed by which actor (e.g., should the network operator develop its own services or outsource them to more focused application providers) or to identify which actor should be the dominant one in the business model (i.e., the operator providing access to its customer base, the content or service provider offering the actual service, or the location-technology vendor offering the enabling positioning equipment).

FUTURE TRENDS

In the new era of 2.5G, 3G, and 4G, location-based services have been recognized as one of the fastest growing areas for novel service provision in the telecommunications sector with great revenue potential. What differentiates them from traditional services is their ability to offer highly personalized, context–sensitive, and timely information to users anytime, anywhere. However, they have not yet matured enough in order to provide the so-much-anticipated killer application mainly due to technical, business, and ethical challenges that have not yet been adequately addressed. All the

participants in the LBS-provision market should first understand and fix their roles within the value chain, then provide the essential guarantees for protecting user privacy, and finally develop new, intelligent ways to manipulate and present location information in order to increase user convenience and satisfaction.

CONCLUSION

We have discussed several aspects of the role of LBS in today's wireless industry. We primarily focused on technological, ethical, and business challenges imposed by LBS and provided directions for further research. User privacy protection, easy-to-use and context-aware service interfaces, sophisticated geospatial-data management techniques, and flexible business models have been identified as the most critical issues that the LBS industry should pay particular attention to in order for LBS to become a success.

REFERENCES

Barkhuus, L., & Dey, A. (2003). Location-based services for mobile telephony: A study of users' privacy concerns. *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003),* Zurich, Switzerland.

Burcea, I., & Jacobsen, H.-A. (2003). L-ToPSS: Push-oriented location-based services. In *Lecture notes in computer science: Vol. 2819. 4th VLDB Workshop on Technologies for E-Services (TES'03)* (pp. 131-142). Berlin, Germany: Springer.

Drane, C., Macnaughtan, M., & Scott, C. (Eds.). (1998, April). *IEEE Communications Magazine, 36*(4), 46-54.

GSM Association. (2003). *Location based services* (Version 3.1.0, Permanent Reference Document No. SE.23). Retrieved August 11, 2005, from http://www.3gpp.org/ftp/tsg_sa/WG2_Serv/TSGS1_07_SophiaAntipolis/Docs/S1-000069.doc

GSM Association Services Expert Rapporteur Group (GSM ASERG). (2000). *Location based services: Service requirements document* (Revision 1.0.0). Retrieved August 11, 2005, from http://www.gsmworld.com/documents/lbs/se23.pdf

Jensen, C. S., Friis-Christensen, A., Pedersen, T. B., Pfoser, D., Saltenis, S., & Tryfona, N. (2001). Location-based services: A database perspective. *Proceedings of the Eighth Scandinavian Research Conference on Geographical Information Science (ScanGIS2001),* Ås, Norway.

Poslad, S., Laamanen, H., Malaka, R., Nick, A., Buckle, P., & Zipf, A. (2001). CRUMPET: Creation of user-friendly mobile services personalised for tourism. *Proceedings of the Second International Conference on 3G Mobile Communication Technologies,* London.

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Rao, B., & Minakakis, L. (2004). Assessing the business impact of location based services. *Proceedings of the 37th IEEE Hawaii International Conference on System Sciences,* Waikoloa, Big Island, Hawaii.

Saltenis, S., & Jensen, S. (2002). Indexing of moving objects for location-based services. *Proceedings of the 18th International Conference on Data Engineering,* San Jose, California.

Swedberg, G. (1999). Ericsson's mobile location solution. *Ericsson Review, 4,* 214-221.

Third-Generation Partnership Project (3GPPP). (2004). *Technical specification group services and system aspects: "Location services (LCS)." Service description: Stage 1, Release 6* (Technical Specification Document No. 3GPP TS 22.071 V6.7.0). Retrieved August 11, 2005, from http://2gpp.

org/ftp/Specs/2004-06/Rel-6/22_series/22071-670.zip

Tsalgatidou, A., Veijalainen, J., Markkula, J., Katasonov, A., & Hadjiefthymiades, S. (2003). Mobile e-commerce and location-based services: Technology and requirements. *Proceedings of the Ninth Scandinavian Research Conference on Geographical Information Science,* Espoo, Finland.

Zipf, A. (2002). User-adaptive maps for location-based services (LBS) for tourism. *Proceedings of the Ninth International Conference for Information and Communication Technologies in Tourism (ENTER 2002),* Innsbruck, Austria.

KEY TERMS

A-GPS: The assisted Global Positioning System uses measurements from fixed GPS receivers scattered throughout the mobile network in order to assist a mobile phone in locating the available satellites and calculating its location.

AOA: The angle-of-arrival method measures the angle of a signal arriving at the antenna of a base station. The intersection of the projection of two calculated angles (from the antennas of two base stations) on the two-dimensional space reveals the location of the mobile phone.

CGI: Each base station in a cellular network has a unique ID that the mobile phone receives when entering the area of the base station. Cell global identity uses this unique ID in order to pinpoint the base station's area of coverage in which the mobile phone is located.

CGI-TA: Cell global identity with timing advance is a positioning method that uses the time needed for a signal to travel from the mobile phone to the base station to compute the distance between the phone and the mobile station. Along with the base station's ID, this method provides a rough estimation of the position of the phone in the base station's area of coverage.

E-OTD: The enhanced observed-time-difference method is similar to OTDOA without the need for base stations to be synchronized (additional elements are used that measure the real-time differences between base stations to correct the measurements).

MMS: The multimedia messaging service is a service giving the capability to a mobile-phone user to send a message containing any combination of images, video clips, text, and audio to another user.

OTDOA: Observed time difference of arrival is an alternative for the TOA method in which the mobile phone measures the time differences between signals from three or more base stations.

PDA: A personal digital assistant is a small, palm-sized mobile device with increased processing and viewing capabilities.

SMS: The short messaging service is a service giving the capability to a mobile-phone user to send a text message to another user.

TOA: The time-of-arrival positioning method is based on measuring the time needed by a signal transmitted by a mobile phone to reach three or more location-measurement units (LMUs). From these measurements, the distance between the phone and the LMU can be calculated as the radius of a circle with the LMU as its center. The intersection of three or more such circles gives the actual position of the mobile phone.

WAP: The wireless application protocol is a protocol for providing Internet-connectivity access to thin-client devices, such as mobile phones.

This work was previously published in the Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 716-721, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Section V
Organizational and Social Implications

This section includes a wide range of research pertaining to the social and organizational impact of mobile computing around the world. Chapters introducing this section analyze mobile virtual communities and consumer attitudes toward mobile marketing, while later contributions offer an extensive analysis of the accessibility of mobile applications and technologies. The inquiries and methods presented in this section offer insight into the implications of mobile computing at both a personal and organizational level, while also emphasizing potential areas of study within the discipline.

Chapter 5.1
Mobile Virtual Communities

Glauber Ferreira
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

Evandro Costa
Federal University of Alagoas, Brazil

INTRODUCTION

The establishment of collective relationships is a native characteristic of individuals. Living in a recluse way cannot be considered part of human nature. Individuals have always been organized in *communities* in which they establish relationships with other individuals, which usually live in one particular area. Proximity among individuals is one of the characteristics that motivate the creation of communities.

Communities are also created when individuals have common interests. Some examples are: religious communities, such as Catholic and Jewish ones; and communities that comprise people having the same job, such as scientific and medical communities. In these examples, the distance among individuals is not an obstacle to the creation of communities, since individuals have common interests. In spite of living in different places, members of these communities have periodic meetings in which collective relationships are established.

The popularization of the Internet after the 1990s along with the well established use of personal computers have allowed the creation of a new form of community, the well known *virtual communities*. They have enabled individuals to communicate through e-mail, forums, instant messaging, and videoconference. People living in different countries have interacted and communicated through Internet enabled personal computers. Distance learning and software users groups such as Linux users and Java developers are some examples of relationships that have been improved by virtual communities.

Mobile virtual communities are the most recent advance in the establishment of collective relationships, mainly due to the progress in mobile devices and wireless communication technologies. Connectivity among wireless mobile devices enables individuals to exchange information and knowledge, anytime and anywhere. These communities are created in an ad hoc way: individuals with common profiles, carrying connected mobile devices, can constitute a community and access/provide information according to their authorization degree. There are various applications of mobile virtual communities, such as workflow management, mobile learning, healthcare communities, personal assistants in academic conferences, and applications for communication among students on campus, among others.

This article introduces the field of mobile virtual communities, describing the main issues that have culminated in the creation of this research area such as the Internet, personal computers, mobile devices and wireless communication technologies. Applications domains of mobile virtual communities and works that support the development of these applications are also presented.

MOBILE VIRTUAL COMMUNITIES

In the book "The Virtual Community," Rheingold (1993) defines virtual communities as social groups whose interaction is mediated by computers. These communities increase the establishment of collective relationships among individuals, since computer-mediated interaction allows creating communities constituted by geographically dispersed people. In order to support interaction among members of these communities, various computational tools are used, such as e-mail, forums, whiteboard, audio/video conference sessions, and instant messaging, among others.

Rheingold (2003) enumerates some characteristics of virtual communities. He defines virtual communities as:

1. Organized around affinities, shared interests, bringing together people who did not necessarily know each other before meeting online.
2. Many-to-many media. Unlike few-to-many (broadcast) or one–to-one (telephone or SMS) media, virtual communities enable groups of people to communicate with many others.
3. Text-based, evolving into text plus graphics-based communications. For decades, online communities were built with nothing more than unformatted text. Web-based media brought inline graphics, animations, video, sounds, formatted text, and links into the conversation.
4. Relatively uncoupled from face–to-face social life in geographic communities. People communicating worldwide about shared interests most often do not live close enough to meet regularly face-to-face.

It is important to point out the relevance of item one for characterizing virtual communities. The absence of shared interests among participants makes unfeasible the constitution of these communities. The similarity among the preferences of individuals is responsible for the establishment of these groups.

Virtual Communities Evolves into Mobile Virtual Communities

The presence of various *portable* computational devices in our everyday lives is incontestable: mobile phones, notebooks, handhelds, smartphones, tablet PCs, and so forth. All of these devices allow the *connectivity* among their owners through wireless technologies such as Wi-Fi, GPRS, WAP, and Bluetooth. This scenario of mobility

and connectivity has increased the establishment of interactions among individuals, allowing the emergence of mobile virtual communities.

Fremaux (2000) considers mobile communities the natural evolution of virtual communities. Mobile communities can be seen as virtual communities to which mobile services are added. In what follows, two important differences between mobile communities and "traditional" Web-based virtual communities are presented (Fremuth, Tasch, & Fränkle, 2003, p. 2):

- Mobile communities can be accessed by mobile devices like mobile phones, smart phones and PDAs. This could lead to a more spontaneous communication in a community.
- Mobile community platforms offer enhanced communication services for their users, made possible by the 2,5 and 3rd generation of mobile networks: *ubiquitous access*, allowing an anytime-anywhere connection to their communities; and *location based services*, through the use of positioning technologies (Hazas, Scott, & Krumm, 2004) such as infrared, GPS, Bluetooth, and Wi-Fi.

Characteristics of Mobile Virtual Communities

In a general way, mobile virtual communities present the following characteristics (Rheingold, 2003):

- Many-to-many, desktop and mobile, always on. Virtual communities and the resources of the Internet are instantly available to people and their software agents wherever people are located—at their desks, in transit, at home.
- Used to coordinate actions of groups in geographic spaces—teenagers swarm in malls, young adults club-hop, activists mobilize on the street.
- Game environments, social arenas, artistic media, business tools, political weapons—like other virtual community media, mobile virtual communities will start with young people as means for entertainment and light social interaction, then diffuse into other institutions.

Two more characteristics are present in applications for mobile virtual communities. (1) These applications are deployed in different *varieties of computational devices*, with different memory size, processing power, and display capability. In this way, such diversity should be considered during the development of software for this domain. (2) This apparent problem is minimized by the use of *information regarding the context* in which individuals are situated. Through such information, applications can be adapted based on the preferences of the user, on the configuration of the device or on the location of the individual.

APPLICATION DOMAINS

Some application domains in which mobile virtual communities have been applied are presented in this section. Characteristics of each application domain are described in order to justify using mobile virtual communities to develop applications for these domains.

Healthcare Communities

Healthcare can be defined as "the prevention, treatment, and management of illness and the preservation of mental and physical well-being through the services offered by the medical and allied health professions" (Dictionary.com, 2006). Beside these professionals, other institutions also offer these services, such as hospitals, non-gov-

ernmental organizations, insurance companies, and so forth (Leimeister, Daum, & Krcmar, 2002). These services are mainly used by patients, who interact with health professionals and institutions that constitute the healthcare systems.

According to some researches, the demand of patients for information increases when they receive diagnosis or treatment (Sheppherd, Charmock, & Gann, 1999). Researches also revealed that patients also need to communicate with other patients in order to exchange experience and receive emotional support, mainly when they are attacked by bad diseases.

Patients participate in activities of self-help groups in order to get information and establish interactions. These groups are an example of communities, since their members establish relationships with each other and have common interests—the discussion about diseases.

Although the participation of patients in these groups is important for helping the treatment of diseases, some problems can complicate the integration of patients in these communities. One of these problems is the incompatibility between the schedules of patients and the meetings of the self-help groups. Another possible problem is the difficulty that patients can have in moving to the meeting place of the self-help groups (Leimeister, Daum, & Krcmar, 2002).

The previously mentioned problems are solved when patients establish collective relationships through applications of mobile virtual communities. These applications enable patients to communicate anytime and anywhere. Besides the resolution of these problems, the use of these applications enables patients to receive contextual information, according to their location, such as the address of the nearest pharmacy/doctor and notifications of the presence of other patients that are located in a near area. It is important to point out that these communities should not substitute the self-help groups; on the contrary, these communities should aid the self-help groups through the expansion of the methods used for interacting.

Mobile Learning

Recent advances in the manufacturing of wireless mobile devices have allowed the expansion of the methods used for distance teaching and learning. Through mobile learning applications, individuals sharing interest in learning any subject can establish communities in which they interact in order to acquire knowledge. Interactions among these individuals can occur regardless of their location, for example, on the bus, in the waiting room, or in the queue for tickets.

Common methods of distance teaching and learning, such as corporate universities and distance undergraduate course, can be benefited from mobile learning applications. The application of the concept of mobility in learning communities constituted at corporate universities enables to accomplish the training of employees not only within the organization limits, but also during the activities performed outside of the organization. Distance undergraduate courses are usually taken by people who have difficulty in adapting their time to the inflexible timetable of learning institutions. To address this problem, theses courses can use the concept of mobility in order to allow students to perform the learning activities regardless of their location, adjusting the time to suit their convenience.

Traditional learning activities can also be benefited from mobile learning applications. The following scenario, presented by Alexander (2004), describes some possible experiences.

For example, suppose a first-year student sees the recent film Master and Commander and becomes interested in the world of eighteenth-century sailing. With no guidance, the student might hit Amazon.com for other novels by Patrick O'Brian, watch a History Channel program about sailing, or conduct a Google search and find a few related

Web pages. Or instead, the college could set up an environment in which the student finds that one history professor regularly teaches 'the great age of sail' in several classes, has Web pages on the 1790s naval wars, and might answer an e-mail or office-hours query; that the library has digital and print resources ready at hand; that several other students share this curiosity and chat about it with IM; and that a staff member sailed on a rebuilt eighteenth-century vessel last summer and would be delighted to discuss the experience (pp. 32-33).

Although all these experiences enrich learning activities, mobile learning applications should not replace attended teaching activities. Even in distance undergraduate courses, a number of lessons must be taken in attended meetings.

Workflow Management

Workflow management systems are frequently used for modeling, monitoring, and controlling the coordinated execution of activities performed in various contexts (Dias, Casanova, & Carvalho, 2003). This application domain is characterized by the frequent interactions among individuals that are responsible for the activities, since these individuals need to collaborate in order to perform the activities.

The tracking of these activities can now be done directly in the place in which they are performed, due to the decrease in the acquisition costs of mobile devices. In this way, tasks can be coordinated even by teams located in different areas and whose members are constantly moving.

Consider, for example, a workflow management system that aids the execution of emergency plans in an oil company, a pipeline operation company and gas distribution companies (Dias, Casanova, & Carvalho, 2003). In this system, mobile devices are used by the emergency team members in order to register the execution of operations, allowing that other correlated operations could be performed after the conclusion of a main operation. For example, in the case of an oil spill, contention barriers must be launched, the oil pumping must be stopped, and cleaning procedures must be executed. Besides the help in the coordination of activities, team members have access to various information that are needed to execute the operations, such as the material resources that should be used in the execution of each procedure, maps, list of authorities to contact, and so forth.

SUPPORTING DEVELOPMENT OF APPLICATIONS

As in other application domains, the development of software for mobile virtual communities involves characteristics that are common in the majority of applications. In this way, it is important to use APIs, software frameworks, and infrastructures that ease the building of these applications. In this section, we present some works that aim to support the development of applications for mobile virtual communities.

In Rakotonirainy, Loke, and Zaslavsky (2000), a multi-agent approach for a high-level model of mobile and distributed systems is introduced, in terms of mobile virtual communities. This model uses concepts from the Reference Model for Open Distributed Processing (RM-ODP) (International Organization for Standardization, 2002) and from a CORBA-like component model (Object Management Group, 1998). Mobile communities are modeled by the composition of *roles* formed to meet an *objective*. Such an objective is expressed in a *policy*, which is a set of rules related to the *activities* performed by the community. The roles constituting the community are described with *component interfaces* that define the *interactions* of the community. The main focus of this work is on the modeling and specification of mobile virtual communities. It does not provide a prototype in

order to help the implementation of applications for mobile virtual communities.

MOOsburg (Carroll, Rosson, Isenhour, Ganoe, Dunlap, Fogarty, et al., 2001) is a community-oriented collaborative environment that models the town of Blacksburg. It provides a range of collaborative tools that provide access to shared content such as whiteboards, message boards, and so forth. In order to move towards a wireless virtual community, MOOsburg++ (Farooq, Isenhour, Carroll, & Rosson, 2002) has been proposed. It is an extension to MOOsburg that allows accessibility from mobile devices such as cellular phones, pagers, and PDAs. MOOsburg++ provides synchronous and asynchronous interaction with people and data. Each piece of data (people, places, things, and collaborative objects) in the environment is represented by a Java object and replicated across all interested clients using the Content Object Replication Kit (CORK) (Isenhour, Rosson, & Carroll, 2001), a toolkit for building Web-accessible interactive distributed applications.

ToothAgent (Bryl, Giorgini, & Fante, 2005) is a prototype of a multi-agent system for virtual communities support. This work proposes a general architecture with independent servers where multi-agent platforms can be installed and where agents can act on behalf of their users. Each server provides services related to the geographical area in which it is located (e.g., a server inside a university could offer the service of selling and buying text-books), and users can contact their personal agents using their Bluetooth-enabled mobile phones or PDAs. Such an architecture is domain-independent since it does not depend on the specific services offered by the server.

CONCLUSION AND FUTURE TRENDS

In this article, we presented a timeline of events that culminate with the constitution of mobile virtual communities. We discussed the first communities organized around common interests, the creation of virtual communities through the popularization of personal computers and the Internet after the 1990s, and finally the constitution of mobile virtual communities through the dissemination of portable computational devices with wireless access.

Some application domains concerned with mobile virtual communities were described. Characteristics and examples of applications in these domains were presented. Also, we discussed some works that support the development of applications for mobile virtual communities, highlighting their strengths and weaknesses.

The mentioned works aid the building of applications for mobile virtual communities since software is not developed from scratch. However, they do not provide a good support for services inherent in mobile virtual community such as authentication of individuals in communities, representation of the interests of individuals, algorithms for identifying individuals which have common interests, and control of access to information available in communities. In this way, the development of a software infrastructure that addresses all these issues is highly desirable for improving the building of applications for mobile virtual communities.

REFERENCES

Alexander, B. (2004). Going nomadic: Mobile learning in higher education. *EDUCAUSE Review*, *39*(5), 28-35.

Bryl, V., Giorgini, P., & Fante, S. (2005). *ToothAgent: A multi-agent system for virtual communities support.* Trento, Italy: University of Trento, Department of Information and Communication Technology.

Carroll, J., Rosson, M., Isenhour, P., Ganoe, C., Dunlap, D., Fogarty, J., et al. (2001). Designing

our town: MOOsburg. *International Journal of Human-Computer Studies*, *54*(5), 725-751.

Dias, F., Casanova, M., & Carvalho, M. (2003). Workflow execution in disconnected environments. In *Proceedings of the XVIII Simpósio Brasileiro de Banco de Dados* (pp. 229-239).

Dictionary.com. (2006). Retrieved, from http://dictionary.reference.com/

Farooq, U., Isenhour, P., Carroll, J., & Rosson, M. (2002). MOOsburg++: Moving towards a wireless virtual community. In *Proceedings of the 2002 International Conference on Wireless Networks*. CSREA Press.

Fremaux, D. (2000). *The next VAS generation*. Telecommunications Online. Retrieved from http://horizontest.bvdep.com/telecom/default.asp?journalid=2&func=articles&page=0009i31&year=2000&month=9

Fremuth, N., Tasch, A., & Fränkle, M. (2003). Mobile communities—New business opportunities for mobile network operators? In *Proceedings of the 2nd Interdisciplinary World Congress on Mass Customization and Personalization*.

International Organization for Standardization. (2002). *ISO/IEC 15414:2002 Information technology—Open distributed processing—Reference model—Enterprise language*. Geneva, Switzerland: International Organization for Standardization.

Isenhour, P., Rosson, M., & Carroll, J. (2001). Supporting interactive collaboration on the Web with CORK. *Interacting with Computers*, *13*(6), 655-676.

Hazas, M., Scott, J., & Krumm, J. (2004). Location-aware computing comes of age. *Computer*, *37*(2), 95-97.

Leimeister, J., Daum, M., & Krcmar, H. (2002). Mobile virtual communities: An approach to community engineering for cancer patients. In *Proceedings of the 10th European Conference on Information Systems* (pp. 1626-1637).

Object Management Group. (1998). *OMG TC document orbos/98-10-18*. Needham, MA: Object Management Group.

Rakotonirainy, A., Loke, S., & Zaslavsky, A. (2000). Towards multi-agent support for open mobile virtual communities. In *Proceedings of the International Conference on Artificial Intelligence* (Vol. I, pp. 127-133). Las Vegas, NV: CSREA Press.

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier*. Boston, MA: Addison-Wesley.

Rheingold, H. (2003). *Mobile virtual communities*. TheFeature.com Archives. Retrieved from http://www.thefeaturearchives.com/topic/Culture/Mobile_Virtual_Communities .html

Sheppherd, S., Charmock, D., & Gann, B. (1999). Helping patients access high quality health information. *British Medical Journal, 319,* 764-766.

KEY TERMS

Communities: A set of individuals living in a particular area and/or sharing common interests. Communities promote establishment of collective relationships among individuals.

Connectivity: Connectivity in this context is the capability of mobile and wireless computational devices to operate in network environments, allowing the interaction of their owners.

Interaction: Interaction in this context is the communication among individuals that constitute communities, virtual communities, or mobile virtual communities.

Mobile Applications: Mobile applications are targeted at mobile computational devices such as

mobile phones, handhelds, smartphones, tablet PCs, PDAs, and so forth. These applications enable users to access their services regardless of their location, since they carry their devices.

Mobile Virtual Communities: Communities constituted by individuals using mobile and wireless computational devices. These communities enable individuals to exchange information and knowledge, anytime and anywhere.

Virtual Communities: Communities enabling individuals to interact through computer-based tools, such as e-mails, forums, whiteboards, instant messaging, audio/video conference sessions, and so forth.

Wireless Applications: Applications that communicate through wireless technologies such as infrared, Bluetooth, Wi-Fi, GPRS, and so forth. These technologies ease the use of these applications since they avoid cumbersome cables.

This work was previously published in Encyclopedia of Networked and Virtual Organizations, edited by G. Putnik and M. Cunha, pp. 944-949, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.2
Mobile Virtual Communities of Commuters

Jalal Kawash
American University of Sharjah, UAE

Christo El Morr
York University, Canada

Hamza Taha
American University of Sharjah, UAE

Wissam Charaf
American University of Sharjah, UAE

INTRODUCTION

Commuting forms an integral part of our lives, whether we are commuting for leisure or business. The use of location-based services and mobile computing has potentials to improve commuting experience and awareness. For instance, printed bus schedules have been only recently complemented with online systems to provide bus timing information for the community of public transport commuters. Commuters can nowadays inquire about bus timings by the use of telephony systems and the Internet. However, the information provided to users is statically produced, just like the still in-use old fashion bus route tables, and does not take into consideration delays and cancellations. The next step in the evolution of these schedules must produce live information, track bus movements, and alert commuters of bus arrivals and timings. The experience of commuting using taxis can also be improved beyond the use of telephony, while the most common way of asking for a taxi continues to be by hand waiving. Such improvements are more crucial for commuters that are not completely aware of their surrounding environment, such as tourists and business visitors.

This article envisions the formation of networked organizations of commuters, through the use of mobile and location-based services. We discuss scenarios and use cases of such organizations and propose an example software implementation for the supporting services.

BACKGROUND

Worldwide, the adoption of smart wireless technologies is taking place at a large scale. For example, about half a billion users carry handheld phones that can run Java and in 2005, mobile manufacturers shipped about 400 million Java enabled phones (Mobile Monday, 2005). There are about 150 wireless operators supporting Java and there are 300 to 400 different phone models that can run Java (Mobile Monday, 2005). This huge and rapid adoption of Java-enabled mobile devices is not fully exploited by the industry, with probably the exception of mobile gaming industry. It is the authors' conjecture that software tools that support the formation of mobile networked virtual organizations and communities is a strong candidate for such exploitation.

Virtual Communities

As early as 1999, Palloff and Pratt (1999) realized the need to redefine the meaning of a "community" due the emergence of the Internet. Preece (2000) defined an online community to consist of: socially interacting *people*, performing special roles or satisfying their needs; a *purpose*, which is the reason behind the community; *policies* to govern people interaction; and *computer systems* that support social interaction.

The proliferation of mobile devices and wireless technologies gave users the ability to practice their roles in online communities while they are on the move. Mobility has tremendous effects on the nature of the tools that enable mobile user participation in a community, such as the human computer interaction (HCI) requirements for mobile devices. Kristofferson and Ljungberg (1999) noted that mobility enforces constraints on HCI so that new interaction styles, characterized by little visual interaction, should be created. Mobile users work in a more context-sensitive environment than classical stationary Internet users. Dix et al. (2000) argued that the participation of a mobile user in a community has an impact on the set of awareness tools that should be used in the community. Mobile users can act to a large extent differently from stationary users.

Some work like the one conducted by Grather and Prinz (2001) focused on the cooperation requirements in a mobile Web-based community and demonstrated the importance of metaphors during cooperation. Luff and Heath (1998) indicated that taking into account the mobility factor in collaboration may result in more innovative approaches to designing collaborative technologies and mobile devices. Few researchers like Watanabe et al. (2000) described the use of mobile phones for awareness support between friends and suggested that collaboration awareness stimulates the need for communication.

For the purpose of this work, a *mobile virtual community* (MVC) consists of user members, the majority of which are practicing their roles using mobile devices, purpose, policies, and technologies supporting interaction among members (El Morr & Kawash, 2007).

MOBILE VIRTUAL COMMUNITIES OF COMMUTERS

Scenarios

This section presents two scenarios, which illustrate the formation of MVCs of commuters, the type of users, and the required supporting technologies. In the following sections, we will see that these MVCs can be captured by a simple collaboration model and an example software implementation that enables such MVCs.

Scenario 1: Live Bus Schedule

Maria is a tourist that has just arrived to Toronto and she decides to visit the CN Tower. She notices the problem of traffic jams in Toronto and she doubts the accuracy of the printed bus schedules.

So, she joins Toronto Bus Users (TBU), an MVC whose purpose is the enhancement of the awareness of bus timings and schedules. TBU allows Maria to specify on her mobile phone a destination in the city and the system shows her a map of her current position with the locations of the nearest bus stops that serve the required destination. TBU also gives her accurate estimation of the times for the next buses serving her destination from the nearby stops. Maria finds that the next bus is expected in 15 minutes, so she decides to spend some time in the nearby gift shop. She asks the system to send her a reminder two minutes prior to the bus arrival. While she is shopping, she receives a notification stating that there will be an unexpected delay of three minutes in the bus arrival, possibly due to a traffic jam. Two minutes before the bus arrives, Maria receives a reminder confirming the arrival time.

Scenario 2: Taxi Now

Don is in Dubai to attend a conference and conduct few business meetings on the side. He is unfamiliar with the operation and procedure of the public transport system in Dubai, so a taxi is his preferred choice for commuting. It is important that Don makes use of his time efficiently during his short visit. He registers in the Dubai Taxi Now (TNow) service, which allows him to ask for a taxi ahead of time specifying the date, time, and location that he requires. Don can receive early reminders of the taxi arrival. This morning, Don asked for a taxi to take him from his hotel to his Jumaira Beach meeting at 10:00 a.m. At 9:50 a.m., Don received a notification on his mobile phone that the Taxi will be two minutes late because traffic is affected by road construction. At 9:52 Don left the room to the hotel lobby, where he used a Wi-Fi connection to TNow to see the taxi cab location on an interactive map.

In the afternoon, Don was having a stroll checking a few shops. He got carried away with shop hopping and lost track of his location. Don used TNow again and invoked the "nearest taxi" function. TNow locates Don and a handful of non-engaged taxi cars in the area. One taxi accepted the request and was given directions for Don's location. Don was informed that a taxi is on the way with an estimated arrival time. While waiting, he requested an interactive map where he can track the approaching taxi.

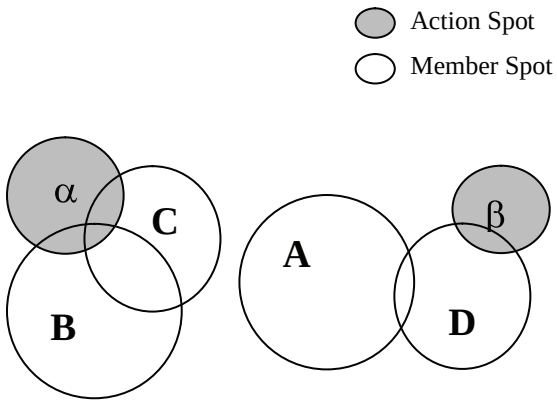
Collaboration Model

The general model of Kawash et al. (2007) can easily capture these scenarios. In this model, a community consists of producer and consumer member spots, and action spots. Producers announce action spots and consumers make use of these announcements. Action spots are community purpose dependant. In the bus schedule scenario, the action spots correspond to bus stop locations; producers are bus operators or some service that tracks and locates buses; and consumers are those community members that make use of the produced announcements. Announcements of action spots, announcing the arrival of a bus to a certain bus stop, are produced upon the departure of a bus from the preceding bus stops. In the TNow scenario, an action spot is the location of a commuter who is requesting a taxi service; commuters are producers; and taxi operators are the consumers.

In this simple model, the community software service tracks users and generates snapshots of the community, where action spots and member spots are represented as circles on the plane. The size of the action spot circle is application dependent. The size of the member spot circle defines the required user level of awareness and is specified by the user or the application (Kawash et al., 2007).

Figure 1 shows a snapshot of a regular activity in some community. In this snapshot, α and β are action spots and A, B, C and D are member spots. Action spot α falls within the level of awareness for both members B and C. Action spot members

Figure 1. A community snapshot based on the model of Kawash et al. (2007)



are candidates for notification of the presence of *α*. Member D should also be made aware of action spot *β*. No action spots fall within the awareness level of member A, so for this snapshot, A does not have to be made aware of any action spot activities.

Sociability

A virtual community must also promote awareness and its success depends on its degree of sociability. In this section, we discuss the awareness and sociability factors for MVCs of commuters. The sociability factors include: lurking, critical mass, policies and governance, personalization, privacy, and user differences.

Awareness has a direct impact on group sociability and social activity. MVCs like TBU aim at increasing commuters' contextual awareness of public transport systems. In the case of TNow, the aim is increasing the awareness of taxi operators. Yet, both examples deliver social benefit to end users motivating social participation. Active participation and persistence are crucial for the success of the community. The communities that are discussed in the previous scenarios may not always require continuous involvement in practicing community roles. The purpose of communities of commuters is long-lived; however, the constituents of these communities sometimes can be extremely volatile, allowing the member body to constantly change over time (as in the case of MVCs directed towards tourists). Yet, the success of these loosely coupled communities crucially depends on active participation from, at the time, constituents of the community.

One possible burden on the sociability of a community is the presence of *lurking*. Lurkers in other forms of virtual communities can form 80% of the community population (Sproull & Faraj, 1997; Nonnecke & Preece, 2000). The MVCs we discuss here can never have the problem of lurkers. There are no opinions when it comes to action spots. This objectivity, which stem from the nature of the model used (Kawash et. al., 2007), makes the type of MVCs presented here, substantially different from other virtual communities.

MVCs for commuters are expected to be restricted to small geographical areas. Members of the same community are likely to be present (permanently or otherwise) in the same city or town, and therefore the size of the community is anticipated to be relatively smaller than worldwide scale virtual communities. Nevertheless, because of the loose social aspects in the above communities, we expect their size to be large enough to reach a *critical mass* that brings the communities to life and make the business model profitable; indeed, less constraints are set for a person to join the community, and therefore users can join regardless of how much they share common social "goals" with other users.

The virtual communities discussed here have simple *policies and governance*. Privacy, trust, and codes of practice are essential but can be easily achieved. The only exposed private information is the physical position of a member, which users are normally willing to expose to a

secure trustworthy system. The codes of practice are also expected to be simple. For instance, it may suffice in a MVC for commuting by taxi to refrain from falsely requesting taxis. This can be enforced through member rating mechanism, which can lead to forcing a member to leave the community upon consistent misconduct.

TBU provides for *personalization***.** The member spot area (radius of the member spot) is a flexible parameter set by the user to draw the line between awareness and disturbance. Members in TBU and TNow can choose to be off-line or invisible, promoting *privacy.*

The *nature of users* can determine the nature of virtual communities. Users differ in their gender, personality, culture, and age. These differences are highly invisible in the MVCs discussed in this article.

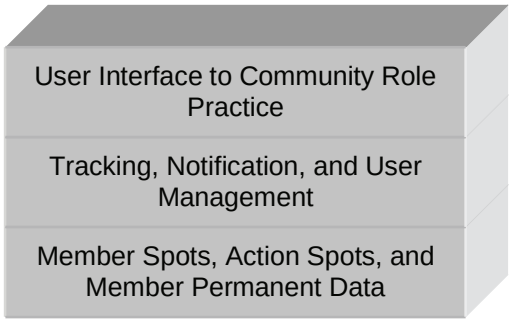
The traVcom Service

The traVcom service is a software implementation that enables MVCs for commuters under the model of Kawash et al. (2007). In particular, it can be used to enable the TBU and TNow services, discussed in the scenarios. This section gives an overview of traVcom; more details on the implementation are elsewhere (Kawash et al., 2005).

traVcom uses a three-tier architecture as depicted in Figure 2. The bottom layer is responsible for storing snapshots of member and action spots and permanent user information, such as personal, mobile device profile, billing information, and preferences. The second layer generates snapshots and analyzes them to generate appropriate notifications. This layer also provides the functionality to manage user permanent information.

The top layer is the gateway to community practice. Here, users can receive notifications, enquire about bus schedules and area maps. Consumers can participate in this community with a range of handheld devices.

Figure 2. Three-tier architecture for traVcom

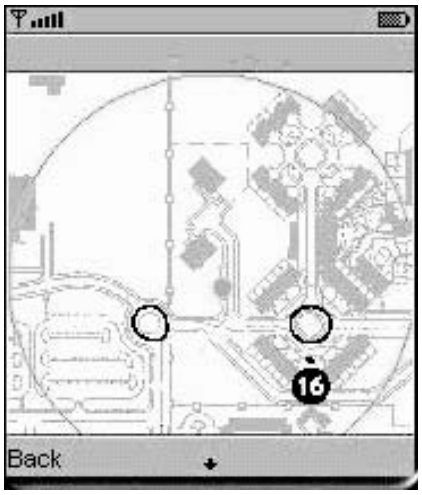


Supported Functions

The mobile user interface is developed as a J2ME MIDlet application. There are several mobile services available for users. We only discuss a few of these here.

- **Position location service:** This service enables users to locate their current positions and locate nearby action spots (bus stop or commuter locations, depending on the MVC) on the map. Figure 3 shows a MIDlet screenshot of the result of invoking this service for a service similar to TNow. It shows the member spot as a small (light) dot with the member's level of awareness as a larger circle. The small dark circle is an action spot for bus stop 16. This can be also used by a commuter to track the bus or taxi as they are approaching his/her position.
- **Arrival-time service:** This service is used to inquire about the time of the arrival of a bus for a specific bus stop or the arrival time of a taxi. The service can work by providing the system with the bus stop number or the location of a member and the system will

Figure 3. Position location service output



estimate the arrival time at the stop or the member's location.

- **Awareness level service:** A bus commuter can change the level of his/her required awareness: the radius within which a commuter needs to be aware of bus stops and bus activity. This service is only available for bus commuters. The level of awareness for taxi drivers is pre-set based on external factors.
- **Mark action spots service:** This function is issued by bus operators or taxi commuters. It is invoked every time a bus departs from a bus stop, marking the next stop, or when a commuter requests a taxi.

Community Snapshot Generation

The traVcom system interacts with the Gateway Mobile Positioning Centre (GMPC) (Programmer's Guide JML API, 2004) using Java Mobile Location (JML) API. traVcom is responsible for handling and establishing all connections between a member's mobile device and the GMPC. Interactions between traVcom and GMPC include requests to determine the current location of the member and whether the member is eligible for notification. Interactions also include a member requesting to be informed about his/her current location and the nearest bus stops within a specified range or taxi tracking.

Community snapshots are generated by determining the locations, member and action spots. Location determination in traVcom is performed as in any typical mobile positioning system (MPS). The traVcom system uses a standardized positioning request, called mobile terminated location request (MT-LR), to locate a specific mobile member at a specific time. A typical MT-LR application works as follows: The location-based application sends a request to GMPC, which authenticates the request and forwards the request to the Serving Mobile Positioning Center (SMPC). Depending on the information provided by mobile network and the GMPC, the SMPC calculates the position for the requested mobile member. A response is sent back to the GMPC which forwards it back to the location-based application.

FUTURE TRENDS

Mobile computing technologies are going through constant improvements. GPS-enabled devices are becoming more affordable and Wi-Fi PDAs and telephones may become the norm in the not too distant future. Location determination technologies (LDT) form an alternative to GPS are being investigated and developed (Jana et al., 2001; Zeimpekis et al., 2002; Wang et al., 2002). Such developments can only enrich MVCs and widen their adoption.

The implementation described in this article can be readily deployed and used without the need to wait for further technological development. For instance, traVcom can be used from rudimentary handheld devices where users can

key in the bus stop number. Unlike existing telephony applications, the times generated by traVcom are dynamically computed rather than returned from a preset schedule. Also, in the taxi commuting scenario, a basic mobile device can be used to enter an address or an intersection of two roads.

However, we realize that further development and adoption of more advanced technological features such as LDT makes an MVCs member's experience extremely richer. This in turn will have its positive impact on the adoption and success of such communities. For instance, the current implementation of traVcom requires bus operators to "mark" hot spots using handheld devices. Nevertheless, when it becomes feasible to equip all busses by tracking mechanism, such hot spots can be implicitly generated by traVcom without any input from the operators.

The current push-style notifications in traVcom are implemented using SMS. Admitting the shortcomings of SMS, new more efficient and reliable notification mechanisms can be an asset to implementations like traVcom.

CONCLUSION

In this article, we have presented MVCs for commuters and an implementation for the supporting services. Such services can be deployed using existing technologies and will lead to the formation of such MVCs. However, the scenarios that are entertained in the article show that advancing exiting technologies or widening the adoption of currently existing technologies can enrich commuters' experience to an extent that would make joining an MVC of commuters inevitable.

The MVCs presented here are based on a very simple, yet powerful model of collaboration. The simplicity of the model contributes to convenient user experience, which leads to highly usable services. As discussed in the article, the sociability factors that normally hinder the formation or continuity of a virtual community are either absent or minimal in the MVCs for commuters discussed here, due to the nature of these communities. For instance, such communities completely hide gender, age, and racial differences; they have no place for lurkers; and they intrinsically provide privacy.

Web-like and e-mail-like applications for the wireless world are so much needed. It is the authors' conjecture that mobile virtual communities will emulate in wireless networks the impact of the Web and e-mail in wired networks.

REFERENCES

Dix, A., Rodden, T., Davies, N., Trevor, J., Friday, A., & Palfreyman, K. (2000). Exploiting space and location as a design framework for interactive mobile systems. *ACM Transactions on HCI, 7,* 285-321.

El-Morr, C., & Kawash, J. (2007). Mobile virtual communities research: A synthesis of current trends and a look at future perspectives. *International Journal of Web Based Communities*, *3*(4), 361-403.

Grather, W., & Prinz, W. (2001). *The social web cockpit: Support for virtual communities.* Paper presented at the Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work.

Jana, R., Johnson, T., Muthukrishnan, S., & Vitaletti, A. (2001). *Location-based services in a wireless WAN using cellular digital packet data (CDPD).* Paper presented at the Proceedings of the 2nd ACM international workshop on Data engineering for wireless and mobile access.

Kawash, J., El-Morr, C., Charaf, W., & Taha, H. (2005). Building mobile virtual communities for public transport awareness. In *Proceedings of the IEE Mobility Conference 2005 (2nd International*

Conference on Mobile Technology, Applications, and Systems). Guangzhu, China.

Kawash, J., El-Morr, C. , & Itani, M. (2007). A novel collaboration model for mobile virtual communities. *International Journal of Web Based Communities, 3*(4), 427-447.

Kristoffersen, S., & Ljungberg, F. (1999). *"Making place" to make IT work: Empirical explorations of HCI for mobile CSCW.* Paper presented at the Proceedings of the international ACM SIGGROUP conference on Supporting group work.

Luff, P., & Heath, C. (1998). *Mobility in collaboration*. Paper presented at the Proceedings of the 1998 ACM conference on Computer supported cooperative work.

Mobile Monday. (2005). Retrieved from http://www.mobilemonday.com

Nonnecke, B., & Preece, J. (2000). *Lurker demographics: Counting the silent.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Palloff, R., & Pratt, K. (1999). *Building learning communities in cyberspace*. USA: Jossey-Bass Publishers.

Preece, J. (2000). *Online communities: Designing usability supporting sociability.* USA: John Wiley & Sons Ltd.

Programmer's Guide JML API 1.1, [*MPS SDK directory*]. (2004). Retrieved from http://api/doc/mps_sdk-api-programmers_guide.pdf

Sproull, L., & Faraj, S. (1997). The Net as a social technology. In S. Kiesler (Ed.), *Culture of the Internet.* Lawrence Erlbaum Associates.

Wang, S., Green, M., & Malkawi, M. (2002). *Mobile positioning technologies and location services.*

Watanabe, S., Kakuta, J., Mitsuoka, M., & Okuyama, S. (2000). *A field experiment on the communication of awareness-support memos among friends with mobile phones.* Paper presented at the Proceedings of the 2000 ACM conference on Computer supported cooperative work.

Zeimpekis, V., Giaglis, G., & Lekakos, G. (2003). A taxonomy of indoor and outdoor positioning techniques for mobile location services. *ACM SIGecom Exchanges* (Vol. 3, pp. 19-27), ACM Press.

KEY TERMS

Community: A community consists of interacting members, playing roles in order to satisfy a social purpose, and governed by social policies.

Virtual Community: A community in which the meeting place is virtual, such as the Internet.

Mobile Virtual Community: A virtual community whose members play their roles while on the move, using mobile and handheld devices, not necessarily hooked to a wired network.

Commuter: A public transport system user.

Mobile Virtual Community of Commuters: A mobile virtual community whose members are commuters and its social purpose is enhancing users' awareness of a public transport system.

Sociability: Sociability represents the level of social interaction support in a community. In a virtual community this is supported by the community purpose, governance structure, roles, policies, and by a set of tools for community members' awareness, privacy, and personalization.

traVcom: A software system that provides services for enabling and supporting the formation of mobile virtual communities of commuters.

This work was previously published in Encyclopedia of Networked and Virtual Organizations, edited by G. Putnik and M. Cunha, pp. 950-956, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of

Chapter 5.3
Wireless Local Communities in Mobile Commerce

Jun Sun
University of Texas – Pan American, USA

INTRODUCTION

In mobile commerce (m-commerce), consumers engage a ubiquitous computing environment that allows them to access and exchange information anywhere and anytime through wireless handheld devices (Lyttinen & Yoo, 2002). While consumers generally sit before personal computers to browse e-commerce websites through the Internet, they are free to move around while connected in m-commerce and can truly be called *mobile consumers.* Compared with stationary consumers in e-commerce, mobile consumers have special information needs regarding their changing environment.

Consumers mainly access information through wireless portals in m-commerce. A lot of these portals provide mobile consumers information specific to where they are. For example, various location-based services have emerged to push information about what is available and occurring nearby to mobile consumers (Rao & Minakakis, 2003). Such wireless portal services overcome the difficulty of searching information with handheld devices, typically cell phones. However, pushing information to users based on where they are may annoy them, because this approach disregards the specific needs and interests of people in context and deprives their control over what they want to know (Barkhuus & Dey, 2003).

In contrast to information pushed by product or service providers, consumers are likely to regard peer-to-peer reference groups as credible sources of product/service information and be open to their informational influence (Miniard & Cohen, 1983). For example, if consumers hear from others that nearby stores offer discounts on certain commodities, they may go to these stores to have a look for themselves. To capitalize on such business opportunities in m-commerce, this article proposes a community portal approach, a so-called *wireless local community* (WLC). As the name suggests, a WLC is a virtual community that allows mobile consumers in a functionally-defined area to exchange information about what is available and occurring nearby with each other through wireless handheld devices.

By far, most virtual communities are built upon the infrastructure of the Internet and they refer to "… groups of people with common interests and needs who come together online… to share a sense of community with like-minded strangers, regardless of where they live" (Hagel & Armstrong, 1997, p.143). Like members in these online communities, WLC members must share something that they are interested in and need in common. Because WLC membership is geographically determined, WLC coverage areas must "supply" what can potentially meet the interests and needs of mobile consumers in them, and such areas may include: shopping plazas, tourist parks, and sports facilities, among others. These functionally-defined areas, which determine the scope, theme, and membership of WLCs, are the settings in which consumer behavior occurs and they constitute the *supply contexts* of local consumers. In this sense, WLCs are context-based virtual communities, in contrast to most on-line communities, which are generally topic-based.

This article first outlines the macro-level conceptual design of the WLC approach and discusses its technical, operational, and economical feasibilities. The success of WLCs, like that of online communities, largely depends on how micro-level implementations can promote member participation and enhance member experience. Based on an understanding of how mobile consumers share contextual information through the mediation of WLCs, this article discusses specific implementation issues.

WLC CONCEPTUAL DESIGN

WLC conceptual design includes an architectural design and an operational design. The architectural design describes the major components of a WLC system, and the operational design identifies all the parties involved in WLC operations and their roles.

As the platform of a context-based virtual community, a typical WLC system has four major components: positioning system, cell phones, wireless network, and WLC server (Figure 1). First, a positioning system is necessary to determine WLC membership by finding out what people are in which supply contexts. Moreover, the location information associated with a message is helpful for readers to understand which part of a supply context it refers to. There are generally two types of positioning systems, network-based and satellite-based (see Roth, 2004), requiring cell phones to be embedded with either triangulation-microchips or GPS-receivers.

New-generation cell phones are not only positioning-enabled, but also data-capable. Users can post and read short textual messages through the interface of cell phones. Moreover, many cell phones have internal digital cameras, allowing users to take pictures/videos of surrounding objects/events to share with others. A WLC server stores textual messages and multimedia attachments posted by members in chronological order, just like an on-line community server. Based on the display capacity of each cell phone, a WLC server can page the messages accordingly. The data communications between cell phones and a WLC server are carried through a wireless network. From this architectural design, Table 1 compares WLCs with traditional on-line communities.

WLC operations involve business partners, hosts and members. *WLC business partners* are businesses that offer financial resource to establish, operate and upgrade WLCs in their areas. They may also assign WLC moderators for member support and help. *WLC hosts* are wireless carriers (or their agents) that provide necessary infrastructure, mainly wireless networks and WLC servers, and technical support for WLC functioning.

WLC members are cell phone users who join particular WLCs at a moment. When a subscriber wants to find out available WLCs in an area, he/she

Figure 1. The architecture of WLC systems

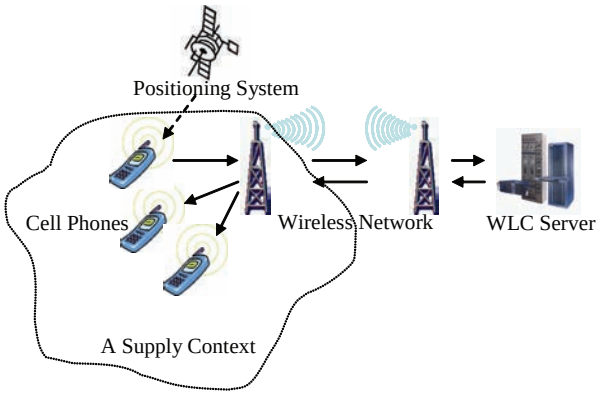


Table 1. Comparisons between two types of virtual communities

Type of Virtual Community	Coverage	User-end Device	Network
Wireless Local Community	Context-based	Cell Phone	Wireless
On-line Community	Topic-based	Personal Computer	Internet

can click the link "*Wireless Local Communities*" on the cell phone display. Through the positioning system, the cell phone obtains user location information and sends it along with the request to the WLC server. The server determines which WLCs are available in that area and displays them on the cell phone. If the subscriber is interested in a particular WLC, he/she can click its link and join it. Depending on the capacity of cell phone, a person may even join multiple WLCs simultaneously. When a member moves out of a supply context, he/she can either exit the WLC immediately or become a "listener" for a while.

A WLC member can share information about his/her part of the supply context with other members. Because the contributions from different members constitute mutually beneficial conjunction of distinct informational elements as resources for all, the sharing of information among WLC members leads to *informational synergy*. While informational synergy can greatly enhance consumer experience and satisfaction of WLC members, WLC business partners may benefit from increased customer patronage as well. For WLC hosts, the main source of revenue is the service contracts with WLC business partners. Therefore, the WLC approach is a win-win solution for all parties involved.

WLC IMPLEMENTATION ISSUES

The success of virtual communities, to a large extent, depends on the active participation of their members (Whittaker, Isaacs, & O'Day, 1997). Micro-level WLC implementation, especially the interface design, must consider the unique characteristics of how WLC members exchange information with each other through the mediation of WLC systems. As mentioned, the mediated

behavior of WLC members is directed towards a common object, their supply context, with the purpose of achieving informational synergy. To study such mediated, purposeful and object-oriented behavior involving multiple actors, activity theory is particularly appropriate.

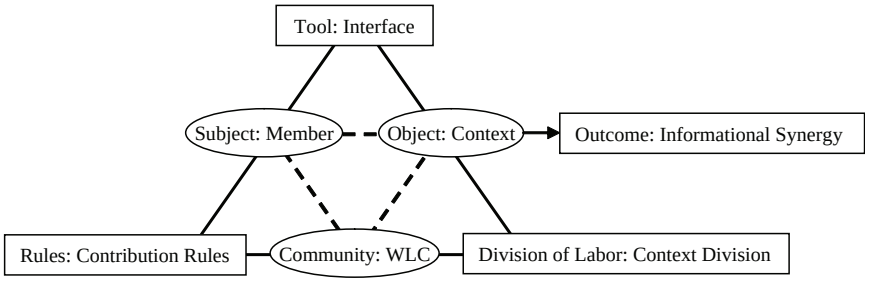
Activity theory (AT) was founded by Russian psychologist Vygotsky in the early 20th century and elaborated by his followers. The basic unit of analysis in AT is human purposeful "activity," rather than specific "action," as in most other psychological theories (Leont'ev, 1978). An activity is composed of a series of actions conducted by one or more individuals for a common purpose. The motivation of an activity provides necessary background to understand specific actions that are situated in that activity. Under this conceptualization, sharing contextual information is a collaborative activity, comprised of individual actions such as posting and reading messages, of WLC members to achieve informational synergy.

Engeström (1987) summarized the relationships in AT with the activity model and we use it to analyze the context-sharing activity of WLC members (Figure 2). In this activity, the subject is a WLC "member," who accesses the information about the *object*, a supply "context," through the mediation of the *tool*, which is the WLC "interface" on a cell phone. The *outcome*, "informational synergy," motivates WLC members to work together as a *community*, so-called "WLC." Because each member shares information about the part of the supply context in his/her proximity, the geographical distribution of WLC members constitutes their *division of labor* in sharing contextual information, and can be denoted as "context division." The *rules* that regulate how WLC members interact with each other through posting messages can be called "contribution rules." To be consistent with the principle of intuitive interface (Bærentsen, 2000), WLC interface design should manifest contribution rules and context division in an intuitive way to WLC members in order to facilitate their context-sharing activity.

Contribution Rules

WLC members post messages either initiatively or responsively. Initiative contributors post messages to share or inquire contextual information, and responsive contributors put comments or answers to the original messages. The privacy of WLC members can be protected by allowing them to choose whether to reveal their usernames or remain "anonymous" when they post messages.

Figure 2. Activity model and context-sharing activity



In sharing contextual information, initiative contributors mainly describe what is interesting nearby. Considering the limited editing and displaying capacity of cell phones, the textual part of messages should be brief. However, multimedia attachments can greatly enrich textual messages. If readers are interested in the attachments, they can download them separately. For example, when a WLC member in a toy store finds some toys interesting, he/she can post a message "Cute toys!" and attach a picture taken with a digital camera embedded in the cell phone. If readers want to have a look at the toys, they can just click the attachment link and view the picture. Readers can respond to messages with comments, such as "interesting," or inquiries for details, such as price.

In asking for information or help, initiative contributors mainly describe their needs. For example, when a shopper is looking for something in a shopping plaza, he/she can post a message asking others for guidance. For another example, when a traveler is lost in a national park, he/she can post a message asking for directions. Other WLC members can respond to these messages if they know the answers. Another important information source is a WLC moderator on duty. As the representatives of business partners, stationed WLC moderators have access to informational and physical resources and they are mainly responsible for answering inquiries and calls-for-help from WLC members. Specifically, WLC moderators can retrieve information from database systems and answer WLC members' questions about their supply contexts. They may also mobilize emergency services (EMS) to provide help to WLC members in urgent situations, such as severe accidents or diseases.

Context Division

Determined by the geographical distribution of WLC members, context division leads to complimentary contextual information sharing. To achieve informational synergy, however, WLC members who read an initiative message must be able to tell which part of the context it refers to. Therefore, WLC members should reveal their locations when they share or inquire contextual information, so that other members can understand the messages in context.

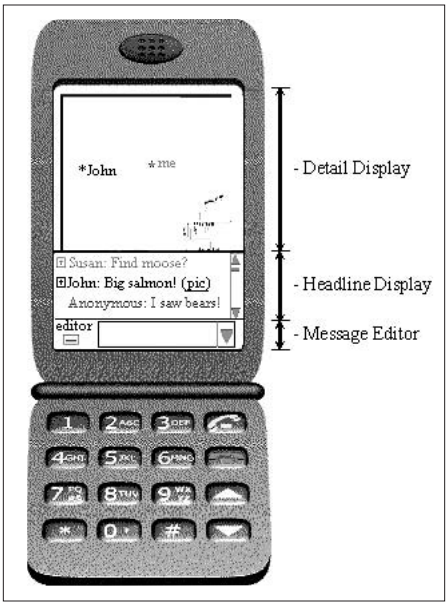
Accordingly, a WLC system can be implemented to reflect context division in the following way. When a WLC member posts an initiative message, his/her cell phone obtains the location information and sends it to the WLC server. At the requests for the message from other members, the server retrieves the corresponding location information from the database and sends it along with the requested message to their cell phones. When they read the message, their cell phones display the associated location on the map of supply context. Members who post responsive messages, on the other hand, are not required to reveal their locations if they are not sharing additional contextual information.

Interface Design

As mentioned, a WLC interface should manifest context division and contribution rules to WLC members in an intuitive way. However, cell phones have much smaller displays compared with those of personal computers, putting a limit on WLC interface design. To meet both the requirement and constraint, the proposed WLC interface design is compact but includes essential components: headline display, detail display and message editor (Figure 3). The headline display lists the textual part of initiative messages, or headlines, in a chronic order. The detail display shows the map of supply context, responsive messages or multimedia attachments. The message editor allows a WLC member to compile short messages.

When a headline is selected, the detail display shows a map indicating the contributor's location relative to the reader's. If there are responsive messages to an initiative message, an "unfold" button will appear before the headline. If a message has

Figure 3. A WLC interface design



a multimedia attachment, an attachment link (e.g., "pic" for picture attachment) will appear at the end. If readers are interested in the comments or attachments, they can click those buttons or links to view them in the Detail Display.

WLC members can compile simple messages in the nessage editor. To facilitate text input, the message editor may have a pull-down menu of commonly-used phrases. For instance, commonly-used phrases for shopping plazas may include "discount," "good bargain," "new styles," and so on. If WLC members just read messages, they can minimize the message editor to leave more space for the headline display and detail display.

Figure 3 illustrates an example of a WLC interface as it appears to a WLC member in a national park. Suppose another member with a username John found big salmon in the nearby water and posted a message about his finding and attached a picture of some salmon he caught. When the reader selects the message (as highlighted) in the headline display, the detail display shows a map indicating John's location (indicated by the red star) relative to the reader's. The reader then knows where it is likely to find big salmon. He/she can click the attachment link "pic" to have a look at the salmon caught by John. There are already some responsive messages to John's original message. The reader can click the "unfold" button on the left side of the message to read them in the detail display.

WLC MEMBER PARTICIPATION

For mediated communications, researchers have found that joint attention and social linkage are necessary conditions for effective information exchange (Clark & Marshall, 1981; Nardi & Whittaker, 2002). In the context-sharing activity, it is through the awareness of contribution rules and context division as manifested by the interface that WLC members can establish joint attention and social linkage with each other.

As suggested for WLC implementation, the interface controls (e.g., unfold button and message editor) indicate the contribution rules to WLC members in how they can exchange information with each other. The map indicates the context division by showing the relative locations of WLC members so that they can have the sense of sharing the same supply context. Studies have shown that exchanging mutually meaningful experience in a shared physical space is an important means of social bonding among people (Nardi & Whittaker, 2002). Researchers have also found that "sharing the same physical environment enables people to coordinate conversational content, by making inferences about the set of objects and events that others in the same environment are likely to know about and want to talk about" (Whittaker, 2003, p. 257). Thus, the sense of sharing the same physical environment through exchanging contextual information helps WLC members establish both

social linkage and joint attention. Socially and cognitively bonded, WLC members are likely to regard each other as "fellow buddies" with whom they can talk about their experience in the same supply context.

In summary, the implementation of WLC systems as suggested should be able to facilitate the context-sharing activity among WLC members. Compared with topic-based online communities, context-based WLCs promote the participation of members through establishing social linkage and joint attention in the process of sharing personal experiences in the same environment. Of course, WLC members usually cannot develop long-term relationships with each other as in online communities. However, the purpose of WLC is to help mobile consumers exchange contextual information, for which long-term relationships are not essential.

CONCLUSION

WLC in m-commerce is a community portal approach that helps mobile consumers to share what they know or want to know about the local supply context with each other. This article discusses the macro-level conceptual design of WLC as well as its micro-level implementation issues. Financially sponsored by business partners, technically supported by hosts and behaviorally participated by members, WLC operations benefit all parties involved. The suggested implementation of WLC systems aims to facilitate the context-sharing activity of WLC members, leading to informational synergy.

To successfully implement WLCs in m-commerce, further technical, behavioral and managerial issues must be addressed. Such issues may include: quality of service (QoS) regarding timely and reliable message delivery over wireless networks, ethical standards and enforcement for appropriate message contribution, specific requirements on WLC implementation and administration for different types of supply contexts, and so on. We hope that this article may enhance further discussions and research in WLC application development.

REFERENCES

Bærentsen, K. B. (2000). Intuitive user interfaces. *Scandinavian Journal of Information Systems, 12*, 29-60.

Barkhuus, L., & Dey, A. (2003). Is context-aware computing taking control away from the user? Three levels of interactivity examined. In *Proceedings of the 5th Annual Conference on Ubiquitous Computing (UbiComp 2003)* (pp. 149-156).

Clark, H., & Marshall, C. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge, UK: Cambridge University Press.

Engeström, Y. (1987). *Learning by expanding.* Helsinki: Orienta-konsultit.

Hagel, J. III, & Armstrong, A. (1997). Net gain—Expanding markets through virtual communities. *The McKinsey Quarterly, 1997*(1), 140-153.

Leont'ev, A. N. (1978). *Activity, consciousness and personality* . Englewood Cliffs, NJ: Prentice-Hall.

Lyttinen, K., & Yoo, Y. (2002). Issues and challenges in ubiquitous computing. *Communication of the ACM, 45*(12), 63-65.

Miniard, P. W., & Cohen, J. B. (1983). Modeling personal and normative influences on behavior. *Journal of Consumer Research, 10*, 169-180.

Nardi, B., & Whittaker, S. (2002). The place of face-to-face communication in distributed work. In P. Hinds & S. Kiesler, (Eds.), *Distributed work* (pp. 83-112). Cambridge, MA: MIT Press.

Rao, B., & Minakakis, L. (2003).Evolution of mobile location-based services. *Communication of the ACM, 46*(12), 61-65.

Roth, J. (2004). Data collection. In J. Schiller & A. Voisard, (Eds.), *Location-based services* (pp.175-205). San Francisco, CA: Morgan Kaufmann Publishers.

Whittaker, S., Isaacs, E., & O'Day, V. (1997). Widening the net: Workshop report on the theory and practice of physical and network communities. *SIGCHI Bulletin, 29*(3), 27-30.

Whittaker, S. (2003). Theories and methods in mediated communication. In A. C. Graesser, M. A. Gernsbacher, S. R. Goldman (Eds.), *Handbook of Discourse Processes* (pp. 243-286). Mahwah, NJ: Lawrence Erlbaum.

KEY TERMS

Context-Sharing Activity: The collaboration among a group of people to share information about their environment with each other through the mediation of information technologies.

Informational Synergy: A mutually advantageous conjunction of distinct information elements as resources for those who share the information with each other.

Mobile Consumer: A person who is free to move around while connected to the wireless network with a handheld device (e.g., cell phone) in mobile commerce.

Supply Context: A functionally-defined area, including what are typically available and occurring in it, that constitutes the settings for people in the area to conduct consumer behavior.

Wireless Local Community (WLC): A type of wireless virtual community that allows mobile consumers within a supply context to exchange information about events that occur and about services and products that are available nearby through handheld devices.

WLC Business Partner: A business that offers the financial resource to establish, operate and upgrade a WLC that covers its supply context.

WLC Host: A wireless carrier (or its agent) that provides necessary infrastructure, mainly wireless networks and WLC servers, and technical support for WLC functioning.

WLC Member: A mobile consumer who joins a WLC at a moment.

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 1204-1209, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.4
From Communities to Mobile Communities of Values

Patricia McManus
Edith Cowan University, Australia

Craig Standing
Edith Cowan University, Australia

INTRODUCTION

The discussion around the impact of information communication technologies in human social interaction has been the centre of many studies and discussions. From 1960 until 1990, researchers, academics, business writers, and futurist novelists have tried to anticipate the impact of these technologies in society, in particular, in cities and urban centres (Graham, 2004). The views during these three decades, although different in many aspects, share in common a deterministic view of the impact of ICT on cities and urban centres. They all see ICT influence as a dooming factor to the existence of cities. These authors have often seen ICT as a leading factor in the disappearance of urban centres and/or cities (Graham; Marvin, 1997; Negroponte, 1995). According to Graham, these views tend to portray ICT impact without taking into consideration the fact that old technologies are not always replaced by newer ones; they can also superimpose and combine into to something else. These views also have generally assumed that the impact of ICT would be the same in all places and have not accounted for geographic differences that could affect the use of information communication technologies.

This article assesses the significance of the theory of consumption value as an explanatory framework for mobile commerce (m-commerce) adoption and use. It discusses whether perceived values can define the characteristics of any discrete "community of use" (group) of m-commerce users. It discusses the significance of online communities and their relation with mobile commerce. We first discuss the impact of ICT in cities. Second, we present the theory of consumption values as a framework to understand mobile commerce use. Then we assess the relevance of communities' values as an explanatory theory to mobile commerce adoption. Finally, we explore the possibility

that consumption values could be mobile-community-binding instruments.

There are a few weaknesses in these deterministic views of the impact of ICT on the development or dooming of cities. Most of them assume that technology impacts exactly the same way everywhere; that is, there is an assumption that a city is the same anywhere on the globe (Graham, 2004). This perspective, also, does not take into account the growth of physical mobility in urban centres (Graham) and the fact that technology does not promote only isolationism (Horan, 2004). Statistics show, for example, that there was a continuous rise in global motor vehicle ownership, from 350 million in 1980 to 500 million in 2001, and a forecast of 1 billion by 2030 (Bell & Gemmel, 2001). Moreover, "in 2001 more mobile phones were shipped than automobiles and PCs" (Clarke, 2001, p. 134). In 200l, out of the 200 million wireless devices sold in the U.S., 13.1 million were personal digital assistants (PDAs) and the other 187 million were mobile phones (Strauss, El-Ansary, & Frost, 2003). It is important, though, not to presume that some level of face-to-face contact is not going to be replaced by electronic technology. Refer, for example, to what is happening with many network-based services like online banking, EDI (electronic data interchange), or the DoCoMo phenomenon in Japan (Graham; Krishnamurthy, 2001). It becomes reasonable to assume that it is very unlikely that ICTs will bring death to the cities. On the contrary, they are deeply entrenched in urbanisation and social economic trends (Graham).

RELEVANCE OF COMMUNITIES

Many works in cultural geography, sociology, and anthropology refer to the mediating role of technologies in structuring the relationship between individuals and their social environment or community (Green, 2002). Community can be defined as "the formation of relatively stable long-term online group associations" (Barkardjiva & Feenberg, 2002, p. 183). Traditionally, the concept of community is associated with many circumstances or factors; however, a common physical location was for many years considered to be a key factor to determine their existence (Graham, 2004). With the development and popularization of ICTs, in particular, the Internet and mobile phones, it is possible to say that the key factor to determine the existence of a community is accessibility (Webber, 2004)

In the social sciences, the concept of community has generated so much discussion that it has already reached a theoretical sophistication (Komito, 1998). However, this theoretical sophistication has not been transferred to the concept of ICT-mediated communities (Komito). The broad interpretation of the community concept in the network environment has many different meanings, ranging from definitions like "norm or values shared by individuals," "a loose collection of like-minded individuals," or "a multifaceted social relation that develops when people live in the same locality and interact, involuntarily, with each other over time" (Komito, p. 97). We consider virtual communities to refer to different types of communities facilitated by information communication technology.

Authors Armstrong and Hagel (1999) were two of the pioneers in using the term virtual community. By virtual community they describe a group of technology enthusiasts in San Francisco. These high-tech enthusiasts created a space in the early days of the Internet prior the World Wide Web. This was and still is a site where people can get together to discuss and exchange cultural information, and today it has migrated to the Web. "The well has been a literate watering hole for thinkers from all walks of life, be they artists, journalists, programmers, educators or activists" (The Well, 2003). Haylock and Muscarella (1999) on the other hand, use the term virtual community when referring specifically to the World-Wide-Web-based communities, but kept their definition

of community quite broad. To them a virtual community is a "group of individuals who belong to particular demographic, profession or share a particular personal interest" (p. 73).

In his 1998 article, Komito discusses extensively the community concept and develops a taxonomy for virtual and electronic communities. He identifies three basic kinds of communities: the moral community (the character of the social relationship is paramount), normative or cognitive community (existence of preset rules of behaviour), and proximate community (interaction happens not because of roles or stereotypes, but because of individuals). A moral community refers to people who share a common ethical system, and it is this shared ethical system that identifies their members. According to Komito, this kind of community is difficult to identify in a computer-mediated communication environment, with the moral purpose of the community being difficult to identify. The normative community is probably the most common type of community associated with ICT. This kind of community is not bound physically or geographically, but is bound by common meaning and culture, such as members being medical doctors, Jews, or jazz aficionados. The individual participants in these communities may never interact with all the other members of this particular community. Authors such as Komito believe that the concepts of community of interest and community of practice borrowed their framework from cognitive communities. Proximate communities have a social emphasis. In this model of community, the interaction between members happens not only in terms of roles or stereotypes, but at the individual level; it is in this kind of community where relationships are developed and conflicts managed (Komito). Although he presented a typology for ICT-mediated communities, Komito concludes that the most useful way of looking at ICT-mediated communities would be to treat the community as a background and concentrate on how individuals and groups deal and adapt to continuously changing environments in terms of social interaction rules. With this in mind, we suggest that a group of individuals who share the same consumption values in relation to mobile services could be members of the same community. The concept of consumption values comes from Sheth, Newman, and Gross' (1991a, 1991b) theory, described next.

THEORY OF CONSUMPTION VALUES: AN ALTERNATIVE FRAMEWORK TO UNDERSTAND MOBILE COMMERCE USE

In reviewing the literature on the adoption and use of technologies, some dominant theoretical frameworks were identified as adaptations or extensions to Rogers' (1962, 2003) diffusion-of-innovation theory or Ajzen's (1991) theory of planned behaviour (TPB). The technology-adoption model (TAM; Davis, Zaner, Farnham, Marcjan, & McCarthy, 1989) is derived from Ajzen and Fishbein's (1980) theory of reasoned action (TRA; which TPB is based upon). Most recently, Venkatesh, Morris, Davis, and Davis (2003) conceptualized the unified theory of acceptance and use of technology (UTAUT). This model is quite comprehensive as it combines TRA, TAM, TPB, the IDT (Innovation Diffusion Theory) model of MPCU (Model of PC Utilization) (personal computer) utilization, the motivational model, and social cognitive theory. However, as the model integrates several theories that focus on user and consumer intention to behave, this model does not concentrate on actual behaviour. For this reason we suggest the utilization of Sheth et al.'s (1991a) theory of consumption values. Although this model has not been directly applied to technology adoption, its unique perspective on consumption values can provide valuable insights to better understand m-commerce-adoption drivers.

Sheth et al. (1991a, 1991b) conceptualized a model to help comprehend how consumers make decisions in the marketplace. They based their

model on the principle that the choices consumers make are based on their perceived values in relation to what the authors called "market choice," and that the perceived values contribute distinctively to specific choices. Because their model examines what the product values are that attract consumers, it can be viewed as a way to understand the attitude toward the product, making this a proactive way to understand m-commerce adoption.

Sheth et al. (1991a) classify five categories of perceived value. Functional values are associated with the utility level of the product (or service) compared to its alternatives. Social value is described as the willingness to please, and social acceptance. Emotional values are those choices made based upon feelings and aesthetics. A common example would be the choice of sports products. Epistemic values can be used to describe the early adopters in the sense that it relates to novelty or knowledge-searching behaviour. Words such as cool and hot are often associated with this value. Finally, the conditional value refers to a set of circumstances depending on the situation (e.g., Christmas, a wedding, etc.). Socioeconomical and physical aspects are included in this value. These five values were conceptualized based on a diversity of disciplines including social psychology, clinical psychology, sociology, economics, and experimental psychology (Sheth et al., 1991a).

This theory has not been used to directly explain adoption; however, its unique conceptualization of product values provides a multidisciplinary approach that would contribute toward the understanding of the actual consumer behaviour in a market choice situation. The limitation of this theory to understanding adoption is that it cannot be used to understand organisational adoption as it does not address influential factors that affect purchase couples or group adoption. Another limitation is that this model cannot be used to understand adoption in cases where the buyer is not the user. Nevertheless, Sheth et al.'s model (1991a) "provides the best foundation for extending value construct as it was validated through an intensive investigation in a variety of fields in which value has been discussed" (Sweeney & Soutar, 2001, p. 205).

The application of Sheth et al.'s model (2001a) would help to provide an understanding of intrinsic influential factors, that is, values about electronic channels such as mobile services (Amit & Zott, 2001; Anckar, 2002; Eastilick & Lotz, 1999; Han & Han, 2001; Venkatesh & Brown, 2001). The theory of consumption values can identify the main value-adding elements in m-commerce or the primary drivers for adopting m-commerce.

Sheth et al. (1991a, 1991b) claim that the main limitation of the theory of consumption value is the fact that it cannot be used to predict the behaviour of two or more individuals. However, this may not be true if the individuals form a group because they share the same perceived values.

COMMUNITIES OF VALUE

The community concept has been used in a number of areas in information systems research. The emergence of networked technologies and the popularization of the Internet have brought a new approach to the study of communities (Bakardjiva & Feenberg, 2000; Haylock and Muscarella, 1999; Komito, 1998). Authors have used the terms online community and virtual community interchangeably. However, one can say that the term virtual community is far broader and may include any technology-mediated communication, whilst online community would be more applicable to the Internet or the World-Wide-Web portion of the Internet. Also, communities of practice have been in the centre of academic journals' and practitioners' publications' attention; however, this community is not dependent on technology. In fact, they have been around for centuries. They can be defined "as groups of individuals informally bound together by shared expertise and passion for a shared enterprise" (Wenger & Snyder, 2000, p. 139). When studying virtual

communities, researchers seek to understand and classify the role that network technology plays in structuring relationships, societies, and their subsets (Armstrong & Hagel, 1999; Bakardjiva & Feenberg; Haylock & Muscarella, 1999). The interest on communities of practice has been driven by researchers who have identified these informal, self-organised nodes. These groups have been identified as beneficial to organisations, and their strength lies in their ability to self-perpetuate and generate knowledge (Wenger & Snyder).

In information systems, studies of communities have helped to better understand systems adoption and usability. In marketing, communities are now an alternative way to segment consumers (Table 1). Mobile technologies have had a profound impact on people's everyday lives to the point of reshaping time and space (Green, 2002). Green explores the impact of mobile technologies in time and space. Underpinning her arguments are concepts such as proximity, mobile work, flexible schedules, and so forth, which depict this new understanding of temporality. In today's life, social relationships have become fragmented, and mobile technologies represent a way to bring continuity back (Green). This new mobile lifestyle is quite prevalent in teenagers. Spero's (2003) white paper points out that the old demographic segmentation of teenagers (ages seven to 10 as tweens, 11 to 13 as young teens, 14 to 16 as teenagers, and 16 and older as young adults) is no longer effective, and a more efficient alternative is segmentation based on mobile lifestyle. These lifestyle traits encompass things like interest, behaviour, upbringing, and eating habits. We propose that identifying communities of mobile service value through the underlying reasons why users perceive those values, from Sheth et al.'s (1991a, 1991b) theory, provides a theoretical framework for understanding mobile service adoption.

CONCLUSION

There are great expectations in relation to the adoption of m-commerce. This article has discussed the utilization of the theory of consumption value (Sheth et al., 1991a, 1991b) as an alternative framework to understand m-commerce adoption and use. The value theory provides deeper explanatory ability as it examines the underlying rationale in the decision-making process. This can more easily be used for predictive purposes. For example, a main driver for teenagers using mobile phones is the relatively low cost of text messaging; however, the motivator for use is the intrinsic social aspect of the service, which caters and builds upon an existing community of use.

Table 1. Examples of communities of use

Community of Use	Lifestyle (Common Traits)	Dominant Perceived value	Issues within the Values	Type of Service
Nomadic Professional	Virtual Office	Functional	Convenience	Micropayment (Parking)
Urban Teens Social Group	Connected Net Generation Sociable	Social	Short Messages	SMS
Postmodern Family	Discontinuous	Functional	Convenience	Voice, SMS

Product and service developers need to examine these deeper factors to come to a sophisticated understanding of adoption-related decisions. Previous theoretical explanations for technology adoption are low in terms of predictive capabilities. This article suggests that the consumer perceived-values approach has significant potential not only in explaining adoption decisions on an individual level, but also across communities of use or practice. These communities exist in the business world as well as society in general.

The concept of community of use represents a more effective way to identify different groups or segments as demographics are no longer reliable. People within the same age group do not necessarily have the same lifestyle and perceive the same values in a service.

The value perceived in a service or product could be what binds groups of individuals in communities, generating what one would call communities of values. The reasons why individuals perceive some values in mobile services can explain group behaviour.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and the Human Decision Process, 50, 179-211.

Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behavior. Englewood Cliffs, NJ: Prentice-Hall.

Amit, R., & Zott, C. (2001). Value creation in e-business. Strategic Management Journal, 22, 493-520.

Anckar, B. (2002). Adoption drivers and intents in the mobile electronic marketplace: Survey findings. Journal of System and Information Technology, 6(2), 1-17.

Armstrong, A., & Hagel, J., III. (1999). The real value of online communities. In D. Tapscott (Ed.), Creating value in the network economy (pp. 173-185). Boston: Harvard Business School Publishing.

Bakardjiva, M., & Feenberg, A. (2002). Community technology and democratic rationalization. The Information Society, 18(3), 181-192.

Bell, G., & Gemmell, J. (2002). A call for the home media network. Communications of the ACM, 45(7), 71-75.

Brown, K. M. (1999). Theory of reasoned action/ theory of planned behaviour. University of South Florida. Retrieved June 21, 2003, from http://hsc.usf.edu/~kmbrown/TRA_TPB.htm

Clarke, I., III. (2001). Emerging value propositions for m-commerce. Journal of Business Strategies, 18(2), 133-148.

Davis, J., Zaner, M., Farnham, S., Marcjan, C., & McCarthy, B.P. (2003, January 7-10). Wireless brainstorming: Overcoming status effects in small group decisions. Paper presented at the 36th Hawaii International Conference on Systems Sciences, Big Island, Hawaii.

Eastlick, M. A, & Lotz, S. (1999). Profiling potential adopters of interactive teleshopping. International Journal of Retail and Distribution Management, 27(6), 209-228.

Fano, A., & Gershman, A. (2002). The future of business services. Communications of the ACM, 45(12), 83-87.

Graham, S. (2004). Introduction: From dreams of transcendence to the remediation of urban life. In S. Graham (Ed.), The cybercities reader (pp. 1-33). London: Routledge Taylor & Francis Group.

Green, N. (2002). On the move: Technology, mobility, and the mediation of social time and space. The Information Society, 18(3), 281-292.

Han, J., & Han, D. (2001). A framework for analysing customer value of Internet business.

Journal of Information Technology Theory and Application (JITTA), 3(5), 25-38.

Han, S., Harkke, V., Landor, P., & Mio, R. R. d. (2002). A foresight framework for understanding the future of mobile commerce. Journal of Systems & Information Technology, 6(2), 19-39.

Haylock, C., & Muscarella, L. (1999). Virtual communities. In C. Haylock & L. Muscarella (Eds.), Net success (chap. 4, p. 320). Holbrook, MA: Adams Media Corporation.

Ho, S. Y., & Kwok, S. H. (2003). The attraction of personalized service for users in mobile commerce: An empirical study. ACM SIGecom Exchanges, 3(4), 10-18.

Horan, T. (2004). Recombinations for community meaning. In S. Graham (Ed.), The cybercities reader. London: Routledge, Taylor & Francis Group.

Jackson, P. B., & Finney, M. (2002). Negative life events and psychological distress among young adults. Social Psychology Quarterly, 65(2), 186-201.

Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. MIS Quarterly, 23(1), 67-94.

Komito, L. (1998). The Net as a foraging society: Flexible communities. The Information Society, 14(2), 97-106.

Krishnamurthy, S. (2001). NTT DoCoMo's I-Mode phone: A case study. Retrieved March 17, 2003, from http://www.swcollege.com/marketing/krishnamur thy/first_edition/ case_updates/ docomo_final.pdf

Levy, M. (2000). Wireless applications become more common. Commerce Net. Retrieved July 5, 2003, from http://www.commerce.net/research/ebusiness-strategies/2000/00_13_n.html

Marvin, S. (1997). Environmental flows: Telecommunications and dematerialisation of cities. Futures, 29(1).

Negroponte, N. (1995). Being digital. London: Hodder & Stoughton.

Rogers, E.M. (2003). Diffusion of innovations (5th ed.). New York: Free Press, A division of Simon & Schuster, Inc. 1230 Avenue. (1962 - 1st ed.).

Ropers, S. (2001, February). New business models for the mobile revolution. EAI, 53-57. Available at http://www.bijonline.com/PDF/Mobile%20Revolution% 20-%20Ropers.pdf

Sheth, J. N., Newman, B. I., & Gross, B. L. (1991a). Consumption values and market choice: Theory and applications. Cincinnati, OH: South-Western Publishing Co.

Sheth, J. N., Newman, B. I., & Gross, B. L. (1991b). Why we buy what we buy: A theory of consumption values. Journal of Business Research, 22, 150-170.

Spero, I. (2003). Agents of change. Teenagers: Mobile lifestyle trends. Retrieved November 28, 2003, from http://www.spero.co.uk/agentsofchange

Strauss, J., El-Ansary, A., & Frost, R. (2003). E-marketing (3rd ed.). Upper Saddle River, NJ: Pearson Education Inc.

Sweeney, J. C., & Soutar, G. N. (2001). Consumer perceived value: The development of a multiple item scale. Journal of Retailing, 77(2), 203-220.

Sweeney, J. C., Soutar, G. N., & Johnson, L. W. (1999). The role of perceived risk in the quality-value relationship: A study in a retail environment. Journal of Retailing, 77(1), 75-105.

Tierney, W. G. (2000). Undaunted courage: Life history and the postmodern challenge. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research (2nd ed., pp. 537-554). Thousand Oaks, CA: Sage.

Venkatesh, V., & Brown, S. A. (2001). A longitudinal investigation of personal computers in homes: Adoption determinants and emerging challenges. MIS Quarterly, 25(1), 71-102.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425-478.

Webber, M. (2004). The urban place and the non-place urban realm. In S. Graham (Ed.), The cybercommunities reader (pp. 50-56). London: Routledge.

The Well. (2003). Retrieved November 25, 2003, from http://www.well.com/aboutwell.html

Wenger, E. C., & Snyder, W. M. (2000, January-February). Communities of practice: The organizational frontier. Harvard Business Review, (January-February), 139-145.

KEY TERMS

DoCoMo: Japanese mobile telecommunication company that is a part of NTT. It is the creator of I-Mode.

EDI (Electronic Data Interchange): A set of computer interchange standards developed in the '60s for business documents such as invoices, bills, and purchase orders. It has evolved to use the Internet.

TAM (Technology-Acceptance Model): Described as an adaptation of TRA customised to technology acceptance. The intention to adopt is affected by two beliefs: perceived usefulness and the perceived ease of use of the new technology.

TPB (Theory of Planned Behaviour): TPB is an extension of TRA. It adds a third dimension—the perceived-behaviour control component—that looks at uncontrolled external circumstances.

TRA (Theory of Reasoned Action): TRA states that the intention to adopt is affected directly by attitudinal components (beliefs about the outcome of the behaviour and beliefs of the consequences of the behaviour) and the subjective norm component (level of importance or desire to please significant others and/or society).

UTAUT (Unified Theory of Acceptance and Use of Technology): This model is quite comprehensive as it combines TRA, TAM, TPB, the DOI model of PC utilization, the motivational model, and social cognitive theory.

This work was previously published in Encyclopedia of Multimedia Technology and Networking, edited by M. Pagani, pp. 336-341, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.5
Economics of Immediate Gratification in Mobile Commerce

Kerem Tomak
University of Texas at Austin, USA

ABSTRACT

In this chapter we attempt to build a bridge between mobile commerce and the emerging field of behavioral economics. We first provide examples from mobile commerce and link them to behavioral economics. We then build a stylized model to assess the impact of hyperbolic discounting on the profit-maximizing behavior of a monopolist firm. We find that the monopolist makes lower profits compared to exponential discounting consumers for low levels of (positive) network externalities. As the network externalities increase, first-period prices increase, second period prices decrease and the profits increase in equilibrium.

INTRODUCTION

Shopping is ubiquitous. Malls and individual shops face the first stage of expansion to the digital environment through fixed wired Internet. Electronic commerce initiates huge investments and leads to controversies as well as financial disappointments since the mid-1990s. From early 2000 onward we are facing a second wave of digital commercial growth. Wireless technologies are enabling individual consumers to access information wherever they are and whenever they want.

Although the use of mobile devices is evolving rapidly, the investigation of mobile consumer behavior is lacking. An increasing number of electronic commerce services for mobile devices coupled with swift adoption rates will enable mobile operators to provide effective customer services and gain competitive advantage. However, this can only be achieved by analogous deeper understanding of mobile users' behavior.

A tool to understand the consumer behavior within mobile context comes from the field of economics. Neoclassical economics approaches the individual as a rational decision maker faced with a series of consumption choices. The cor-

responding model of human behavior is called "Homoeconomicus," who is endowed with perfect rationality, self-interest, and knowledge. In reality humans are largely driven by their emotions, and emotions are often irrational. They also perform altruistic acts like charity, volunteerism, lending a helping hand, parenting, and even giving one's life for one's country. These all fall contrary to the assumption of self-interest. They perform self-destructive acts like substance abuse, negative addiction, negative risk-taking, procrastination, inability to complete projects, masochism, and suicide. They are also highly ignorant about all their affairs; they can be expert in only a few topics at a time (Laibson, 2001). In parallel to the technology achievements in wireless communications, maybe relatively less rapidly, our understanding of the "homoeconomicus" is expanding toward a complementary economic perspective of the homosapiens. As we discuss in the next section, behavioral economics provides novel concepts using traditional tools. Our goal in this chapter is to discuss the viability of some of the mobile business models through the lens of behavioral economics.

IMPACT OF MOBILE TECHNOLOGY

In this section we provide an overview of the mobile commerce technologies that we believe impact consumers' decision making. We start with a definition of mobile commerce.

Definition: Mobile commerce is defined as all activities related to a (potential) commercial transaction conducted through communications networks that interface with wireless (or mobile) devices.

The most salient feature of mobile commerce is the availability of ubiquitous access to information whenever and wherever it is needed. Using a mobile device a customer can watch streaming video and complete financial transactions while on the road. Digital content is enriched when ubiquity is coupled with location and time-specific knowledge.

Constant access to information can increase efficiency and lower supplier costs for critical decision making. Examples include Siemens' wireless extension to SAP Business Warehouse backend system, UPS' tracking shipments using wireless devices, and Office Depot's logistics management system using custom wireless handheld units.

Coordination costs for buyers can also decrease. CitiGroup customers receive daily bank balance updates via SMS messages, and major brokerage firms such as Charles Schwab and Merill Lynch provide wireless access to aggregated account information. In this chapter we are interested in buyer-side impact of mobile technologies.

Although the number of mobile users is expanding, as Table 1 shows, the percentage of consumers using mobile channels to make purchases is very low, according to an AT Kearney study. According to Forrester, there is an upward trend on the expected sales of mobile devices by 2005. Interestingly, the interest in 3G applications focuses on financial and payment solutions after e-mail applications, according to a Taylor Nelson Sofres survey. These are all indications of increased use of mobile devices in the future for payment purposes.

BEHAVIORAL ECONOMICS OF MOBILE TECHNOLOGY

Instant gratification is key to the use of mobile devices. Mobile services that deliver context-dependent content to users fulfill the instant gratification behavior that consumers seek. According to a Jupiter report, consumer interest in purchasing items using a wireless device is not a priority, with only 7% expressing interest in conducting transactions via a wireless phone. The report adds

Table 1. Services used by Internet-enabled mobile phone users globally

Country (as % of IE-mobile users)	E-mail	Banking	Purchasing	Games
Asia	10%	2%	3%	3%
Brazil	11%	7%	1%	2%
Europe	10%	3%	1%	3%
Japan	77%	4%	12%	5%
North America	27%	6%	3%	7%
Worldwide	19%	3%	3%	3%

Source: AT Kearney, August 2002

Table 2. mCommerce Sales Predictions, 2001 - 2005

Device	2001	2002	2003	2004	2005
Sales closed on devices (in billions)					
PDA	0.0	0.1	0.5	1.4	3.1
Cell Phone	0.0	0.0	0.0	0.1	0.3
Sales influenced by devices (in billions)					
PDA	1.0	5.6	14.4	20.7	24.0
Cell Phone	0.0	0.0	0.1	0.3	1.3

Source: Forrester Research, January 2002

Table 3. Current mobile phone users' Interest in 3G applications

Application	W Europe	E Europe	USA
On 6-point interest scale, 6 = high interest, and 1 = low interest			
E-mail	4.5	4.7	4.3
Payment Authorization/ Enablement	3.4	3.8	3.0
Banking/ Trading Online	3.5	3.4	3.2
Shopping/ Reservations	3.0	3.1	2.9
Interactive Games	2.0	2.2	2.4

Source: Taylor Nelson Sofres, May 2002

that mobile commerce will be driven by a desire for instant gratification (www.jup.com).

In an attempt to increase the use of mobile devices for purchase, Alon USA LP, which operates Fina gas stations and 7-Eleven outlets in the Southwest, has established an "m-commerce" system using existing cellular telephone technology and already-installed point-of-sale systems. The company is using mobile-commerce payment technology developed by Cellenium Inc. in

Englewood Cliffs, NJ, that will let any cellular telephone, including aging voice-only models, conduct a mobile transaction.

Each transaction is funneled through Alliance Data Systems Inc., which already provides transaction services to Alon and other gas station and convenience store operators. Alliance Data, Cellenium, and Alon have formed a partnership called Cellerate to manage, market, and promote their mobile-commerce system. The Cellerate software also keeps track of customers' premium points and, in Fina's case, can offer instant gratification by automatically controlling a voice-activated vending machine to provide a customer with a free soda.

There is also instant gratification through the consumption of digital products on mobile devices. Recent mobile purchase history of the customers shows that they want to buy downloadable features and extras like ring tones, games, and the ability to send digital photos. In order to satisfy this demand Handango sells digital content for mobile devices as well as software for handhelds. Nokia and MasterCard, banking on customers' desire for convenience, have done run trials of a quick-pay system that attaches to a cell phone. These efforts imply that cell phones are about instant gratification and making a social statement.

O'Donoghue and Rabin (1999, 2000) say that due to preference for immediate gratification, people under-indulge in activities that involve immediate costs and delayed rewards (for example, putting off an unpleasant but necessary task) but over-indulge in activities with immediate rewards and delayed costs (for example, overeating). Based on Strotz (1956) and Pollak (1968), O'Donoghue and Rabin (1999; 2000) distinguished between two types of consumers — (i) sophisticates, people who know that their preferences may reverse due to immediate gratification, and (ii) naives, people who don't realize that their preferences may reverse due to immediate gratification. Naives exhibit immediate gratification behavior with respect to both immediate costs (procrastinate costs) and immediate benefits ("preproperate" benefits). Surprisingly, though sophisticates are able to tackle procrastination, they exacerbate immediate gratification behavior with respect to immediate benefits (O'Donoghue and Rabin, 1999).

Demand for instant gratification raises the issue of payment mechanisms available for related purchases. In the next section, we discuss economic characteristics of mobile payments.

MOBILE PAYMENTS AND CONSUMPTION

According to Celent, a financial services research and consulting firm, by 2004 there will be 60 million mobile payment users generating sales of $50 billion. A joint survey by Visa International and Boston Consulting predicts that combined e-commerce and m-commerce volumes will grow from $38 billion in 2002 to $128 billion in 2004.

There are increasingly more sophisticated devices that are developed together with new applications that take advantage of color screens, keyboards, and longer battery life. Introduction of these applications will drive the use of new payment opportunities that bridle the capabilities of wireless devices. Note that while these have been developed and are mostly also commercially available, their usage is indeed quite limited.

A rich example of mobile payment solutions can be found in Finland, as the country has the highest mobile phone penetration rate in Europe. Dynexco, a Finnish company, has launched a payment solution called DNX MobileMoney. A customer with a DNX account can transfer funds from his or her bank account and pay for purchases of goods or transfer funds to other DNX accounts in real time. Payment is based on text messages sent by a GSM phone or via the Internet (www.dnxmobiiliraha.com).

Sonera Shopper is another mobile payment solution. A customer opens a Shopper account

and transfers money to it from his or her bank account. He or she can pay for purchases at merchants who have joined the system by sending a text message. The customer can also pay for purchases out of his or her credit card account (Visa, Eurocard, MasterCard) instead of his or her Shopper account. In that case the customer's credit card number must be entered into the Shopper system and the customer decides when sending a text message which way he or she wants to pay (www.sonera.fi).

E-Pay sells branded services to merchants. At the moment these merchants include some restaurants and ski resorts. Also in this solution, the customer first registers for a service and has his or her own account opened. After that, he or she can transfer money to this account and pay for purchases and services via mobile phone.

Some purchases can also be aggregated to the customer's monthly mobile phone bill. Purchase of logos, ring tones, or chocolate bars from vending machines are included on the mobile bill at the end of the month. Similarly, using a service called Parkit, one can also pay for parking in some Finnish cities by calling a parking area service number. The parking fee will be included on the customer's telephone bill, credit card bill, or a separate bill, or the customer can pay for parking by Sonera Shopper.

Outside Finland one of the most widespread mobile phone payment applications is the Germany-based paybox, which was launched in May 2000. This service enables the customer to purchase goods and services and make bank transactions via mobile phone. The value of purchases

Figure 1. Mobile payments and their timing vis a vis consumption

Payment options used in existing Mobile Payment solutions		Required before payment	High level description payment process	Currently used technologies
	Prepaid	• Stored (reloadable) valuecard • PIN (to reload card)	• Select product/service • Select "mobile payment" • Authorise transaction (using PIN or password) • Make payment (money deducted from value stored card) • Payment party executes settlement	• Stored-value cards in combination with dual slot phone and smart-card reader
	"Direct" from Credit or Debit account	• Pre-standing agreement • User has to give bank account number or credit card number to payment party • PIN code / password	• Select product/service • Select "mobile payment" • Authorise transaction (using PIN or pw) • Payment party forwards bank account number or credit card number to the merchant`` • Bank/credit card company deducts money from account and makes payment to vendor	• Dual-slot phone in combination with smartcard and smart-card reader • Internet based • Call back system
	Phone bill paid	• Pre-standing agreement which allows payment party to charge the subscriber's (phone)bill	• Infrared: • Vending machine communicates with mobile phone (infrared) • Choose product/service • Authorise payment with button click • Purchase costs charged to phone bill	• Infrared (till bluetooth is available) connection between mobile and Point of Sale
			• Premium rate number: • Call premium rate number • Select product • Network calls vending point to authorise the sale • Purchase costs charged to phone bill	• Premium rate number

Source: *Arthur D. Little*

or credit transfers is debited from the customer's bank account (www.paybox.net).

In Spain a mobile payment solution called Mobipay is available that can be used for payments at real or virtual POS or vending machines. Person-to-person payments and paying for invoices are possible. Mobipay activates through existing payment means, that is, normal or virtual credit, debit, or prepaid cards (www.mobipay.com).

In Norway a customer can sign up for and open his or her own Payex account at Payex's website (www.payex.no) or he or she can send a text message. Before using their Payex account, customers must transfer money into it. Certain purchases can be paid by Payex via Internet.

In all the examples above, the payments are either done in real time or aggregated to the end of the month. The following table from a study by Arthur D. Little characterizes the current mobile payment solutions with respect to the timing of payments.

Economic impact of such a separation in timing of payments and consumption cannot be fully explained using neoclassical economic theory but as the following section explains, behavioral economics can help complement the insights that can be gained from the classical theory of consumption and payments.

HYPERBOLIC DISCOUNTING

Hyperbolic discounting is a way of accounting in a model for the difference in the preferences an agent has over consumption now vs. consumption in the future. For a and g scalar real parameters greater than zero, under hyperbolic discounting events t periods in the future are discounted by the factor (1 + at)-g/a. The expression "hyperbolic discounting" describes the "class of generalized hyperbolas." This formulation comes from a 1999 working paper of C. Harris and D. Laibson, which cites Ainslie (1992) and Loewenstein and Prelec (1992). In dynamic models it is common to use the more convenient assumption that agents have a common discount rate applying for any t-period forecast, starting now or starting in the future.

One reason hyperbolic preferences are less convenient in a model is not only that there are more parameters but also that the agent's decisions are not time-consistent as they are with a constant discount rate. That is, when planning for time two (two periods ahead), the agent might prepare for what looks like the optimal consumption path as seen from time zero; but at time two his preferences would be different (About.com, 2003).

In a simple model of a two-period monopoly firm, we compare the profits and prices for two cases. Our benchmark case is the standard exponential discounting that we assume both firms and consumers adopt. In the case of hyperbolic discounting we fix the parameter in a specific form of hyperbolic discounting:

$$\frac{1}{1+\alpha t}$$

In both cases second-period sales of the monopoly firm face positive network externalities from the first period. This represents the mobile firms' customer base and its impact on the use of (mobile) technology at a later stage.

Box 1.

Variable	Description
p_1	First period price
p_2	Second period price
$\prod$	Profit
e	Level of network externality
δ	Exponential discount factor
α	Hyperbolic discount parameter

In order to build our model, we use the following notation (See Box 1).

We assume that the consumers are distributed uniformly along the [0,1] interval. The firm knows the distribution of the consumers but not their exact location. In the first period the net consumer surplus is v1 = u – p1. In the second period the net consumer surplus with hyperbolic discounting is:

$$v_2 = \frac{1}{1+\alpha}(u - p_2 + e(1-u))$$

with exponential discounting, it is given by $v_2 = \delta(u - p_2 + e(1-u))$

For the hyperbolic discounting case, we find the marginal consumer who is indifferent between consumption in either periods by equating the net consumer surpluses from each period and solve for u:

$$u_1^* = \frac{e + (1+\alpha)p_1 - p_2}{e+\alpha}$$

Similarly, the marginal consumer indifferent between buying or not buying in the second period is given by

$$u_2^* = \frac{p_2 - e}{1-e}$$

The derived demand functions are then given by

D1 = 1 – u1*
D2 = u1* – u2*

Thus the profit function of the monopoly firm is simply.

$$\Pi = p_1 D_1 + \delta\, p_2 D_2$$

The maximization problem we solve to find the optimal prices and profit level is the following:

$\max_{p_1, p_2} \Pi$
$D_1 \le 1$
$D_2 \le 1$
$u_2^* \ge 0$
$p_1, p_2 \ge 0$

The lagrangian that corresponds to the problem above is:

$$\ell = \Pi - \lambda_1(D_1 - 1) - \lambda_2(D_2 - 1) - \lambda_3(-u_2^*)$$

Finally, the system we solve is given by (See Box 2).

The only feasible solutions to this system are given below.

Case 1: $\lambda_1 > 0$, $\lambda_2 = \lambda_3 = 0$
The solution in this case is (See Box 3).

For this solution to yield positive prices and demand, the following conditions need to hold:

$$\delta < \frac{\alpha}{1+\alpha}, e < \min\{\frac{\alpha - (1+\alpha)\delta}{1-(1+\alpha)\delta}, 1\}$$

Case 2: $\lambda_1 = \lambda_2 = \lambda_3 = 0$
This is the interior solution, which yields (See Box 4).

Case 3: $\lambda_1 = 0$, $\lambda_2 > 0$, $\lambda_3 > 0$

$$\lambda_2 = e\,\delta - \frac{e+\alpha}{1+\alpha}$$
$$\lambda_3 = \frac{e+\alpha}{1+\alpha} - (1-e)\,\delta$$
$$p_1 = \frac{e+\alpha}{1+\alpha}$$
$$p_2 = e$$
$$\Pi = e\delta$$
$$D_1 = 0$$
$$D_2 = 1$$

Box 2.

$$\frac{d\ell}{dp_1} = \frac{d\Pi}{dp_1} - \lambda_1\left(-\frac{(1+\alpha)}{e+\alpha}\right) - \lambda_2\left(\frac{(1+\alpha)}{e+\alpha}\right)$$

$$\frac{d\ell}{dp_2} = \frac{d\Pi}{dp_2} - \lambda_1\left(\frac{1}{e+\alpha}\right) - \lambda_2\left(-\frac{1}{e+\alpha} - \frac{1}{1-e}\right) - \lambda_3\left(-\frac{1}{1-e}\right)$$

$$\lambda_1(D_1 - 1) = 0$$

$$\lambda_2(D_2 - 1) = 0$$

$$\lambda_3(-u_2^*) = 0$$

$$p_1, p_2, \lambda_1, \lambda_2, \lambda_3 \geq 0$$

Box 3.

$$\lambda_2 = \frac{(-1+e)(e+\alpha) + (1+\alpha)(-1+e+2e^2-3\alpha+5e\alpha)\delta + (-1+e)^2(1+\alpha)^2\delta^2}{2(1+\alpha)^2}$$

$$p_1 = \frac{e+\alpha+(1-e)(1+\alpha)\delta}{2(1+\alpha)}$$

$$p_2 = \frac{e+e^2-\alpha(1+3e)+(1-e)^2(1+\alpha)\delta}{2(1+\alpha)}$$

$$\Pi = \frac{(e+\alpha)^2 + 2(1+\alpha)(e+e^2-\alpha(1+3e))\delta + (1-e)^2(1+\alpha)^2\delta^2}{4(1+\alpha)^2}$$

$$D_1 = \frac{e+\alpha-(1-e)(1+\alpha)\delta}{2(1+\alpha)}$$

$$D_2 = 1$$

Box 4.

$$p_1 = \frac{(1+\alpha)\delta((2-e)(e+\alpha) + (-1+e)e(1+\alpha)\delta)}{(e+\alpha-(-1+e)(1+\alpha)\delta)^2}$$

$$p_2 = \frac{e+\alpha}{e+\alpha+(1-e)(1+\alpha)\delta}$$

$$\Pi = \frac{(1+\alpha)\delta(\alpha(1+e(-1+\delta))+e\delta)}{(-1+e)(-1+e+2(1+\alpha)(1+e+2\alpha)\delta + (-1+e)(1+\alpha)^2\delta^2)}$$

$$D_1 = \frac{(1+\alpha)\delta(1+\alpha(2-\delta)-\delta)}{(-1+e+2(1+\alpha)(1+e+2\alpha)\delta + (-1+e)(1+\alpha)^2\delta^2)}$$

$$D_2 = \frac{(1+\alpha)(1-e-(1+e)(1+\alpha)\delta)}{(-1+e)(-1+e+2(1+\alpha)(1+e+2\alpha)\delta + (-1+e)(1+\alpha)^2\delta^2)}$$

For this system to yield a feasible solution:

$e > \frac{e+\alpha}{\delta(1+\alpha)} > (1-e)$ *and* $e > \frac{1}{2}$ has to hold.

Case 4: $\lambda_1 = \lambda_2 = 0$, $\lambda_3 > 0$
This yields the following (See Box 5).

For this to yield a feasible solution,

$e < \frac{1}{\delta(1+\alpha)}$ has to hold.

Following figures show the cases for which the exponential discounting parameter is set at d = 0.9 and the hyperbolic discounting parameter is a = 0.2. For this example, we see that the profits when consumers are believed to have hyperbolic discounting are lower for low levels of network externalities. As the network externality effect increases, the profits also increase. This may be due to the fact that the monopoly can benefit from those consumers who value first-period consumption over the second period by charging them higher than the exponential discounting case for high levels of network externalities. This is also seen in Figure 3, where for high levels of first-period price is higher in the hyperbolic discounting case than the exponential discounting.

The monopoly can then add to the profits by charging less in the second period in order to avoid the Coase conjecture, which predicts market failure in the second period for such a monopoly firm. This can be easily seen in Figure 3, where first-period price under exponential discounting decreases as network externalities increase but the second-period price remains at its highest possible rate. The neoclassical monopolist tries to charge lower prices in the first period to attract consumers in the hopes of charging them a higher price in the second period. In this case the market share in the first period is ½, whereas the second-period market share is 0. This implies that the monopoly firm sells only in the first period,

Box 5.

$$\lambda_3 = \frac{(-1+e)(e+\alpha)+(1+\alpha)(2e^2-\alpha+3e\alpha)\delta+(-1+e)e(1+\alpha)^2\delta^2}{2(1+\alpha)(e+\alpha)}$$

$$p_1 = \frac{e+\alpha+e\delta(1+\alpha)}{2(1+\alpha)}$$

$$p_2 = e$$

$$\Pi = \frac{(e+\alpha+e(1+\alpha)\delta)^2}{4(1+\alpha)(e+\alpha)}$$

$$D_1 = \frac{1}{2} - \frac{e(1+\alpha)\delta}{2(e+\alpha)}$$

$$D_2 = \frac{1}{2} + \frac{e(1+\alpha)\delta}{2(e+\alpha)}$$

Figure 2. Profits and first period price of a monopoly firm with and without hyperbolic discounting of the consumers. Alpha represents the hyperbolic discounting parameter

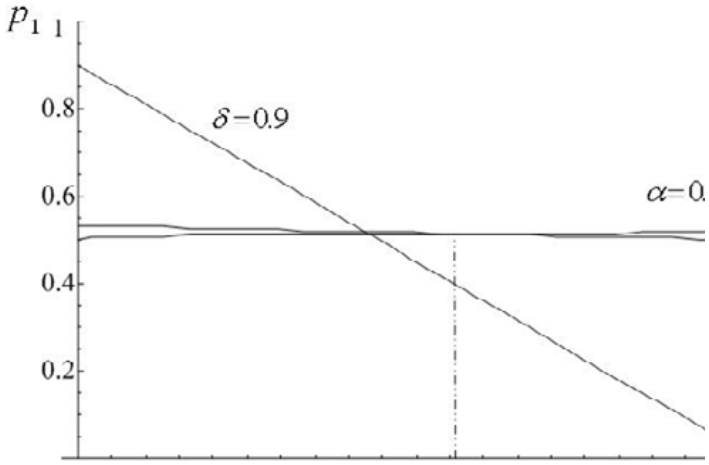
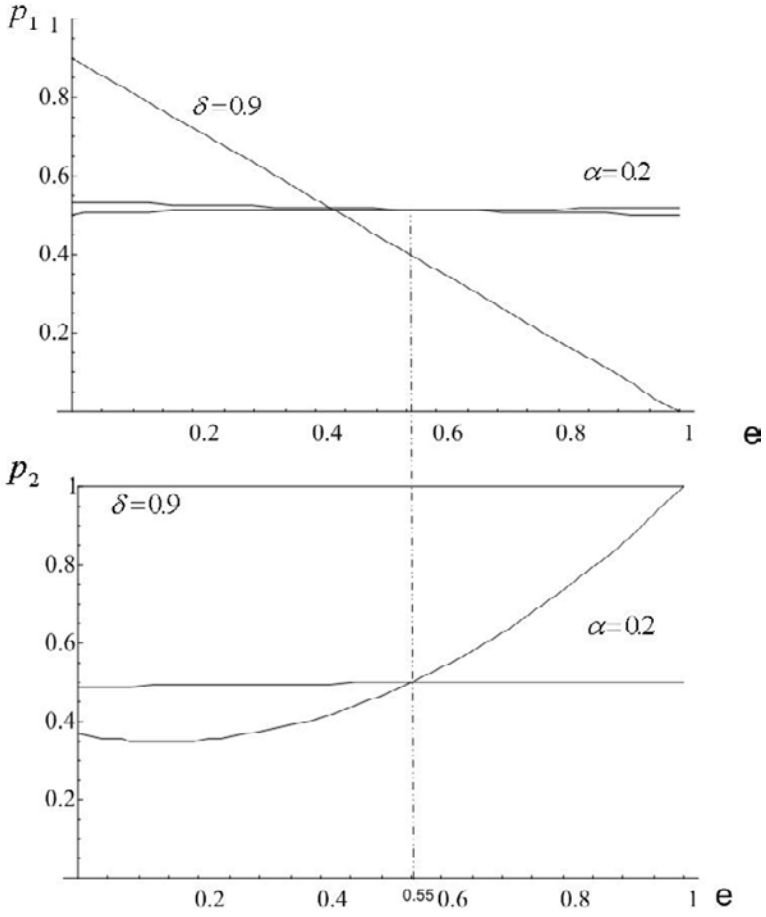


Figure 3. First and second period prices with and without hyperbolic discounting of the consumers. Alpha represents the hyperbolic discounting parameter



as the consumers expect to be charged a higher price in the second period.

The outlook changes once we introduce hyperbolic discounting. The first-period market share becomes

$$\frac{1}{2} - \frac{0.855e}{0.9 + e}$$

which is decreasing but positive in e, converging to 0.05, while the second-period market share is

$$\frac{0.9 + 2.71e}{2(0.9 + e)}$$

which is increasing in e, converging to 0.95. Hence, by taking instant gratification, or present biased preferences, into account, the monopoly can benefit from smoother pricing in the first period and gradually increasing second-period pricing.

CONCLUSION

Behavioral economics provides new perspectives to understand various aspects of consumers' consumption and payment behavior. In this chapter we highlight some of the aspects that we believe can help technology companies form market strategies, especially in the mobile commerce area.

Mobile devices provide a new frontier for firms to reach consumers. They enable companies to better comprehend consumers' purchasing behavior by tracking their spending and consumption patterns in real time. We show that this understanding may help firms make more profits and better position themselves in the marketplace.

Mobile payments and consumption inherit characteristics that can be explained using concepts from behavioral economics. Instant gratification, mental accounting, and hyperbolic discounting are a few that we focus on in this paper. We build a stylized model that compares exponential to hyperbolic discounting within a network externalities framework. We find that when consumers are assumed to have present biased preferences, which is usually the case for instant gratification, as the literature suggests, a monopolist may make more profits and charge more strategically to keep all the consumers purchasing his or her services.

Although we do not mention it in this chapter, the wealth of the consumer, and hence the size of the payment, is as important as the timing of the payments: buying a latte is no pain at all, buying a restaurant meal is a minor pain, buying a computer is a major pain, and buying a car is a massive pain. Consequently the use of mobile payments will be confined to medium- to low-value items until/unless mobile phones are accepted by the consumers as payment instruments.

On the technology side there are emerging payment tools such as Bluetooth-enabled point of sale devices. Global wireless access to any media (voice, data, video) mobile services from/to wherever you may be (homes, offices, hotels, airports, in the air, or at the beach) and for any device (cell phones, PDAs, Internet-aware appliances, ATMs, POS devices, Kiosk, PCs, laptops, and so forth) is already available. Bluetooth, WAP, DSL, and cable modems that integrate seamlessly, Personal Area Networks (PAN), devices with long-distance high-bandwidth wired/wireless Internet, and public telephone network access make it possible.

Bluetooth's advantage is that it is much less expensive to implement. Thus it can be used in various POS devices. A supermarket in Sweden, ICA Ahold, completed a successful test of wireless Bluetooth payments enabled by Ericcson phones in 2000. Customers used their mobile telephones to make purchases, check their account balances, and receive special offer information. Bluetooth sends wireless signals between devices equipped with a Bluetooth chip on the 2.45 GHz ISM band. Depending on the strength of the signal, compatible

Bluetooth devices can communicate at distances of up to 80 meters, although distances of up to 10 meters are more common. Lack of standards is slowing the wide adoption of Bluetooth payment systems. Security is also a concern, since Bluetooth can transmit messages over relatively long distances, which poses a greater threat to payment information since it can be intercepted en route.

Radio Frequency Identification Device (RFID) is another technology solution that has a wide application and direct impact on the payment systems. Since 1997 this technology has been used in ski passes in Switzerland and in Swatch watches, some of which can store credit, as well as more recently in London Underground electronic tickets.

A retail outlet using RFIDs can allow consumers to walk out of the store while charging the card they set up previously. RFIDs prevent theft, help guarantee quality, and provide absolute 100% precision about what stock remains in the food store and when products are close to sell-by dates. They also mean a consumer can pay for products and services ranging from bottles of wine to travel tickets using a card that never leaves their pocket. This will obviously increase the separation between payments and consumption further, making payments more transparent and the pain less apparent. One can foresee the negative impact on the level of debt the consumers might accumulate in the United States.

There are several dimensions over which this work can be extended. We use a very simple model of hyperbolic discounting. The model can be extended to include a more generalized form of hyperbolic discounting function, and instead of two periods, multiple periods can be considered. Mental accounting can also be an important avenue to explore. For initial work in this area, see Balasubramanian, Dutta, and Tomak (2003) or Balasubramanian and Tomak (2003).

Finally, behavioral economics provides new policy guidance to financial and governmental institutions that look into regulating or deregulating competition in mobile telecommunications markets. This is especially important when financial debt in the U.S. has reached new heights.

A cross-cultural study to assess the international differences in consumption and payments as well as present biased preferences can be extremely interesting. For instance, a Finland-U.S. comparison would potentially reveal major differences, not only at the consumer level, but also at the legislative and policy levels. Unlike in Finland, in the U.S. personal bankruptcy is a right that consumers can exercise whereas in Finland "only death" can free one from his or her accumulated debt.

Considering these implications of payment systems and understanding payments and consumption in this new area of mobile technology-based consumption may increase social welfare and ensure ignorance will never be a bliss for the future generations.

REFERENCES

About.com. http://economics.about.com/library/glossary/bldef-hyperbolic-discounting.htm

Ainslie, G. (1992). Picoeconomics. Cambridge, MA: Cambridge University Press.

Ariely, D., & Silva, J.D. (2002). Payment method design: Economic and psychological aspects of payments.

Balasubramanian, S., Dutta, R., & Tomak, K. (2003). Pricing of digital content when consumers maintain mental accounts.

Balasubramanian, S., & Tomak, K. (2003). Strategic implications of mental accounting.

Camerer, C., Babcock, L., Loewenstein, G., & Thaler, R. (1997). Labor supply of New York City cab drivers: One day at a time. Quarterly Journal of Economics, 111, 408-441.

Dutta, R., Jarvenpaa, S., & Tomak, K. (2003). Impact of feedback and usability of online payment processes on consumer decision making.

Harris, C., & Laibson, D. (1999). Instantaneous gratification.

Heath, C., & Soll, J. (1996). Mental accounting and consumer decisions. Journal of Consumer Research, 23, 40-52.

Henderson, P., & Peterson, R. (1992). Mental accounting and categorization. Organizational Behavior and Human Decision Processes, 51, 92-117.

Hirst, D., Joyce, E., & Schadewald, M. (1994). Mental accounting and outcome contiguity in consumer-borrowing decisions. Organizational Behavior and Human Decision Processes, 58, 136-152.

Kahneman, D., Frederickson, B., Schereiber, C., & Redelmeier, D. (1993). When more pain is preferred to less: Adding a better ending. Psychological Science, 4(6), 401-405.

Kahneman, D., & Knetsch, J. (1992). Valuing public goods: The purchase of moral satisfaction. Journal of Environmental Economics and Management, 22, 57-70.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 2, 263-291.

Laibson, D. (2001, February). A cue-theory of consumption. Quarterly Journal of Economics, 66(1), 81-120.

Loewenstein, G., & Prelec, D. (1991). Negative time preference. American Economic Review: Papers and Proceedings, 82(2), 347-352.

Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. Quarterly Journal of Economics, 107, 573-597.

Loewenstein, G., & Prelec, D. (1993). Preferences for sequences of outcomes. Psychological Review, 100, 91-108.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. American Economic Review, 89(1), 103-124.

O'Donoghue, T., & Rabin, M. (2000). The economics of immediate gratification. Journal of Behavioral Decision Making, 13(2), 233-250.

Pollak, R. (1968). Consistent planning. Review of Economic Studies, 35, 201-208.

Prelec, D., & Loewenstein, G. (1997). Beyond time discounting. Marketing Letters, 8(1), 97-108.

Prelec, D., & Loewenstein, G. (1998). The red and the black: Mental accounting of savings and debt. Marketing Science 17(1), 4-28.

Prelec, D., Loewenstein, G., & Zellamayer, O (1997, October). Closet tightwads: Compulsive reluctance to spend and the pain of paying. Proceedings of the Association for Consumer Research Annual Conference, Denver, CO.

Prelec, D., & Simester, D. (2001). Always leave home without it: A further investigation of the credit-card effect on willingness to pay. Marketing Letters, 12(1), 5-12.

Ross, W., & Simonson, I. (1991). Evaluations of pairs of experiences: A preference for happy endings. Journal of Behavioral Decision Making, 4, 273-282.

Soman, D. (2001a, March). Effects of payment mechanism on spending behavior: The role of rehearsal and immediacy of payments. Journal of Consumer Research, 27,460-474.

Soman, D. (2001b). The mental accounting of sunk time costs: Why time is not like money. Journal of Behavioral Decision Making, 14, 169-185.

Strotz, R. (1956). Myopia and inconsistency in dynamic utility maximization. Review of Economic Studies, 23, 165-180.

Thaler, R. (1980). Toward a positive theory of consumer choice. Journal of Economic Behavior and Organization, 1, 39-60.

Thaler, R. (1985). Mental accounting and consumer choice. Marketing Science, 4(3), 199-214.

Thaler, R. (1999). Mental accounting matters. Journal of Behavioral Decision Making, 12, 183-206.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and rationality of choice. Science, 211, 453-458.

This work was previously published in Advances in the Economics of Information Systems, edited by K. Tomak, pp. 206-226, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 5.6
Consumer Perceptions and Attitudes Towards Mobile Marketing

Amy Carroll
Victoria University of Wellington, New Zealand

Stuart J. Barnes
University of East Anglia, UK

Eusebio Scornavacca
Victoria University of Wellington, New Zealand

ABSTRACT

Mobile marketing is an area of m-commerce expected to experience tremendous growth in the next 5 years. This chapter explores consumers' perceptions and attitudes towards mobile marketing via SMS through a sequential, mixed-methods investigation. Four factors were identified and proven as all having a significant impact on mobile marketing acceptance—permission, content, wireless service provider (WSP) control, and the delivery of the message, which guided the development of a revised and empirically tested model of m-marketing consumer acceptance. The findings also suggest that marketers should be optimistic about choosing to deploy mobile marketing, but exercise caution around the factors that will determine consumer acceptance. The chapter concludes with a discussion about directions for future research.

INTRODUCTION

One area of m-commerce that is expected to experience tremendous growth is global wireless advertising. It has been predicted that the mobile marketing industry will grow from $4 billion to $16 billion from 2003 to 2005 (Ververidis & Polyzos, 2002). Mobile marketing provides new revenue streams and the opportunities for subsidized access, along with the potential for

customers to experience more convenient and relevant content value, sponsored by advertising (Barnes & Scornavacca, 2004). It is expected that 33% of cellular service provider's revenue will be coming from advertising and from payments and commissions from mobile commerce activities (Ververidis & Polyzos, 2002).

Wireless marketing allows effective targeting and tailoring of messages to customers to enhance the customer-business relationship (Barnes & Scornavacca, 2004). Studies on this new advertising medium indicate that mobile advertising campaigns can generate responses, which are as high as 40% compared with a 3% response rate through direct mail and 1% with Internet banner ads (Jelassi & Enders, 2004). Despite this phenomenal marketing potential, there has been very little research on mobile marketing and particularly through its most successful application, short message service (SMS) (Barnes & Scornavacca, 2004). According to GSM Association, cell phone users send more than 10 billion SMS messages each month, making SMS the most popular data service (Dickinger, Haghirian, Murphy, & Scharl, 2004). Conceptual frameworks and models identified in the literature provide insight into the critical success factors of m-commerce marketing; however, very few of these studies have empirically tested or generated models from a consumer's perspective (Barnes & Scornavacca, 2004; Dickinger et al., 2004; Scornavacca & Barnes, 2004).

The aim of this chapter is to explore consumers' perceptions and attitudes towards mobile marketing via SMS, and to empirically test Barnes and Scornavacca's (2004) m-marketing acceptance model. The following section provides a background to mobile marketing and identifies some of the prominent models in the m-business literature. It also examines the factors believed to influence consumer acceptance of mobile marketing. The third section discusses the methodology, while the fourth and fifth sections provide the results of the study and a revised model for mobile marketing acceptance. The chapter concludes with a discussion about the future for SMS mobile marketing, and directions for further research.

BACKGROUND ON MOBILE MARKETING

Mobile marketing can be defined as "Using interactive wireless media to provide customers with time and location sensitive, personalized information that promotes goods, services and ideas, thereby generating value for all stakeholders" (Dickinger et al., 2004). This definition includes an important concept of adding value not just for the marketing party, but also for the consumer. The literature shows a variety of technological platforms such as wireless application protocol (WAP), SMS, and multimedia message service (MMS) that are available to support mobile marketing applications (Barnes & Scornavacca, 2004; Dickinger et al., 2004).

SMS is the most popular mobile data application to date, showing phenomenal usage with 580 million mobile messaging users sending over 430 billion messages worldwide in 2002 (TTI, 2003). Text message services have been hugely popular for interpersonal communication, allowing users of all ages to exchange messages with both social and business contacts (Dickinger et al., 2004; Xu, Teo, & Wang, 2003). Xu, Teo, and Wang (2003) identified three consistent success indicators for SMS messaging. The first factor is the cost effectiveness and interoperability of the wireless infrastructure, the second is the high penetration of mobile phones (ubiquitous penetration levels of over 80% in some countries), and the third is the relatively low cost of the SMS messaging service.

Countries such as Japan, New Zealand, Germany, and the UK have cost-effective and interoperable wireless structures, a high penetration of mobile phones, and a relatively low cost for the SMS messaging service have experienced

remarkable success with the SMS application (Barnes & Scornavacca, 2004). The success that SMS has had as a messaging service provides a potentially huge SMS messaging customer base which could lend itself as a SMS mobile marketing customer base, making it an attractive opportunity for marketers (Kellet & Linde, 2001).

One of the main challenges and opportunities for mobile advertising companies is to understand and respect the personal nature of the usage of mobile phones (Barnes & Scornavacca, 2004; Barwise & Strong, 2002; Jelassi & Enders, 2004; Heinonen & Strandvik, 2003).

Consumer Acceptance of Mobile Marketing

The acceptance of a mobile marketing message is likely to be influenced by the consumer's acceptance of the mobile medium, the relevance of the content, and the context of the marketing message (Barnes & Scornavacca, 2004; Dickinger et al., 2004; Enpocket, 2003; Heinonen & Strandvik, 2003). Messages that are concise, funny, interactive, entertaining, and relevant to the target group usually achieve higher levels of success (Dickinger et al., 2004; Jelassi & Enders, 2004). The recent m-business literature offers a couple of frameworks that investigate user acceptance of SMS based mobile marketing (Barnes & Scornavacca, 2003; Dickinger et al., 2004).

The guiding model used for this research is the conceptual model of permission and acceptance developed by Barnes and Scornavacca (2004). This model was selected as it looks at a small subset of factors identified in the literature, which are believed to be the most important variables influencing consumer acceptance.

Barnes and Scornavacca (2004) believed that user permission, wireless service provider control (WSP), and brand recognition are the three most important variables that could influence consumers' acceptance of mobile marketing.

Among those, user permission was believed to be the most important variable, the main reason for this being that most consumers are fearful of SMS mobile marketing becoming like e-mail marketing, that is, with high levels of spam. WSP control is found to increase the probability of user acceptance to mobile marketing. This was supported by the fact that users are likely to have high levels of trust with their WSP (Enpocket, 2002b; Ericsson, 2000).

The model also puts forward eight propositions of varying levels of acceptance according to the different combinations of factors. Table 1 presents Barnes and Scornavacca's (2004) hypothesized acceptability of SMS marketing messages based on high and low levels of permission, WSP control, and brand trust. This model is yet to be empirically tested with primary data.

These propositions provide a starting point in further exploring the factors that could contribute to consumer acceptance of mobile marketing.

METHODOLOGY

The chosen strategy of inquiry for this research is sequential exploratory mixed methods. Sequential procedures are ones in which the researcher uses the findings of one method to elaborate on or expand with another method (Creswell, 2003; Green, Caracelli, & Graham, 1989). The objectives of the sequential exploratory approach for the purpose of this study is to use two qualitative focus groups to explore the perceptions of mobile marketing, focusing on the main variables believed to influence mobile marketing acceptance, and then elaborate on this through experimental research in which the findings of the initial phase will be used. The empirical data will hopefully confirm what has been identified from the literature and the findings from the focus groups.

Table 1. Scenarios for m-marketing acceptance (Barnes & Scornavacca, 2004)

User's Permission	WSP Control	Brand Trust	Acceptance
High	High	High	High Acceptance
High	High	Low	Acceptable
High	Low	High	Acceptable
High	Low	Low	Acceptable
Low	High	High	Low Acceptance
Low	High	Low	Low Acceptance
Low	Low	High	Low Acceptance
Low	Low	Low	Not Acceptable

Focus Groups

The samples for the focus groups were purposely selected based on convenience sampling, availability, and profiling. Participants for both groups were in the age range 20–28 reflective of one of the major target groups for SMS mobile marketing. Four participants were selected for focus group A and five participants for focus group B. The participants in focus group A had a greater knowledge of mobile commerce technologies and applications than the participants in focus group B, which was purposely achieved in order to canvas a range of experiences and provide differing viewpoints. The participants in this study were students of a university in New Zealand as well as professionals working in the local central business district.

Interviews were based on open-ended questions and triggers. Video recording was used to tape the focus group discussions, with additional notes being taken by the facilitator. The advantages of using a focus group was that a range of ideas and perceptions were derived and the dynamics of the group provided a rich understanding of the research problem. These focus groups generated new propositions that were tested in the survey questionnaire phase.

Data analysis for the focus groups involved initially transcribing interviews and sorting the data into groups of information based on various topics. The transcriptions were then read over to look for ideas, depth, and credibility of the information from participants; thoughts were noted down in the margins of the transcript (Creswell, 2003). A coding process was then carried out where the data was organized into clusters before any meaning was derived from it (Rossman & Rallis, 1998). The themes and categories identified from the analysis are the major findings of the qualitative phase, and have been shaped into a general description of the phenomenon of mobile marketing acceptance (see the results section for details). Reliability measures were used to check for consistency of themes and patterns, while validity measures (triangulation, member checking, bias discussion, and peer debriefing) were used to determine the accuracy of the findings (Creswell, 2003).

Survey Questionnaire

This phase involved the use of a cross-sectional survey questionnaire to test the acceptance of mobile marketing messages against 16 various propositions that were formulated from the results of the focus groups. The advantage of using a survey in this study was the economy and rapid turnaround of data collection that a survey provides. Surveys are also advantageous in their ability to make inferences about consumer behaviour for given populations based on a sample (Babbie, 1990).

A survey questionnaire was chosen due to its cost effectiveness, data availability, and convenience. Seventy-eight participants for the quantitative phase of the research were selected using random convenience sampling with eight members of the sample being nonrespondents.

The instrument used in the survey was a modified version of the permission and acceptance model of mobile marketing developed by Barnes and Scornavacca (2004) with four variables: permission, WSP control, content, and delivery of the message. Sixteen propositions were formulated around these variables that were tested with a 4-point Likert scale ranging from "unacceptable" to "accept enthusiastically."

The data that was collected from the surveys was entered into an Excel spreadsheet, and statistical calculations were carried out. The 16 propositions were then placed in a table with the expected and actual levels of acceptance that were found for each proposition (see Tables 2 and 3). Tabular analysis was conducted in order to analyze the change in SMS mobile marketing acceptance through the various combinations of the set of variables (permission, WSP control, content, and delivery). The results from the quantitative phase were then compared against previous literature in order to provide further insight of the findings.

To avoid possible threats to validity, caution was taken when the results of this experiment were generalized to other populations and environments, when conducting statistical analysis on the data, and when the definitions and boundaries of the terms were defined.

RESULTS FROM THE FOCUS GROUPS

While focus group A was more knowledgeable in the area of mobile commerce, mobile technologies, and the potential of mobile marketing; both focus groups had only ever experienced mobile marketing through their wireless service providers. To some extent the participants' experience of receiving marketing messages from their service provider influenced their individual perceptions and perceived importance of varying factors contributing to consumer acceptance. The results of both focus groups were consistent with little disparity between the two.

Factors identified in the focus groups as having a significant impact on consumer acceptance of mobile marketing were permission to receive mobile marketing messages, control of the wireless service provider, relevance of the content, timeliness and frequency of the messages, simplicity and convenience of the messages, the brand or company sending the message, the control of the marketing from the consumer, and the privacy of the consumer. Consistent with Barnes and Scornavacca's (2004) model, permission and WSP control were perceived to have a heavy bearing on the acceptance of a mobile marketing message; however, brand was found to have little or no impact on acceptance than the likes of content, and time and frequency of the messages. The emerging there are classified as follows:

- Permission: Permission raised the most discussion in each focus group, and it was concluded by the participants as the most important success factor. Participants stated that consumers should have to "opt in" before they receive mobile marketing messages of

any kind, and have the option to "opt out" at any stage.

- Wireless service provider (WSP) control: Although there was great emphasis on permission, it was also strongly felt that there needed to be a degree of filtering from the service provider. As participant A stated, "there has to be some sort of protection; they can't just open it up to anyone—if companies want to market to customers they should have to go through Vodafone." The idea was raised that if participants had just one company to go to which was linked to their service provider, then there would be just one point of contact allowing consumers to easily "opt in" and "opt out" rather than tracking down several different companies. Participants agreed that it should be evident in the message that it is being filtered by the service provider and legitimate.
- Personalization and content: It was agreed that permission regarding time of day, frequency, and content would also be critical to the acceptance of mobile marketing. Both focus groups agreed that content and its relevance would play a key role in the acceptance of a mobile marketing message, with some participants arguing this as the most important factor. It was believed that marketers should make use of the technology and the advantages it provides over traditional forms of marketing and the Internet, looking to add value other than just advertising. Other ideas discussed in the focus groups were to tie content with location, timing, and ensure that the format of the message works with the limitations of the phone.
- Frequency: Participants agreed that there would be a limit to the number of mobile marketing messages they wished to receive, and there should be some control over the number of messages they are receiving depending on what good or service was being marketed or the industry (e.g., food/flowers). Both focus groups agreed that if consumers were to be hounded by marketing messages, it may result in switching providers, or deleting messages without reading them.
- Time: Participants raised the issue of time playing an important role in the acceptance of mobile marketing messages. It was believed that it is important for consumers to receive marketing messages at times suitable for them, and consumers are able to not only give permission to receive messages but also choose the times they wish to receive them.
- Brand: As far as the brand or company that was marketing was concerned, the general feeling among both focus groups was that as long as the marketing messages were being filtered by the service provider it would not matter too much who it was from; however, if it was third party, they would be annoyed right away. The majority of participants argued that it would be the more well-known brands or brands that the individual consumer recognizes. However, some consumers may prefer to receive messages from a little boutique shop down the road and there should be a way smaller companies can afford mobile marketing. Again if the brand or company doing the marketing was to go through the wireless service provider, this would result in an even higher level of trust. Focus group B believed that consumers should be able to select which companies and brands they receive messages from to a very specific point.
- Technology/Ease of use: A number of important issues were raised with regard to the mobile technology and convenience of the marketing message, some of which have already been pointed out in the previous sections. The main point raised that falls under this section is that marketing mes-

sages should not be a hassle for consumers to receive, they should work with the limitations of the phone, and there should be a manageable way to deal with them.

REVISED MODEL AND SURVEY RESULTS

Four conceptual factors emerged as having the most influence on consumer acceptance based on the tabular analysis and findings of the focus groups. Similar topics were merged as conceptualized themes and then these themes were analyzed according to the number of times they were mentioned in the focus groups, whether these comments were implying that they were important factors and whether the participants explicitly stated them as being one of the most important factors.

Table 2 presents 16 new propositions based on varying combinations of the identified factors, ranked according to the importance of factors: (1) permission, (2) WSP control, (3) content, (4) delivery, and also the number of factors which are low (0, 1, 2, 3, or 4).

The results obtained in the survey demonstrated that propositions 6, 7, 8, 11, 12, 13, 14, 15, and 16 were supported, while propositions 1, 2, 3, 4, 5, 9, and 10 were not found to be supported by the data collected. Tables 3 and 4 show the revised propositions with the expected and actual levels of acceptance for mobile marketing. Notice that the second table actually shows the propositions

Table 2. Revised model with the sixteen scenarios for marketing acceptance

Proposition	Permission	WSP Control	Content	Delivery	Expected Acceptance Level
1	High	High	High	High	Accept Enthusiastically (4)
2	High	High	High	Low	Acceptable (3)
3	High	High	Low	High	Acceptable (3)
4	High	Low	High	High	Acceptable (3)
5	Low	High	High	High	Acceptable (3)
6	High	High	Low	Low	Accept reluctantly (2)
7	High	Low	High	Low	Accept reluctantly (2)
8	High	Low	Low	High	Accept reluctantly (2)
9	Low	High	High	Low	Accept reluctantly (2)
10	Low	High	Low	High	Accept reluctantly (2)
11	Low	Low	High	High	Accept reluctantly (2)
12	High	Low	Low	Low	Unacceptable (1)
13	Low	High	Low	Low	Unacceptable (1)
14	Low	Low	High	Low	Unacceptable (1)
15	Low	Low	Low	High	Unacceptable (1)
16	Low	Low	Low	Low	Unacceptable (1)

Table 3. Revised model ranked according to expected results

		WSP Control	Content	Delivery	Expected Acceptance Level		Rank	Actual level
1	High	High	High	High	Accept Enthusiastically (4)	3.16	1	Acceptable
2	High	High	High	Low	Acceptable (3)	1.60	8	Accept reluctantly
3	High	High	Low	High	Acceptable (3)	1.99	3	Accept reluctantly
4	High	Low	High	High	Acceptable (3)	2.29	2	Accept reluctantly
5	Low	High	High	High	Acceptable (3)	1.91	4	Accept reluctantly
6	High	High	Low	Low	Accept reluctantly (2)	1.50	9	Accept reluctantly
7	High	Low	High	Low	Accept reluctantly (2)	1.70	5	Accept reluctantly
8	High	Low	Low	High	Accept reluctantly (2)	1.63	7	Accept reluctantly
9	Low	High	High	Low	Accept reluctantly (2)	1.43	10	Unacceptable
10	Low	High	Low	High	Accept reluctantly (2)	1.41	11	Unacceptable
11	Low	Low	High	High	Accept reluctantly (2)	1.66	6	Accept reluctantly
12	High	Low	Low	Low	Unacceptable (1)	1.41	12	Unacceptable
13	Low	High	Low	Low	Unacceptable (1)	1.30	14	Unacceptable
14	Low	Low	High	Low	Unacceptable (1)	1.30	15	Unacceptable
15	Low	Low	Low	High	Unacceptable (1)	1.39	13	Unacceptable
16	Low	Low	Low	Low	Unacceptable (1)	1.19	16	Unacceptable

reshuffled in order to demonstrate their rank of acceptance according to the results.

Overall, consumer acceptance of mobile marketing messages was much lower than expected. Over 50% of respondents answered unacceptable to more than 10 out of the 16 scenarios put forward to them, with the average number of scenarios answered as unacceptable being 9. On the other hand, nearly 70% of the respondents did not answer "accept enthusiastically" to anything, and of the 30% who did give this response for at least one scenario, more than 80% only gave this response for one or two of the questions (Tables 3 and 4).

Of all the propositions the highest level of acceptance for mobile marketing was as expected for proposition 1. However, it can be seen that even where consumers have given permission, the content of the message was relevant, the delivery appropriate, and the message had come through

Table 4. Revised model ranked according to actual results

		WSP Control	Content	Delivery	Expected Acceptance Level		Rank	Actual level
1	High	High	High	High	Accept Enthusiastically (4)	3.16	1	Acceptable
4	High	Low	High	High	Acceptable (3)	2.29	2	Accept reluctantly
3	High	High	Low	High	Acceptable (3)	1.99	3	Accept reluctantly
5	Low	High	High	High	Acceptable (3)	1.91	4	Accept reluctantly
7	High	Low	High	Low	Accept reluctantly (2)	1.70	5	Accept reluctantly
11	Low	Low	High	High	Accept reluctantly (2)	1.66	6	Accept reluctantly
8	High	Low	Low	High	Accept reluctantly (2)	1.63	7	Accept reluctantly
2	High	High	High	Low	Acceptable (3)	1.60	8	Accept reluctantly
6	High	High	Low	Low	Accept reluctantly (2)	1.50	9	Accept reluctantly
9	Low	High	High	Low	Accept reluctantly (2)	1.43	10	Unacceptable
10	Low	High	Low	High	Accept reluctantly (2)	1.41	11	Unacceptable
12	High	Low	Low	Low	Unacceptable (1)	1.41	12	Unacceptable
15	Low	Low	Low	High	Unacceptable (1)	1.39	13	Unacceptable
13	Low	High	Low	Low	Unacceptable (1)	1.30	14	Unacceptable
14	Low	Low	High	Low	Unacceptable (1)	1.30	15	Unacceptable
16	Low	Low	Low	Low	Unacceptable (1)	1.19	16	Unacceptable

the WSP, it was found on average to be only acceptable, with just 31% of respondents accepting this message enthusiastically. Thus disproving proposition 1. Alternatively on average the lowest level of acceptance (unacceptable) was found where there was a low level of all these factors. Only 9 out of the 70 participants answered anything other than unacceptable for this question. This result was expected and consistent in proving proposition 16.

Permission and delivery of the message were the two variables that were found to equally have the most influence on the participant's level of acceptance, while content was found to be the next most important factor with control of the WSP having the least amount of impact on the level of acceptance. Participants were more likely to accept messages that had a lower level of WSP control or irrelevant content than messages that they had not given permission for or that came at an

inappropriate time or frequency. This was shown again in Table 4, rows 12–15, where participants found scenarios 13 and 14 more unacceptable, despite having high levels of WSP control and content, respectively, than scenarios 12 and 15 where there were higher levels of permission and appropriate delivery, respectively.

It is interesting to note that consistent with the propositions, the level of acceptance declined with the number of factors that had low levels, except in the case of proposition 2, which was expected to generate the second highest level of acceptance and in actual fact dropped down to position 8. Where all other factors were high, yet the delivery of the message was inappropriate, more than 50% of respondents found this message unacceptable, compared to just 26% of respondents who considered a message with low levels of WSP control unacceptable.

Looking at the other rankings of propositions from their expected to actual perceived influence on acceptance, just three propositions stayed in the same ranked position. However, of the propositions that did get shuffled in rank, nine of these moved only within one or two ranks, with just three propositions moving three places or more. Participants found all messages that had three or more factors with low levels to be completely unacceptable. This was consistent with the expected results, and supported the propositions 12, 13, 14, 15, and 16. Messages that had only high levels of WSP control or relevant content were found to be 10% less unacceptable than messages with only high levels of permission or appropriate delivery—thus supporting the theory that permission and delivery of the messages are perceived to be the most important factors.

DISCUSSION

The findings indicated a number of factors that are critical to the acceptance of mobile marketing by consumers. While the empirical testing showed that some factors are more important than others in influencing the overall level of acceptance, it was found that all factors played a significant role.

Consistent with the literature explicit permission was found to be essential (Barnes & Scornavacca, 2004; Enpocket, 2003; Godin et al., 1999). The wireless channel is relatively protected and spam free with consumers having little experience with mobile marketing. Due to the personal nature of the phone, and experiences with unsolicited spam via e-mail users were weary of receiving marketing to their cell phones, and a number of privacy issues were raised in the focus groups. Another finding that emerged from the study was the importance of delivery with the marketing message. Literature has suggested that frequency and time are linked to targeting, where users are happy to receive messages at a higher frequency so long as the relevance to them is maintained (Enpocket, 2002b). This was supported by the empirical testing where it shows messages with a low level of relevant content yet appropriate delivery were found to be much more acceptable than messages with a low level of relevant content and inappropriate delivery (a higher frequency). While participants in the focus groups made a point of saying that it is useless receiving any messages containing content that is irrelevant, there are a number of possible reasons why the respondents may have found delivery to be more important. If a consumer receives a message that is irrelevant to them once in a blue moon, and it does not come at a disturbing time, they may not be that bothered by it. On the other hand, if they were messages on something that was relevant to them but were receiving these messages continuously and at interruptive times, it is likely to be more unacceptable.

It was interesting to see that the control of the WSP had the least impact on consumer acceptance in the survey results, conflicting with the results of the focus groups where participants expressed their strong opinions towards the importance of

WSP control. The results may in fact indicate that where consumers receive messages they find disturbing or intrusive, they would rather it had not come from the service provider they trust. The focus groups indicated this, stating that they trust their service provider's judgment and would expect them to behave responsibly. Consumer attention seems more likely to divert to the filter when they are receiving unsolicited messages that they find disturbing.

Despite literature showing mobile marketing to be a successful tool in building brand awareness, and an important factor in consumer acceptance (Dickinger et al., 2004; Enpocket, 2002a), the study revealed that the brand being marketed may have very little impact. Consumers are more likely to care whether a brand has been accepted by their service provider and has come through a filter, than about their level of trust between two different brands. Despite having a high trust in a brand, consumers are still doubtful of the bona fide of these messages when they have come direct. They are also less likely to care about the brand that is being marketed to them than whether the content is relevant. The importance that is placed on brand is likely to increase when all other factors are high, and there is more choice in the market. Currently there are a limited number of brands being marketed through the mobile phone in New Zealand and more attention to brand is likely to arise in the future where consumers receive similar messages, with all other factors being equal, from competing brands.

CONCLUSION

This research highlights the importance of consumer perceptions and acceptance levels of mobile marketing. The literature showed the powerful marketing potential that mobile marketing can offer companies through its anytime and anywhere nature, yet limited research looking at consumers' perceptions and acceptance of mobile marketing has been carried out. This study set out to overcome the apparent gap in the literature, and through the use of both qualitative and quantitative methodology, a model has been adopted, explored, developed, and empirically tested and validated.

This study suggests that marketers should be optimistic about choosing to deploy mobile marketing; however, exercise caution around the factors that will determine consumer acceptance. While consumers can see the potential in the mobile medium, they are weary of receiving unsolicited messages they do not want. Obtaining user trust and permission will be the main challenge faced by marketers, and future research should focus on ways to overcome these challenges. Consumers are more likely to trust messages coming from their service providers than anywhere else, so it is important that service providers provide a high level of filtering and protection as reassurance for their users. Trust and permission are necessary factors of consumer acceptance; however, they should not be seen as the only objectives. Attention needs to be focused around the relevance of the content and the timeliness and frequency of the delivery of marketing messages.

The research showed that simply focusing on contextual, content, or permission/control factors in isolation is unlikely to result in a high or even moderate level of acceptance. Instead, marketers need to take into account all these factors and how varying combinations of these factors will impact consumer acceptance.

The permission and acceptance model, which has been developed and tested in this research, provides a foundation for further SMS mobile marketing research to be built upon. Academics can refer to this model as a guide for further understanding of consumer acceptance to mobile marketing, while practitioners may find this model useful in providing direction for mobile marketing strategies. The device media aspects discussed in the focus groups may also provide an indication as to what new technologies and mobile devices

will be of significance in meeting consumers' needs for the future.

The generalizability of this study is limited by it being conducted only in New Zealand as well as the lack of further qualitative interviews to further elaborate on the initial quantitative analysis. This cross-sectional study only looked at consumer acceptance at one point in time, and little is known about the sample frame that was used for the survey questionnaire. Furthermore the sample of the participants for the quantitative phases was only a small number which leaves possibility for self-selection bias. Longitudinal research testing consumer perceptions and acceptance over a set amount of time, and taking into account demographics when testing consumer acceptance levels would provide some deeper insight into these areas.

REFERENCES

Babbie, E. (1990). Survey research methods (2nd ed.). Belmont, CA: Wadsworth.

Barnes, S.J., & Scornavacca, E. (2004). Mobile marketing: The role of permission and acceptance. International Journal of Mobile Communications, 2(2), 128–139.

Barwise, P., & Strong, C. (2002). Permission-based mobile advertising. Journal of Interactive Marketing, 16(1), 14–24.

Creswell, J. (2003). Research design qualitative, quantitative, and mixed methods approaches (2nd ed.). Thousand Oaks, CA: Sage Publications.

Dickinger, A., Haghirian, P., Murphy, J., & Scharl, A. (2004). An investigation and conceptual model of SMS marketing. Paper presented at the 37th Hawaii International Conference on System Sciences, HI.

Enpocket. (2002a). The branding performance of SMS advertising. Retrieved March 13, 2003, from www.enpocket.co.uk

Enpocket. (2002b). Consumer preferences for SMS marketing in the UK. Retrieved March 13, 2003, from www.enpocket.co.uk

Enpocket. (2003). The response performance of SMS advertising. Retrieved March 12, 2003, from www.mda-mobiledata.org

Ericsson. (2000). Wireless advertising. Stockholm: Ericsson Ltd.

Godin, S., Hardcover, p., 1 edition (May 1, & 0684856360., S. S. I. (1999). Permission Marketing: Turning strangers into friends, and friends into customers.

Green, J.C., Caracelli, V.J., & Graham, W.F. (1989). Toward a conceptual framework for mixed method evaluation designs. Educational Evaluation and Policy Analysis, 11(3), 255–274.

Heinonen, K., & Strandvik, T. (2003, May 22–23). Consumer responsiveness to mobile marketing. Paper presented at the Stockholm Mobility Roundtable, Stockholm, Sweden.

Jelassi, T., & Enders, A. (2004, June 14–16). Leveraging wireless technology for mobile advertising. Paper presented at the 12th European Conference on Information Systems, Turku, Finland.

Kellet, K., & Linde, A. (2001). EMS, MMS, & the future of mobile messaging, white paper. Retrieved , from, www.magic4.com.

Rossman, G.B., & Rallis, S.F. (1998). Learning in the field: An introduction to qualitative research. Thousand Oaks, CA: Sage.

Scornavacca, E., & Barnes, S.J. (2004, March). Raising the bar: Barcode-enabled m-commerce solutions. Paper presented at the Austin Mobility Roundtable, Austin, TX.

TTI. (2003). Mobile messaging: Which technologies and applications will succeed? Retrieved July 5, 2004, from www.telecomtrends.net

Ververidis, C., & Polyzos, G. (2002). Mobile marketing using location based services. Paper presented at the First International Conference on Mobile Business, Athens, Greece.

Xu, H., Teo, H.H., & Wang, H. (2003, January 7–10). Foundations of SMS commerce success: Lessons from SMS messaging and co-opetition. Paper presented at the 36th Hawaii International Conference on System Sciences, Big Island, HI.

This work was previously published in Unwired Business: Cases in Mobile Business, edited by S. Barnes and E. Scornavacca, pp. 109-123, copyright 2006 by IRM Press (an imprint of IGI Global).

Chapter 5.7
An Empirical Examination of Customer Perceptions of Mobile Advertising

Su-Fang Lee
Overseas Chinese Institute of Technology, Taiwan

Yuan-Cheng Tsai
Da-Yeh University, Taiwan

Wen-Jang (Kenny) Jih
Middle Tennessee State University, USA

ABSTRACT

A two-stage approach is employed in order to examine the influencing factors of consumer behaviors in the context of mobile advertising. The first stage of the study evaluates the correlation relationship of consumer motives for receiving mobile advertising and their attitudes toward mobile advertising. It also investigates the relationship between consumer intentions for receiving advertisements on their cellular phones and their subsequent actions once the mobile advertising was received. A negative sentiment was revealed by cellular phone users toward mobile advertising, a signal that current practices of mobile advertising are ineffective and require a careful reevaluation on the part of mobile commerce firms. The second stage of the research validates a Fishbein and Ajzen's Theory of Reasoned Action model. It is found that positive actions on the received advertisements are significantly influenced by strong intentions; strong intentions are influenced significantly by favorable attitudes, and favorable attitudes are influenced significantly by strong motives. Implications for e-commerce application developers and marketers are discussed.

INTRODUCTION

The convergence of the Internet and wireless communications has led to the development of

an emerging market for mobile e-commerce, or m-commerce. As the business impact of e-commerce has been witnessed in almost every facet of the business arena, the advancement of wireless Internet access capabilities is adding to the convenience and flexibility of the online shopping process. This growing trend of m-commerce has been confirmed by numerous industry research reports. For example, Malhotra and Segars (2005) reported that the global market for mobile commerce is predicted to reach $20 billion in 2006. Web-enabled wireless devices allow users to search, communicate, and purchase products from anywhere at any time. These convenient features are contributing to e-commerce's growth in the knowledge economy, as attention and time are becoming scarce resources for consumers (Hague, 2004).

As wireless technologies and standards for security, bandwidth, and interoperability continue to advance, the impact of online shopping via wireless communication devices is becoming a crucial issue for marketers as they are striving to design their organizations' marketing and other strategic initiatives. This new development also is posing a new challenge for information system personnel, as they often are called upon to implement enabling system capabilities to support innovative business initiatives. Different from wired communication networks, wireless networks are relatively more limited in processing power, transmission bandwidth, user interface (e.g., screen size) and security protection. Advancements in all these areas, however, have been made in order to improve the technical capabilities of wireless communication as a viable vehicle for serious business innovation. Further, information system personnel need to be guided by an integrated framework that addresses the relationships among technology, user, and application domain. However, most existing literatures on m-commerce are anecdotal reports that center on industrial development. Systematic empirical investigation into various aspects of m-commerce development is relatively limited. Clarke (2001) points out this problem, saying, "Despite tremendous interest in the melioration of m-commerce, there is little, if any, research that examines how to develop a comprehensive consumer-oriented mobile e-commerce strategy" (p. 134).

This study is a response to this calling. Our concern is with business practices and theory development in mobile commerce. The objective is to obtain a theory-based understanding of an important aspect of mobile commerce: mobile advertising. Basing our study on a well-established theory facilitates a systematic inquiry of a newly emergent phenomenon such as mobile advertising. Such an inquiry not only provides better understanding of mobile advertising, but it also generates additional new evidence for further validation of the theory.

The use of wireless communication services is becoming a global phenomenon. Cellular phones increasingly are becoming an essential vehicle for business and personal communications, as well. Mobile phone users are being targeted by companies that seek to incorporate Internet-enabled operations into their advertising approaches. These companies must develop their business strategy based on an in-depth understanding of the distinct characteristics of their customers.

Guided by the Theory of Reasoned Action proposed by Fishbein and Ajzen (1975), the study examined the consumer motives (beliefs), attitudes, intentions, and actions associated with e-commerce advertising through Web-enabled cellular phone services. A two-staged, empirical study was conducted in order to investigate consumer perceptions of and reactions to mobile advertising via cellular phones. The purpose of the first stage is to develop a theoretical framework by analyzing the survey data using factor analysis and canonical correlation analysis. This framework then was validated in the second stage using structured equation modeling.

In the next section, we present contrasting views over the future of e-commerce, unique

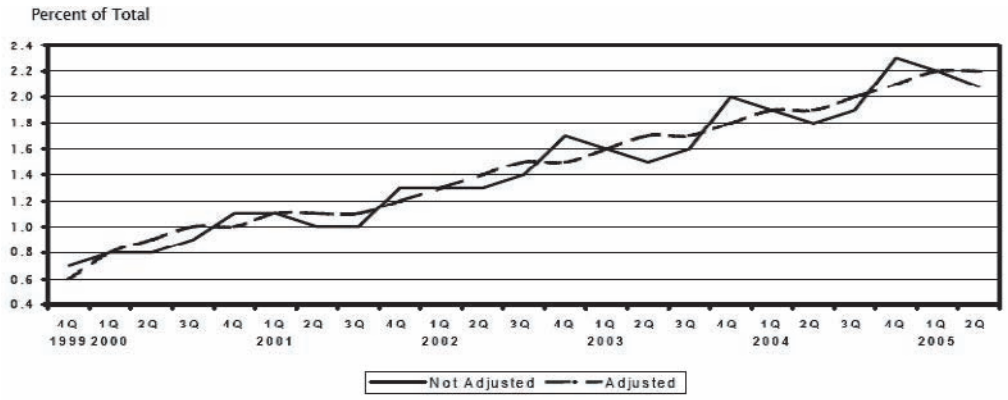
features of m-commerce, and consumers' attitudes toward advertising. We also introduce Fishbein and Ajzen's (1975) Theory of Reasoned Action and elaborate on its implications for mobile advertising. The research hypotheses designed to answer our research questions then are presented. In the section on research method, we describe the research framework and the approach to analyze the collected data. This is followed by the findings in the study. In addition to summarizing the research, the last section also documents the limitations that may negatively affect the validity as well as the generalization ability of the research, and presents our suggestions for future research.

CONTRASTING VIEWS OF ELECTRONIC COMMERCE

Despite a concern expressed in a Jupiter Research report about the likely slowing down of online sales in the near future (Bhatnagar, 2005), most literature on electronic commerce (e-commerce) is still optimistic about the continuing growth of this relatively new business environment. This optimism is fueled partly by economic data from such sources as the Census Bureau of the Department of Commerce. For example, a recent announcement from the Department of Commerce reported that the estimate of U.S. retail e-commerce sales for the second quarter of 2005 represented an increase of 7.2% from the previous quarter of 2004 (Scheleur, King, & Shimberg, 2005). During the period from the fourth quarter of 1999 to the third quarter of 2004, retail e-commerce sales as a percent of total quarterly retail sales generally have exhibited a growing trend though not a smooth line on the chart (see Figure 1).

The optimistic prospect also is supported by various views about e-commerce's unique value propositions. The more reserved view contends that the economic impact of the Internet is limited to improvement of consumer convenience and expanded choices rather than in more direct areas such as higher productivity and lower prices (Litan & Rivlin, 2001; Porter, 2001). The more enthusiastic view, however, holds that the Internet virtually has become a new platform for business activities and, thereby, is transforming many ways in which businesses interact with stakeholders. Joines, Scherer, and Scheufele (2003), for example, emphasize the strategic value of this new platform for developing advertising strategies, recognizing the fact that few of the Internet-based businesses are making profits. Among the unique features of the Internet-based businesses are interactivity,

Figure 1. Estimated quarterly U.S. retail e-commerce sales as a percent of total quarterly retail sales: 4th quarter 1999-2nd quarter 2005



(Source: http://www.census.gov/mrts/www/data/pdf/05Q2.pdf)

rich content, wide reach, personalization, convenience, and online shopping capabilities (e.g., online payment) (Joines et al., 2003; Turban, King, Lee, & Viehland, 2004). The difference between dissimilar views, however, appears to lie mainly in the prospect of rate of growth rather than in the strategic role that e-commerce is playing in all aspects of business activities.

UNIQUE CONSIDERATIONS OF M-COMMERCE

M-commerce generally is defined as the use of wireless communications networking technology as the primary interaction vehicle between buyers and sellers of products or services. Currently, the Web-enabled cellular phone is the most popular device used by customers of m-commerce. This definition accommodates a number of slightly different definitions found in the literature. For example, Siau and Shen (2002) defined m-commerce transactions as those conducted via mobile devices using wireless telecommunication networks and other wired e-commerce technologies. In Wen and Gyires (2002) study, m-commerce was defined as an extension of e-commerce beyond the static terminal of the PC/TV to Web-enabled mobile and other wireless communication devices. As wireless communication technology continues to advance in many directions (e.g., bandwidth, security, user interface, pricing strategy, etc.), substantial growth potential of m-commerce in the near future has been predicted by both practitioners and academicians (Zhang, Yuan, & Archer, 2002).

Innovative business strategies are required to leverage the unique features of wireless communications in order to offer unique and appealing customer value. Contrasted with the traditional, wired telecommunication networks, a wireless communication infrastructure is relatively less expensive to construct in terms of capital requirement and time frame. This cost advantage is applicable to wide-area, metropolitan-area, and local-area network installations (White, 2004). Wireless communication devices are also more tightly tied to the users than to desktop personal computers or fixed line-based telephones. This personalization capability has allowed m-commerce companies to bring customers more into their major business processes, such as new product development, in an attempt to produce outcomes that may enhance customer satisfaction and loyalty (Napier, Judd, Rivers, & Adams, 2003; Varshney & Vetter, 2002). In addition, when equipped with wireless cards and Web-browsing capability, wireless devices such as laptop computers or even cellular phones can be used to access internal as well as external information resources with little concern of wiring for network connection.

Researchers have identified major advantages of m-commerce that are attributable to these unique features of wireless communications. For example, Wen and Gyires (2002) indicated that the key ingredients of m-commerce were portability, connectivity, usability, and ubiquity. Clark (2001) suggested four value propositions of m-commerce that set m-commerce apart from conventional e-commerce: ubiquity, localization, personalization, and convenience. Frolick and Chen (2004) indicated that m-commerce contributes to overall business operations through real-time interactions with customers and immediate dissemination of decision support information to employees. Malhotra and Segars (2005) identified six unique capabilities that may help wireless Web become killer applications: immediacy, constancy, personalization, ubiquity, timeliness, and context. Balasabramanian, Peterson, and Jarvenpaa (2002) emphasized three valuable characteristics of mobile commerce: location sensitivity, time criticality, and user control. In explicating major differences between m-commerce and e-commerce, Zhang et al. (2002) contended, "M-commerce is not simply a new distribution channel, a mobile Internet or a substitute for PCs. Rather, it is a new aspect of consumerism and a much more powerful way to communicate with customers" (p. 83). Rather than treat m-commerce merely

as an extension of e-commerce, a new way of thinking has been called for in order to unleash the value of m-commerce associated with the role of mobility (Clark, 2001; Nohria & Leestma, 2001). From a strategic perspective, the potential of m-commerce can be realized only through the development of a mobile-specific business strategy (Clark, 2001).

CONSUMER ATTITUDE TOWARD ADVERTISING

Consumer attitude toward advertising is characterized by consumers' favorable or unfavorable evaluations of advertising through mobile devices and whether it is evaluative or affective in nature or plays an important role in determining their intention and behavior when exposed to a specific advertising in a specific environment (Fishbein & Azjen, 1975; Mackenzie, Lutz, & Belch, 1986). The positive as well as negative effects of consumer attitude toward advertising have been researched extensively n advertising and marketing. For example, Mitchell and Olson (1981) found that consumers' attitudes toward advertising affected their brand attitudes and purchasing willingness through their emotional feelings over the advertising itself. In general, consumers' attitudes toward advertising reflect the degree to which they identify with the advertising (Mackenzie et al., 1986; Shimp, 1981). In a study conducted to investigate the recall effect of outdoor advertisements, Donthu, Cherian, and Bhargava (1993) found that better recall tended to be exhibited by the respondents with more positive attitudes toward advertising in general.

In contrast with the positive effect of consumer attitudes toward advertising advocated in the early research, the negative aspect has been revealed by more recent studies. The shifting began in the 1970s (Zanot, 1984) and became ever more significant in the 1980s and 1990s (Alwitt & Prabhaker, 1994; Mittal, 1994). The driving forces include such factors as increased awakening of consumerism, risk perception, self-defense, and the excessiveness of advertising activities (Mackenzie et al., 1986). These factors have been confirmed by more recent studies that investigated newer media as well as traditional media. For example, in examining six traditional mass communication media (television, radio, printed magazine, printed newspaper, yellow pages, and direct mail), Elliot and Speck (1998) identified three phenomena associated with negative perception of advertising. First, the excessiveness of the amount of advertising was a matter of perception rather than objective data. Second, consumers often look at advertising as an annoyance that interferes with the content reception. Third, consumers often decide to regain control or to avoid exposure to an unwelcome advertisement. A recent study conducted by Tsang, Ho, and Liang (2004) also revealed the negative aspect of consumer attitude toward advertising in the context of mobile commerce.

THEORY OF REASONED ACTION

Originated in the field of social psychology, the Theory of Reasoned Action (TRA) was developed by Fishbein and Azjen (1975) in the 1970s. The purpose of the theory is to predict and understand the factors influencing an individual's behavior in a specific context. The theory and its subsequent variation, the Theory of Planned Behavior (TPB), have been applied to research in a variety of fields. In the field of management information systems, for example, Harrison, Mykytyn, and Riemenschneider (1997) examined small business executives' intentions to adopt information technology for the purpose of establishing or enhancing sustainable competitive advantage. Mathieson (1991) compared TPB and Technology Acceptance Model for predicting an individual's intention to use an information system. They concluded that TPB

provided more useful information for information systems development. Mykytyn, Mykytyn, and Harrison (2005) used TPB to examine the integration of intellectual property concepts into information systems education.

TRA provides a theoretical foundation for the linkage among four constructs: behavior, intention, attitude, and belief. Through exposure to an object, people link the object with its attributes with varying strengths. "The totality of a person's belief serves as the informational base that ultimately determines his attitudes, intentions, and behaviors" (Fishbein & Ajzen, 1975, p. 14). Influenced by belief, attitude refers to the favorable or unfavorable feelings or evaluations a person holds of an object or a behavior. Behavioral intention refers to "the strength of a person's conscious plans to perform the target behavior" (Mykytyn et al., 2005, p.6). In TRA, intention is hypothesized to be the best predictor of a person's behaviors, which are observable acts of the person. As applied to the context of mobile advertising, the theory suggests that mobile phone users' beliefs or motives regarding mobile advertising affect their positive or negative attitudes toward mobile advertising; that positive attitudes may lead to strong intentions for the received advertisement; and that positive actions taken by the users upon receiving the advertisement (e.g., immediate reading vs. delayed reading) are, in turn, affected by the strong intentions. This study seeks to understand better consumers' responses to an emerging marketing practice, mobile advertising, using TRA as the theoretical guidance.

RESEARCH QUESTIONS

Motivated by the lack of systematic research about the use of mobile communication devices as an advertising medium, this study attempts to answer two general research questions:

1. How do cellular phone users perceive the advertisement received over Web-enabled cellular phones?
2. Can favorable behaviors be attributed to favorable intentions, positive attitudes, and favorable motives or beliefs in the context of mobile advertising?

Given the unique characteristics of the users and the technologies in mobile commerce, it is a great challenge on the part of e-commerce companies to be creative and to devise truly appealing advertising strategies. If the findings based on empirical data confirm the applicability of TRA in this context, then e-commerce companies would be strongly encouraged to first seek ways to help consumers develop favorable beliefs regarding their advertisements and to strengthen the favorable motives and attitudes toward mobile advertising.

Two-Stage Research Design

This study employed a two-stage research approach. The first stage investigated the correlation relationship between motives and attitudes toward mobile advertising and between intentions and behaviors. A theoretical model was formulated as the outcome of this stage. The second stage assessed the applicability of the TRA to mobile advertising by examining the cause-effect relationships between the constructs contained in the theoretical model. The following two hypotheses were formulated to test the correlation relationships addressed in the first stage:

H_1: Consumer attitudes toward mobile advertising are not significantly related to their motives for receiving the advertisement on their Web-enabled cellular phones.

H_2: Consumer behaviors on mobile advertising are not significantly related to their intentions to receive the advertisement.

In order to assess consumers' general perceptions of mobile advertising, two descriptive statistics (means and standard deviations) were obtained to address the first general research question. To respond to the second general research question, correlation relationships were examined using canonical correlation analysis after applying factor analysis to compress the number of variables. The findings from these analyses led to the formulation of the theoretical model that then was validated in the second stage. The statistics software program SPSS (Version 10.0) was used for the analyses in Stage 1. The second stage used AMOS, a structural equation modeling software program, to validate the resultant model. The analysis assessed the causal effect of strong motives on positive attitudes, positive attitudes on favorable intentions, favorable intentions on positive behaviors, and positive attitudes on positive behaviors. The following four research hypotheses, stated in positive forms, were tested in the second stage:

H_3: Strong motives lead to positive attitude.
H_4: Positive attitudes lead to strong intention.
H_5: Positive attitudes lead to positive action.
H_6: Strong intentions lead to positive action.

The questionnaire consisted of six sets of questions that were devised to gather data on motives, attitudes, intentions, behaviors, cellular phone usage, and demographical data. The first set had four questions that asked about consumer motives for receiving advertisements on cell phones. These motives represented their beliefs in the potential benefits of the mobile advertising services. The second set of seven questions addressed consumer attitudes toward mobile advertisements. The three questions in the third set asked about the intentions for the received advertisements. The five questions in the fourth set covered the consumer actions taken on the received advertisements. The last two sets of questions gathered demographical and usage experience data.

A total of 400 questionnaires were distributed in order to gather data from three types of mobile phone users (college students, college employees, and business practitioners) in Taiwan during the months of June and July of 2004. Similar to people in many countries in the more developed world, Taiwanese consumers have found cellular phones to be essential communication tools in their daily lives (Jih & Lee, 2004). The findings of the study, therefore, may have significance for building a more generalized theory in mobile commerce. Some returned questionnaires were discarded because of incomplete or apparently casual responses, resulting in 358 effective responses that were used for data analysis. The structure of effective samples consisted of 33% males and 67% females. The majority (95.5%) of the respondents had usage experience with mobile phones for at least one year and, therefore, can be considered experienced users for the purpose of this study. In general, there were more young consumers than their older counterparts: 15.5% ages 21 and under; 46.9% ages 22 to 30; 27.1% ages 30 to 39; and 11.9% ages 40 and over.

The reliability and validity aspects of the survey questionnaire were assessed to ensure overall adequacy. Factor analysis was performed to assess the dimensionality of the research constructs. When the questions representing each construct were analyzed separately, the analysis revealed only one factor (eigenvalue > 1) for each of the six model constructs (reception motives, positive attitudes, negative attitudes, intentions, positive behaviors, and negative behaviors), an evidence of unidimensionality of the set of questions that represented the construct. In addition, each construct had a fairly high factor loading and extracted variance. An adequate convergent validity of the questionnaire, therefore, was concluded.

The discriminant validity of the questionnaire is another important indicator of the questionnaire adequacy. According to Fornell and Larcker (1981), a questionnaire's discriminant validity is adequate if the individual extracted variance of

each of the two constructs exceeds the square of the correlation coefficient between the two constructs. The result in Table 1 indicates proper discriminant validity.

For reliability assessment, the Cronbach's α values were used as the reliability measures. Both Nunnally's (1978) and Cuieford's (1965) criteria were considered. Nunnaly's (1978) criterion calls for basing the reliability assessment on the threshold value being at least 0.7. Cuieford (1965), however, contended that for an inquiry highly exploratory in nature, the Chronbach's α values greater than 0.7 can be considered high levels of reliability; those between 0.35 and 0.7 can be considered acceptable; and only those with Cronbach's α values less than 0.35 should be discarded. All constructs in our questionnaire had the Cronbach's α values above 0.5, an indication of acceptable reliability. These analyses established the overall adequacy of the questionnaire.

FINDINGS

Consumers' Reactions to Mobile Advertising

In general, the study found that mobile advertising was not receiving an enthusiastic welcome from the cellular phone users in Taiwan. This lack of consumer interest could be witnessed from the low average scores on motives, attitudes, intentions, and actions. On a Likert scale of 1 to 5 with 1 standing for strongly disagree and 5 for strongly agree, the strongest motive was for information acquisition with an average of 2.75. This result indicated that consumers currently did not have much desire to receive advertisements on their cellular phones.

Similar responses were gathered on consumers' attitudes toward mobile advertising. The statements describing mobile advertising as annoying, excessive, and offensive received average scores of 3.41 or higher, an indication of generally unfavorable consumer attitudes toward mobile advertising. This result was consistent with the findings reported in a number of previous studies

Table 1. Discriminant validity evaluation

Construct	Motives	Positive Attitudes	Negative Attitudes	Intentions	Positive Behaviors	Negative Behaviors
Motives	0.68					
Positive Attitudes	0.57	0.61				
Negative Attitudes	0.12	0.12	0.64			
Intentions	0.51	0.56	0.13	0.76		
Positive Behaviors	0.36	0.36	0.06	0.43	0.59	
Negative Behaviors	0.01	0.02	0.01	0.01	0.04	0.69

Note: Measures on the diagonal are extracted variance percentages. The rest are squared correlation coefficients.

(Alwitt & Prabhaker, 1994; Mittal, 1994; Tsang et al., 2004; Zanot, 1984).

Behavior intention measures the strength of a person's conscious plans to perform the target behavior, which in this study measures consumer reception of mobile advertising. The TRA suggests that intention is the best predictor of a person's behavior. Our data found the current state of consumer intentions of receiving mobile advertising less than optimistic. The averages of intention measures for using mobile advertising were 2.54 for purchasing information, 2.35 for enjoyment, and 2.25 for forwarding to friends. This finding offered an alert to mobile commerce marketers and suggested that being sensitive in their advertising practices to customer perception is imperative in engaging customers.

What did the consumers do when they received an advertisement on their cellular phones? The highest ranked action was to keep it aside for later browsing until they had a chance to do so (3.19). The second highest ranked action was Immediate Reading (2.90). It appeared that more people were putting off reading the mobile advertisements than those who read them upon receipt to their mobile phones, a sure sign of lack of consumer enthusiasm toward mobile advertising.

Factor Analyses

The questionnaire employed multiple questions in order to measure each research construct; there were four questions for motives, seven for attitudes, three for intentions, and five for actions. Two constructs — attitudes and actions — were analyzed using factor analysis in order to reduce their number of questions. Two tests were performed on each of the two constructs in order to evaluate the correlation between the observed values (Bartlett Sphericity) and its sampling adequacy (Kaiser-Meyer-Olkin coefficient). For the consumer attitudes, the χ^2 value was 744.487 (p value < 0.001), and the KMO coefficient was 0.775. The tests indicated the existence of correlation between observed values and adequacy of factor analysis. For the consumer actions, the χ^2 value yielded from the Bartlett Test was 763.782 ($p < 0.001$), and the KMO coefficient was 0.798. Both of these two analyses evaluations suggested the adequacy of conducting factor analysis on the consumer actions.

The principal component analysis was first used to extract two factors for the consumer attitudes, with accumulated extracted variance 62.98%. The varimax procedure of the orthogonal rotation approach then was performed to facilitate convenient labeling of the resulting latent variables. The reliability measures of both latent variables were more than 0.70, indicating adequate reliability by Nunally's (1978) standard (Table 2). Based on the variables (questions) included in each of the latent variables, the two factors represented positive attitudes and negative attitudes correspondingly.

As the same factor analysis procedures were applied on the consumer actions, two latent variables (Immediate Reading/Keeping and Delayed Reading) were produced with accumulated extracted variance 63.57% (Table 3). The reliability measures were 0.65 and 0.54, respectively. Although these Cronbach's α values would be considered low by the stricter Nunally's (1978) standard, they are acceptable by Cuieford's (1965) standard. Given the exploratory nature of the measures, we accepted the less strict Cuieford's (1965) rule and used this result in the subsequent canonical correlation analysis.

Canonical Correlation Analyses

The first canonical analysis assessed the correlation relationship between the consumer motives for and attitudes toward receiving mobile advertising (H_1). As shown in Table 4, one set of canonical factors was identified. The four motives for receiving mobile advertising (information acquisition, enjoyment of browsing, novel attraction, and discount deals) were significantly related with

Table 2. Factor analysis of consumers' attitudes toward mobile advertising

Factors	**Questions/Variables**	**Factor Loadings**		**Reliability Coefficient Cronbach's α**
		1	2	
Information Helpful and Interesting (positive attitude)	Mobile advertisements are interesting to me.	0.85	-0.11	0.78
	Receiving mobile advertisements is enjoyable.	0.84	-0.21	
	Mobile advertisements are a great source of timely information.	0.68	-0.01	
	Mobile advertisements are trustworthy.	0.67	-0.21	
Advertising Offensive and Annoying (negative attitude)	Mobile advertisements are offensive.	-0.23	0.80	0.71
	Mobile advertisements are annoying.	-0.23	0.78	
	Mobile advertisements are excessive and out of control.	0.02	0.77	
Eigenvalues		3.01	1.40	
Explained Variance %		42.98	20.01	
Accumulated Explained Variance %		62.98		

Table 3. Factor analysis of consumers' action on mobile advertising

Factors	**Questions/Variables**	**Factor Loadings**		**Reliability Coefficient Cronbach's α**
		1	2	
Immediate Reading and Keeping	Keeping the advertisement	0.81	0.01	0.65
	Reading the entire advertisement	0.79	0.22	
	Reading the advertisement right away	0.68	0.01	
Delayed Reading	Putting off reading the advertisement until more time available	-0.05	0.86	0.54
	Putting off reading until too many advertisements have piled up	0.20	0.78	
Eigenvalues		1.95	1.23	
Explained Variance %		39.02	24.55	
Accumulated Explained Variance %		63.57		

Table 4. Canonical correlation between motives for receiving and attitudes toward mobile advertisements

Motives for Receiving Mobile Advertising	**Canonical Factor**	**Attitude Toward Mobile Advertising**		**Canonical Factor**
	χ^1			λ_1
Information acquisition	0.736	Information Helpful and Interesting	Positive Attitude	0.993
Enjoyment of browsing	0.842			
Novel attraction	0.897	Advertisements Offensive and Annoying	Neg. Attitude	-0.452
Discount deals	0.802			
Extracted Variance %	0.675	0.595		
Redundancy	0.395	0.348		
ρ^2	0.585			
ρ	0.765			

the two latent variables of attitudes toward mobile advertising (information helpful and interesting, advertisements offensive and annoying) at the significant level of 0.01. Hypothesis 1 (Consumers' attitudes toward mobile advertising are not significantly related to their motives for receiving the advertisement on their Web-enabled cellular phones) was rejected.

The canonical analysis between the intentions for receiving mobile advertising and the actions taken on the received mobile advertising also yielded one set of canonical variates. Three intention variables (looking forward to receiving, basing purchasing decisions on the advertising, and likely to forward to friends) were significantly related with two action variables (read and act immediately, keep for later browsing) at the significance level of 0.01 (Table 5). The first action variable represents positive actions, and the second represents negative or passive actions. Based on this result, Hypothesis 2 (Consumers' behaviors on mobile advertising are not significantly related to their intentions to receive the advertisement.) also was rejected. Figures 2 and 3 graphically depict the results of the two canonical correlation analyses.

Validation of the Theoretical Model

The results of data analysis in the first stage provided evidence for rejection of Hypotheses 1 and 2 and thereby argued for the existence of correlation relationships between consumer motives for and consumer attitudes toward receiving mobile advertising, and between consumer behaviors and their intentions. These results were taken one step further in the second stage of data analysis. Whereas the emphasis of the first stage was on the correlation relationship, the concern of the analysis in the second stage was with the causal relationships among the involved constructs. A software program designed to evaluate causal effects between research variables, AMOS, was used to validate the theoretical model formulated

Table 5. Canonical correlation analysis between intentions and behaviors

Intentions for Receiving Mobile Advertising	**Canonical Factor**	**Actions on Mobile Advertising**	**Canonical Factor**
	χ^1		λ_1
Looking forward to receiving	0.89*	Read and act immediately (positive action)	0.99*
Basing purchasing decisions on the advertising	0.90*		
Likely to forward to friends	0.81*	Keep for later browsing (passive action)	0.16
Extracted Variance %	0.752	0.512	
Redundancy	0.324	0.220	
ρ^2	0.431		
ρ	0.656		

as a result of the analyses conducted in the first stage.

Four hypotheses (H_3 through H_6) were tested in the second stage. In addition to the three hypotheses (H_3, H_4, and H_6) that reflected the relationships suggested by the TRA, H_5 (Positive attitudes lead to positive actions.) was included as an alternative model. This alternative model hypothesized that attitudes bypassed intentions to influence actions directly. The rejection of H_5 would strengthen the validity of and increase our confidence in the TRA.

The results summarized in Table 6 confirmed the applicability of the TRA to mobile advertising. All hypotheses but H_5 were significant with all model fitness indexes (GFI, AGFI, NFI, CHI, RMESA, and the χ^2 value per degree of freedom) passing all evaluation criteria. These results indicated positive causal effects of strong consumer motives on positive attitudes toward receiving mobile advertising, positive attitudes on strong intentions, and strong intentions on positive actions. The rejection of H_5 indicated that attitudes did not bypass intentions to directly influence behaviors, a finding consistent with the proposition of the TRA that attitudes influence behaviors through intentions (Figure 4). The findings from this analysis also signal to the mobile commerce marketers that cellular phone users often have distinct characteristics, and their purchasing behaviors often are influenced by a multitude of factors (Clark, 2001).

CONCLUSION

Inspired by Fishbein and Ajzen's (1975) Theory of Reasoned Action, this study was conducted to examine the influencing factors of consumer behaviors in the context of mobile advertising. The first stage of the study investigated the correlation between consumer motives for receiving and consumer attitudes toward mobile advertising. It also investigated the correlation between consumer intentions for receiving advertisements on their cellular phones and their actions taken

Figure 2. Canonical correlation path diagram between motives and attitudes

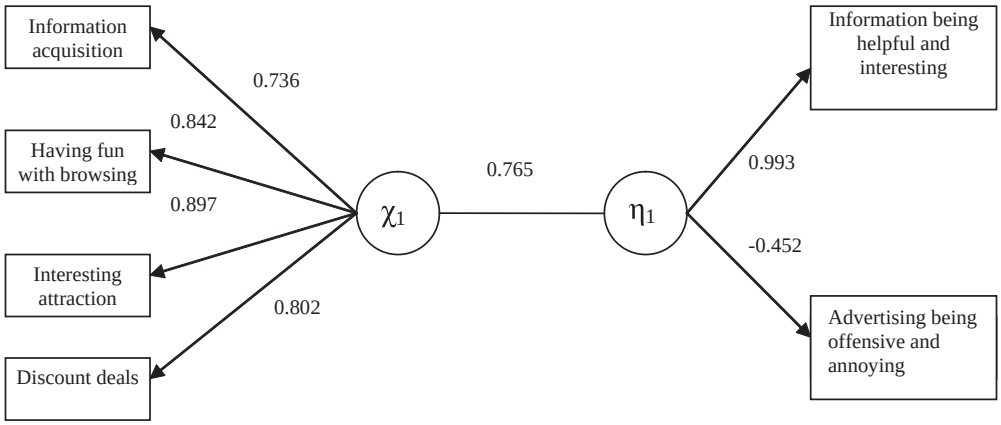
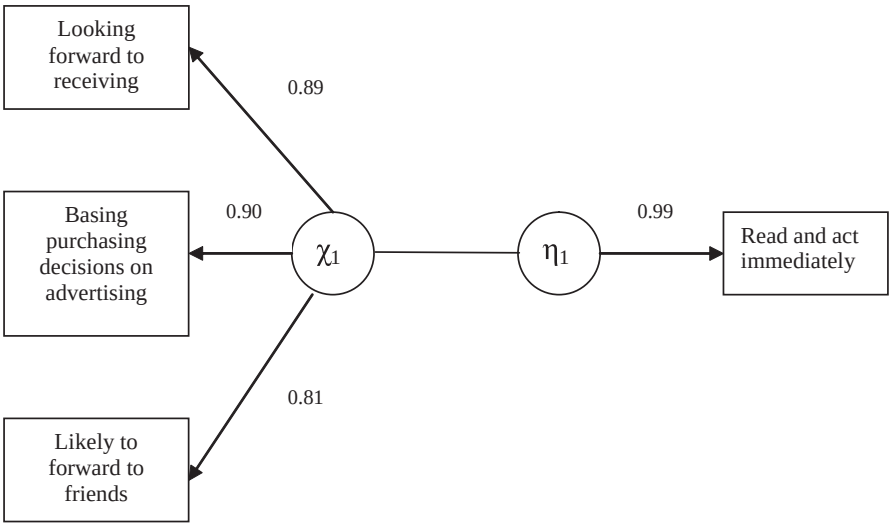


Figure 3. Canonical correlation path diagram between intentions and behaviors



on the mobile advertising that they received. A negative sentiment was revealed toward mobile advertising: all statements expressed in positive and active terms received below average ratings, and the statements expressed in negative and passive terms received above average ratings. The findings suggest that mobile advertising is still an underutilized aspect of mobile commerce, implying that current practices of mobile advertising are not effective and require a careful reevaluation in order to identify more innovative measures. The second stage of the research, however, uncovered some encouraging messages for marketers (Figure 4).

Using Fishbein and Ajzen's (1975) Theory of Reasoned Actions model as the theoretical foundation, it was found that positive actions on the received advertisements were significantly influenced by strong intentions; strong intentions were significantly influenced by favorable atti-

Table 6. Results of model validation

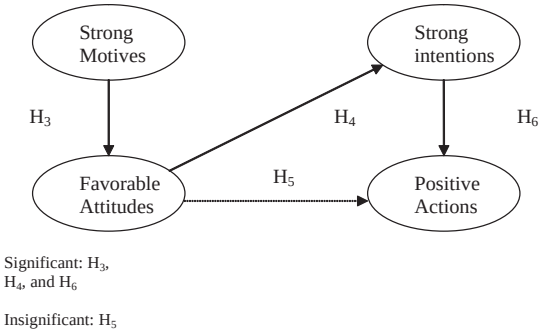
Path	Hypothesis	Expected Sign	Revised Theoretical Model	
			Standardized Structural Coefficient	**t-value**
Strong Motives -> Positive Attitudes	H_3	+	0.946	15.605*
Positive Attitudes -> Strong Intentions	H_4	+	0.901	14.952*
Positive Attitudes -> Positive Actions	H_5	+	0.040	0.209
Strong Intentions -> Positive Actions	H_6	+	0.846	4.233*
Fitness Indexes	Criteria		Actual Fitness Value	Evaluation of Fitness
GFI	> 0.9		0.953	Good
AGFI	> 0.9		0.916	Good
NFI	> 0.9		0.954	Good
CFI	>0.95		0.970	Good
RMSEA	< 0.08		0.070	Good
χ^2 / df	< 3		2.748	Good

**: $p < 0.01$*

tudes; and favorable attitudes were significantly influenced by strong motives for receiving mobile advertisements. Mobile commerce marketers should be inspired by these findings. With the continued advancement of wireless communication technologies and the associated nontechnical measures such as proper trust management, a substantial improvement in consumer perception and acceptance of mobile advertising in the future are not out of reach. However, mobile advertising professionals must incorporate unique features of mobile communication devices and the characteristics of the users in the advertising campaign design. It is vital to find value-adding approaches in order to transform negative beliefs and weak motives into positive expectations.

The research findings reported in this article should not be overly generalized due to several limitations. The first limitation is caused by somewhat low reliability measures of some constructs. Using more questions for the construct may alleviate this problem, but it would do so at the cost of increasing the length of the survey instrument. Increasing questionnaire length may result in a reduced response rate. The second limitation is associated with the exclusive use of Taiwanese data in the analyses. The generalizability ability of the research findings may be limited due to the use of single culture participants. The Web-enabled mobile commerce is emerging as a global phenomenon, with new participants joining from many parts of the world. More theory-based research conducted in different cultural contexts allows for cross-cultural comparison and contribute to theory development in the context of mobile commerce.

Figure 4. The theoretical model



When interpreted from a customer-oriented perspective, the findings of this study suggest that mobile advertising companies should exercise due sensitivity with customer perception when designing mobile advertising strategies and that new mobile advertising should be rolled out as an evolutionary rather than a revolutionary process (Malhotra & Segars, 2005). This observation is echoed by a similar warning voiced by Balasabramanian et al. (2002), who said, "Numerous failed business ventures attest to the fact that managers emphasize technologies over consumers at their own peril. This is particularly relevant in the context of m-commerce, where the technologies employed are usually of a cutting-edge nature and involve substantial investment, but the benefits to consumers are often nebulous" (p. 359).

REFERENCES

Alwitt, L.F., & Prabhaker, P.R. (1994). Identifying who dislikes television advertising: Not by demographics alone. *Journal of Advertising Research, 34*(6), 17–29.

Balasabramanian, S., Peterson, R.A., & Jarvenpaa, S.L. (2002). Exploring the implications of m-commerce for markets and marketing. *Academy of Marketing Science Journal, 30*(4), 348–361.

Bhatnagar, P. (2005). *Online sales to lose steam in '05.* Retrieved January 19, 2005, from http://money.cnn.com/2005/01/19/news/economy/online_sales/index.htm

Clark, I., III. (2001). Emerging value propositions for m-commerce. *Journal of Business Strategy, 18*(2), 133–148.

Cuieford, J.P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw Hill.

Donthu, N., Cherian, J., & Bhargava, M. (1993). Factors influencing recall of outdoor advertising. *Journal of Advertising Research, 33*(3), 64–72.

Elliott, M.T., & Speck, P.S. (1998). Consumer perceptions of advertising clutter and its impact across various media. *Journal of Advertising Research, 38*(1), 29–41.

Fishbein, M., & Azjen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equations models with unobservable variables and measurement error. *Journal of Marketing Research, 18,* 39–50.

Frolick, M.N., & Chen, L.D. (2004). Assessing m-commerce opportunities. *Information Systems Management, 21*(2), 53–61.

Haque, A. (2004). Mobile commerce: Customer perception and its prospect on business operation in Malaysia. *Journal of American Academy of Business, 4*(1/2), 257–262.

Harrison, D.A., Mykytyn, P., Jr., & Riemenschneider, C.K. (1997). Executive decisions about adoption of information technology in small business: Theory and empirical Tests. *Information Systems Research, 8*, 171–195.

Jih, W.J.K., & Lee, S.F. (2004). An exploratory analysis of relationships between cellular phone users shopping motivators and life styles indicators. *Journal of Computer Information Systems, 44*(2), 65–74.

Joines, J., Scherer, C.W., & Scheufele, D.A. (2003). Exploring motivations for consumer Web use and their implications for e-commerce. *The Journal of Consumer Marketing, 2*0(2/3), 90–103.

Litan, R., & Rivlin, A.M. (2001). Project the economic impact of the Internet. *The American Economic Review, 91*(2). 313–317.

Mackenzie, S.B., Lutz, R.J., & Belch, G.E. (1986). The role of attitude toward the ad as a mediator of advertising effectiveness: A test of competing explanations. *Journal of Marketing Research, 23*(2), 130–143.

Malhotra, A., & Segars, A.H. (2005). Investigating wireless Web adoption patterns in the U.S. *Communications of the ACM, 48*(10), 105–110.

Mathieson, K. (1991). Predicting user intentions: Comparing the technology acceptance model with the theory of planned behavior. *Information Systems Research, 2*(3), 173–191.

Mitchell, A.A., & Olson, J.C. (1981). Are product attribute beliefs the only mediator of advertising effects on brand attitude? *Journal of Marketing Research, 18*(3), 318–332.

Mittal, B. (1994). Public assessment of TV advertising: Faint praise and harsh criticism. *Journal of Advertising Research, 34*(1), 35–53.

Mykytyn, P.P., Jr., Mykytyn, K., & Harrison, D.A. (2005). Integrating intellectual property concepts into MIS education: An empirical assessment. *Decision Science Journal of Innovative Education, 3*(1), 1–28.

Napier, H.A., Judd, P.J., Rivers, O.N., & Adams, A. (2003). *E-business technologies*. Boston: Course Technology.

Nohria, N., & Leestma, M. (2001). **A moving target: The mobile-commerce customer.** *Sloan Management Review, 42*(3), 104–125.

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Porter, M.E. (2001). Strategy and the Internet. *Harvard Business Review, 79*(3), 63–78.

Scheleur, S., King, C., & Shimberg, M. (2005, August). Quarterly retail e-commerce sales 2nd quarterly 2005. *U.S. Census Bureau News*, CB05-114. Retrieved August 19, 2005, from http://www.census.gov/mrts/www/data/pdf/05Q2.pdf

Shimp, T.A. (1981). Attitude toward the ad as a mediator of consumer brand choice. *Journal of Advertising, 10*(2), 9–15.

Siau, K., & Shen, Z. (2002). Mobile commerce applications in supply chain management. *Journal of Internet Commerce, 1*(3), 3–14.

Tsang, M.M., Ho, S.C., & Liang, T.P. (2004). Consumer attitudes toward mobile advertising: An empirical study. *International Journal of Electronic Commerce, 8*(3), 65–78.

Turban, E., King, D., Lee, J, & Viehland, D. (2004).

Electronic commerce: A managerial perspective (3rd ed). Upper Saddle River, NJ: Prentice-Hall.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7*(3), 185–197.

Wen, H.J., & Gyires, T. (2002). The impact of wireless application protocol (WAP) on m-commerce. *Journal of Internet Commerce, 1*(3), 15–28.

White, C.M. (2004). *Data communications & computer networks: A business user's approach*. Boston: Course Technology.

Zanot, E. (1984). Public attitude toward advertising. *International Journal of Advertising, 3,* 3–15.

Zhang, J.J., Yuan, Y., & Archer, A. (2002). Driving forces for m-commerce success. *Journal of Internet Commerce, 1*(3), 81–106.

This work was previously published in Information Resources Management Journal, Vol. 19, Issue 4, edited by M. Khosrow-Pour, pp. 39-55, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 5.8
Effects of Consumer-Perceived Convenience on Shopping Intention in Mobile Commerce: An Empirical Study

Wen-Jang (Kenny) Jih
Middle Tennessee State University, USA

ABSTRACT

Wireless communication and Internet services are converging to provide an unprecedented level of convenience for online shopping. Although the concept of consumer-perceived convenience has been extensively discussed in marketing and consumer behavior literature, there still is a lack of empirical validation in the context of mobile commerce. This study was conducted to examine the effect of convenience on customers' intention of shopping via their mobile communication devices. The primary data collected from college students in Taiwan were analyzed to examine the relationship between perceived convenience and shopping intention. The result shows a significant relationship between the two variables, and a positive effect of convenience perception on shopping intention. The findings have practical implications for mobile commerce strategists by providing more understanding of the mobile commerce success factors from a consumer behavior point of view.

INTRODUCTION

The convergence of the Internet and wireless communications has led to the development of an emerging market for mobile e-commerce, or m-commerce. As the business impact of e-commerce can be witnessed in almost every facet of the business arena, the advancement of wireless Internet access capabilities is adding to the flexibility of the online shopping process (Haskin, 1999). Specifically, Web-enabled wireless devices allow users to search, communicate, and purchase products and services from anywhere at any time. These convenient features are welcomed by today's busy customers and are helping to make

e-commerce grow even further.

Wireless communications technology has received much attention in both voice and data communication markets. A marketing research firm called iSuppli predicts that the global wireless market will increase from the $520 million of 2004 to $430 million by 2010 (Focus on Internet News and Data, 2006). Telecom trends estimates that almost 100 million people are m-commerce users today, and their numbers are expected to double in the near future (Fitchard, 2004). Lewis (1999) predicts that, as the average cost of wireless usage will drop substantially in the next several years, wireless Internet devices will outnumber wired devices. Wireless business forecast (2005) predicts that U.S. wireless customers will expand from the current 175 million to 200 million by 2008. Portio Research, a British research firm, estimates that a half of the world population will become mobile phone users by the year 2009 (Wu, 2006). China currently adds 3 million to 4 million cellular phone users each month. By the end of January 2006, its cellular phone population has reached 400 million, the largest in the world (Focus on Internet News and Data, 2006). Although these specific forecast numbers don't match, as is typical with other types of forecasts, it appears clear that, as wireless technologies and standards for security, bandwidth, and interoperability continue to advance, the impact of online shopping via wireless communication devices is bound to become a crucial issue for information system professionals as they strive to support their organizations' marketing and strategic initiatives.

Most of the existing literature on mobile commerce developments are anecdotal reports that center on either technological advancement (e.g., Olla, Atkinson, & Gandceha, 2003) or business activities of technological service providers. Systematic empirical investigation into major aspects of m-commerce development to support theory building in this field is relatively limited. This problem was pointed out by Clarke (2001), saying that "Despite tremendous interest in the melioration of m-commerce, there is little, if any, research that examines how to develop a comprehensive consumer-oriented mobile e-commerce strategy (p. 134)." In attempting to furnish a theoretical basis for academic research, Clarke (2001) proposed four value propositions for m-commerce applications: ubiquity, convenience, localization, and personalization. Zhang, Yuan, and Archer (2002) also suggested three driving forces to account for m-commerce success: technology innovation, evolution of a new value chain, and active customer demand. Two related themes stand out in these researches regarding m-commerce: the importance of integrated business strategies that truly accommodate the unique features of mobile communication devices, mobile phone users and the significance of consumer-perceived convenience provided by the mobile devices.

The concept of service or product convenience as a research construct has primarily been discussed in the marketing and consumer behavior literature (for example, Berry, Seiders, & Grewal, 2002; Brown, 1990; Gross & Sheth, 1989; Ng-Kruelle, Swatman, Rebme, & Hampe, 2002; Seiders, Berry, & Gresham, 2000). Although mostly conceptual and speculative in nature, the literature on the significance of convenience consistently argues for the positive impact of product and service convenience on customers' shopping and the satisfaction resulted from the use experience (Brown, 1989; Berry et al., 2002; Litan & Rivlin, 2001). Little research has been reported, however, about the effort that empirically investigates the impact of service or product convenience on various aspects of customer behaviors, such as shopping intention. The need for research regarding the significance of convenience that is conducted in the context of m-commerce is especially important, given the unique features and appeals of wireless communication products and services. The primary purpose of this study is to help bridge this gap by investigating the perception of cellular phone users concerning

the effects of m-commerce service convenience perception on the intention of shopping via the Internet-enabled cellular phone. The relationship between perception of specific cellular phone service features and convenience perception is also examined.

The remainder of the article first briefly reviews the existing academic literature on consumer-perceived convenience as well as some distinctive characteristics of m-commerce. The literature review serves as the basis for the formulation of research hypotheses. The paper then describes the research method of the study, including questionnaire design, collection of the research data, and the statistical techniques employed to analyze the data. The results of data analysis and our interpretations as related to hypotheses testing are presented in the subsequent section. The final section summarizes the research findings, highlights the implications for practice and research, and also proposes some promising directions for future research.

LITERATURE REVIEW AND HYPOTHESES

Distinctive Characteristics of M-Commerce

Briefly, m-commerce can be defined as the use of wireless communications networking technology as the primary interaction vehicle between buyers and sellers of products or services. Currently, the Web-enabled cellular phone is the most popular device used by the customers of m-commerce. This definition is based on a number of slightly different definitions found in the literature. For example, Siau & Shen (2002) defined m-commerce transactions as those conducted via mobile devices using wireless telecommunication networks and other wired e-commerce technologies. In O'Dea's (2000) study, m-commerce was defined as an extension of e-commerce beyond the static terminal of the PC/TV to anytime, anyplace, anywhere on mobile and other wireless devices. As the wireless communication technology continues to advance along many directions (e.g., bandwidth, security, user interface, pricing strategy, etc.) (White, 2002), the substantial growth potential of m-commerce in the near future has been predicted by both practitioners (Fitchard, 2004) and academicians (Zhang et al., 2002).

Innovative business strategies must be developed to leverage the unique features of wireless communications in order to offer unique and appealing customer value. Contrasted with the traditional, wired telecommunication networks, a wireless communication infrastructure is relatively less expensive to construct in terms of capital requirement and time frame. This cost advantage is applicable to wide-area, metropolitan-area, and local-area network installations (White, 2004). Wireless communication devices are also more tightly tied to the service users than desktop personal computers or fixed line-based telephones. This personalization capability has allowed m-commerce companies to more closely connect customer with their major business processes, such as new product development, in the attempt to enhance customer satisfaction and loyalty (Napier, Judd, Rivers, & Adams, 2003; Ng-Kruelle et al., 2002). In addition, when equipped with wireless cards and Web browsing capability, user wireless devices such as laptop computers or even cellular phones, can be used to access internal as well external information resources with little concern of wiring for network connection.

Researchers have identified major advantages of m-commerce that can be derived from these unique features of wireless communications. For example, Wen and Gyires (2002) indicated the key ingredients of m-commerce to be portability, connectivity, usability, and ubiquity. Ng-Kruelle et al. (2002) listed six advantages of m-commerce: ubiquity, reachability, security, convenience, localization, and personalization. Clarke (2001) pointed out four value propositions of m-commerce that

set m-commerce apart from conventional e-commerce: ubiquity, localization, personalization, and convenience. Frolick and Chen (2004) indicated that m-commerce contributes to overall business operations through real time interactions with customers and immediate dissemination of decision support information to employees. Thayer (2002) emphasized the advantage of expanded contact points with customers. In explicating major differences between m-commerce and e-commerce, Zhang et al. (2002) contended that "M-commerce is not simply a new distribution channel, a mobile Internet or a substitute for PCs. Rather, it is a new aspect of consumerism and a much more powerful way to communicate with customers (p. 83)." Rather than treating m-commerce merely as an extension of e-commerce, a new way of thinking has been called for in order to unleash the value of m-commerce associated with the role of mobility (Clarke, 2001; Nohria & Leestma, 2001). From a strategic perspective, the potential of m-commerce can be realized only through the development of mobile-specific business strategy (Clarke, 2001). Viewed from customers' point of view, the technical capability of mobility essentially forms the basis of convenience.

The Concept of Customer Convenience

Convenience is an important value proposition to customers in the e-commerce business. Merriam-Webster's online dictionary defines convenience as, "something (as an appliance, device, or service) conducive to comfort or ease; fitness or suitability for performing an action or fulfilling a requirement." While the first definition links to a psychological dimension and the second refers to problem solving, both definitions suggest the subjective and perceptive nature of the concept.

In business literature, convenience is typically viewed as a multidimensional construct. It first appeared in the business literature as Copeland (1923) defined convenience goods as a class of consumer products that were intensively distributed and required minimal time and physical and mental effort to purchase. Some later definitions of convenience also focused on resources such as time and effort required of the consumer in shopping for a product (Brown, 1990). Other researchers, however, expanded the concept of convenience to incorporate non-shopping activities. For example, Yale and VenKatseh (1986) identified six aspects of convenience: time utilization, accessibility, portability, appropriateness, handiness, and avoidance of unpleasantness. However, this framework was criticized for the lack of theoretical underpinning and means of measurement (Berry et al., 2002; Brown, 1989; Gehrt & Yale, 1993). In the context of Internet-enabled commerce, the five-dimension framework of convenience proposed by Brown (1989) appears to be both inclusive and measurable: time dimension, place dimension, acquisition dimension, use dimension, and execution dimension. Some researchers even contend that convenience is the most critical benefit of the Internet. Economists Litan et al. (2001), for example, suggested that "Much of the benefit from the Internet is likely to show up in improved consumer convenience and expanded choices, rather than in higher productivity and lower prices (p. 317)," as a conclusion of a team research conducted to examine the economic impact of the Internet.

The notion of convenience perception also receives much attention in the field of information systems. Studies in the technology acceptance model (TAM) (Davis, 1989; Gefen & Straub, 2000), for example, examine the impact of perceived ease-of-use on intended adoption of information technology. In a simulation study conducted to validate a theoretical explanation of the effects of perceived ease-of-use on IT adoption, Gefen et al. (2000) distinguish between the extrinsic vs. intrinsic aspects of IT characteristics with regards to IT adoption. Citing purchasing a book through a Web site as an example, they clarify that the purchasing itself is a task extrinsic

to the IT because the IT serves only as an interface of an integrated system. The entire system typically consists of many other components such as shipping and payments handling system modules. Conversely, using the same Web site to inquire about a book represents an intrinsic IT task because the Web site provides a complete application associated with the actual service. Their data support the proposition that perceived ease of use only affects intended use of tasks that are intrinsic to the IT. In light of unique characteristics of mobile commerce such as personalization and localization, however, our study adopts Brown's framework described above for the operational definition of convenience perception. The question items devised to measure the extrinsic and intrinsic characteristics of perceived ease of use in Gefen et al. (2000) appear to indicate that Brown's framework seems to define convenience perception in a broader sense than perceived ease of use in TAM.

In summary, the literature in consumer behavior indicates that convenience is a multi-dimensional and context-dependent perception. An empirical investigation of its impact on customers' shopping behavior must treat convenience as a composite variable and be conducted in a specific context, such as mobile commerce in the case of this research. In addition, since the perception is subjective in nature, it may be measured differently between the mobile commerce companies and their customers.

Research Hypotheses

In order to investigate of role of customer perception of convenience in the context of m-commerce, three research hypotheses are established for statistical testing. First of all, due to its subjective nature, it is assumed that the value of convenience may be affected by individual differences. In Bergada's (1990) study, consumers' perception of convenience was found to vary with their demographical characteristics as well as their shopping patterns. Specific individual characteristics that are capable of affecting convenience perception include: time inclination (Gagliano & Hathcote, 1994; Luqmani, Yavas, & Quraeshi, 1994), tolerance of time pressure (Landy, Rastegary, Thayer, & Colvin, 1991), empathy (Aaker & Williams, 1998), and experience (Brucks, 1985; Kumar, Kalwani, & Dada, 1997). The first hypothesis is set up to explore the influence of demographical characteristics on the perceived customer convenience.

H_1: The customer perception of convenience is significantly related to m-commerce customers' demographical characteristics.

Secondly, the convenience features of m-commerce are provided by specific product/service items offered by the wireless communications service providers and the m-commerce Web sites. Given the subjective nature of customer convenience perception, although service providers usually strive to incorporate customer-friendly features in their offerings, all specific product/service items may not be equally associated with customers' convenience perception. The second research hypothesis is formulated to test the relationship between the product/service features and the customers' convenience perception:

H_2: M-commerce customers' convenience perceptions are significantly correlated with product/service features.

Finally, the concept of convenience has strategic and tactical implications in marketing. Brown (1989) proposed a five-dimension convenience model (time dimension, place dimension, acquisition dimension, use dimension, and execution dimension) and demonstrated its value for marketing decision analysis. Berry et al. (2002) also developed a decision model that centered on service convenience. This model identified five classes of service convenience: decision convenience,

access convenience, transaction convenience, benefit convenience, and post-benefit convenience. Viewed from the life cycle perspective, each of these conveniences may contribute to customers' shopping or re-shopping decision. In an empirical study, Anderson and Srinivasan (2003) found that consumer's convenience motivation was a major factor affecting the impact of e-satisfaction on e-loyalty. As indicated by Anderson (1972), when properly integrated into marketing decisions, the concept of convenience may become a powerful enabling tool. This leads to the following research hypothesis:

H_3: M-commerce customers' shopping intention is significantly affected by their convenience perception.

RESEARCH METHODOLOGY

Collection and Analysis of Research Data

In order to investigate the effect of convenience on m-commerce customers' shopping intention, a questionnaire survey was conducted using young cellular phone users in Taiwan as the convenience sample. The questionnaire contained three sets of questions that were devised to collect primary data from the research sample about the three research constructs: the perception of Internet services offered to Internet-enabled cellular phone users, the perception of product/service convenience associated with m-commerce, and the demographical distribution of the respondents. A total of 23 cellular phone services were compiled from the providers' company Web pages, covering most of the popular services offered by wireless communication service providers. The development of the convenience feature perception questions was based on an expanded version of Brown's (1989) model. In addition to the original five dimensions (time, place, acquisition, use, and execution), three questions representing Web-based service are included to account for the specificity of m-commerce. Sixteen questions in total were developed to measure these five dimensions of the convenience construct. In the attempt to determine how basic demographical characteristics affected the usage of service items as well as convenience perceptions, we include three questions regarding gender, age, and usage experience in the questionnaire. Shopping intention is measured with a question that asks cellular phone users to indicate their general shopping intention via cellular phones. With the exception of demographical characteristics, the questions are not necessarily all mutually exclusive and are evaluated using the Likert scale with 1 indicating very unimportant and 5 very important.

As a pilot test, the questionnaire was administered to one hundred cellular phone users to evaluate the adequacy of the questionnaire. The reliability measures of the questionnaire evaluated by Cronbach's α values as well as the feedbacks from the questionnaire respondents were used to refine the questionnaire subsequently. According to Nunnally (1978), a data collection instrument that has a Cronbach's αhigher than 0.7 is considered to be highly reliable. The evaluation of the construct validity of the questionnaire was based on Kerlinger's (1986) suggestion: The correlation coefficients between the individual question item scores and the total score were used as the construct validity measures. As shown in Table 1, the questions in all categories have Cronbach's α values higher than 0.8, and the item-total correlation coefficients are all close to 0.7 or above. The former value indicates that the questionnaire is reliable and the latter suggests a good validity of the data collection instrument.

College students were used as the convenience sample in this study for both the pilot test and the formal survey because they represented the most active group of cellular phone users as well as m-commerce customers in Taiwan (Jih & Lee, 2004; Lin, Chen, & Lin, 2001). Another reason is that,

Table 1. Measuring convenience: Item-total correlation and dimension reliabilities

Convenience Dimensions	Question Number	Item-Total Correlation	Cronbach's α Value
Use Dimension	1	0.56	0.8058
	2	0.66	
	3	0.62	
	4	0.66	
	5	0.48	
Time Dimension	6	0.82	0.9023
	7	0.83	
	8	0.77	
Place Dimension	9	0.70	0.8218
	10	0.70	
Shopping (Execution) Dimension	11	0.69	0.8347
	12	0.74	
	13	0.68	
Service Dimension	14	0.74	0.8805
	15	0.80	
	16	0.76	

although the widespread use of the cellular phone for shopping is still at the initial stage of innovation diffusion, the experience of the current college students will grow in synchronization with the maturing process of the technology itself. College students' perception of m-commerce will serve as a good referencing and thereby facilitates sound business decision-making on the part of m-commerce companies. Students of a variety of majors in six colleges participated in the study. A total of 400 copies of the questionnaire were distributed. Of which, 370 were deemed effective responses. The high response rate was achieved because the questionnaires were distributed in class and the students were encouraged to respond on an anonymous and voluntary basis. The effective respondents consist of 43.2% of males and 56.8% of females; 48.1% with ages 17-20, 41.6% with ages 21-25, and 10.3% with ages outside these typical college student age ranges.

The research hypotheses were tested using t-test, analysis of variance (ANOVA) canonical correlation analysis, and regression analysis. The difference of convenience perception between different demographical groups (H_1) was tested using t-test and ANOVA. Canonical correlation analysis was used to test the correlation relationship between specific service items and convenience factors (H_2). Regression analysis was employed to determine the impact of convenience factors on shopping intention (H_3).

RESULTS OF DATA ANALYSIS

Service and Convenience Preferences

An issue of practical concern regarding the Internet services via wireless communications

and the perception of convenience features in m-commerce is how they are ranked by customer preference. Mean scores are used to provide the rankings. Among the 23 specific service items offered by most service providers, the five most welcome ones are emergency service, short message, e-mail, medical information, and transportation acquisition service. Among the 16 convenience perception items reviewed by the cellular phone users, the five most desirable ones are portability of user device, lightweight and compactness of user device, convenience of information search, transaction or information search not limited by location, and service on demand. These responses are not surprising. Other than psychological and other non-technical reasons, these top-ranked desires reflect, to a certain degree, the problem of crowded traffic on the densely populated island as well as the phenomenon that most people, especially the young generation, appear to be extremely busy in this fast-paced world. The services provided by the wireless communications technology and m-commerce service providers are increasingly becoming an essential part of many people's lives (www.find.org.tw/news/).

Factor Analysis of Internet Services and Convenience Perception

Before the research hypotheses were tested, factor analysis was performed to compress the number of Internet service variables from 23 to 4 (Table 2) and the number of convenience perception variables from 16 to 2 (Table 3). The four latent variables that represent the observed Internet service variables and the two representing observed convenience perception were then used to test the research hypotheses. In addition, the Bartlett's sphericity test was computed to validate significant correlation between the observed variables, and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was obtained to further establish the adequate use of factor analysis on the data. A significantly high χ^2 value indicates significant correlation between the observed variables and a high KMO value (>= .80) indicates high shared-variance and low uniqueness in variance. Both evaluation criteria measures signify that the data are appropriate for factor analysis. Both sets of variables were analyzed using principal component analysis to extract the factors and varimax rotation to achieve a simplified factor structure.

As summarized in Table 2, the result of factor analysis produced four factors for Internet services on cellular phone: life-enhancement services, value-added services, entertainment features, and basic services. The accumulated variance of these four factors is 58.358% with the overall reliability 0.9162. The reliability measures of the four factors are 0.8946, 0.8357, 0.8019, and 0.5338, respectively. Nunnally (1978) suggests a Cronbach's α Value 0.7 as the cutoff point for acceptable reliability. A less strict criterion for reliability evaluation is suggested by Cuieford (1965). This criterion contends that a Cronbach's α Value 0.7 or higher indicates highly reliable, that between 0.35 and 0.7 indicates acceptable, and that below 0.35 unacceptable. Due to the exploratory nature of the questions, Cuieford's criterion was adopted to accept the fourth factor, basic service, in our analysis. The χ^2 value from the Bartlett's test is 4075.727 at the p value < 0.01 and the KMO measure is 0.906. Both measures suggest that the data is appropriate for factor analysis.

The result of factor analysis of convenience perception is shown is Table 3. The two factors produced are labeled transaction convenience and operational convenience. The accumulated variance extracted by these two factors is 60.935%, with the overall reliability measure 0.9401. The reliability measures for the individual factors are 0.9069 and 0.8789, respectively. These are high reliability measures even by the more strict,

Table 2. Factor analysis of Internet services for cellular phone users

Factors	Variables	Factor Loading				Cronbach's α Values
		1	2	3	4	
Life-Enhancement Services	Ticket Shopping	0.774	0.136	-0.007	0.108	0.8946
	Medical Information	0.741	0.138	0.138	0.063	
	Service Reservation	0.677	0.106	0.183	0.166	
	E-Learning Service Use	0.647	0.434	-0.003	-0.051	
	Transportation Service Acquisition	0.639	0.032	0.080	0.218	
	Employment Information	0.634	0.147	0.309	-0.130	
	Online Banking	0.621	0.161	0.176	-0.095	
	Discount Coupon	0.618	0.158	0.186	0.137	
	Transportation Information	0.614	0.243	0.250	0.056	
	Emergency Service Use	0.613	-0.145	-0.004	0.303	
	News	0.567	0.352	0.127	-0.168	
	E-mail	0.480	0.175	0.368	0.276	
Value-Added Services	Horoscope	0.034	0.834	0.299	0.107	0.8357
	Psychological Testing	0.057	0.830	0.283	0.102	
	Food Menu Information	0.443	0.627	0.145	-0.006	
	Online Shopping	0.493	0.598	0.105	0.103	
	Lottery Shopping	0.328	0.492	0.265	-0.259	
EntertainmentFeatures	Game	0.058	0.067	0.852	0.115	0.8019
	Entertainment Information	0.239	0.200	0.752	0.010	
	Fellowship and Social Interaction	0.122	0.332	0.646	-0.111	
	E-book	0.315	0.336	0.642	-0.009	
Basic Services	Short Message	0.196	-0.085	-0.010	0.799	0.5338
	Ring Pattern Download	0.101	0.413	0.177	0.636	
Eigenvalue		8.337	2.410	1.433	1.242	
Explained Variance (%)		36.246	10.480	6.230	5.401	Overall Reliability 0.9162
Accumulated Explained Variance (%)		58.358				

Table 3. Factor analysis of convenience perception of cellular phone users

Factors	Variables	Factor Loadings		Cronbach's α Values
		1	2	
Transaction Convenience	Immediate payment for shopping	0.821	0.110	
	Individual password for shopping payment	0.802	0.141	0.9069
	Multiple means of payment for online shopping	0.722	0.182	
	localization service	0.649	0.442	
	Transaction inquiry on holidays	0.611	0.521	
	24-hour-based online Inquiry	0.596	0.540	
	Any-time Internet connection	0.588	0.316	
	Service not limited by location	0.568	0.528	
	Convenience of information search	0.547	0.541	
Operational Convenience	Portability of user device	0.009	0.830	0.8789
	Light weight and compactness of user device	0.166	0.792	
	Ease of operation	0.400	0.647	
	Reduction of information search time	0.567	0.593	
	Transaction or information search not limited by location	0.563	0.580	
	Multimedia-based communications	0.437	0.544	
	Service on demand	0.530	0.532	
Eigenvalue		8.518	1.231	Overall Reliability: 0.9401
Explained Variance		53.240%	7.695%	
Accumulated Explained Variance		60.935%		

Nunnally's (1978) standard. The χ^2 value from the Bartlett's test is 3850.51 at $p < 0.01$. The KMO coefficient is 0.94. Both measures indicate that the data is also appropriate for factory analysis.

Results of Hypotheses Testing

In order to test the hypothesis H_1 (the perception of convenience is significantly related to m-commerce customers' demographical characteristics), a t-test was conducted with each of the two

convenience perception factors as the dependent variables and gender and age as the independent variables. T-test was also conducted with wireless Internet service factors as the dependent variable and gender and age as the independent variables to determine the effect of gender and age on the evaluation of Internet services offered to cellular phone users. The hypothesis is accepted according to the results of the analysis:

1. Females have significantly higher perception of both transaction convenience and operational convenience than males at $p < 0.05$.
2. The perception of older users of both transaction convenience and operational convenience are significantly higher than their younger counterparts at $p < 0.01$.

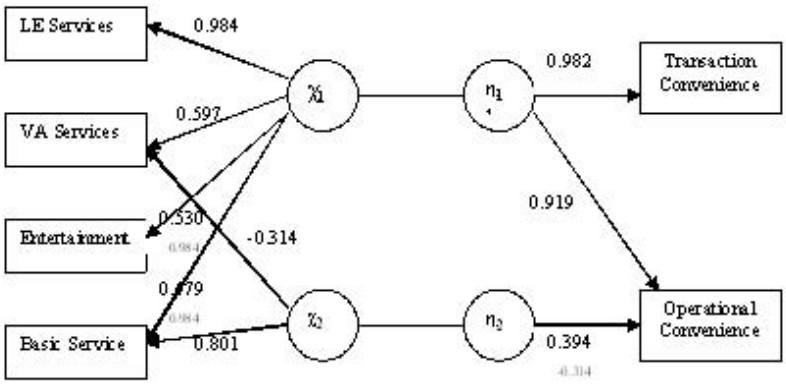
With regard to the effect of gender and age on the evaluation of wireless Internet services, it is found that while females have significantly higher evaluation of life-enhancement services at $p < 0.05$, males' evaluation of entertainment services are significantly higher than females at $p < 0.01$. Age is also found to be a significant factor regarding the evaluation of wireless Internet services: Older cellular phone users' evaluations are higher than their younger counterparts on life-enhancement services and lower on entertainment services, both at $p < 0.05$.

The second hypothesis, H_2: M-commerce customers' convenience perceptions are significantly correlated with product/service features, was tested using canonical correlation analysis. Two sets of canonical variates were produced as the result (Table 4 and Figure 1). The first set of canonical variates significantly relates all four service categories (life-enhancement services, value-added services, entertainment features, and basic features) with both types of convenience perception at $p < 0.01$ (canonical correlation coefficient $\rho = 0.692$). The second set of canonical variates significantly relates two service categories (value-added services and basic services) only with operational convenience at $p < 0.05$ (canonical correlation coefficient $\rho = 0.227$). According to these analysis results, those cellular phone users who favor life-enhancement services, value-added services, entertainment features, and basic features tend to place more emphasis on all types of convenience offered by

Table 4. Canonical analysis of relationship between convenience perceptions and service categories

Service Categories	Canonical Variates		Convenience Perceptions	Canonical Variates	
	χ_1	χ_2		η_1	η_2
Life-Enhancement Services	0.984	-0.164	Transaction Convenience	0.982	-0.189
Value-Added Services	0.597	-0.314			
Entertainment Services	0.530	0.028	Operational Convenience	0.919	0.394
Basic Services	0.479	0.801			
Percentage of Variance Extracted	46.0	19.2	Percentage of Variance Extracted	90.4	9.6
Redundancy	0.221	0.099	Redundancy	0.433	0.005
ρ^2	0.479	0.052			
Canonical Correlation	0.692	0.227			

Figure 1. Path diagram of significant relationship between service features and convenience perceptions



m-commerce businesses. In addition, those who emphasize basic services and neglect value-added services tend to favor the aspect of convenience associated with handling and operation of the cellular phone. These results lead to the acceptance of the second hypothesis.

The impact of convenience perception on shopping intention stated in the third hypothesis, H_3: M-commerce customers' shopping intention is significantly affected by their convenience perception, was tested using regression analysis**.** The results show that the type of convenience perception labeled Transaction Convenience has a significant impact on customers' intention to shop with mobile commerce companies using their cellular phones (regression coefficient = 0.497, $p < 0.01$). The fact that operational convenience does not exhibit a significant impact on shopping intention (regression coefficient = 0.172, $p = 0.139$) suggests that it may take more than just commonly available features to attract customers' attention in the mobile commerce business. The high F-value ($F = 53.920$) indicates that, in general, customers' shopping intention is significantly affected by their perception of convenience offered by mobile commerce businesses ($p < 0.01$). The third hypothesis, "M-commerce customers' shopping intention is significantly affected by their convenience perception." is accepted.

The results of hypotheses testing, stated in the alternative form, are summarized below,

H_1: The customer perception of convenience is significantly related to m-commerce customers' demographical characteristics. (Accepted)

H_2: M-commerce customers' convenience perceptions are significantly correlated with product/service features. (Accepted)

H_3: M-commerce customers' shopping intention is significantly affected by their convenience perception. (Accepted)

CONCLUSION

Strategic deployment of information technology requires integrating unique capabilities of technological tools with innovative customer-centered business processes. The convergence of Internet-based services and wireless communications creates technological business possibilities, which, if properly harnessed, have the potential to transform

Table 5. Regression analysis of the impact of convenience perception on shopping intention

Convenience Perception	Regression Coefficient	t-value	p-value
Constant Item	0.561	2.164	0.031*
Transaction Convenience	0.497	4.540	0.000**
Operational Convenience	0.172	1.485	0.139
R^2	0.227		
F	53.920		0.000**

**: $p < 0.05$; **: $p < 0.01$*

a company's competitive advantage. The advancement of wireless communication technology has allowed for multimedia messages and data being smoothly and securely exchanged with little regard for geographical distance or time consideration. The capability to transmit voice and data over the same network connection and the convenience provided through such feature as location-based service offer virtually unlimited possibilities for innovative businesses in designing product and service offerings.

When contrasted with traditional electronic commerce using desk-top personal computers, one of the most cited attractions of using mobile devices as a consumer shopping vehicle is convenience (Frolic et al., 2004; NG-Kruelle, et al., 2002; Seager, 2003; Siau et al., 2002). A subjective perception that typically varies between different people and across different contexts, convenience perception can significantly influence consumer behavior in various stage of the shopping process (Anderson, 1972; Brown, 1989; Gehrt et al., 1993).

In light of the unique business value of wireless communication applications and the important role of convenience perception, this study was conducted to empirically investigate the impact of convenience on customers' shopping intention in the context of m-commerce. The primary data regarding customers shopping on the Internet via cellular phones was collected using a survey questionnaire. The results of data analysis revealed a positive correlation between convenience perception and demographical data (gender and age). Females were found to value convenience more than males. Older m-commerce customers were found to value convenience more than their younger counterparts. A positive correlation relationship also exists between the convenience perception and the user evaluation of wireless Internet services. Those who appreciate the use of wireless communication services also tend to value the convenience of shopping in m-commerce. Most notably, the study showed that customers' intention of shopping on the Internet via their cellular phones was positively affected by their perception of convenience features offered by m-commerce businesses, wireless communication service provider, and vendors of user devices. In other words, convenience offering should be viewed as an importance element in an m-commerce company's business strategy.

The findings have implications for practicing functional managers as well as for information system professionals. The major implications of the research findings for practicing functional managers are twofold. The impact of product and service convenience in consumers' shopping decision-making has been well-documented in marketing and consumer behavior literature (Brown, 1989; Berry et al., 2002). This study demonstrates that the concept can be even more important in the context of m-commerce. Faced with rapid proliferation of offerings in the cyber

space, consumers are only attracted to and retained by the companies that consciously build convenience into their Web sites and the entire customer relationship management program.

The second implication presented to practitioners by the research findings is associated with the way a convenient customer interface may be designed. Through regression analysis, the study found that, although both categories of convenience (transaction convenience and operational convenience) have significant impact on customers' shopping intention, transaction convenience appears to have more influence than operational convenience. In other words, it tends to be the transaction convenience, rather than the operational convenience, features that provide differentiating value.

Information system professionals must take into consideration the importance of consumer perceptions of the mobile commerce offerings and design a website that is both technically versatile in processing capability and convenient in its user interface. Traditionally, user-friendliness of user interface primarily requires ease of operation and ease of learning. In mobile commerce, however, integrating transaction convenience with operational convenience is essential to winning customer attention in the vast cyber business market space.

Due to several research limitations mentioned next, the findings reported herein must be interpreted and applied with due caution on the part of the reader. The use of college students in Taiwan as the source of research data may restrict the external validity of the study. The difference between college students and other age groups must be accounted for. As in many other survey research projects, this study assumes that the questionnaire respondents fill out the survey instrument seriously. In addition, the convenience perception factor in different consumption cultures may play a different kind of role in m-commerce.

The results of this study shed some light on an important characteristic of business applications of wireless communications technology, convenience. An inter-disciplinary research field, m-commerce is still in its infancy in many ways and requires more systematic inquiries being conducted from different angles. First of all, this study operationalizes the concept of convenience based on Brown's (1989) definition of convenience. A different framework may be used to determine if significant difference would result from different definitional frameworks of convenience. For example, the model of service convenience proposed by Berry et al. (2002), which characterizes consumer's time and effort perceptions in terms of decision convenience, access convenience, transaction convenience, benefit convenience, and postbenefit convenience, may also be empirically validated for comparison. Secondly, as mobile communication devices are increasingly used as an avenue of advertisement, it is important to know how users perceive, through an independent research, this new mode of advertisement. Thirdly, since convenience usually interacts with other factors, such as service characteristics and individual differences, in affecting user perceptions, researches that investigate compound effects of these relevant factors would contribute to formulation of effective business strategy for m-commerce. Another interesting and important area of research involves cross cultural comparative studies. Currently, European and Asian customers are ahead of American customers in using the cellular phone as a shopping tool. The results of this study may be validated in different cultures to allow for more general conclusions to be drawn. Finally, information system researchers may examine possible impacts of technological capabilities, such as screen display or bandwidth, on user perception of convenience and shopping intentions. The best ways in which commerce contents, such as product display or promotion messages, ought to be presented on the small screen for relatively impatient consumers also deserve more research.

REFERENCES

Aaker, J. L., & Williams, P. (1998). Empathy versus pride: The influence of emotional appeals across cultures. *Journal of Consumer Research, 25,* 241-261.

Anderson, R. E., & Srinivasan, S. S. (2003). E-satisfaction and e-loyalty: A contingency framework. *Psychology & Marketing, 20*(2), 123-138.

Anderson, W. T. Jr. (1972). Convenience orientation and consumption behavior. *Journal of Retailing, 48,* 49-71.

Bergadaa, M. (1990). The role of time in the action of the consumer. *Journal of Consumer Research, 17,* 289-302.

Berry, L. L., Seiders, K., & Grewal, D. (2002). Understanding service convenience. *Journal of Marketing, 66*(3), 1-17.

Brown, L. G. (1989). The strategic and tactical implications of convenience in consumer product marketing. *Journal of Consumer Marketing, 6*, 13-19.

Brown, L. G. (1990). Convenience in services marketing. *Journal of Service Marketing, 4,* 53-59.

Brucks, M. (1985). The effect of product class knowledge on information search behavior. *Journal of Consumer Research, 12,* 1-16.

Clark, I. III. (2001). Emerging value propositions for m-commerce. *Journal of Business Strategy, 18*(2), 133-148.

Copeland, M. T. (1923). Relation of consumers' buying habits to marketing methods. *Harvard Business Review,* 282-289.

Cuieford, J. P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw Hill.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-339.

Fitchard, K. (2004). The two m's of commerce. *Telephony, 245*(8), 26.

Focus on Internet News and Data. (2006). Retrieved on 03/17/2006 from http://www.find.org.tw/find/home.aspx?page=news&id=4185

Focus on Internet News and Data. (2006). Retrieved on 03/01/2006 from http://www.find.org.tw/find/home.aspx?page=news&cal=□□□□&p=1

Frolick, M. N., & Chen, L. (2004). Assessing m-commerce opportunities. *Information Systems Management, 21*(2), 53-61.

Gagliano, K., & Hathcote, B. J. (1994). Customer expectations and perceptions of service quality in retail apparel specialty stores. *Journal of Services Marketing, 8*(1), 60-69.

Gefen, D., & Straub, D. (2000). *The relative importance of perceived ease of use in is adoption: A study of e-commerce adoption*. Retrieved from http://jais.isworld.org/articles/1-8/article.htm

Gehrt, K. C., & Yale, L. J. (1993). The dimensionality of the convenience phenomenon: A qualitative reexamination. *Journal of Business and Psychology 8*(2), 163-180.

Gross, B. L., & Sheth, J. N. (1989). Time-oriented advertising: a content analysis of United States Magazine Advertising, 1890-1988. *Journal of Marketing 53,* 76-83.

Haskin, D. (1999). *Analysts: Smart phones to lead e-commerce explosion*. All-NetDevices. Retrieved from http://www.allnetdevices.com/news/9911/991103ecomm/991101ecomm.html

Jih, W. J., & Lee, S, F. (2004). Relationship between online shoppers' motivation and life style

indicators. *Journal of Computer Information Systems, XLIV*(2), 65-73.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: McGraw-Hill.

Kumar, P., Kalwani, M. U., & Dada, M. (1997). The impact of waiting time guarantees on consumer waiting experiences. *Marketing Science, 16*(4), 295-314.

Landy, P. J., Rastegary, H., Thayer, J., & Colvin, C. (1991). Time urgency: The construct and its measurement. *Journal of Applied Psychology, 76*(5), 644-657.

Lewis, T. (1999). Ubiner: The ubiquitous Internet will be wireless. *IEEE Computer, 32*(10), 56-63.

Lin, S., Chen, Y. J., & Lin, T. T. (2001). A study of college students' usage of and satisfaction with mobile phones--The cases of Taipei University and Chiao-Tung University. The *7h Internet Conference – Taiwan*. Tanet.net.

Litan, R., & Rivlin, A. M. (2001). Project the economic impact of the Internet. *The American Economic Review, 91*(2), 313-317.

Luqmani, M., Yavas, U., & Quraeshi, Z. A. (1994). A convenience-oriented approach to country segmentation: Implications for global marketing strategies. *Journal of Consumer Marketing, 11*(4), 29-40.

Napier, H. A., Judd, P. J., Rivers, O. N., & Adams, A. (2003). *E-business technologies*. Boston: Course Technology.

Ng-Kruelle, G., Swatman, P. A., Rebme, D. S., & Hampe, J. F. (2002). The price of convenience: Privacy and mobile commerce. *Quarterly Journal of Electronic Commerce, 3*(3), 273-285.

Nohria, N., & Leestma, M. (2001). A moving target: The mobile-commerce customer. *Sloan Management Review, 42*(3), 104-115.

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

O'Dea, F. (2000). Mobile-commerce: Trend in wireless interactions. *Anderson Consulting Report.*

Olla, P., Atkinson, C., & Gandceha, R. (2003). Wireless systems development methodologies: An analysis of practice using actor network theory. *Journal of Computer Information Systems, XXXXIV*(1), 102-119.

Seager, A. (2003). M-commerce: An integrated approach. *Telecommunications International, 37*(2), 36-38.

Seiders, K., & Berry, L. L. (1998). Service fairness: What it is and why it matters. *Academy Management Executive, 12*(2), 8-21.

Siau, K., & Shen, Z. (2002). Mobile commerce applications in supply chain management. *Journal of Internet Commerce, 1*(3), 3-14.

Solomon, M. R. (1986). The missing link: Surrogate consumers in the marketing chain. *Journal of Marketing, 50*, 208-218.

Thayer, G. (2002). M-commerce: Long trek to the promised land. *Pen Computing, 9*(45), 17.

Wen, H. J., & Gyires, T. (2002). The impact of wireless application protocol (WAP) on m-commerce security. *Journal of Internet Commerce, 1*(3), 15-27.

White, C. (2004). *Data communications and computer networks* (3rd ed.). Boston: Course Technology.

Whitt, W. (1999). Improving service by informing customers about anticipated delays. *Management Science, 45*(2), 192-207.

Wireless Business Forecast. (2005). Wireless in the driver's seat. *Wireless Business Forecast, 13*(4), 24, 1.

Wu, C. H. (2006). *Portio research: Half of world population will be pan-pacific mobile phone users in ten years*. Retrieved from http://www.find.org.tw/find/home.aspx?page=news&id=4117

Yale, L., & VenKatseh, A. (1986). Toward the construct of convenience in consumer research. *Advances in Consumer Research, 13*, 403-408.

Zhang, J. J., Yuan, Y., & Archer, N. (2002). Driving forces for m-commerce. *Journal of Internet Commerce, 1*(3), 81-106.

This work was previously published in International Journal of E-Business Research, Vol. 3, Issue 4, edited by I. Lee, pp. 33-48, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 5.9
Factors Influencing Segmentation and Demographics of Mobile-Customers

Anne-Marie Ranft
University of Technology, Australia

ABSTRACT

This chapter addresses important factors for consideration when readying a mobile commerce business for global business, addressing both regional differentiation in demographics that influence classifications of customer segments, and differentiation in demographics within a region. Globally, not all customer segments have regular access to mobile commerce facilities, and even for those that do, other demographic factors can impede their potential as mobile-customers. When starting from an Anglo-centric perspective, it is vital to have awareness of global differences in culture, language, payment options, time zones, legal restrictions, infrastructures, product needs, and market growth that could either improve or inhibit mobile-customer uptake, and in the worst case, result in unexpected litigation.

INTRODUCTION

Mobile-customers should be considered as one of the most significant elements for a mobile commerce enterprise. Mobile-customers of the enterprise are those customers that use mobile devices—the most common ones being mobile phones, personal digital assistants (PDAs), and notebook PCs. Mobile commerce products can include: physical devices, applications, and accessories; access to the mobile infrastructure; and unrelated products and services marketed, bought, and sold using a mobile device as the communication tool.

Internet-based e-commerce interactions are generally categorised by the broad segments of Consumer (C), Business (B), and Government (G), and then decomposed into the relevant market segments. However, when undertaking global commerce, regional factors providing

differentiation in demographics can alter classifications of customer segments, and differences in demographics can occur within a region. A market segment that exists in Australia, the United States, or the United Kingdom may not exist in some regions. It should also be noted that market segments based on Internet e-customer demographics may not necessarily be directly applicable for mobile-customers.

Before targeting a product or service to a particular market segment and location, these issues should be considered to maximise mobile-customer uptake and prevent unexpected litigation.

FACTORS INFLUENCING GLOBAL DIFFERENCES

First, the question of regional mobile-customer segment sizes will be discussed with reference to the *digital divide*, then other differentiating factors will be listed, followed by a list of possible strategies to consider when designing global mobile commerce products and marketing.

Digital Divide—Historical Factors

The first issue to be addressed is one of whether potential mobile-customers for a segment even exist in the targeted regions.

"Visions of a global knowledge-based economy and universal electronic commerce, characterised by the 'death of distance' must be tempered by the reality that half the world's population has never made a telephone call, much less accessed the Internet" is the caveat noted by the Organisation for Economic Cooperation and Development (OECD, 1999). The OECD uses the term "digital divide" to describe the inter- and intra-country inequalities in access to information and communication technologies by both individuals and businesses due to socio-economic and geographic differences (OECD, 2001). They provided statistics that highlight the differences between OECD and non-OECD countries (see Table 1).

They further noted that the higher growth rate in telecommunication access for non-OECD countries is especially due to rises in China, but there was insignificant African growth during that period.

Within a geographic region, different demographic factors also contribute to a reduction in potential mobile-customers.

Uptake of mobile commerce in some regions is still biased towards the business and professional consumer sectors, especially mobile phone ownership in the Asian region.

It should be noted that many *developing* nations suffer from lack of suitable telecommunication infrastructure; access to a reliable electrical source for re-charging of mobile devices and permanent

Table 1. Potential access to mobile commerce and eCommerce - summary

	OECD		non-OECD	
	1990	1998/2000	1990	1998/2000
Fixed & mobile telecommunications access paths per 100 inhabitants	41.1	72.1	2.7	7.8
Internet hosts per 1,000 inhabitants	23	82	0.21	0.85

Data source: Organisation For Economic Co-Operation And Development 2001

OECD countries – there are 30 member countries, who are mainly in the European and North American regions; as well as the United Kingdom, Australia and New Zealand.

housing can be limited for lower socio-economic groups. For instance, in my experience I have found that many Indian businesses have access to high-speed Internet lines and mobile connectivity, but require their own generators to back up the state power supply.

Overall, it can be concluded that there are two major groups of potential mobile-customer segments not currently available due to this *digital divide* factor—consumer and business segments whose geographical demographics are characterised by lack of telecommunication and other infrastructure, and consumer and business segments whose socio-economic demographics make mobile commerce unviable.

In many regions, especially Asia and Africa, consumer and small business sectors in lower socio-economic groups have the double barriers of no infrastructure and un-affordability, with the result that much of their population cannot today be counted as potential mobile-customers for C2C, B2C, and G2C segments.

Digital Divide—Transition Factors

The last few years have seen an enormous increase in the number of mobile phone connections in all global regions. This is shown in Table 2.

A common trend noted globally is the increase in the proportion of mobile subscribers to fixed telephone line customers. Some customer segments, especially youth segments in rental accommodation, may no longer see the necessity for a fixed line. Logistically, the resources required for installation of new mobile infrastructures in rural or undeveloped regions may be less than that required for new fixed-line infrastructures.

In Australia, the Australian Communications Authority's "Telecommunications Performance Report 2003-04" tabled that the number of mobile phone services had exceeded the number of fixed telephone services operating by June 2004. The number of mobile phone services grew by 15.4% over the period, with a growth in prepaid services, which by then made up 43% of mobile services

Table 2. mobile phone connections - summary

Region	1998 (1000s)	2003 (1000s)	CAGR (%) 1998-03 [a)]	Per 100 inhabitants 2003	As % of total telephone subscribers 2003
Africa	4,156.9	50,803.2	65.0	6.16	67.3
Americas	95,066.8	288,219.9	24.8	33.80	49.8
Asia [b)]	108,320.6	543,153.4	38.1	15.03	52.4
Europe [c)]	104,382.0	441,234.9	33.4	55.40	57.5
Oceania	5,748.5	17,256.3	24.6	54.45	57.2
World	317,674.8	1,340,667.7	33.4	21.91	53.9

Notes:

[a)] The compound annual growth rate (CAGR) is computed by the formula:

$[(P_v / P_0)^{(1/n)}]-1$

where P_v = Present value

P_0 = Beginning value

n = Number of periods

The result is multiplied by 100 to obtain a percentage.

[b)] by end 2003, Hong Kong and Taiwan had exceeded a rate of 100% phones per inhabitant.

[c)] by end 2003, Italy and Luxembourg had exceeded a rate of 100% phones per inhabitant.

Data source: International Telecommunication Union, 2004: Cellular subscribers.

(Australian Communications Authority, 2004).

Customer segments in regions with limited fixed-line infrastructures may now, for the first time, have access to modern telecommunications. Of particular interest is the increase in the size of the potential mobile-customer segments in regions that until 2002 were limited in the infrastructure required to support mobile commerce, thus enabling the creation of an emerging market segment.

The UN's International Telecommunications Union industry report, "Trends in Telecommunications Reform 2004-2005," has been reported by the press to state that globally, 2004 revenue from mobile services is expected to be higher than revenue from fixed telephone line services. China, India, and Russia were stated to have the highest rate of increase (Australian IT, 2004).

In India, a press release from the Telecom Regulatory Authority of India stated that during 2004, approximately 19.5 million mobile subscribers were added, giving a total of 48 million mobile subscribers (an increase of 68%). The number of mobile subscribers now exceeds that of fixed-line subscribers, who only experienced a small increase in numbers over the same period (Telecom Regulatory Authority of India, 2005).

Influence on Demographic Factors for This New Segment

The demographics of this emerging segment, especially of those located in less developed regions, may differ from early adopters of mobile commerce and Internet users in these regions by factors including:

- more likely to use a pre-paid account and less likely to own a credit card or have access to other e-commerce payment methods;
- mobile devices more likely to be limited to mobile phones, rather than business-oriented devices such as PDAs;
- wider geographic location—that is, rural areas without fixed-line telephony and Internet may now have access to mobile telephony infrastructure;
- wider age spread—that is, may be used for communication between many generations of a family structure; parents may purchase a mobile phone for their children to enable a sense of security, and conversely, adult children may purchase a mobile phone for their elderly parents to satisfy the same objective;
- may have attained lower levels of education and literacy;
- less likely to speak or read English, or even to be fluent in their own national language;
- may be less familiar with current communication technologies; and
- small businesses, especially in the rural sector, may now have access to mobile telephony, thus facilitating the potential for the deployment of new business and agricultural techniques.

Location Differentiation

Some differences affect all potential mobile-customers in a specific location, be it geographical region or individual country/province.

Geography

- Time differences in mobile-customers' time zones, established business hours, and public and religious holidays could affect peak and off-peak system processing loads, with implications for the scheduling of system downtimes for maintenance or upgrades, and the staffing of call centres and other customer services.
- Seasonal and climate differences affect the marketability and usability of some products.

- Metropolitan vs. rural locality can impact the availability and quality of communications and product delivery infrastructure, unless the product can be delivered via the mobile device. Many Asian and African rural areas lack communication and other infrastructures, and even remote locations well serviced by satellite communications, such as the Australian outback or Antarctic bases, can have poor or expensive product delivery services.

Products and Services

- Suitability for use in global locations must be considered. Is there a need for the product or service? What use is a service to send payment details to a parking meter if few customers in the region own a car? Will the product actually work? This is especially an issue for electrical goods such as chargers for mobile devices or other items purchased via mobile commerce which may not be compatible with local equipment.
- Accuracy and knowledge of locality is important for some products, especially location-based services that interact with and require a global positioning system (GPS) infrastructure in the region.
- Social acceptance of products needs to be understood. Is the product attractive to the locations' typical mobile-customer needs, social values, and religious beliefs, or even legal?
- Equipment and availability for mobile commerce may differ for some customer segments. A business traveller expecting global availability of Wi-Fi "hotspots" for PDA or PC connection may be disappointed when travelling in less technically developed regions, and there are some regions that are not yet reachable by commercial GPS satellites.
- Handset types required depend on whether the local networks offer Global System for Mobile communication (GSM), Code-Division Multiple Access (CDMA) of which there are many variations, Personal Digital Cellular (PDC), or Third-Generation/Universal Mobile Telecommunications System (3G/UMTS) services. The Japanese network types are fairly unique to Japan; few commercially available handsets can be used both in Japan and other countries. Despite the availability of Japans' NTT DoCoMo iMode service in many countries, including Australia, the applications available and handsets required do differ between the individual countries of implementation.
- The number of potential customers who are visiting a region affect the viability of services that are aimed at the visitor, for example local directories, tourism guides, or special communication roaming deals such as SingTel's "Local Direct Dial" in Singapore (SingTel, 2005).

Product Content and Interface Presentation

- Language and keyboard/screen character sets differ. This is especially important to remember if mobile-customers are sought in China or Japan. Emerging mobile-customer segments may require mobile devices and applications to be designed using the local language for the interfacing component.
- Marketing promotions should be sensitive to customers' varying social backgrounds and local legislation regarding content.

Financial

- Credit card ownership is not ubiquitous in some regions. While customers can use other forms of payments such as invoicing, COD,

or local debit cards for national purchases, credit cards are the most widely acceptable payment method for international mobile commerce. In many Western European countries, Spain and Germany in particular, most consumers use debit rather than credit cards, limiting their global mobile-customer potential (Barclays, 2001; Forrester Research Technology, 2004). Some Asian countries such as South Korea, Japan, and Hong Kong have high credit card ownership (Lafferty Cards International, 2004a), while most others do not.

On an optimistic note for global mobile commerce, data shows that credit card ownership is growing/is projected to grow strongly in the Asia Pacific (Visa, 2004a), especially Indian credit card use (Gupta & Dasgupta, 2004), as well as Central and Eastern Europe, and Middle East regions (Visa, 2004a). To overcome this limitation, billing options that integrate with the customers' mobile account should be considered, whether it is a pre-paid or post-paid account.

- Cash payments are preferred by consumers in some regions. Visa notes that over 90% of transactions in the Asia Pacific region are made in cash (Visa, 2004b). Some European countries such as Greece are still cash oriented (Lafferty Cards International, 2004b). Again, these customers could be catered for by billing options that integrate with the customers' mobile account, which may well be a pre-paid account.
- Currencies for transactions—can customers pay in their own currency, only in the major currencies, or only in the currency of the mobile commerce business?
- Taxes—VAT, GST, state, and other sales taxes may or may not be payable on transactions depending on where the mobile commerce site is located and the location of the customer.

Legal

- Forbidden products both create and limit mobile commerce opportunities in some regions. There may be a large potential market for prohibited goods, especially in countries such as Saudi Arabia where alcohol and a range of other goods are forbidden (Department of Foreign Affairs and Trade, 2004) for a mobile commerce enterprise willing to engage in a high-risk venture. Otherwise, such products should not be included when targeting consumers in those regions to avoid causing offence, litigation, or censorship.

  Mobile services and content that does not meet local legislative requirements could cause the loss of a mobile operator's license.
- Privacy regulations differ greatly across the world in regard to data collection and management, and unsolicited marketing. In Australia the Privacy Act applies to businesses with an annual turnover of more than $3 million and all businesses of certain types (Office of the Federal Privacy Commissioner, 2005). There are no significant data protection laws in the U.S. at this point. Member countries in the European Union have some of the strictest data protection laws in the world which attempt to control their citizens' data stored in non-member countries too (European Commission, 2005).

Customer Differentiation

Within a location, individual customer demographic and lifestyle differences may alter the identification and classification of customer segments from standards in the mobile commerce's home location.

Demographic

- Age group usage may differ especially in locations where older groups have limited literacy. Younger groups may embrace internationalism and be confident using a wide range of services, including those marketed in the English language, while older groups may be more conservative and prefer using brands and services that reflect their own culture. Younger groups may be more confident using their mobile telephone for more than just telephony and are enthusiastic users of Short Message Service (SMS).
- Education is especially important in developing locations, where generally only the better educated have an opportunity to earn sufficient income to acquire the neccessary infrastructure.
- Gender may affect customer segments in locations where females in lower socio-economic groups are less educated.
- Family lifecycle stage groups may differ in relative segment sizes. For instance, the relative size of the European "adult with no dependents" demographic is larger than that in many Asian countries.
- Metropolitan vs. rural locality differentiation is covered above. In some regions, education and financial infrastructures may also be limited in rural and remote areas.
- Language used may be different to the national language. Many regions comprise many ethnic language groups, especially India. English is more likely to be understood by the higher socio-economic groups.

Lifestyle

- Time consciousness—mobile devices are more likely to be used in the course of performing business functions when timing of communications is critical, or of a personal nature when the customer has limited time for family and social activities. Different cultures experience a difference in expectations of what is considered "on time" or not.
- Moral attitudes vary greatly, especially for sexuality. Various "adult" services of a sexual nature are marketed heavily to mobile customers in some regions, but could cause the loss of an operator's license if marketed or offered in a region with strict legislation controlling mobile content.
- Personal values differ between cultures, which should be taken into account when marketing and designing features. Is the target society one that values concepts of individuality or social and family group membership? Is there prestige associated with acquisition of new mobile and other technologies?
- Attitude to adoption of new technology may differ between different segments within a region. Japanese youth are well known for their enthusiastic embrace of mobile telephones, individualizing accessories, and mobile services offered in particular by their iMode system.

Firmographics

- Size does matter. Globally, smaller businesses are less likely to use the latest technologies (OECD, 2001). Small businesses in developing locations are even less likely due to infrastructure issues listed above.
- Industry sector is shown to affect Internet use (OECD, 2001). Predominately subsistence-level agricultural communities may not require mobile commerce.

STRATEGIES FOR THE DESIGN OF PRODUCTS AND MARKETING

Strategies for the products, services, and marketing delivered by the mobile commerce business require tailoring for the targeted mobile-customer segments.

First, the customer segments should be identified according to the global differentiations outlined. Next, a decision should be made whether to create individual products, services, and marketing for different segments, or create a common suite to be used for all.

Factors indicating individual suites include:

- Significant differences in deliverable products and services, and customer differentiation, especially in legal restrictions, currencies, language, and social values.
- Economic justification for developing multiple products, services, and marketing campaigns.

Factors indicating a common suite include:

- Uniformity in products, services, and customer demographics.
- Uneconomic to develop multiple products, services, and marketing campaigns.

Then, the targeted mobile-customer segments should be guided to the appropriate site by strategies such as:

- Marketing and linking the mobile commerce product or service from an established mobile commerce portal, perhaps run by the telecommunication operator in the region or from a relevant Internet site in the region.
- Marketing via traditional channels such as print advertising in the region.

And finally, the mobile-customers should be provided with good "quality of service" regardless of their time zones and other differences. Consistent service availability and customer support should be provided to the most profitable customer segments at least, and ideally, to all.

CONCLUSION

While the recent arrival of mobile telecommunications infrastructures in most regions of the world has created a vast number of potential mobile-customers, mobile commerce businesses should be aware of the many geographical, legal, and demographic differences summarised in the following diagram before attempting to trade internationally, or deliver products and services developed outside their region to the local market.

Shrinkage of the digital divide for business and medium-high socio-economic groups across international boundaries, especially in the Asian region and within developed countries, is enabling the potential for even more growth in the size and variety of mobile-customer segments.

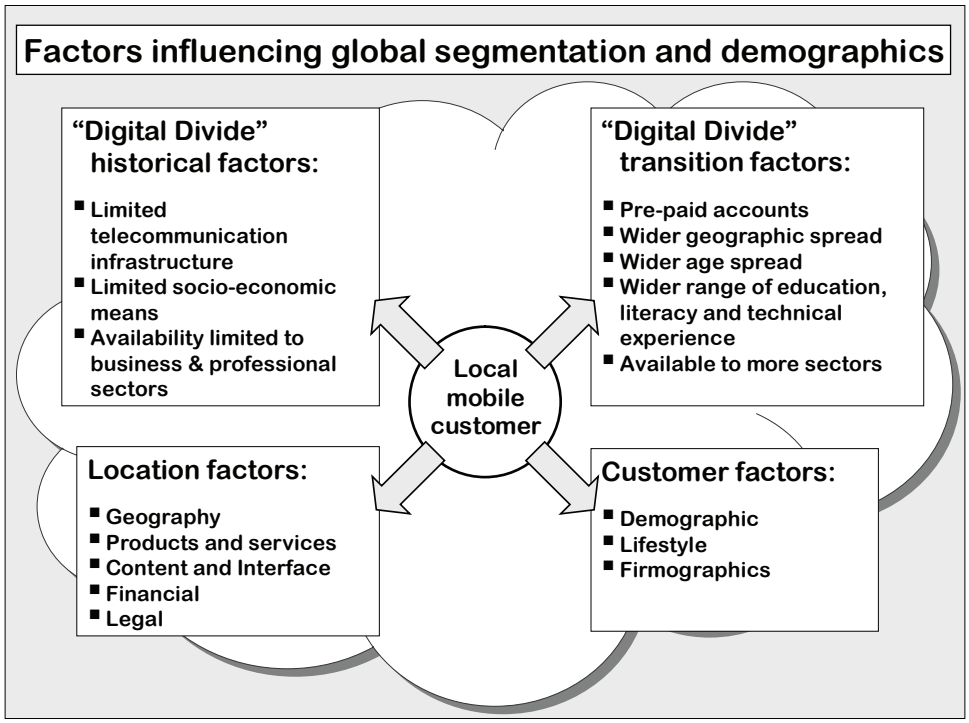
The recent emergence of new potential mobile-customers outside the established socio-economic and urban-located demographic groups requires more careful tailoring of products, services, and billing options than for the more established segments.

Benefiting from this expected growth can only be achieved by ensuring the mobile commerce's products and services, interface design, and marketing; customer service is tailored to satisfy the targeted market segment, being either the established or emerging mobile-customer segments.

REFERENCES

Australian Communications Authority. (2004, December). *Media release 95: Growth in mobiles*

Figure 1. Summary of influencing factors



and wireless broadband highlight year in telecommunications. Retrieved January 22, 2005, from http://internet.aca.gov.au

Australian IT. (2004, December 14). *Mobile revenue to outstrip landlines.* Retrieved January 10, 2005, from http://www.australianit. news.com.au

Barclays. (2001). *International growth*. Retrieved September 10, 2004, from http://www.investor.barclays.co.uk

Department of Foreign Affairs and Trade. (2004). *Department of Foreign Affairs and Trade, Saudi Arabia country brief.* Retrieved September 10, 2004, from http://www.smartraveller.gov.au

European Commission. (2005). *Information society—Telecommunications, privacy protection*. Retrieved January 23, 2005, from http://europa.eu.int

Forrester Research Technology. (2004, August). *Forrester's consumer technographics.* Retrieved September 10, 2004, from http://www.forrester.com

Gupta, N. S., & Dasgupta, S. (2004). Dragon fire's no match for India's credit card club. *The Economic Times* (April 8). Retrieved September 10, 2004, from http://economictimes. indiatimes.com

International Telecommunication Union. (2004). *Mobile cellular, subscribers per 100 people 2003.* Retrieved January 21, 2005, from http://www.itu.int

Lafferty Cards International. (2004a, August). *Korean card use declines.* Retrieved September 10, 2004, from http://www.lafferty.com

Lafferty Cards International. (2004b, August). *Olympian leap forward for Greek cards.* Retrieved September 10, 2004, from http://www.lafferty.com

OECD (Organisation For Economic Cooperation and Development). (1999). *The economic and social impact of electronic commerce: Preliminary findings and research agenda.* Retrieved September 11, 2004, from http://www.oecd.org

OECD. (2001). *Understanding the digital divide.* Retrieved September 11, 2004, from http://www.oecd.org

Office of the Federal Privacy Commissioner. (2005). *Private sector—business.* Retrieved January 22, 2005, from http://www.privacy.gov.au

SingTel. (2005). *Visiting Singapore*. Retrieved January 23, 2005, from http://home.singtel.com

Telecom Regulatory Authority of India. (2005, January 9). *Press Release no. 6/2005.* Retrieved January 22, 2005, from http://www.trai.gov.in

Visa. (2004a). *Visa Asia Pacific.* Retrieved September 10, 2004, from http://corporate.visa.com

Visa. (2004b). *CEMEA.* Retrieved September 10, 2004, from http://corporate.visa.com

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 655-665 , copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.10
Identified Customer Requirements in Mobile Video Markets—A Pan-European Case

Torsten Brodt
University of St. Gallen, Switzerland

ABSTRACT

Due to a significant cost advantage, mobile multicasting technology bears the potential to achieve extensive diffusion of mobile rich media applications. As weak performance of previous mobile data services suggests, past developments have focused on technology and missed customer preferences. Mobile multicasting represents a radical innovation. Currently, little insight on consumer behaviour exists regarding such services. This chapter presents results of qualitative and quantitative field research conducted in three countries. It provides a continuous customer integration approach that applies established methods of market research to the creation of mobile services. Means-end chain analysis reveals consumers' cognitive reasoning and conjoint analysis drills down to the importance of service attributes. Desire for self confidence and social integration are identified key motivators for consumption of mobile media. Services should aim for technological perfection and deliver actual and entertaining content. Interestingly, consumers appreciate reduced but tailored contents and price appears not to be a superseding criterion.

INTRODUCTION

After its first years of existence, the still emerging mobile telecommunications industry is undergoing a period of fundamental change. Since previously high growth rates of voice revenues started to decrease, the industry is looking for additional sources of revenue, such as mobile data services. However, the development of marketable services proves to be far more challenging than the one of stable, high-quality voice services.

Immature technologies are often blamed to be the reason for bad performances. Undoubtedly, the technological development is dynamic and, in fact, we argue that the intense focus on technology push has been one key factor of the

misfortune with mobile data services, as it detracts from customer needs. Furthermore, since vertical integration in the mobile telecommunication industry is low, product development is often organized in cooperative forms (Hagedoorn & Duysters, 2002). Coping with the complexity of innovation network management additionally detaches actors from actual customer needs.

Based on this, we see a need for a thorough understanding of the consumer behaviour side of mobile data services. Numerous studies have addressed issues of adoption and diffusion of mobile data services with the aim to identify diffusion barriers (e.g., Pedersen & Ling, 2003; Pousttchi & Schurig, 2004). However, such research seldom results in operational recommendations for companies on how to align their services with customer needs. We chose to focus on a specific range of services that exploit the investments in larger bandwidths and to develop a thorough understanding of the relations between service characteristics and fulfilment of customer needs and desires.

Since mobile multicasting services are based on a new technology and address a new market, they are termed a radical innovation (Veryzer, 1998). Thus, customer preferences can hardly be drawn from existing resources. By participating in the European "mobile multicasting service development and field trial project" MCAST (www.mcast.info), we were able to conduct the necessary market research.

Within a new product development process, customer integration is best realized after a first internal clarification of product ideas and possibilities, and subsequently after the technical engineering phase before market introduction (Gruner & Homburg, 2000). For this purpose, we integrated qualitative and quantitative methods to explore and formally describe customer needs. In the early stage we aimed to decrease uncertainty by conducting focus groups. We complemented the results by conducting individual laddering interviews following the means-end chain framework (Gutman, 1982). With both methods we were able to obtain a complete set of service characteristics and the underlying cognitive reasoning. In the later stages of development, we conducted a prototype-based adaptive conjoint analysis to quantify relative importance and the preferred levels of service characteristics. These analyses were conducted in Switzerland, Israel, and Greece.

We claim three major contributions to extant research. First, our results provide information on what consumers expect of mobile video services and which reasons drive these expectations. Second, our results quantify the relative importance of service attributes, for example price vs. context dependency. Third, we provide a methodology on how customer needs for break-through mobile service innovations can be obtained. This enables a customer-centric development of radical innovations.

BACKGROUND—MOBILE MULTICASTING

MCAST's multicasting technology enables cellular operators to use shared channel resources for broadcasting video and any other data over 2.5G and 3G networks. MCAST also yields a seamless roaming to WLAN networks. Therefore, MCAST aims at supporting cellular operators to establish affordable flat-fee services for end users and increase operators' revenues per channel resource, allowing economic delivery of media to an unlimited number of cellular and WLAN devices.

Current Technology Constraints

Currently, rich media content can be delivered over cellular networks using unicasting (one-to-one) technology. This has two major shortcomings: high delivery cost and limited cell capacity. Delivery cost is high, since each mobile terminal

accesses a content server for on-demand content. When users view rich media content, their mobile terminals consume an excessive amount of bandwidth. This results in very high by-the-minute or by-the-packet charges. Due to limited cell capacity, unicasting of rich media can only support a limited number of subscribers at any given time. As the number of online users increases, additional bandwidth is required. Current technology performance, therefore, allows only poor service levels and implies lost revenues.

Challenge, Solution, and Opportunity

Multicasting technology is based on a one-to-many broadcast concept. It enables the delivery of identical content simultaneously to an unlimited number of subscribers. This allows services to scale to almost any number of users while having a manageable and limited impact on available bandwidth per cell. For the end user, multicasting represents a convenient way of accessing rich media content. In this sense, from a user as well as business model perspective, multicasting is believed to be a successful bearer for rich media content over 2.5G and 3G cellular networks.

Since there is currently no competing or ready-to-market technology that can provide multicasting services over 2.5G or 3G cellular networks, the MCAST research project moves on the forefront of technological development (Heitmann, Lenz, & Zimmermann, 2003; Northstream, 2002), and it will contribute to the ongoing standardization process of multicasting in the 3rdGeneration Partnership Project (3GPP). Alternative technologies like DVB-H required substantial investments in new network infrastructure, and others like unicasting have an operational cost disadvantage. With its technological characteristics, multicasting is particularly suitable for rich media content (e.g., video, audio, gaming). Major market research institutions forecast the market potential of video services to nearly double that of audio services (e.g., Müller-Veerse, 2001) and a take up in 2005/06 (e.g., de Lussanet, 2003; Ovum Research, 2002). Based on this, our research focuses primarily on the delivery of video clips to mobile handsets.

EARLY-STAGE IDENTIFICATION OF CUSTOMER REQUIREMENTS

The early and qualitative part of customer integration employs focus groups to determine critical customer requirements as well as individual in-depth interviews following the means-end chain (MEC) methodology (Gutman, 1982) to understand the cognitive structures of decisions and the social motivation for requirements.

Focus Groups and In-Depth MEC Interviews: Background and Methodology

The focus group research was structured according to a theoretical concept for comprehensive and customer-driven product and service design: the OIL product design concept (Schmid, 2002). According to this, an evaluation of product expectations has to consider the levels of organizational design, interaction design, and logic design.

The organizational design level supplies the structural basis for the product design task. It answers the question of *who* and *what* is involved in the product use. Thus, in the case of a customer-oriented design of MCAST services, user groups and content categories must be determined. The interaction design concentrates on the processes and interactions between the relevant elements defined in the organizational level. It thus answers the question of *how* the product will be integrated into everyday life. The logic design examines *why* users use a specific innovation. Based on this understanding of the decision process, the product's language and communication strategy can be designed.

The succeeding means-end chain (MEC) approach (Gutman, 1982) generates an understanding of customers cognitive structure. The MEC concept is partitioning this cognitive structure in three layer—that is, service attributes, needs, and values. In the market research and service design literature, the qualitative MEC analysis has been increasingly an object of scientific debate (e.g., Aschmoneit & Heitmann, 2002; Grunert & Grunert, 1995; Herrmann, 1996a, 1996b; Wansink, 2000). It is based on two assumptions: (1) values, defined as desirable end states of existence, are dominant in the formation of selection structures; and (2) people deal with the variety of services by forming classes to reduce decision complexity.

For the formation of classes, consumers consult perceived and anticipated consequences of their actions or decisions. They associate positive consequences, namely benefits, with certain decisions (Reynolds & Gutman, 1988). Personal values allocate a positive or negative valence to these consequences (Rokeach, 1973). Thus, a correlation between the concrete and abstract characteristics of a service, the functional and psychological consequences, as well as the instrumental and target values is assumed (Gutman, 1997). Since consumers form classes to simplify their decision-making process, relatively few values are connected to a larger number of consequences and attributes. In this hierarchy, the importance of values determines the importance of consequences and attributes (Rosenberg, 1956).

Values represent beliefs about oneself and the reception of oneself by others. They are understood as universal, object-, and situation-independent convictions about desirable end states of life (Schwartz, 1994). The MEC framework is used to reveal the connections between time-stable values and product attributes directly relevant to decision making.

To obtain such results, the laddering technique with individual in-depth interviews is employed (Reynolds & Gutman, 1988). Research has shown that, on average, after 10 to 15 interviews, the number of additionally obtained consumer needs is decreasing radically (Griffin & Hauser, 1993). The technique reveals links between attributes, consequences, and values. The mentioned interactions between the obtained constructs were counted and entered into an implication matrix (not shown), a quantitative, tabular summary of the laddering interviews. This matrix provides the basis for the graphical representation in the form of a hierarchical value map (HVM), which displays the chains between values, benefits, and attributes, and their strengths (Herrmann, 1996a; Reynolds & Gutman, 1988).

Focus Groups and In-Depth MEC Interviews: Results and Implications

In total seven focus groups were conducted, three in Switzerland and four in Israel. Participants were selected from two mobile operators' customer databases according to a screener questionnaire to find high-volume customers with strong interest in innovative services. Each group consisted of five to eight participants; discussions lasted 60 to 90 minutes. The identified issues relate to: (1) relevance and entertainment qualities of content; (2) speed, visual quality, and reliability of technology; and (3) customizability of the service. We spare a detailed discussion of the focus group results and provide an exemplary overview of key requirements mentioned in two of the groups in Table 1.

For the MEC analysis, 30 innovators and early adopters were selected in Switzerland. The sample consisted of students and employees between the ages 20 to 40 of companies offering financial and consulting services. Interviews lasted between 30 minutes and one hour.

The obtained constructs complement on the one hand the requirements identified within the focus groups. On the other hand the MEC approach allows structuring of the cognitive reasoning in

an HVM (see Figure 1), summarizing service characteristics at the bottom, service benefits in the middle, and associated personal values at the top layer. For MCAST, two key paths of end user reasoning can be identified. One relates to information and self-confidence, and the other is associated with social integration:

- **Self-Confidence:** Being informed and deriving a personal opinion are among the main benefits associated with the reception of rich media content on a mobile device. That is, end users seek news content, enabling them to feel up to date at any point in time. Three characteristics led to this benefit—the immediacy, the usefulness of the content, and the "anytime-and-any-place" characteristic. A service that follows this reasoning should not only provide updated information, but also ensure the contextual relevance of information.
- **Social Integration:** Consumers feel multicasting services may support them in achieving this goal by providing a basis for social interaction and the development of a personal opinion. While the latter greatly depends on the reliability of the service and the independence of the presented information, the support for social interaction also depends on the entertainment characteristics of the service.

The identified cognitive pathways provide guidance for service development. Immediacy, the relevance of content, and entertainment qualities should especially be taken into consideration. Winning companies will include the benefits of "Feeling Informed," "Support for Social Interaction," and "Forming Personal Opinions" to address the beliefs of consumers.

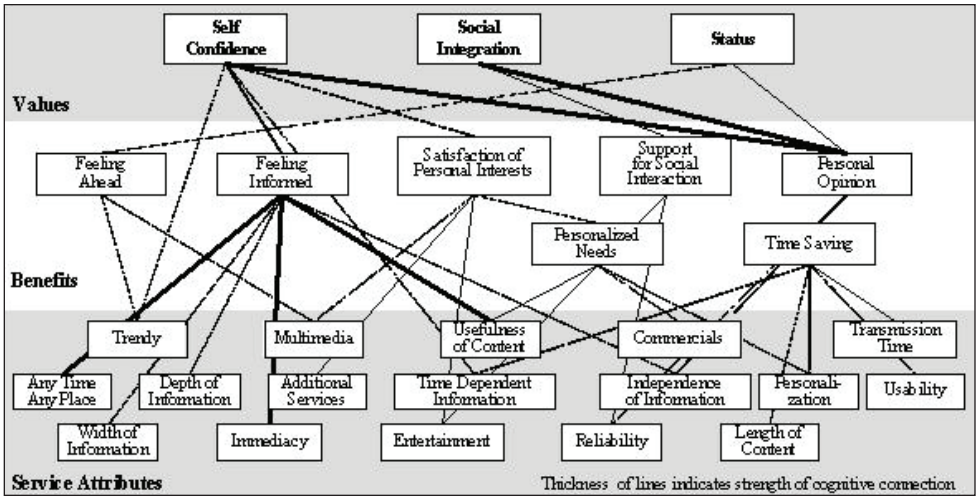
LATER STAGE CUSTOMER REQUIREMENT ANALYSIS

The preceding analyses show that customers consider a wide range of characteristics when

Table 1. Exemplary Focus Group Results - Key Requirements

	Group 1	Group 2
Content	• Good Editing of Content	• Availability/Quality of content
	• Up-to-date Content	• Up-to-date Content
	• Local Content	• Width of Content
	• Fun	• Independence of Content
Technology	• Reliability of the Service	• Battery Consumption
	• Saving Functionality	• Screen Size
	• Picture Quality / Resolution	• Size of Device
	• Sound Quality	• Rapid and Secure Transmission
Service	• Personalization	• Any time and any where
	• No Advertisement	• No Advertisement
	• Forwarding	• Forwarding
	• Easy to Operate	• International Roaming
	• Price	• Price
	• Customizability of Content	• Customizability of Content

Figure 1. Hierarchical Value Map (HVM)



evaluating a mobile multicasting service, which bears still too much complexity for service design. Therefore, we employed an adaptive conjoint analysis (ACA), a sophisticated customer research approach (Green, Krieger, & Wind, 2001; Hauser & Rao, 2002) to determine the weights of characteristics.

Adaptive Conjoint Analysis (ACA): Background and Methodology

The ACA allows identification of the relevance of service attributes and their levels—that is, it reveals the relative importance of different service attributes. The generated database allows the running of price sensitivity analysis for different product scenarios and an estimation of purchase probabilities (Johnson, 1991). Compared to other types of conjoint analyses, the ACA enables a dynamic adoption of a questionnaire according to given answers to preceding questions. This allows generation of robust results also for complex product offerings with a high number of attributes (Huber, Wittink, Johnson, & Miller, 1992; Orme, 1999) and ensures suitability for Web-based survey design (Dahan & Hauser, 2002).

Before implementing the ACA, the attributes under investigation have been reduced in an additional iteration step to 13 attributes. The objective of this step was not only to fulfil the conjoint requirements (of attribute independence, relevance, objective exclusiveness), but also to select attributes in conjunction with technological capabilities and business relevance. Accordingly, defined attribute levels are shown in Table 2.

The ACA was programmed using SSI Web of Sawtooth Software. The ACA questionnaire was hosted online and complemented by supplementary questions on general mobile usage behaviour and content requirements. This survey was conducted in three countries. In Switzerland 125 individuals have been invited from an academic database. Participants were required to be heavy users of mobile services. They were informed about the multicasting service by use of an animated prototype and in-depth information pro-

Table 2. Attribute-Level Matrix

Attribute	Level 1	Level 2	Level 3
Length of Content	• Max. 30 sec.	• Max. 1 minute	• Max 2 minutes
Number of Clips per Day	• 5	• 10	• 15
Premium Content	• Available	• Not Available	•
Subscription Fee (€)	• 3	• 6	• 9
Forwarding	• Via MMS	• Not Possible	•
Ensured Transmission	• Retransmission	• Clips Lost	•
Supplemental Internet Service	• All Clips Online	• Missed Clips Online	• No Clips Online
Advertisements	• Yes	• No	•
Notification on Missed Clips	• No	• Per SMS	• Per MMS
Number of Content Categories	• 5	• 10	• 15
Number of Clips in MCAST Inbox	• 3	• 5	• 7
International Roaming	• Available	• Not Available	•
Location Based Content	• Available	• Not Available	•

vided with an interactive CD-Rom. Participants were then asked to answer the online survey. After data was cleaned to ensure data robustness, 103 data sets were used for analysis. In Greece and Israel the participants have been recruited from the project partners' databases. The main difference was that users in these countries had the opportunity to use the service in the life network for a duration of four weeks. After data cleaning, the analysis contained 67 participants in Greece and 97 in Israel.

Adaptive Conjoint (ACA): Results and Implications

Questions complementary to the actual ACA asked for the general background of participating individuals (e.g., demographics, telecommunication behaviour). Among others, these questions confirmed the interest in news and location-specific content. Furthermore, video content proves to be the most preferred format.

The following section selectively documents the quantitative conjoint results. Data reveals that in all cases, five service attributes influence almost 50% of the consumer decision (see Figure 2). Not surprisingly, price concerns yield a high score of 13.8% in Israel, 10.6% in Greece, and 11.7% in Switzerland. However, attribute scores are rather evenly distributed, and given an adequate price span and a flat fee pricing model, price appears to be not an overriding decision criterion for consumers.

Figure 2 depicts the attribute importance in consumer decision making for the three countries. Certain similarities can be identified—for example, the high weight of the item "Ensured Transmission," which points to the importance of a technically flawless service. Since multicasting services are broadcast to a group of subscribers once and simultaneously, it might happen that a few subscribers do not receive the content due to handset unavailability or interrupted transmission. Users are concerned to lose out on these clips and therefore strongly require the notification and the back-up through supplemental Internet services.

Taking a closer look at single attributes (here we chose the data for Switzerland) and their levels reveals interesting aspects of the willingness to consume mobile data services. As documented in Figure 3, subscribers rather prefer a reduced num-

Figure 2. Cross Country Comparison of Attribute Importance from ACA

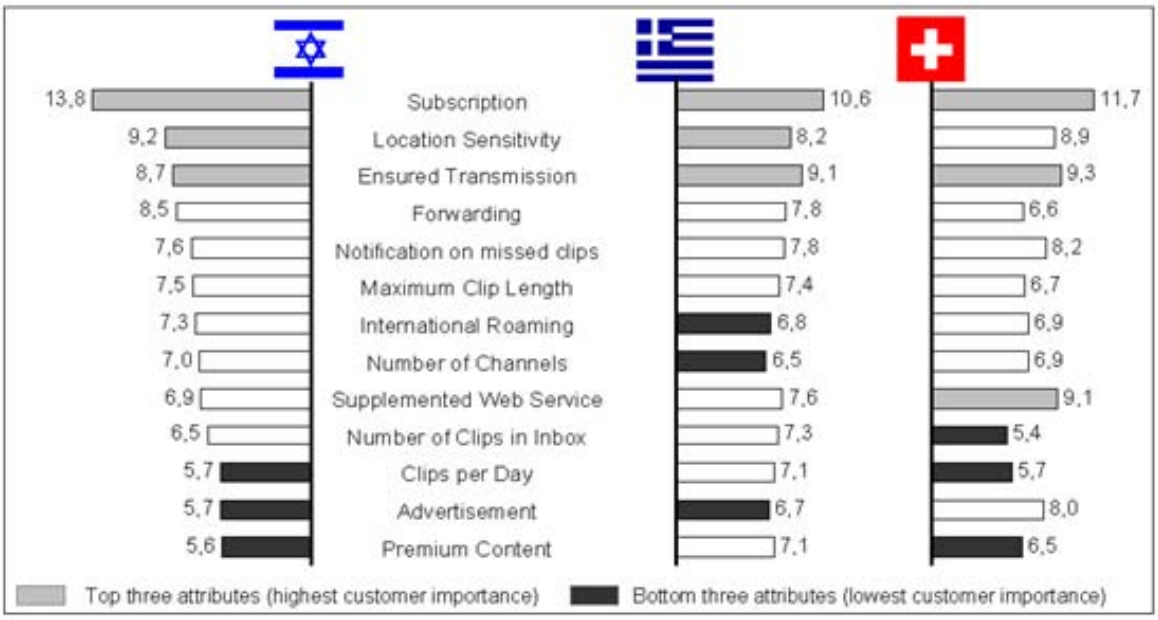
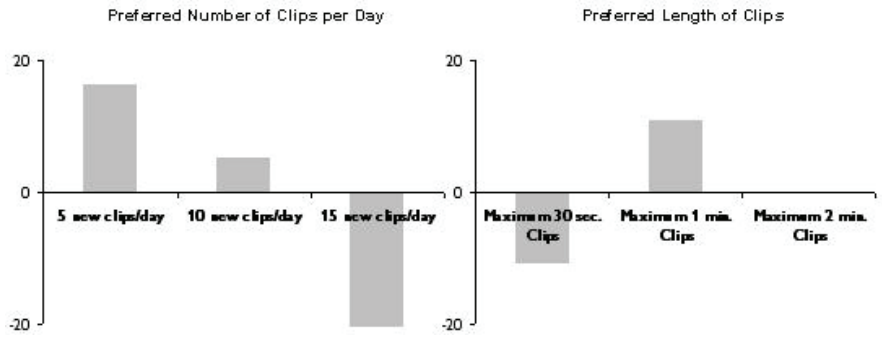


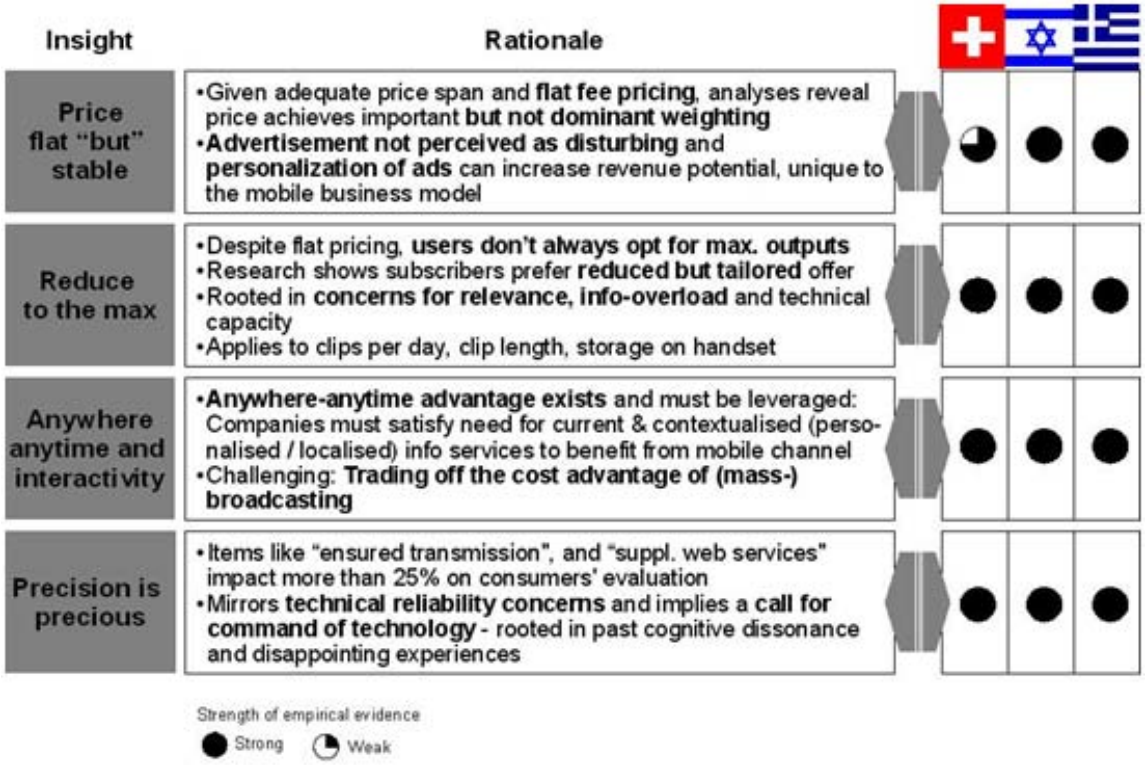
Figure 3. Attribute Levels for Clip Length and Number of Clips per Day (Switzerland)



ber of (five) clips per day combined with a maximum clip-length of one minute. This behaviour relates to the concern about content relevance, but also about technical capacity (e.g., transmission speed and memory capacity) mentioned during the preceding qualitative surveys.

Analysing the findings of the three country-specific surveys on an aggregate level reveals four main patterns of consumer behaviour regarding mobile video services. As shown in Figure 4, these patterns relate to (1) the proven existence of a willingness to pay, if the price is controllable, preferably a flat fee. (2) The second pattern describes the users' preference for a reduced but tailored mobile video offer; that is, despite flat

Figure 4. Evidence for Behavioural Patterns in Observed Countries



pricing, users do not always opt for maximum of outputs. This behaviour is rooted in concerns for relevance, information overload, and technical capacity, as also shown in the MEC-analysis. (3) The very advantage of mobile technology of delivering services "anywhere-anytime" is also a valuable selling point for mobile video services. That is, companies must develop intelligent means to satisfy the need for current and contextualised (personalised/localised) services, without destroying the scale effects of mass-broadcasting. (4) Precision is precious—this pattern represents the users' concerns about technical reliability rooted in past cognitive dissonance and disappointing experiences, and it implies a call for command of technology.

CONCLUSION AND BUSINESS BENEFITS

By reporting insights in terms of methodology and identified customer preferences regarding mobile-rich media services, we address the lack of customer knowledge in marketing practice and research in the mobile media industry. While dealing with the development of a leading-edge multicasting technology, we deployed a set of sophisticated tools for customer integration along the development process. For customer research science, we show a methodology, on how customer needs for break-through mobile service innovations can be obtained in a way that generates results, which can be easily communicated within single companies and across innovation networks. With the growing importance of cooperative product development, investigations on the latter, such

as a joint customer integration and its qualities, will be an area for future research.

For management, our quantitative empirical results imply precise insights for superior mobile multicasting service design. Additionally, the identified cognitive reasoning of consumers provides input for general communication and marketing strategies. We show that most importantly, management needs to master the doubts on technology performance, and that mobile content must be tailored. The latter point complicates the marketing challenge as it trades off the multicasting cost advantage. For marketing and communication strategy, we have identified that the consumers' desire for self-confidence and social interaction should be addressed.

NOTE

An earlier version of this chapter appeared in: Cunningham, P., & Cunningham, M. (Eds.). (2004). *E-adoption and the knowledge economy: Issues applications, case studies* (Vol. 1, pp. 50-58). Amsterdam: IOS Press.

REFERENCES

Aschmoneit, P., & Heitmann, M. (2002). Customer-centred community application design. *The International Journal on Media Management, 4*(1), 13-21.

Dahan, E., & Hauser, J. R. (2002). Product development—Managing a dispersed process. In R. Wensley (Ed.), *Handbook of marketing* (pp. 179-222). Thousand Oaks, CA: Sage Publications.

de Lussanet, M. (2003). *Limits to growth for new mobile services.* Cambridge, MA: Forrester Research.

Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces, 31*(3), 56-73.

Griffin, A., & Hauser, J.R. (1993). The voice of the customer. *Marketing Science, 12*(1), 1-17.

Gruner, K. E., & Homburg, C. (2000). Does customer interaction enhance new product success? *Journal of Business Research, 49*(1), 1-14.

Grunert, K. G., & Grunert, S. C. (1995). Measuring subjective meaning structures by the laddering method: Theoretical considerations and methodological problems. *International Journal of Research in Marketing, 12*(3), 209-225.

Gutman, J. (1982). A means-end chain model based on consumer categorization processes. *Journal of Marketing, 14*(6), 545-560.

Gutman, J. (1997). Means-end chains as goal hierarchies. *Psychology and Marketing, 14*(6), 545-560.

Hagedoorn, J., & Duysters, G. (2002). External sources of innovative capabilities: The preferences for strategic alliances or mergers and acquisitions. *Journal of Management Studies, 39*(2), 167-188.

Hauser, J. R., & Rao, V.R. (2002). *Conjoint analysis, related modeling, and applications.* Unpublished manuscript, MIT Sloan, USA.

Heitmann, M., Lenz, M., & Zimmermann, H.-D. (2003). *Preliminary user needs analysis for MCAST.* St. Gallen, Switzerland: MCM Institute.

Herrmann, A. (1996a). *Nachfrageorientierte produktgestaltung: Ein ansatz auf basis der "means end"—theorie.* Wiesbaden, Germany.

Herrmann, A. (1996b). Wertorientierte produkt— und werbegestaltung. *Marketing ZFP, 18*(3), 153-163.

Huber, J., Wittink, D. R., Johnson, R. M., & Miller, R. (1992). Learning effects in preference

tasks: Choice-based versus standard conjoint. In *Proceedings of the Sawtooth Software Conference,* Ketchum, ID (pp. 232-244).

Johnson, R. (1991). Comment on adaptive conjoint analysis: Some caveats and suggestions. *Journal of Marketing Research, 28,* 223-225.

Müller-Veerse, F. (2001). *UMTS report—an investment perspective [online].* Retrieved August, 2002, from http://www.durlacher.com

Northstream. (2002). *The competitive landscape of mobile video on demand [online].* Retrieved February, 2002, from http://www.northstream.se/21/

Orme, B. (1999). *ACA, CBC, or both?: Effective strategies for conjoint research: Sawtooth software.* Sequim, WA.

Ovum Research. (2002). *Ovum forecast: Global wireless markets.* London.

Pedersen, E., & Ling, R. (2003). Modifying adoption research for Mobile Internet service adoption: Cross-disciplinary interactions. In *Proceedings of the 36th Hawaii International Conference on System Sciences 2003,* Hawaii (pp. 534-544).

Pousttchi, K., & Schurig, M. (2004). Assessment of today's mobile banking applications from the view of consumer requirements. In *Proceedings of the 37th Hawaii International Conference on System Sciences 2004,* Hawaii (pp. 184-191).

Reynolds, T. J., & Gutman, J. (1988). Laddering theory, method, analysis, and interpretation. *Journal of Advertising Research, 28*(1), 11-31.

Rokeach, M. J. (1973). *The nature of human values.* New York: The Free Press.

Rosenberg, M. J. (1956). Cognitive structure and attitudinal affect. *Journal of Abnormal and Social Psychology, 22,* 368-372.

Schmid, B. (2002). *Kommunikations—und medienmanagement.* Unpublished manuscript, St. Gallen, Switzerland.

Schwartz, S. H. (1994). Are there universal aspects in the structure and content of human values? *Journal of Social Issues, 50*(4), 19-45.

Veryzer, R. W. (1998). Discontinuous innovation and the new product development process. *Journal of Product Innovation Management, 15,* 304-321.

Wansink, B. (2000). New techniques to generate key marketing insights. *Marketing Research, 12*(2), 28-36.

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 754-764, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.11
Special Features of Mobile Advertising and Their Utilization

Jari Salo
University of Oulu, Finland

Jaana Tähtinen
University of Oulu, Finland

INTRODUCTION

Mobile advertising, or m-advertising, refers to ads sent to and presented on mobile devices such as cellular phones, PDAs (personal digital assistants), and other handheld devices.[1] M-advertising can be seen as a part of m-commerce (e.g., Mennecke & Strader, 2003), which is seen as radically different from traditional commerce (e.g., Choi, Stahl, & Whinston, 1997). Thus, it can be argued that m-advertising is also different. M-advertising enables the advertiser not only to send unique, personalized, and customized ads (Turban, King, Lee, Warkentin, & Chung, 2002), but also to engage consumers in discussions and transactions with the advertiser.

Any retailer can make use of m-advertising. Thus this study focuses on the brick-and-mortar retailers' use of m-advertising in Finland. In Finland, mobile phone subscriptions reached 84% of the population at the end of the year 2002 (Ministry of Transport and Communications Finland, 2003), and more than 30% of the users under 35 years and over 20% of all users have received m-advertising in the form of SMS (www.opas.net/suora/mob%20markk%20nous.htm). However, there are no commercial solutions available for the MMS type of m-advertising. Therefore, the empirical setting of this study is a service system SmartRotuaari, which is a part of a research project (see Ojala et al., 2003; www.rotuaari.net) offering the retailers an infrastructure and a service system for context-dependent m-advertising in the city of Oulu in Northern Finland.

This study focuses on permission-based m-advertising. In Finland, that is the only form of m-advertising that is legal. Firstly, we will discuss the features of m-advertising that make it unique.

Secondly, we will present some empirical results from the SmartRotuaari case. Based on the recognized features, we study which of them retailers utilized in their m-ads, as well as those remaining unused. The aim is to find out how well the uniqueness of m-advertising was portrayed in the m-ads. The study concludes by suggesting how retailers could improve the use of m-advertising in order to fully harness its power.

DESCRIPTION OF MOBILE COMMERCE

Based on existing research and the empirical data gathered for this study, we suggest a framework that describes the factors that influence the success of retailers' use of m-advertising. The factors are related to the media or advertising channel itself and its special features, and to the receiver of the messages—that is, the individual customer and her/his goal in using the mobile device.

Factors Influencing the Success of Permission-Based M-Advertising

Because of the special features, m-advertising can and should be used to deliver ads which are different from the traditional ones. The special features include: the personal nature of the device, the interactivity that the device enables, and the context dependency that the infrastructure enables. The features influence the type of content that permission-based m-advertising should offer to the consumer in order to be perceived as valuable and/or entertaining. The value of the content is also related to the individual's needs and reasons for using the media, such as media goals (Juntunen, 2001). A person may use a mobile device to receive information, but also for the purpose of personal entertainment. Both these goals influence the expectations she/he has for the mobile ads. Unless the consumer perceives permission-based m-advertising positively, she/he can deny the company or any company the permission to send ads to her/him. Thus it is vital for a m-advertiser to be aware of the special features and the requirements that the features set for the content of the ads, as well as for the segmentation or almost individual targeting of the ad. In the sections below we will take a closer look at each of the features depicted in Figure 1.

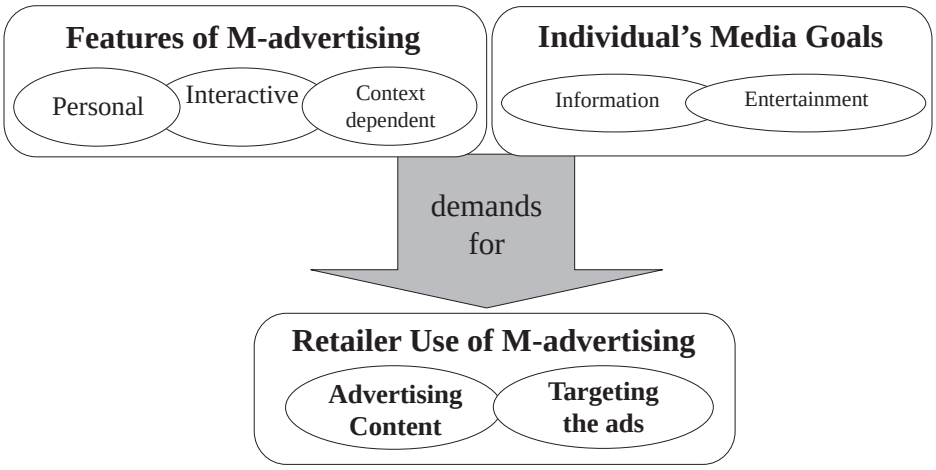
Personal Nature of M-Advertising

M-advertising is as personal as personal selling. Mobile devices, especially mobile phones, are highly personal devices, with personally selected or even self-composed ringing tones, individually tailored covers or general appearance, and additional decorations, not to mention the 'content' of the phone, including information on personal friends as well as a personal calendar. Moreover, the users wear their device almost everywhere and at all times. Thus the personal nature of the device is transferred to the information that is sent and received through the device (see also Barwise & Strong, 2002). Therefore, m-advertising is not for the masses, but for individuals.

Interactive Nature of M-Advertising

The mobile device allows m-advertising to be highly interactive—that is, the parties can act on each other, on the communication medium, and on the messages (Liu & Shrum, 2002). A customer may reply to an ad by phoning; sending an SMS, MMS, or an e-mail; or logging into the advertiser's Web page by using the mobile device. In addition, a customer may distribute the ad to her/his friends. Such viral marketing is very beneficial for the advertiser, as the customer forwarding the ad her/himself becomes the sender of the message and therefore the message gains in credibility.

Figure 1. Possible features influencing the success of permission-based m-advertising



Context Awareness of M-Advertising

The first context to be taken into account is the device to which the advertisement is distributed. Unless the message is tailored to the terminal, the receiver will face problems in receiving and understanding the message. Even if such problems are avoided, the devices have a relatively small screen size, limited screen colors, and limited battery time. However, the technology used in building m-advertising systems enables context awareness. The context may be location, time, and/or weather. For example, the m-advertising service is able to locate the user's mobile device and send an ad only when the customer walks by the retailer's shop.

An Individual's Media Goals

An individual's goals are often referred to as a person's cognition of what s/he is pursuing in a particular situation and to an associated inner state of arousal (e.g., Eysenck, 1982; Pervin, 1989). Thus an individual's media goal is her/his cognition of the processing goal s/he is pursuing when attending to her/his mobile device (see Juntunen, 2001), which in this case is the medium for m-ads. Depending on what type of goal the receiver is trying to achieve by using a mobile device also affects her/his processing of the ads. If the user's media goal is information, the customer will be more interested in ads that provide her/him relevant information on products/services or companies.

On the other hand, if the customer's goals are more towards entertainment, s/he will enjoy ads that are entertaining and provide experiential satisfaction through aesthetic pleasure, emotional stimulation, or social experience (see also Barwise & Strong, 2002). A consumer may wish to achieve both kinds of goals at the same time, and the relative importance of the types may change according to the situation that s/he is in.

In the above, we have discussed the features that make m-advertising a unique form of advertising, as well as the ways users are using their mobile devices. Together they place m-advertising closer to personal selling than traditional advertising—having the same message sent to many receivers with limited control over the context. Since m-advertising is so personal, it sets new kinds of demands for the advertising planning.

We will now move into considering how to plan m-advertising—that is, targeting and planning the content of the ad.

Targeting the Ads

It is possible to target m-advertising if the retailer can make use of the user-specific information that is added to the m-advertising service system. This can be done through two different, but complementary ways. Firstly, each user, when granting the permission to send ads, also fills in a user profile that can include demographic details, user's current mood (e.g., is s/he hungry, looking for fun, shopping), and areas of personal interest (e.g., fashion, food, hunting). All this can be done directly form the user's mobile device. Secondly, the retailer may use existing data from the company's customer relationship management (CRM) database, which can be connected to the mobile user's personal customer number.

Moreover, the system may obtain up-to-date weather information from a local weather station via Internet. It is thus possible to send ads of sunglasses only when the sun is actually shining. In addition, time can be used in targeting. In the morning restaurants can send special breakfast offers or in the evening they can send discount coupons for a dinner if there are seats available.

A well-planned execution of m-advertising can be more effective than, for example, direct mail (which is often left unopened) or television advertising, although the number of receivers that see the ad is considerably smaller. Based on the targeting options, the retailer can send ads that match with the mobile user's personal interests and current needs, making sure that the customer will only receive ads that s/he is willing to. This is extremely important in permission-based m-advertising, since spam messages annoy the receiver (see also Barwise & Strong, 2002; Edwards, Li, & Lee, 2002). Therefore, the advertiser can reach high view-through rates by targeting the ad successfully. In SMS m-advertising, 81% of all trialists viewed all messages before deleting them and 77% did that as soon as they received the ad (Barwise & Strong, 2002). At the same time, this means that the same ad should only be sent to each customer once during a campaign. If the campaign contains repetition, the m-ads have to be different each time they are being sent to the same consumers, otherwise they can annoy the consumer.

Advertising Content

As for the content of ads, the advertiser in any type of advertising has to decide what is being said and how to say it. Both these decisions affect the success of m-advertising as well. Kalakota and Robinson (2002) suggest that m-ads work best if customers receive concrete benefits from it, such as retail alerts, coupons, special offers, and m-tickets. However, Barwise, and Strong (2002) found six types of ads used in SMS permission-based m-advertising, ranging from messages directed to long-term effects (like brand building) to messages attempting to engage the receiver in immediate interaction with the advertiser (competitions, votes). By applying the information given by the consumer and/or information retrieved from the CRM databases, the advertiser can also provide quick and timely information (i.e., news that interests the receiver). The existing research being scarce, we do not know which type of ads are the most effective ones.

The style of the ad is also an important issue to be considered. Duchnicky and Kolers (1983) suggest that reading from mobile devices may take more time and effort than reading from a desktop computer. Because of that, and also due to the space limitations, the copy should be kept short and the use of graphics or photos is encouraged (see Edens & Cormick, 2000). Humor and surprises in the design of the ad create positive feelings toward the advertisement and may lead to viral marketing, especially among the younger receivers (Barwise & Strong, 2002). Furthermore,

we assume that the personal nature of the mobile devices as well as the context specificity and novelty of m-advertising will lead consumers towards high involvement. In such situations the contrast effect appears to stimulate consumers to process the advertising even more (De Pelsmacker, Geuens, & Anckaert, 2002).

IMPACT OF MOBILE COMMERCE ON THE ORGANIZATION

The empirical part of the study is derived from the SmartRotuaari service system. The system provides a functional framework for large-scale field trials for the purpose of empirical evaluation of technology, new mobile services, customer behavior, and retailers' use of the services (for more details see Ojala et al., 2003; www.rotuaari.net). The retailers use a Web portal to send ads, which are then delivered through a WLAN network to consumers' mobile devices, in this trial the PDAs.

Retailers' Use of Permission-Based M-Advertising

The data consists of 42 m-ads that were sent to trial users (186 persons) by 12 retailers (shops, bars, restaurants, cafes) during the first field trial of the SmartRotuaari.[2] Thus, the retailers had their first experiences of m-advertising during this trial, and they had not received any special training to guide their m-ad design decisions. Thus, this data provides a great opportunity to study how retailers that are not advertising professionals apply the uniqueness of mobile channel.

The ads were analyzed using content analysis, as it is the standard analytical tool for advertising studies (e.g., Kassarjian, 1977; Kolbe & Burnett, 1991). As suggested by Kassarjian (1977), four coders (A, B, C, and D) analyzed the commercials. However, due to confidentiality of the data, the authors served as coders as well. The authors provided the coders A and C with instructions and a brief training before they commenced the task. Since the number of ads was relatively small, all disagreements between the two pairs of coders were solved through discussion (see Kassarjian, 1977). Since the coders were able to agree on all the decisions, no measure of interjudge reliability was calculated (see Perreault & Leigh, 1989)

All m-ads used the company location as the focal point from which the distance that triggered the sending of an ad was measured. However, there were huge differences in the way the retailers used the location awareness. The distance used varied from 75 meters to 3,000 meters. The diameter of the town centre in Oulu is below the 3 kilometres, so the use of the highest distance in the location awareness does not aid the targeting of the ads. The time awareness according to certain hours of the day (e.g., opening hours, lunch hours) was used in only 18 ads, although it could have been used in every ad, so that customers would receive ads only during the opening hours.

The most-used feature of the consumer that the retailers used in targeting was age. Only some clothing shops and a few restaurants did not use it. The bars and pubs especially targeted the ads towards either the younger or the more mature customers. The mood information was used in 65% of all the ads. Especially the clothing shops and cafeterias selected customers in shopping mood, and bars and restaurants people who were hungry, thirsty, seeking company, or in a mood to party. As for the consumers' interest areas, only 14 ads included certain interest areas as criteria for targeting. None of the 42 ads used the local weather as a criterion for sending the ad.

We also analysed the content of the ads. Fifty-five percent of the ads contained either photos (people, product, or the interior of a restaurant) or graphics. The copy length ranged from 0 to 31 words. As the ads were received on PDAs, even the longest copy was readable, but it did not provide an aesthetic pleasure. From the ads, there was only one that was classified as brand building,

but this is easily explained by the fact that all the advertisers were retailers, and thus most of the ads concentrated on describing the shop or the restaurant (e.g., what type of food was served). Many ads (40%) included their contact address (only three with phone number), although it was possible for the consumer to use a mobile map to locate the company. Moreover, 45% of the ads contained information on opening hours, which explains the fact that many advertisers did not use the option of sending the ad during the opening hours only. One-third of the ads included price information or special offers, thus responding to the consumer's relevant information needs. Moreover, only three ads addressed the receiver in the copy by asking them a question ("Are you hungry?") or by welcoming them to the cafe.

The retailers used very traditional profiling criteria such as the age of the customer. We can also argue that although mood or interest areas are not really a criterion that can explicitly be used in, for example, magazine or newspaper advertising, it is used implicitly when choosing the magazine (e.g., interior magazines) or placing the ad under the 'entertainment' section. Also in the content of the ad, traditional newspaper advertising was clearly the point of reference when retailers designed the m-ads. How to fit the message and the format into the context of m-advertising is a question also to be solved by advertising agencies (e.g., Kiani, 1998; Kunoe, 1998).

CONCLUSION

This study on retailers' usage of permission-based mobile advertising underlines the notion that mobile advertising is different from any other form of advertising. In addition, the retailers, using m-advertising for the first time, are not able to apply the unique features of m-advertising. Thus, both the receivers and the senders of mobile advertising messages have to learn how to use this new channel and how to fully make use of the opportunities it offers for speedy, personal, and interactive advertising communication with the consumer (see also Pura, 2002). The features of m-advertising (personal, interactive, context dependent) and individual media goals (information and/or entertainment) should be the basis to start m-advertising activities and campaigns. Therefore, the m-advertising should be personal, thus requiring a certain amount of knowledge about the receivers of the m-ads. The message of the advertisement, as well as the way it is expressed, should be carefully designed to match the needs of the target person. Moreover, m-advertising should fit into the marketing communication mix, enabling interactivity. In time, we are sure that m-advertising will move more and more towards m-crm and constant interaction between buyer and seller.

REFERENCES

Barwise, P., & Strong, C. (2002). Permission-based mobile advertising. *Journal of Interactive Marketing, 16*(1), 14-24.

Choi, S. Y., Stahl, D. O., & Whinston, A.B. (1997). *The economics of electronic commerce*. Indianapolis, IN: Macmillan Technical.

De Pelsmacker, P., Geuens, M., & Anckaert, P. (2002). Media context and advertising effectiveness: The role of context appreciation and context/ad similarity. *Journal of Advertising, 31*(2), 49-61.

Duchnicky, R. L., & Kolers, P. A. (1983). Readability of text scrolled on visual display terminals as a function of window size. *Human Factors, 25*(1), 683-692.

Edens, K. M., & McCormick, C. B. (2000). How do adolescents process advertisements? The influence of ad characteristics, processing objective, and gender. *Contemporary Educational Psychology, 25*(2), 450-463.

Edwards, S. M., Li, H., & Lee, J.-H. (2002). Forced exposure and psychological reactance: Antecedents and consequences of the perceived intrusiveness of pop-up ads. *Journal of Advertising, 31*(4), 83-95.

Eysenck, M. (1982). *Attention and arousal, cognition and performance.* New York: Springer-Verlag.

Goldsborough, R. (1995, May 8). Hong Kong trams keep ads rolling. *Advertising Age, 66,* 36.

Hume, S. (1988, April 11). New medium is semi success. *Advertising Age, 59,* 22-24.

Juntunen, A. (2001). *Audience members' goals of media use and processing of advertisements.* Unpublished doctoral dissertation, Helsinki School of Economics and Business Administration, Finland.

Kalakota, R., & Robinson, M. (2002). *M-business. The race to mobility.* New York: McGraw-Hill.

Kassarjian, H. (1977). Content analysis in consumer research. *Journal of Consumer Research, 4*(1), 8-18.

Kiani, G. R. (1998). Marketing opportunities in the digital world. *Internet Research: Electronic Networking Applications and Policy, 8*(2), 185-194.

Kunoe, G. (1998). On the ability of ad agencies to assist in developing one-to-one communication. Measuring "the core dialogue." *European Journal of Marketing, 32,* 1124-1137.

Liu, Y., & Shrum, L. J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. *Journal of Marketing, 31*(2), 53-64.

Mennecke, B. E., & Strader, T. J. (2003). *Mobile commerce: Technology, theory and applications.* Hershey, PA: Idea Group Publishing.

Ministry of Transport and Communications Finland. (2003). *Ensimmäisen aallon harjalla. Tekstiviesti-, WAP- ja MMS-palveluiden markkinat 2000-2004.* [On the first wave]. Publication of the Ministry of Transport and Communications Finland. Retrieved October 14, 2004, from http://www.mintc.fi/www/sivut/dokumentit/julkaisu/julkaisusarja/2003/a192003.pdf

Ojala, T., Korhonen, M., Aittola, M., Ollila, M., Koivumaki, T., & Tahitinen, J. (2003, December 10-12). SmartRotuaari—Context-aware mobile multimedia services. *Proceedings of the 2nd International Conference on Mobile and Ubiquitous Multimedia* (pp. 9-18). Norrköping, Sweden. Retrieved from http://www.ep.liu.se/ecp/011/005/

Perreault, W. D., & Leigh, L. (1989). Reliability of nominal data base on qualitative judgments. *Journal of Marketing Research, 26*(3), 135-148.

Pervin, L. (1989). Goals concepts: Themes, issues, and questions. In L. A. Pervin (Ed.), *Goal concepts in personality and social psychology* (pp. 473-479). Hillsdale, NJ: Lawrence Erlbaum.

Pura, M. (2002). Case study: The role of mobile advertising in building a brand. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce: Technology, theory and applications* (pp. 291-308). Hershey, PA: Idea Group Publishing.

Turban, E., King, D., Lee, J., Warkentin, M., & Chung, H.M. (2002). *Electronic commerce: A managerial perspective.* Upper Saddle River, NJ: Prentice-Hall.

KEY TERMS

Mobile Advertisement: All advertisements sent to mobile and wireless devices.

Mobile Advertising: All advertising activities conducted via mobile and wireless devices.

Mobile Commerce: All commerce conducted via mobile and wireless devices.

Mobile Marketing: All marketing activities conducted via mobile and wireless devices.

Permission-Based Mobile Marketing and Advertising: All marketing activities conducted with permission of the consumer via mobile and wireless devices.

ENDNOTES

1 Mobile advertising can be used to refer to advertisements that move from place to place, (i.e., in busses, trucks, trains, etc.) (e.g., Hume, 1988; Goldsborough, 1995).

2 The SmartRotuaari service system consists of several mobile services, which are tested and studied in field trials. So far, m-advertising service is one possessing the most commercial potential.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 1035-1040, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.12
Personalization and Customer Satisfaction in Mobile Commerce

HY Sonya Hsu
Southern Illinois University, USA

Songpol Kulviwat
Hofstra University, USA

INTRODUCTION

The advancement of wireless technology facilitates both consumers' activities and business transactions. With the rapid proliferation and widespread use of mobile devices, including mobile phones, personal digital assistants (PDAs), and handheld computers, mobile commerce or m-commerce is widely considered to be a driving force for the next generation of electronic commerce (e-commerce). According to Jupiter Research, the m-commerce industry is expected to be US$22 billion globally by 2005. However, to date many promising technologies—especially m-commerce applications—have failed with the notable exceptions of i-Mode service and short messaging service (SMS).

Popular "i-Mode", produced by NTT DoCoMo of Japan, is a service that enables wireless Web browsing and e-mail from mobile phones. The "i-Mode service" has been the first successful commercial introduction of 3G (third-generation) mobile applications. It exceeded expectations and acquired over 30 million profitable users in a three-year period (Cohen, 2002).

One of the main goals of most operators might be building customer satisfaction and loyalty by providing one or more 'killer apps' to them. One way is to integrate customer relationship management (CRM) into the development of mobile services' applications. Some firms have tried to target these applications to their customers on an individualized basis. "Personalization" may be the way to achieve that. Specifically, personalization can be regarded as the use of technology and user/customer information to match multimedia content with individual needs with the goal of producing user satisfaction. Personalization can

be presented by an IP services framework that allows operators and subscribers through self-service provisioning approaches to control the types of service and applications they want and are willing to buy.

The purpose of this article is to develop a deeper understanding of personalization, with an emphasis on those factors that lead to customer satisfaction and/or delight. Specifically, this article presents factors contributing to consequences derived from using personalized applications and services in m-commerce.

BACKGROUND

In their pilot study, Ho and Kwok (2003) applied the technology acceptance model (TAM) originated by Davis (1989) to their m-commerce study. They utilized four constructs to predict the service subscribers' intention to switch: number of generalized messages, perceived ease of use of general advertisements, perceived usefulness of personalized message, and privacy issues about personalized advertisements.

This article extends the thrust of Ho and Kwok's research to incorporate the effect of personalization on customers' satisfaction and delight that could contribute to CRM. Customers' satisfaction and delight are derived from expectancy theory, and they are discussed by Oliver (1981), Oliver, Rust, and Varki (1997), Spreng, Mackenzie, and Olshavsky (1996), and Verma (2003).

Expectancy: Satisfaction and Delight

Expectancy theory is used to frame the evaluation of mobile services users. Oliver (1981) defined expectation to include two components: the probability of occurrence (e.g., the likelihood that a personalized cell service will be available) and an evaluation of the occurrence (e.g., the degree to which the personalization level is desirable or undesirable). The disconfirmation/confirmation paradigm of satisfaction is based on expectancy theory. It can be an emotional response to the comparison of the performance received and the products' normative standards. When the performance and expectations are at variance with each other, there is a discrepancy. This discrepancy could be either *positive* (when performance exceeds the expectations), which often causes satisfied state, or it could be *negative*, when performance is worse off than expected (Oliver, 1981). In other words, the consumer would be satisfied if perceptions match expectations or if confirmations are reached. Consistent with Spreng et al. (1996), satisfaction arises when consumers compare their perceptions of the performance of a good and/or service to both their desires and expectations. As such, satisfaction is a subjective judgment and may imply mere fulfillment.

Delight is a positively valence state reflecting high levels of consumption-based affect. The feeling of delight is experienced when the customer is pleasantly surprised in response to an experienced disconfirmation. It is the feeling state containing high levels of joy and surprise (Westbrook & Oliver, 1991). Further, Oliver et al. (1997) proposed and confirmed that delight is a function of surprising consumption, arousal, and positive effect or a function of surprisingly unexpected pleasure. They empirically confirmed that delight is a "mixture" of positive effect and arousal or surprise. It is associated with the level of arousal intensity. Moreover, it is a reaction experienced by the customer when he or she receives a service and/or a good that does not simply evoke a feeling of satisfaction, but also provides an unexpected value or unanticipated additional pleasure. In other words, delight occurs when the outcome is unanticipated or surprising. It can be marked by pleasurable, unforgettable, and memorable feelings in a service encounter or a product purchase (Verma, 2003). It is thought to be the key to customer loyalty and loyalty-driven profit (Oliver et al., 1997) and is known as the highest level of expectation-disconfirmation paradigm.

Technology Acceptance Model (TAM)

From Davis' (1989) TAM model, ease of use (EOU), and perceived usefulness (PU) of a technology are factors that either directly or indirectly increase a person's intention to adopt an innovation. While *perceived usefulness* is the degree to which a person believes that using a particular technology/system would enhance the outcome performance, *perceived ease of use* is the extent to which a person believes that using a particular technology/system will be free of effort (Davis, 1989). TAM could be helpful in predicting the usage of personalized applications and services. Greer and Murtaza (2003) adapted the TAM model to study issues that impact the valuation of Web personalization as well as factors that determine customer use of Web personalization. Ho and Kwok (2003) adapted Davis' (1989) EOU and supported the effect of using a generalized message on changing a service provider. They also used "PU of personalized service" to test the importance of personalization in mobile commerce. They found support for both. Most importantly, the PU of personalized service was the most effective factor, together with ease of locating generalized message and the amount of generalized message that affected the decision to change to a new service (Ho & Kwok, 2003).

MAIN THRUST OF THE ARTICLE

Usually when there are too many generalized messages, customers lose their motivation to read, retrieve, or even locate a useful message. In addition, the amount of space available on the mobile screen limits the amount of options and information. Given this, personalization is considered to be the key factor for success/failure of mobile devices and services. Information and services must become increasingly tailored to individual user preferences and characteristics in order to accommodate limited space and scarce airtime. Personalization is viewed as including "recognition of a customer's uniqueness" (Surprenant & Solomon, 1987, p. 87), use of a customer's name, and response to customer needs (Goodwin & Smith, 1990).

Message Format

Carlson et al. (1998) characterized medium *richness* as the capacity to convey information. It is further defined as the ability to provide immediate feedback to customers' consumption of media. Rich information can be produced by giving immediate feedback, having a variety of available communication cues, understandable/common language, and foremost, personalization of the medium (Carlson et al., 1998).

Media richness theory postulates that media selection depends on the uncertainty of the task at hand (Kumar & Benbasat, 2002). Both media richness theory and the TAM model have illustrated their relationships with task orientation. Also, social presence theory postulates a particular communication task based on the degree of necessary social presence that links a selection of media (Kumar & Benbasat, 2002). Originally, it referred to the degree to which a medium allows a user to establish a personal connection with the other users. Social presence seems to be moving towards a task orientation at an individual level in the latter theoretical development, such as the para-social concept from Kumar and Benbasat (2002). Para-social is a combination product of social presence and media richness. This article focuses on the PU of personalized messages that employ a task orientation, while two different formats of messages (text and multimedia) were drawn from media richness theory.

Personalization

Personalization can be defined as the use of technology and user/customer information to

customize multimedia content so as to match with individual needs and ultimately produce user satisfaction (Zhang, 2003). Personalization is primarily regarded with sending the right message to the right person at the right time. The main goal behind personalization is to make any medium's usage easier and enhance any channel communication between customers and service providers.

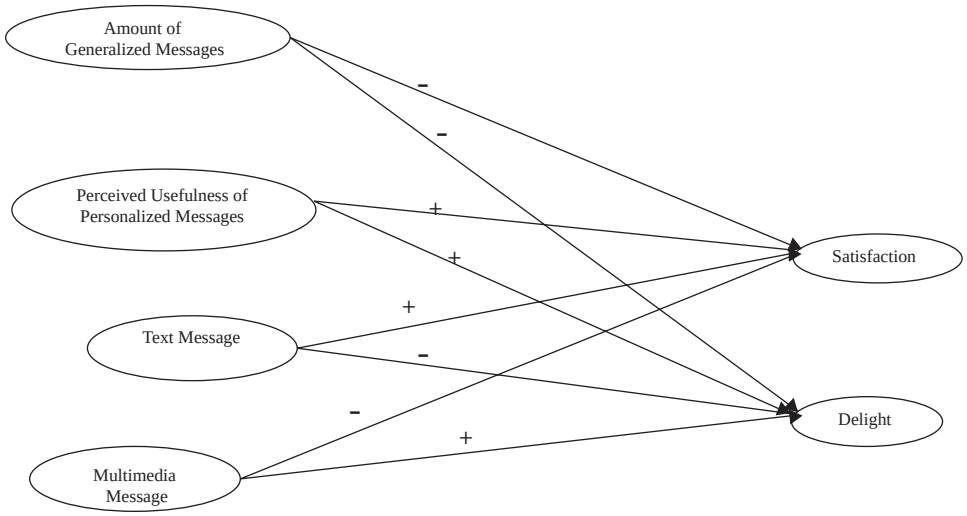
Personalization translates individual profiles into unique presentations. The individual profiles can be built upon user preferences, the quality of his or her senses, user location/environment, contexts, users' network, and terminal capabilities. Morris-Lee's (2002) study on personalization of brochures indicated that personalization helped increase interest and involvement. The more personalized features are, the greater the possibility of increased costs (Greer & Murtaza, 2003). Hence, these increasing costs hopefully should produce greater customer satisfaction and retention, thus a greater return. This is a very important point because, for example, a 5% increase in customer retention costs can translate into a 25%-125% increase in company profitability (Reichheld et al., 2001). Also, personalization of service has been found to have a positive impact on customers' evaluations of service encounters (Surprenant & Solomon, 1987).

Figure 1 presents the research model of personalization in m-commerce. The model is developed based on the TAM (Davis, 1989) and expectancy theory (Oliver, 1981). Specifically, this research model extends Ho and Kwok's (2003) research framework. Further, our model integrates customer expectancy as an endogenous variable. The model includes six sets of variables: (1) number of generalized messages, (2) perceived usefulness, (3) text messages, (4) multimedia message, (5) satisfaction, and (6) delight.

FUTURE TRENDS

As predicted, text message predictor had a positive association with the dependent variable. Number of generalized messages was negatively related to satisfaction. Multimedia message was not a significant predictor of satisfaction and was deleted in the second model. On the other hand, the analyzes of "delight," "multimedia message" contributed the most importance to the

Figure 1. Research Model



equations and was positively related to delight. Text message and PU of personalized message had positive associations with delight, while the number of generalized messages had a negative contribution to the equation (Hsu, Bruner, & Kulviwat, 2005).

According to Santos et al. (2003), even though satisfaction and delight are two different constructs, each serves a dimension of confirmation (satisfaction) at one end and disconfirmation (delight) on the other end. Increasing literature has being drawn in the difference between consumer satisfaction and delight (Kumar & Olshavsky, 1997; Oliver et al., 1997). To compare satisfaction with delight, Oliver et al. (1997) see customer delight as being fundamentally different from customer satisfaction. Compared to satisfaction, delight seems more abstract and more extreme in terms of affection. While satisfaction may be induced by avoiding problems or may meet standard/minimum requirement, delight requires more than that. Oliver et al. (1997) empirically confirmed the distinction between the satisfaction and delight constructs, with delight being a higher level of satisfaction. In fact, customer delight is associated with a strong and positive emotional reaction to a product or service. Thus, both practitioners and scholars should manage customer delight as a separate goal from satisfaction.

Te'eni, Sagie, Schwartz, Zaidman, and Amichai-Hamburger (2001) used three dimensions to define media richness further; these are interactivity, adaptiveness, and channel capacity. Beyond just a different format from a text message, future researchers may look into a deeper understanding of multimedia messages that convey information for possible customers' delight in addition to satisfaction. Delightedness can be marked with pleasurable, unforgettable, and memorable where customer loyalty is rooted (Verma, 2003; Oliver et al., 1997). With the limitation of student population, future research may investigate some professional group that has reasons to use mobile commerce and/or some population that has more disposable income at hand. Another limitation that can also be addressed in the future is the sophistication of multimedia services and the maturity of users. In other words, future researchers may look into some markets that have rolled out the mMode of AT&T, Mobile Web of Verizon, and/or Sprint's PCS.

From mobile application point of view, "personalization" can be more sensitive to users' needs, for example, location-based application as in www.mobull.usf.edu. Local merchants ally to deliver a personalized text message—such as sales, promotion advertisement, coupons—from a Web site to a wireless device based on personal preferences that are set up by each individual. Location-based services utilize location information to provide specialized contents to mobile users (Varshney, 2003). Explicit user permissions should be obtained before "pushing" any advertising contents to particular users (Varshney, 2003). Push and pull advertisement, of course, relates to the issues of privacy and sharing of user information. Therefore, the "trust" matter may surface between a group of local merchants and individual consumers.

CONCLUSION

This article identifies the same situation as in Ho and Kwok (2003) that the amount of generalized message had a negative effect on customer satisfaction. Personalized message is more likely related to customer satisfaction and delight. The TAM model and expectancy theory were drawn as the foundation of this research model. Media richness explains the division between text message and multimedia message, whereas TAM contributes the perceived usefulness of personalized message.

Beyond personalization, this article attempts to merge the media richness theory with expectancy theory. Specifically, it explains the relationships between text/multimedia message and customer's

satisfaction/delight. The article concludes that consumers would like to have a richer media to experience a "delightful" emotion. Consistent with the principle of media richness (Carlson et al., 1998): the more complex media format, the more information can be delivered in a message. If managers would like to increase effectiveness and/or efficiency of mobile services, text message alone would not be sufficient for market differentiation to gain competitive advantage. With personalization, multimedia formats can be a supplement tool to increase the interaction with consumers when launching advertising campaigns. The richer the media, the more effective it is in communication.

REFERENCES

Carlson, P. J., & Davis, G. B. (1998). An investigation of media selection among directors and managers: From "self" to "other" orientation. *MIS Quarterly, 22*(3), 335-362.

Cohen, A. S. (2002). *Unlocking ARPU through killer mobile data networks.*

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-339.

Goodwin, C., & Smith, K. L. (1990). Courtesy and friendliness: Conflicting goals for the service provider? *Journal of Services Marketing, 4*(1), 5-21.

Greer, T. H., & Murtaza, M. B. (2003). Web personalization: The impact of perceived innovation characteristics on the intention to use personalization. *Journal of Computer Information Systems, 43*(3), 50-55.

Ho, S. Y., & Kwok, S. H (2003). The attraction of personalized service for users in mobile commerce: An empirical study. *ACM SIGecom Exchanges, 3*(4), 10-18.

Hsu, S. H., Bruner G. C. II, & Kulviwat, S. (2005). Personalization in mobile commerce. Managing modern organizations with information technology. In M. Khosrow-Pour (Eds.), *Proceedings of the Information Resources Management Association International Conference,* San Diego (pp. 1103-1106). Hershey, PA: Idea Group Publishing.

Kumar, A., & Olshavsky, R. (1997, October 11). Distinguishing satisfaction from delight: An appraisal approach. *Proceedings of the Annual Conference of the Association for Consumer Research,* Tucson, AZ.

Kumar, N., & Benbasat, I. (2002). Para-social presence and communication capabilities of a Web site: A theoretical perspective. *e-Service Journal.*

Morris-Lee, J. (2002).Custom communication: Does it pay? *Journal of Database Marketing, 10*(2), 133-138.

Oliver, R. L., Rust, R. T., & Varki, S. (1997). Customer delight: Foundations, findings, and managerial insight. *Journal of Retailing, 73*(3), 311-336.

Oliver. R. L. (1981). Measurement and evaluation of satisfaction processes in retail settings. *Journal of Retailing, 57*(Fall), 25-48.

Reichheld, F. F., & Teal, T. (2001). *The loyalty effect: The hidden force behind growth, profits, and lasting value.* Boston: Harvard Business School Press.

Santos, J., & Boote, J. (2003). A theoretical exploration and model of consumer expectations, post-purchase affective states and affective behavior. *Journal of Consumer Behavior, 3*(2), 142-157.

Spreng, R. A., Mackenzie, S. B., & Olshavsky, R. W. (1996). A reexamination of the determinants of consumer satisfaction. *Journal of Marketing, 60,* 15-32.

Surprenant, C. F., & Solomonm, M. R. (1987). Predictability and personalization in the service encounter. *Journal of Marketing, 51*(2), 86-96.

Te'eni, D., Sagie, A., Schwartz, D. G., Zaidman, N., & Amichai-Hamburger, Y. (2001). The process of organizational communication: A model and field study. *IEEE Transactions on Professional Communication, 44*(1), 6-21.

Varshney, U. (2003). Wireless I: Mobile wireless information systems: Applications, networks, and research problems. *Communications of the Association for Information Systems, 12,* 155-166.

Verma, H. V. (2003). Customer outrage and delight. *Journal of Services Research, 3*(1), 119-133.

Westbrook, R. A., & Oliver, R. L. (1991). The dimensionality of consumption emotion patterns and consumer satisfaction. *Journal of Consumer Research, 18*(June), 84-91.

Zhang, D. (2003). Delivery of personalized and adaptive content to mobile devices: A framework and enabling technology. *Communications of AIS, 12*(13), 183-204.

KEY TERMS

Customer Delight: The feeling of delight is experienced when the customer is pleasantly surprised in response to an experienced disconfirmation.

Customer Satisfaction: Based on the consumption, consumer would be satisfied if perceptions match expectations or if confirmations are reached.

Expectance Theory: Oliver defined expectation to include two components: the probability of occurrence and an evaluation of the occurrence. The discrepancy of confirmation could be either positive or negative.

Media Richness: Media richness theory postulates that media selection depends on the uncertainty of the task at hand. The more complex media format, the more information can be delivered in a message.

Mobile or M-Commerce: Both consumers' activities and business transactions are facilitated by the advancement of wireless technology including cellular phones, wireless PDAs, or any hand-held units.

Personalization: Can be regarded as services of the use of technology and user/customer information to customize multimedia content aiming to match with individual needs and ultimately deliver customers' or users' satisfaction.

Technology Acceptance Model (TAM): From Davis' TAM model, ease of use (EOU) and perceived usefulness (PU) of a technology are factors that either directly or indirectly increase a person's intention to adopt an innovation.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 914-918, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.13
Cross-Cultural Consumer Perceptions of Advertising via Mobile Devices:
Some Evidence from Europe and Japan

Parissa Haghirian
Sophia University, Japan

Maria Madlberger
Vienna University of Economics and Business Administration, Austria

ABSTRACT

This chapter tries to answer the question on how international consumers differ in their perceptions of mobile advertising (m-advertising). In order to answer this research question a survey among mobile phone users was carried out in Austria and in Japan. These two countries have been selected because they show a high dissimilarity in culture in historical terms but also in the degree of technological development and usage. Both countries experienced a similar economic development and show comparable living standards. Furthermore, Japan and Austria are leading markets for advanced mobile technology in their geographic areas. On the basis of the literature review, variables picturing perceptions of m-advertising are defined, hypotheses in comparing consumer perspectives in the respective countries are presented, and results of the MANOVA analysis are shown and discussed. Finally, the paper presents theoretical and managerial implications, limitations, and recommendations for future research on this issue.

INTRODUCTION

Permanent Internet access enabled by mobile devices like mobile phones or PDAs is becoming more and more widely used. Mobile technologies open up new challenges for companies which want to benefit from mobile commerce (m-commerce). One of the most important commercial applications in this respect is marketing. Marketing instruments using mobile devices (m-marketing)

allow innovative forms of customer relationships and interaction. They can lead to the development of numerous m-commerce-based services (Venkatesh, Ramesh, & Massey, 2003). In m-commerce, a mobile phone can serve as a "portable entertainment player, a new marketing tool for retailers and manufacturers, a multi-channel shopping device, a navigation tool, a new type of ticket and money, and a new mobile Intranet device" (Funk, 2004, p. 2).

Compared with computer-based e-commerce, m-commerce is a very new area of development. As a consequence, m-commerce applications have been subject to academic research to a much lesser extent. Many potential applications of m-commerce are still under researched. One of them is the application of mobile devices for advertising purposes. One important exception is the empirical study by Okazaki (2004) who investigated Japanese consumers' perceptions of wireless ads. Beyond that, there is only little knowledge about how consumers react to advertising via mobile devices. This gap becomes even wider when this issue is addressed on an international basis. By now, no findings on cross-country differences in the context of m-advertising are available. In contrast, online advertising accessed via desktop computers is well researched. For example, Web-based research in cross-cultural context revealed that users' cultural background strongly influences their perception of visible advertising design elements and colors (Del Galdo & Neilson, 1996; Marcus & Could, 2000). World Wide Web advertisers react to this by creating locally oriented Web sites (Cyr & Trevor-Smith, 2004).

The current trend in e-commerce toward globalization may be observed in m-commerce as well. This makes a cross-cultural investigation on consumers' perceptions of m-commerce applications a critical research issue. The research project described in this chapter has been carried out in order to analyze to what extent consumers differ in their perceptions of advertising via mobile devices across different cultures. In order to achieve comparable results, the study focuses on push marketing activities in the form of text advertising messages sent to consumers.

MOBILE-ADVERTISING AND ITS TECHNOLOGICAL DIMENSIONS

Together with the development from seller markets to buyer markets in the consumer goods sector, technological innovations were a key driver for a paradigm shift in marketing theory in the 1980s and 1990s (Dwyer, Schurr, & Oh, 1997; Groenroos, 1994; Gummesson, 1987). Although there are critical arguments on this approach as well (Brodie, Coviello, Brookes, & Little, 1997; Fournier, Dobscha, & Mick, 1998), this marketing practice has gained importance. Information technology plays a key role in this development as data warehousing and data mining are necessary sources of information for obtaining knowledge about the customer (Parvatiyar & Sheth, 2000).

In general, advertising is defined as "any paid form of non-personal presentation and promotion of ideas, goods or services by an identified sponsor" (Kotler, 2003, p. 590). Advertising via mobile devices or m-advertising is defined as *the usage of interactive wireless media (such as mobile phones and pagers; cordless telephones; personal digital assistants; two-way radios; baby crib monitors; wireless networking systems; GPS-based locators; and maps) to transmit advertising messages to consumers with the overall goal to promote goods and services.* M-advertising can be carried out on the basis of a number of technologies. Besides Web-based approaches that apply mobile Internet, messaging-based push advertisements can be used. Since the target consumer can be clearly identified by the advertiser, these advertising messages may include time and location sensitive, personalized information that can be transmitted via text messages or via e-mail on the mobile Internet.[1] There are different synonyms for m-advertising, such as wireless advertising

(Krishnamurthy, 2003) or wireless advertising messaging (Petty, 2003).

Information systems are also vital in order to address each consumer on an individual basis (Peppers, Rogers, & Dorf, 1999). The opportunities of one-to-one marketing on the Web can be extended in m-commerce to context-specific marketing on allowing a higher degree of individualization. In particular, online activities can be closely linked with off-line activities. Some examples of potential m-advertising measures illustrate that, in practice, more innovative campaigns are possible. Advertising messages can be sent considering the *location* of the recipient, for example, containing a coupon with a price reduction on a certain product to a consumer who is in a particular shop. In addition, consumers' shopping needs can be accommodated with location-dependent offers and promotions (Stafford & Gillenson, 2003), like advertisements showing the menu of the day at a nearby restaurant or allowing access to a branded online game while waiting for the train at a railway station. All these advertising activities can also be used to create perceivable benefits for the consumers.

Like classical Web-based advertising, m-marketing activities allow personalization and interactivity. But m-advertising also has some distinctive features that enhance as well as limit advertising opportunities for marketers and lead to considerable differences compared to Web-based advertising. Besides the different optical appearances of m-advertising messages due to screen size, the linkage between online and off-line activities becomes more relevant. The recipient's context serves as an integrative part of the communication as messages can be adapted to the consumer's current location and time.

Personalization of Advertising Message Content

Marketing activities performed via mobile devices provide potential for personalization, because the transmission tools usually carry the user's assigned identity (Lee & Benbasat, 2003). Marketers can so use consumer feedback to customize their messages and offerings and collect information about consumers' preferences to improve future products and services (Stewart & Pavlou, 2002). The advantages of doing so are obvious. Potential customers can be addressed in a very individual way and relationships with the user improve because users are generally receptive to advertising that is personalized and relevant to their lifestyle (DeZoysa, 2002). Advertising can be carried out very precisely and with a clear focus on the target group (Varshney & Vetter, 2002). Using mobile devices to transmit messages to consumers also enables marketers to collect information on their current location. Consequently, advertising activities can be adapted to time and location-related consumer interests. In Japan, Internet-based services like city maps or train schedule information are commonly downloaded via mobile phones.

Information that is transferred in the context of m-commerce can thus be related to three situational aspects: (1) location, (2) time and location, and (3) context.

Location-Related Information

This takes into account where the recipient is situated during message transmittal. Consequently, consumers' shopping needs can be accommodated with location-dependent offers and promotions (Stafford & Gillenson 2003). In contrast to many Web-specific advertising instruments, which are limited to desktop computers, m-advertising allows their application at any location, for example, shops, pubs, cafes, public transportation, and other locations where a personal computer usually is not available. This allows a considerably improved customization of the advertising message. For example, a company can send an advertising message with a coupon for a price reduction on a certain product to a consumer who stays in a certain store. Such promotion campaigns

have been carried out successfully via Web-based coupons where consumers could redeem coupons obtained on the Web in the physical store (Madlberger, 2004).

Time and Location-Related Information

This takes time-specific settings into account. In these settings, a firm can transfer information to remind recipients of a happening in the near future, for example, an event or a time-dependent service (e.g., a dinner at a restaurant). Hence, this kind of information might encourage the recipient to move to a specific location at a certain point of time.

Customer-Context-Specific Information

This can be related to time or the recipient's location, but it is primarily focused on the actual situation of a recipient. In a setting, in which an individual is waiting or being bored (e.g., waiting for a train, waiting at a hospital), he/she might be more likely and willing to grasp information or access the Internet than during a period of activity. In such a situation, the perception of an advertising message might be higher.

Limitations of M-Advertising

Although m-advertising offers attractive and innovative opportunities, it also has important *limitations*, which make m-advertising rather impractical in its current form. These limitations imply that today's application opportunities are still far away from the aforementioned scenarios. Most limitations are due to technical attributes of the mobile devices. In order to be portable, mobile devices today have limited processing power, low bandwidth, and unfavorable input/output devices. It is expected that many of these drawbacks will be overcome in some years, but screen size will remain limited (Lee & Benbasat, 2003) and will be an obstacle to extensive m-advertising messages. Beyond the mobile device's limitations, today's technology is also characterized by limited capacity, for example, the maximum length of SMS texts or network operating systems. Design and content of m-advertising messages are therefore restricted to constraints in data volume and visual presentation.

CROSS CULTURAL PERCEPTIONS OF M-ADVERTISING

Mobile Development in Japan and Europe

In Japan, mobile phones started to gain popularity among young consumers as early as in the mid-1990s. In 1999, market leader NTT DoCoMo launched its mobile Internet-based i-mode service. The i-mode service allows mobile phone users constant access to the World Wide Web and enables subscribers to view Web pages via their mobile phones. Furthermore, they can send and receive mobile e-mails and can be directly addressed with advertising messages. As of the end of 2002 the proportion of mobile Internet users among mobile phone owners was 79.2 %. This was the highest percentage worldwide (Ministry of Public Management, Home Affairs, Post, and Telecommunications [MPHPT], 2003). Mobile phones also have quickly become a new advertising tool for more than 100 Japanese retailers and manufacturers that use mobile Internet as an instrument to target customers with discount coupons, to conduct surveys, or offer free samples (Funk, 2004).

The Austrian mobile phone market shows one of the largest penetration rates in Europe. In 2002, 6.8 million mobile phone users were registered (83.6%); in March 2004 penetration reached a level of 89.7% (Telekom Austria, 2004). GPRS and the Universal Mobile Telecommunications System (UMTS) (the European pendant to the Japanese

mobile Internet) were introduced in 2003 (Merrill Lynch, 2002). In March 2004, the number of Austrian GPRS users increased to 840.000 (Telekom Austria, 2004). The most popular non-voice-based service, however, is short message service (SMS), which is a part of the older Global System for Mobile Communications (GSM) standard. Basic SMS messaging, which counts for almost 10% of mobile telecommunications revenue, is not or to a very small extent related to mobile Internet. The frequent usage of SMS in Austria is mainly due to its usability, whereas other services like e-mail download and the usage of mobile Internet applications are considerably less applied. One major reason is consumers' lack of technology knowledge (Gutmann & Sochatzky, 2003).

Perceptions of M-Advertising in Japan and Austria

An important goal of any advertising activity is the achievement of certain reactions by the recipients. In order to get insights into how customers react to these campaigns, it is necessary to measure the mechanisms that drive consumers' reactions. Advertising research has revealed that the success of an advertising campaign strongly depends on how the customer reacts to a message. Effectiveness of advertising campaigns depends on numerous constructs; the most important ones are attitude toward advertising and attitude toward an advertising message (Gardner, 1985; Lutz, 1985; MacKenzie & Lutz, 1989; Moore & Hutchinson, 1983). On the basis of literature research in empirical results concerning Web-based advertising (Ducoffe, 1995, 1996) we derived four more variables picturing the effectiveness of m-advertising. These constructs are entertainment, informativeness, irritation, and credibility and will be discussed in the following.

Entertainment of M-Advertising

Feelings of enjoyment evoked by advertisements positively influence people's attitude toward the advertisement (Shavitt, Lowrey, & Haefner, 1998). Entertainment fulfills the consumers' needs for "escapism, diversion, aesthetic enjoyment or emotional release" (McQuail, 1983). It can be used to involve customers more deeply and make them more familiar with the advertised product or service (Lehmkuhl, 2003). In Japan, mobile communication providers have very strongly promoted mobile Internet as a means of entertainment for many years. Japanese consumers regard their mobile phones not as mere communication tools anymore, but as portable entertainment players (Haghirian, Dickinger, & Kohlbacher, 2004). In contrast, in Europe and the United States, mobile Internet-based services are chiefly positioned as a convenient service for business professionals (Funk, 2004). As Johansson & Nonaka (1996) point out, advertising in Japan is more fantasy oriented but less logic. Advertising messages are often implicit, intuitive, and rather emotional. This is also true for m-advertising. Consequently, we assume that Japanese perceive m-advertising messages as more entertaining than Austrians.

H1: Japanese perceive m-advertising as more entertaining than Austrians.

Informativeness of M-Advertising

Information is considered a very valuable issue in m-marketing because recipients react very positively to advertising transferring incentives (Varshney, 2003). Marketers generally want to convey information via advertising messages (Gordon & De Lima-Turner, 1997). When it comes to m-advertising, the consumers want the message's content to be tailored to their interests (Robins, 2003). They prefer messages that are relevant for them (Milne & Gordon, 1993). In contrast to Europeans, Japanese prefer information to flow freely (Hall & Hall, 1987). Information plays an important role in Japanese society. A larger quantity of information is collected and transmitted within the Japanese society than in

a Western society. Japanese are avid information gatherers, hence information exchanged refers to all kinds of data, including information that would not be relevant in Western countries (Johansson & Nonaka, 1996). Therefore we assume that m-advertising messages are considered a source of information to a higher degree by Japanese.

H2: Japanese perceive m-advertising as more informative than Austrians.

Irritation of M-Advertising

Advertisements might also evoke negative feelings. One important effect is irritation. If people feel indignity when being addressed by advertisements, their attitudes can be negatively influenced (Shavitt et al., 1998). A typical reaction is ignoring the message. Like any advertising message, m-advertising may provide an array of information that can confuse the recipient (Stewart & Pavlou, 2002). Moreover, as it is sent to a consumer's mobile phone, it can be perceived as an intrusion into his/her privacy. Many consumers are still uncomfortable with mobile business and are skeptical whether such business models are feasible and secure (Siau & Shen, 2003). M-advertising might affect users' feeling of being watched or recorded by organizations or other individuals (Rust, Kannan, & Peng, 2002). This leads to feelings of insecurity. Privacy concerns differ across cultures. Japanese are generally considered members of a collectivistic culture, where also information about individuals is frequently and openly shared. Hence, people share information that would be considered very private by Western standards (Hall & Hall, 1987). In contrast, Austrians belong to an individualistic culture where personal information is not freely distributed (Hall & Hall, 1987; Hofstede, 1980). Hence, we conclude that Austrians will be more easily irritated by m-advertising messages intruding into their lives than Japanese consumers are.

H3: Japanese perceive m-advertising as less irritating than Austrians.

Credibility of M-Advertising

Advertising credibility refers to "consumers' perception of the truthfulness and believability of advertising in general" (MacKenzie & Lutz, 1989, p. 51). An advertisement's credibility is particularly influenced by the company's credibility and the bearer of the message (Goldsmith, Lafferty, & Newell, 2000; Lafferty, Goldsmith, & Newell, 2002). In Japan, companies use social group allegiances to create value-added options for customers. They believe that the best way to perform advertising is to present a buyer who is satisfied with the product. Thus, they try to establish a mutual supportive relationship between buyer and seller (Johansson & Nonaka, 1996). This concept is strongly based on Japanese groupism and collectivistic features of Japanese society. Hence, Japanese are in general more trustful than their Western counterparts (Downes, Hemmasi, Graf, Kelley, & Huff, 2002). Companies they buy from are considered trustful partners. We thus conclude that Japanese perceive m-advertising messages as more credible than Austrians.

H4: Japanese perceive m-advertising as more credible than Austrians.

Perceived Advertising Value

Ducoffe (1995) argues that advertising value is a measure for advertising effectiveness and "may serve as an index of customer satisfaction with the *communication products* of organizations" (p. 1). The perceived value of advertising is "a subjective evaluation of the relative worth or utility of advertising to consumers" (Ducoffe, 1995, p. 1). Japanese retailers generally try to create value with their m-advertising messages, mainly because conveying service and product information to consumers readily and on time is a crucial aspect

of advertising in Japan (Schneidewind, 1998). M-advertising messages contain information about bargains and new products or carry incentives to increase customers' convenience. We thus conclude that Japanese perceive m-advertising as more valuable than Austrians.

H5: Japanese perceive m-advertising as more valuable than Austrians.

Attitude Toward M-Advertising

Attitudes are "mental states used by individuals to structure the way they perceive their environment and guide the way they respond to it" (Aaker, Kumar, & Day, 1995, p. 254). An attitude toward an advertisement is defined as consumers' "learned predisposition to respond in a consistently favorable or unfavorable manner toward advertising in general" (MacKenzie & Lutz, 1989, p. 54). As it is known from the *theory of reasoned action* (TRA) (Ajzen & Fishbein, 1980) and the *theory of planned behavior* (Ajzen, 1991), attitudes have a considerable impact on behavior (Churchill & Iacobucci, 2002). A major influencing factor on attitude toward an advertisement is the general attitude toward the advertising medium (Larkin, 1979). A positive attitude toward mobile phones also reflects on attitude toward m-advertising. In Japan, 45% of mobile consumers state that their mobile phone is essential in their lives (NTT Docomo, 2001). Mobile phones play an important role in Japanese everyday life. The consumers show an extraordinarily positive attitude toward their mobile phones (Haghirian et al., 2004). The situation in Europe is different. People use their mobile phones chiefly for communication and to a lesser extent for handling contents, and mobile phones are also an integrative part of everyday life.

H6: Japanese show a more positive attitude toward m-advertising than Austrians.

RESEARCH METHODOLOGY

In order to analyze differences in users' attitudes and the mentioned antecedents, we conducted an empirical survey in Japan and in Austria. The study focused on messaging–based, push mobile advertisements, such as SMS and MMS. We included only the owners and users of mobile phones in the survey. In order to reflect general differences in user perceptions, we carried out both surveys with undergraduate students. This was done in order to obtain homogenous samples concerning socio-demographic structure and in order to cover a very relevant target group of mobile phone users.

In Japan, data collection was conducted in summer 2004. The respondents were undergraduate business students of two different Japanese universities. Out of 450 questionnaires handed out, 420 were returned; 367 of them provided usable answers for this investigation. In Austria, data collection took place in fall 2003. In an Austrian university, 408 undergraduate business students were surveyed. Out of 550 questionnaires handed out, 448 were returned; 408 of them provided usable answers for the investigation. Table 1 provides an overview of the demographic distribution of the Japanese and Austrian respondents. As data shows, there are differences in the gender and age structure between the responding undergraduates in the two countries.

In the survey, a standardized questionnaire was developed in English and then translated into German and Japanese by native speakers. After a back-translation into English and a comparison of the two English versions, two pre-tests (Austria: 30 students, Japan: 35 students) were conducted and adaptations were integrated into the questionnaires.

The scales for informativeness, entertainment, irritation, and advertising value were derived from the Web-based advertising scales of Ducoffe (1996). The scale measuring attitude toward m-advertising was based on Alwitt and

Table 1. Demographic attributes of the investigation samples

	Austrian sample (n=408)		Japanese sample (n=367)	
Age of Respondent	**Female**	**Male**	**Female**	**Male**
18-20 years	6.4%	1.0%	16.3%	37.1%
21-25 years	39.0%	34.8%	13.0%	28.3%
older than 26 years	7.2%	11.6%	1.9%	3.3%
Total	**52.6%**	**47.4%**	**31.3%**	**68.7%**

Table 2. Cronbach alphas of scale items

Measures	**Scale Origin**	**Items**	**Alpha Japan**	**Alpha Austria**
			n=367	**n=408**
Entertainment	Ducoffe, 1996	6	.86	.84
Informativeness	Ducoffe, 1996; Lastovicka, 1983	7	.78	.88
Irritation	Lastovicka, 1983; Ducoffe, 1996	5	.62	.65
Credibility	MacKenzie and Lutz, 1989	4	.79	.77
Attitude toward advertising	Alwitt and Prabhaker, 1992	8	.76	.72
Perceived advertising value	Ducoffe, 1996	2	.83	.90

Prabhacker's (1994) scale measuring consumer attitudes toward TV ads. The credibility scale based on Mackenzie and Lutz's (1989) scale for measuring advertisement credibility. All measures were assessed via 5-point Likert-type scales ranging from "strongly agree" (1) to "strongly disagree" (5). Sample questions can be found in the appendix. Table 2 provides an overview of the reliabilities (Cronbach's alphas) of the investigated items. All variables, except the irritation scale in both samples, show alpha levels above .7.

The factor analysis was performed by main component analysis with Varimax rotation. Only factors with eigenvalues < 1 were further used.

STUDY RESULTS

The analysis of the hypotheses developed in the previous section was conducted via MANOVA tests. Table 3 summarizes the results of the comparative analysis of the Austrian and Japanese sample.

Entertainment of m-advertising is perceived more positively by Japanese than by Austrians, hence Hypothesis 1 is supported by the data (F=20.51). Like mobile phones in general, also advertisements received via them are considered a source of entertainment to a larger extent in Japan. Concerning the informativeness of m-advertising, the MANOVA results indicate no significant difference between Japanese and Austrian students, Hypothesis 2 is therefore rejected (F=.86). Although Japanese advertisers send a large amount of consumer-relevant information via m-advertising messages, the surveyed recipients do not perceive them as more informative. Hypothesis 3 indicates that Japanese are less irritated by m-advertising messages than Austrians. In this respect, Austrian and Japanese students differ significantly (F=132.2). But in contrast to Hypothesis 3, it is the Japanese students who perceive m-advertising

Table 3. Hypotheses tests via MANOVA

	F-Ratio	Country	Mean	Standard Deviation	Hypothesis
Entertainment (H1)	20.51**	Japan	3.96	.87	Supported
		Austria	4.2	.80	
Informativeness (H2)	.86	Japan	3.7	.91	Rejected
		Austria	3.7	.86	
Irritation (H3)	132.2**	Japan	2.1	.84	Rejected
		Austria	2.9	.92	
Credibility (H4)	.355	Japan	4.0	.80	Rejected
		Austria	3.9	.81	
Attitude toward m-advertising (H5)	50.62**	Japan	3.8	.83	Supported
		Austria	4.2	.88	
Advertising value of m-advertising (H6)	30.03**	Japan	3.7	.93	Supported
		Austria	4.1	.85	

**p<0.001, 1= Strongly Agree, 5= Strongly Disagree

more irritating than Austrian students. Therefore we reject Hypothesis 3. In the light of the high group orientation of Japanese (Hofstede, 1980), these results are rather unanticipated. One explanation could be a larger number of messages received by Japanese students that influences the degree of irritation. In the context of credibility, the analysis shows that Japanese students score slightly lower on this variable. But the difference is not significant, hence Hypothesis 4 is not supported either (F=.355).

Both dependent variables, advertising value, and attitude toward advertising are rated more positively by the Japanese respondents. Therefore we accept Hypothesis 5 (F=30.03) and Hypothesis 6 (F=50.62) according to the MANOVA analysis. M-advertising messages are obviously regarded as more valuable and are therefore being appreciated to a higher extent by Japanese students. This result is consistent with the observation obtained by Haghirian et al. (2004) who are stating that Japanese consumers generally perceive mobile phones and their impact on daily life as very positive.

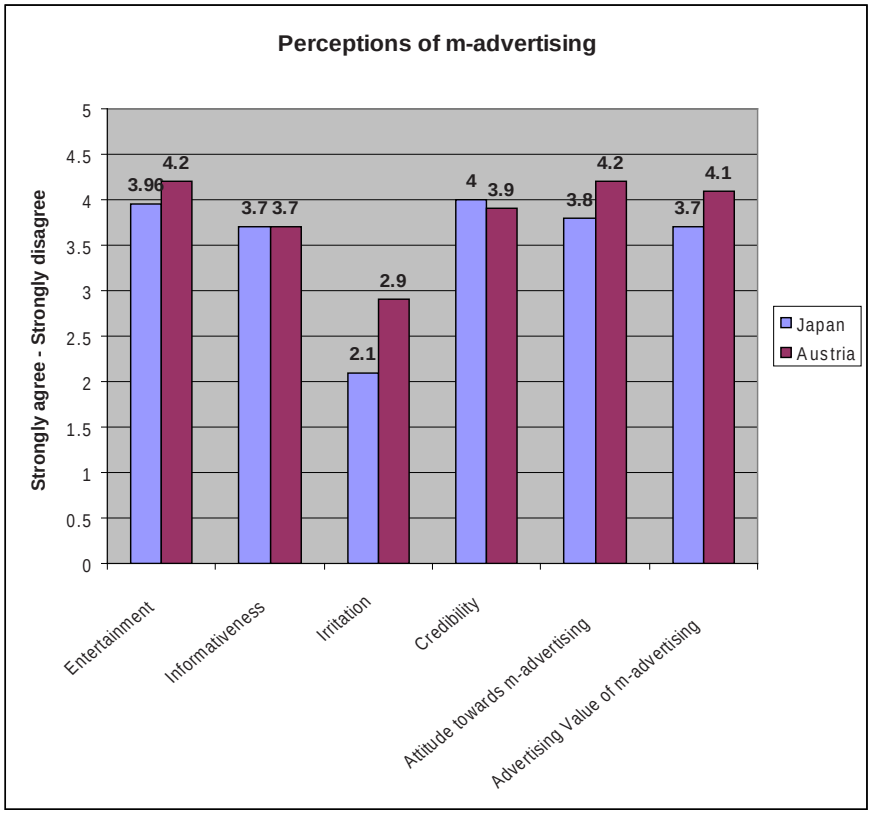
DISCUSSION

This investigation is a pioneer in analyzing cross-cultural differences in the perception of m-advertising. Figure 1 shows the differences between perceptions of m-advertising of the two respective markets.

In general, the study reveals that Japanese and Austrian students do not rate m-advertising very positvely. Looking at attitude toward m-advertising, it can be observed that Japanese students' display show a mean agreement of 3.8. The Austrian sample even shows a mean of 4.2, hence being located between the two most negative answer categories. The difference between these two means is highly significant (p = .01). Almost the same result is achieved for advertising value of m-advertising. Here, Japanese's mean value is 3.7, Austrians' mean value is 4.1. Again, this difference is highly significant at a .01 level. Obviously, m-advertising is not popular among students in both countries.

Having a closer look at the antecedents, we can see that also here the two groups have different

Figure 1. Perceptions of m-advertising in Japan and Austria (Means)



perceptions. There are significant differences in terms of entertainment and irritation. Japanese students regard m-advertising as more entertaining than Austrian students do. This finding might be consistent with Japanese strong orientation on emotions and entertainment what might explain their better attitude toward m-advertising. On the other hand, as cultural research has shown that Japanese are more liberal concerning privacy, their relatively negative perception of irritation might show a weaker influence on advertising value and attitude toward m-advertising than among the Austrians.

But there are probably additional differences that could explain the different perceptions between the two samples. They can be separated into technical issues, socio-demographic issues, and legal issues.

Technology

The two countries are in different stages of mobile technology adoption and diffusion. These differences occur both on the supply side and the demand side. Concerning the supply of m-advertising messages, in Japan, m-advertising is a popular and frequently used instrument for addressing consumers. Mobile Internet, which can largely contribute to a value added to m-advertising, is commonplace in Japan but not widespread in Austria. Hence, m-advertising is mainly carried out by SMS and does not offer any value that exceeds textual information. M-advertising plays a minor role in Austria.

From the demand side's point of view, user adoption of mobile technologies is extraordinarily high in Japan. Like in other Asian countries,

consumers are very technology oriented. They integrate the mobile Internet into their daily activities very intensively. For example, it is common to download information about train schedules or baseball results and navigate through urban Tokyo using a mobile Internet map. Users regard mobile Internet and its services as a very important and helpful asset and therefore regularly use it. In Austria, mobile technology adoption is at an early stage of diffusion. As long as technologies are still being introduced into a market, their benefits are often nebulous to consumers (Balasubramanian, Peterson, & Jarvenpaa, 2002). Many Austrian consumers are less educated on the functions and usability of mobile devices. This results in a lower willingness to use these services. In addition, many Austrian mobile phone users stick to their phones for several years. As a consequence, they have hardware devices that do not support advanced, Internet-based applications. These differences might have, in general, a strong influence on attitudes and perceived value of m-advertising.

Socio-Demographic Issues

The two samples differ from each other concerning the socio-demographic structure. In the Japanese sample, there are considerably more respondents younger than 20 years (53.4%) than in the Austrian sample (7.4%). In contrast, the largest portion of the Austrian respondents was the age group of 21 to 25 years (73.8% compared with 41.4% in Japan). Respondents older than 26 years accounted for 18.6% in the Austrian sample, but only 5.2% in the Japanese sample. In the light of the observation that young users are early technology adopters, these differences could be another explaining factor for the better attitudes and value perceptions of Japanese. In previous research, age turned out to show a moderating influence on consumers' acceptance of wireless finance services (Kleijnen, Wetzels, & De Ruyter, 2004). Similarly, the Japanese sample contains more male users, whereas the Austrian sample consists of more female respondents. Hence the effect of gender could also be observed, although this influence is considered to become less relevant with increasing technology adoption. But, as Nysveen, Pedersen, and Thorbjornsen (2005) show, gender affects the relevance of antecedents of the intention to use mobile chat services, hence there might also be significant differences in the respect of m-advertising.

Legal Issues

Finally, legal regulations form different conditions for m-advertising in both countries. In Europe, marketers must obtain consumers' explicit agreement before they may send advertising messages to personal communication media (the same is true for e-mail, fax, or telephone advertising). In contrast, Japanese companies need not rely on consumers' explicit agreement to provide them with m-advertising messages. The legal regulations can hinder fast distribution of m-advertising in Europe. Consequently, consumers are less used to m-advertising, which can negatively influence their perceptions of this way of advertising.

FUTURE TRENDS AND IMPLICATIONS FOR PRACTITIONERS

Low Overall Perception of M-Advertising

M-advertising is currently not a popular marketing instrument. Although it allows context-sensitive messages and a high level of customization, it obviously does not provide much added value to the surveyed students. As this effect is observed in two culturally different countries, this problem seems to be located at a higher level. It seems that being able to send customized advertising mes-

sages to consumers, which was and is considered one of the most profitable aspects of m-advertising, has turned out to be a major hindrance in consumers' acceptance of mobile devices as a carrier of product and service information. The following aspects could be obstacles to a more positive perception of m-advertising.

Privacy Concerns

Customer privacy has always been a critical issue in marketing. But it has experienced a greater significance in recent years with the rise of Internet-based commercial transactions (Rust et al., 2002). Most consumers are still quite uncomfortable with the concept of mobile business and they are skeptical whether these businesses are feasible and secure (Siau & Shen, 2003). Privacy issues are therefore very important when using mobile devices in addressing the consumers. Before receiving advertising messages via a mobile device, consumers need to empower a marketer to send promotional messages in certain interest categories to them. Typically, this is done by asking the consumer to fill out a survey indicating his or her interest when registering for a service. After that, the marketer can match advertising messages with the interests of the consumer (Krishnamurthy, 2001). Although this procedure of permission marketing is obligatory by law in European countries, the benefit and value of this approach should be made clear to the users. In order to convince users to subscribe to advertising newsletters, firms should clearly point at the benefits that are associated with such m-advertising campaigns. This implies certainly the necessity to provide real value added to the customers.

Mobile Phones as Private Items

Originally meant to connect the world of business, the mobile phone has been increasingly applied by private households and therewith entered the domestic sphere. Accordingly, the mobile phone has changed its identity: It has lost its internal coherence and its connotations of being a mobile technology (Fortunati, 2001). Consumers regard their mobile phone a very private item. Mobile technologies are considered "personal" technologies, attached to a particular body or person (Green, Harper, Murtagh, & Cooper, 2001). Consequently, individuals are very sensitive towards receiving messages from unknown persons or organizations. Data control by unknown individuals can easily lead to annoyance among receivers (Whitaker, 2001).

Frequency of Messages

The number of advertising messages received via mobile devices is an important factor influencing the advertising value for the consumer (Haghirian & Dickinger, 2004). Ducoffe (1995) states that informativeness and entertainment of the advertising information should decline with repetition, because the information will be learned by the audience and thereby lessening its value. As the quantity of promotional messages rises, the attitude of the individual towards the promotional vehicle also worsens and leads to tedium from consumers' point of view (Ha, 1996; Tellis, 1997).

Implications

The analysis shows that the Japanese sample rates m-advertising significantly better than the Austrian respondents. In Japan, where the adoption of mobile technology is by far more advanced and also the number of m-advertising messages received per customer is higher, the perception of m- advertising seems significantly better. This may lead to the conclusion that mobile devices as communication channels for marketers will become more popular once consumers become more familiar with the underlying technology.

Future Trends

If firms can derive valuable services from these future opportunities, m-commerce will become a significant part of the advertising industry. Cyriac Roeding, the European chair of the Mobile Marketing Association indicates that "as bandwidth increases, advertisers will have to be innovative in their campaigns to overcome the limitations of handsets with small screens" (DeZoyza, 2002).

However, m-advertising is currently still in its infancy. Both the supply side and the demand side of m-advertising are not yet ready for a broad usage of m-advertising. As the study in Japan shows, an advanced technological basis is not sufficient for a wide acceptance of m-advertising. Technology is an enabler but not a guarantee for a positive development of m-advertising. The future development of m-advertising should be regarded in a larger context, especially in the light of e-commerce. Unlike e-commerce, m-commerce did not experience the extreme phases of hype and disappointment in the early 21st century. This is due to the later development of m-commerce. Hence it is not surprising that many people are skeptical about another trend in the new economy after the disastrous failures in the dot-com world. But if conclusions from e-commerce development are drawn, one could also assume a more positive future in m-commerce. Like the e-commerce history shows, technology acceptance and the implementation of intelligent Internet-based business models requires some time. Firms have to develop business models that follow the classical rules of business governance, consumer behavior, and profit generation. The same will be true for m-commerce, which is just a part of e-commerce. With increasing convergence of media, the boundaries between different devices—computers, mobile devices, TV sets—will lose relevance.

Consumers will thus only accept this form of advertising, if they can get a concrete added value. One of the key potentials for m-commerce will be the advanced possibilities of technology usage (see the *M-Advertising and its Technological Dimensions* section). As the results of the study shows the design of the advertising message is by no way trivial and message characteristics need to be developed carefully. If companies decide to send out m-advertising messages, they should be both entertaining and informative. Marketers can not only rely on the fact that an advertising message sent via a mobile device will be read and remembered automatically. The mobile device may be an attention getter, but an attention getting device that is unrelated to the message will not attract consumers interested in the message or the product (Ogilvy, 1963).

CONCLUSION

Like any empirical investigation, this study has several limitations, which calls for further research in this area. These limitations address a lack of comparability of the investigated samples, cultural differences in interpreting the survey items, possible biases in response styles, and differences in socio-demographic respondent structure. From a technological point of view, the study's results cannot be generalized to m-commerce in total as only selected technologies and mobile devices have been considered. Hence, in order to gain deeper insights into attitudes and m-advertising value, other technologies, especially Internet-based approaches, should be considered. In addition, environmental conditions should be addressed in further research. Such issues are legal regulations, technology diffusion, and costs related to sending and receiving m-advertising messages.

Considering future research, this study offers several research avenues. First, empirical research should address the role of demographics in attitudes toward m-advertising. As earlier research in e-commerce and m-commerce has shown, there are considerable differences between men and women as well as across different age

groups. Second, the relevance of informativeness of a m-advertising message should be revisited. In contrast to Web advertising, m-advertising can provide valuable, time and location-oriented information for consumers. Future research must also clarify cross-cultural perceptions on informativeness of m-advertising. Another issue is the "evergreen" discussion of standardization vs. adaptation of global advertising activities. The impact on standardization or adaptation of global m-advertising needs to be further investigated in order to develop normative recommendations for advertisers and international marketing researchers.

REFERENCES

Aaker, D. A., Kumar, V., & Day, G. (1995*). Marketing research*. New York: Wiley.

Ajzen, I. (1991). The theory of planned behaviour. *Organizational Behavior & Human Decision Processes, 50*(2), 179-211.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice Hall.

Alwitt, L. F., & Prabhaker, P. R. (1994). Identifying who dislikes television advertising: Not by demographics alone. *Journal of Advertising Research, 34*(6), 17-29.

Balasubramanian, S., Peterson, R. A., & Jarvenpaa, S. L. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science, 30*(4), 348-361.

Brodie, R. J., Coviello, N. E., Brookes, R. W., & Little, V. (1997). Towards a paradigm shift in marketing? An examination of current marketing practices. *Journal of Marketing Management, 13*(5), 383-406.

Churchill, G. A. J., & Iacobucci, D. (2002). *Marketing research; Methodological foundations.* South-Mason, OH: Western Publishing.

Cyr, D., & Trevor-Smith, H. (2004). Localization of Web design: An empirical comparison of German, Japanese, and United States Web site characteristics. *Journal of the American Society for Information Science and Technology, 55*(13), 1199-1208.

Del Galdo, E., & Neilson, J. (1996). *International user interfaces.* New York: Wiley.

DeZoysa, S. (2002, February). Mobile advertising needs to get personal. *Telecommunications International.*

Downes, M., Hemmasi, M., Graf, L. A., Kelley, L., & Huff, L. (2002). The propensity to trust: A comparative study of United States and Japanese managers. *International Journal of Management, 19*(4), 614-621.

Ducoffe, R. H. (1995). How consumers assess the value of advertising. *Journal of Current Issues and Research in Advertising, 17*(1), 1-18.

Ducoffe, R. H. (1996). Advertising value and advertising on the Web. *Journal of Advertising Research, 36,* 21-36.

Dwyer, F. R., Schurr, P. H., & Oh, S. (1987). Developing buyer-seller relationships. *Journal of Marketing, 51*(2), 11-27.

Fortunati, L. (2001). The mobile phone: An identity on the move. *Personal and Ubiquitous Computing, 5*(2), 85-98.

Fournier, S., Dobscha, S., & Mick, D. G. (1998). Preventing the premature death of relationship marketing. *Harvard Business Review, 76*(11), 43-51.

Funk, J. L. (2004). Key technological trajectories and the expansion of mobile Internet applications. *Info—The Journal of Policy, Regulation and Strategy for Telecommunications, 6*(3), 208-215.

Gardner, M. P. (1985) Does attitude toward the ad affect brand attitude under a brand evaluation set? *Journal of Marketing Research, 22,* 192-198.

Goldsmith, R. E., Lafferty, B. A., & Newell, S. J. (2000). The impact of corporate credibility and celebrity credibility on consumer reaction to advertisements and brands. *Journal of Advertising, 29*(3), 43-54.

Gordon, M. E., & De Lima-Turner, K. (1997). Consumer attitudes towards Internet advertising. *International Marketing Review, 14*(5), 352-375.

Green, N., Harper, R. H. R., Murtagh, G., & Cooper, G. (2001) Configuring the mobile user: Sociological and industry views. *Personal and Ubiquitous Computing, 5*(2), 146-156.

Groenroos, C. (1994). From marketing mix to relationship marketing: Towards a paradigm shift in marketing. *Marketing Decision, 32*(2), 4-20.

Gummesson, E. (1987). The new marketing. Developing long-term interactive relationships. *Long Range Planning, 20*(4), 10-20.

Gutmann, A., & Sochatzky, C. (2003). *Mobile applications for teenagers.* Unpublished masters thesis, Vienna University of Economics and Business Administration, Austria.

Ha, L. (1996). Observations: Advertising clutter in consumer magazines: Dimensions and effects. *Journal of Advertising Research, 36*(4), 76-84.

Haghirian, P., & Dickinger, A. (2004). *Identifying success factors of mobile marketing.* ACR Asia-Pacific 2004, Association of Consumer Research, 28-29.

Haghirian, P., Dickinger, A., & Kohlbacher, F. (2004, November). Adopting innovative technology—A qualitative study among Japanese mobile consumers. In *Proceedings of the 5th International Working with e-Business Conference (WeB-2004),* Perth, Australia.

Hall, E. T., & Hall, M. R. (1987). *Hidden differences; Doing business with the Japanese.* New York: Anchor Books, Doubleday.

Hofstede, G. (1980). *Culture's consequences.* Beverly Hills, CA: Sage.

Johansson, J. K., & Nonaka, I. (1996). *Relentless—The Japanese way of marketing.* New York: Harper Business.

Kleijnen, M., Wetzels, M., & De Ruyter, K. (2004). Consumer acceptance of wireless finance. *Journal of Financial Services Marketing, 8*(3), 206-217.

Kotler, P. (2003). *Marketing management.* Upper Saddle River, NJ: Pearson Education.

Krishnamurthy, S. (2001). A comprehensive analysis of permission marketing. *Journal of Computer Mediated Communication, 6*(2). Retrieved from http://www.ascusc.org/jcmc/vol6/7issue2/krishnamurthy.html

Krishnamurthy, S. (2003). *E-Commerce management.* Mason, OH: Thomson Publishing.

Lafferty, B. A., Goldsmith, R. E., & Newell, S. J. (2002). The dual credibility model: The influence of corporate and endorser credibility on attitudes and purchase intentions. *Journal of Marketing Theory and Practice, 10*(3), 1-12.

Larkin, E. F. (1979). Consumer perceptions of the media and their advertising content. *Journal of Advertising, 8*(2), 5-48.

Lastovicka, J. L. (1983). Convergent and discriminant validity of television commercial rating scales. *Journal of Advertising, 12*(2), 14-23.

Lee, Y. E., & Benbasat, I. (2003). Interface design for mobile commerce. *Communications of the ACM, 46*(12), 49-52.

Lehmkuhl, F. (2003, January 6). Küsse und machotests. *FOCUS.*

Lutz, R. J. (1985). Affective and cognitive antecedents of attitude toward the ad: A conceptual framework. In L. F. Alwitt & A. A. Mitchell (Eds.), *Psychological processes and advertising effects: Theory, research and application* (pp. 54-63). Hillsdale, NJ: Lawrence Erlbaum Associates.

MacKenzie, S. B., & Lutz, R. L. (1989). An empirical examination of the structural antecedents of attitude toward the ad in an advertising pretesting context. *Journal of Marketing, 53,* 48-65.

Madlberger, M. (2004). *Electronic retailing.* Wiesbaden, Germany: Deutscher Universitätsverlag.

Marcus, A., & Could, E. W. (2000). Cultural dimensions and global Web user-interface design. *Interactions, 7*(4), 33-46.

McQuail, D. (1983). *Mass communication theory: An introduction.* London: Sage.

Merrill Lynch. (2002). *Wireless matrix—3Q02, quarterly update on global wireless industry metrics.* Author.

Milne, G., & Gordon, M. E. (1993). Direct mail privacy—Efficiency trade-offs within an implied social contract framework. *Journal of Public Policy & Marketing, 12*(2), 206-216.

Ministry of Public Management, Home Affairs, Post, and Telecommunications (MPHPT). (2003). *2003 white paper: Information and communications in Japan.* Retrieved September 29, 2003, from http://www.johotsusintokei.soumu.go.jp/whitepaper/eng/WP2003/2003-index.html

Moore, D. L., & Hutchinson, J. W. (1983). The effects of ad affect on advertising effectiveness. In R. P. Bagozzi & A. M. Tybout (Eds.), *Advances in consumer research* (Vol. 10, pp. 526-531). Ann Arbor, MI: Association for Consumer Research.

NTT DoCoMo. (2001). *Docomo report current trends in mobile phone usage among adolescents* (Company Report). Author.

Nysveen, H., Pedersen, P. E., & Thorbjornsen, H. (2005). Explaining intention to use mobile chat services: Moderating effects of gender. *Journal of Consumer Marketing, 22*(5), 247-256.

Ogilvy, D. (1963). *Confessions of an advertising man.* New York: Ballantine Books.

Okazaki, S. (2004). How do Japanese consumers perceive wireless ads? A multivariate analysis. *International Journal of Advertising, 23*(4), 429-454.

Parvatiyar, A., & Sheth, J. N. (2000). The domain and conceptual foundations of relationship marketing. In J. N. Sheth & A. Parvatiyar (Eds.), *Handbook of relationship marketing* (pp. 3-38). Thousand Oaks, CA: Sage.

Peppers, D., Rogers, M. & Dorf, B. (1999). Is your company ready for one-to-one marketing?, *Harvard Business Review, 77*(1), 151-160.

Petty, R. D. (2003). Wireless advertising messaging: Legal analysis and public policy issues. *Journal of Public Policy & Marketing, 22*(1), 71-82.

Robins, F. (2003). The marketing of 3G. *Marketing Intelligence & Planning, 21*(6), 370-378.

Rust, R. T., Kannan, P. K., & Peng, N. (2002). The customer economics of Internet privacy. *Journal of the Academy of Marketing Science, 30*(4), 455-464.

Schneidewind, D. (1998). *Markt und Marketing in Japan—Shin Hatsubai*. Munich, Germany: Verlag C. H. Beck.

Shavitt, S., Lowrey, P., & Haefner, J. (1998). Public attitudes towards advertising: More favourable than you might think. *Journal of Advertising Research, 38*(4), 7-22.

Siau, K., & Shen, Z. (2003). Building customer trust in mobile commerce. *Communications of the ACM, 46*(4), 91-94.

Stafford, T. F., & Gillenson, M. L. (2003). Mobile commerce: What it is and what it could be. *Communications of the ACM, 46*(12), 33-34.

Stewart, D. W., & Pavlou, P. A. (2002). From consumer response to active consumer: Measuring the effectiveness of interactive media. *Journal of the Academy of Marketing Science, 30*(4), 376-396.

Telekom Austria. (2004, March 29). [Press release]. Mobilkom Austria Group.

Tellis, G. J. (1997). Effective frequency: One exposure or three factors? *Journal of Advertising Research, 37*(4), 75-80.

Varshney, U. (2003). Location management for mobile commerce applications in wireless Internet environment. *ACM Transactions on Internet Technology, 3*(3), 236-255.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7,* 185-198.

Venkatesh, V., Ramesh, V., & Massey, A. P. (2003). Understanding usability in mobile marketing. *Communications of the ACM, 46*(12), 53-56.

Whitaker, L. (2001). Ads unplugged. *American Demographics, 23*(6), 30-34.

ENDNOTE

[1] Mobile Internet can be understood as free access to the Internet via mobile devices by means of mobile telecom operators or wireless devices, but it can also denote limited Internet access that is restricted to selected Web sites supported by the mobile telecom operator. In the paper at hand, mobile Internet is related to free Internet access to any Web sites.

APPENDIX 1. SAMPLE QUESTIONNAIRE ITEMS

Message content
Informativeness: Advertising on the mobile Internet is a good source of information.
Entertainment: Advertising on the mobile Internet is entertaining.
Irritation: I do not always understand advertising on the mobile Internet.
Credibility: Advertising on the mobile Internet is believable.
Perceived value of m-advertising
Advertising on the mobile Internet is important.
Attitude toward m-advertising
In general, advertising on the mobile Internet presents a true picture of the product advertised.

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, and J. Day, pp. 215-232, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.14
Do Mobile CRM Services Appeal to Loyalty Program Customers?

Veronica Liljander
Swedish School of Economics and Business Administration, Finland

Pia Polsa
Swedish School of Economics and Business Administration, Finland

Kim Forsberg
Intrum Justitia Finland, Finland

ABSTRACT

Not until very recently has mobile phone technology become sophisticated enough to allow more complex customized programs, which enable companies to offer new services to customers as part of customer relationship management (CRM) programs. In order to enhance customer relationships and to be adopted by customers, new mobile services need to be perceived as valuable additions to existing services. The purpose of this study was to investigate the appeal of new mobile CRM services to airline customers. An empirical study was conducted among loyalty program customers (frequent flyers) of an airline that was considering using MIDlet applications in order to add new mobile services to enhance customer relationships. The results show that customers do not yet seem to be ready to fully embrace new mobile applications. Although the services appeared to slightly improve customers' image of the airline, the services did not seem to enhance their loyalty towards it. However, customers who already used sophisticated mobile services, such as the Mobile Internet, had a significantly more positive attitude towards the proposed services. Thus the success of mobile CRM seems closely linked with customers' readiness to use existing mobile services. Before engaging in costly new investments, companies need to take this factor into serious consideration.

INTRODUCTION

During the last two decades the marketing community has witnessed a transfer from transaction-based marketing strategies to an emphasis on creating interactive relationships between the company and its customers (cf. Grönroos, 2000). With the overall aim of increasing customer retention and managing customer relationships for profit, CRM has become an essential part of many companies' marketing strategies. One of the newest tools to improve individual services to customers is mobile technology. Because of the rapid development in mobile technologies, it has recently become a noteworthy tool in CRM strategies, and therefore marketing strategies will need to be developed to suit this new channel (Akhgar, Siddiqi, Foster, Siddiqi, & Akhgar, 2002; Balasubramanian, Peterson, & Jarvenpaa, 2002; Helenius & Liljander, 2005). However, so far little is known about how companies intend to incorporate mobile technologies into CRM and about the effects it will have on customer retention (Crosby & Johnson, 2001; Okazaki, 2005). The mobile channel will be of particular interest to companies that already have a loyal customer base that has trusted the company with personal information. This is the case in customer loyalty programs, which have been shown to positively affect customer retention and customer share development (Verhoef, 2003). Loyalty programs already use online services for loyalty program details, such as customer services for members and information on accumulated benefits (Lam & Chan, 2003).

One new software solution for customized relationship programs is known as Mobile Information Device Profile (MIDP). Programs subscribing to this standard are called MIDlets and are coded in Java, which by the end of 2007 will be included in most mobile devices in Western Europe (Riivari, 2005). The easy-to-use universal nature of MIDlet applications offers both corporate programmers and individual end users a convenient way to create their own mobile programs to serve company- and user-specific needs.

Given the scarcity of empirical research on mobile CRM and the availability of new applications, the purpose of our study is to investigate how mobile CRM services, developed for a MIDlet application, are perceived by the loyalty program customers of an airline. More specifically, we study the appeal of the proposed mobile services to customers, their intention to adopt the services, and whether the services would improve customers' perceived image of the airline and enhance customer loyalty. The goal of CRM is to build a competitive advantage that distinguishes the brand from competitors and creates stronger customer loyalty (Crosby & Johnson, 2001). Since few studies have combined CRM and mobile services research (notable exceptions being Lin & Wang, 2006; Mort & Drennan, 2005), the current study contributes to the extant mobile service literature by offering a CRM perspective on mobile commerce and by investigating customers' attitudes towards mobile CRM.

The paper is structured as follows. First, the concept of mobile CRM and its benefits to customers are discussed. Second, the empirical study is introduced and the results are presented in the form of descriptive statistics. The paper ends with a discussion of the results, limitations, suggestions for future research directions, and managerial implications.

MOBILE CRM

Relationship marketing and CRM are frequently used interchangeably, but equally often CRM refers to a company's technology solutions for managing relationships, such as direct mail, loyalty cards, and e-commerce (Payne & Frow, 2005; Verhoef, 2003). A common conceptualization of CRM is still lacking: it has been described as a process, strategy, philosophy, capability, and technology (Zablah, Bellenger, & Johnston,

2004). Thus CRM can be viewed in a broad or a narrow sense, as a holistic approach to managing relationships, or the implementation of a specific technology solution project (Payne & Frow, 2005). CRM is clearly more than a technology, but in practice it is often associated with the use of databases and technological applications (Payne & Frow, 2005; Shah & Murtaza, 2005). A distinction is often made between operational, analytical, and collaborative CRM applications (Crosby & Johnson, 2001). Operational e-CRM includes customer service applications (Fjermestad & Romano, 2003), which is the focus of the present study. Thus our study employs a narrow CRM approach, by investigating the potential positive consequences for a company of implementing mobile CRM.

We define mobile CRM as *customer relationship management of any kind including interactive communication between an organization and a customer using a mobile device* (cf. Helenius & Liljander, 2005; Lam & Chan, 2003; Mort & Drennan, 2005). The special characteristics of mobile CRM in contrast to CRM in general are its temporal and spatial autonomy. Mobile devices include a large number of wireless mobile communication tools, such as regular cell phones, smart phones, pagers, PDA's, and notebooks, the most common device being some sort of cell phone with more or less sophisticated data transmission capabilities. One of the technological solutions enhancing mobile CRM that is applicable to cell phone devices is MIDlets—the technological solution investigated in our study.

MIDlet Applications as a Gateway to Mobile CRM

The rapid technical progress has led to new ways of processing data and of serving the mobile consumer. Our chosen example is MIDlet applications (Adjari, 2001), which bring the mathematical and information processing functions of a small computer into a mobile phone. Through their mobile phones customers can manage information and launch applications in the same way as when using the fixed Internet. Applications are invisible to customers, who only evaluate the services as they are offered through the application. To use a company's services through a MIDlet application, customers need to download it to a mobile device. Applications can be provided for free by a company as part of its CRM program, or they can be offered to customers as a value-added component at a price.

The success of a mobile CRM strategy depends on how well the application is designed, the design of the interface and services, as well as customers' evaluation of the service content in relation to any additional costs of using it. Although in the past consumers have felt cautious about using mobile services (Anckar & D'Incau, 2002), in the future mobile applications are expected to have an important impact on customer acquisition and retention, by offering additional services and benefits to customers (Kannan, Chang, & Whinston, 2001; Riivari, 2005; Varshney & Vetter, 2001). We will next discuss the benefits of mobile CRM using MIDlet applications in the context of frequent flyer customers.

Benefits of the Mobile Channel to Customers

The perceived relative advantage of a new technology such as added benefits in comparison to other service modes is essential for customer adoption (cf. Walker, Craig-Lees, Hecker, & Francis, 2002). Several benefits have been mentioned in relation to mobile technologies. Often cited as the main characteristic and added value to customers of mobile services is the possibility of accessing services whenever and wherever required (Heinonen, 2004, 2006; Sugai, 2005; Turban, King, Lee, Warkentin, & Chung, 2002). Mobile value arises in particular from spontaneous and immediate service needs (Anckar & D'Incau, 2002; Pura, 2005). In CRM the mobile channel can be

used to actively communicate with customers wherever they are, offering them access to the same services as through the fixed Internet or through personal contact.

Another advantage often mentioned is that companies can provide location-specific information and service to customers (Jukic, Sharma, Jukic, & Parameswaran, 2002; Turban et al, 2002; Wang & Cheung, 2004), for example, informing customers of the nearest physical touch point for the company's services.

The quality and usefulness of mobile services have received less attention than time and place benefits but are important for customer satisfaction and loyalty (cf. Chae, Kim, Kim, & Ryu, 2002; Nordman & Liljander, 2004). There are few studies on the relationship between e-CRM features and customer service evaluations (Feinberg & Kadam, 2002). However, research has shown that mobile services are evaluated on similar dimensions as e-services, while taking into account the limitations of the technology (Chae et al., 2002; Lin & Wang, 2006; Nordman & Liljander, 2004). Our study includes customer evaluations of mobile service content and usability; comfort and security; and mobile feedback services.

Service content and usability (SCU) can be viewed as intangible benefits (Money, Tromp, & Wegner, 1988) or as mobile life quality enhancers (Mort & Drennan, 2005). They are of particular importance for customer satisfaction with utility services (Chae et al., 2002), such as airline travel and frequent flier services. Mobile CRM could offer completely new services to customers, such as entertainment services or enhancements of existing offerings by adding a new wireless dimension to them. One example would be improving the usability of the main product, for example, by offering updated flight information to air travelers. Until the launch of new application technologies such as MIDlets, such opportunities and intangible benefits for enhancing customer relationships have not been widely available. Our study examines SCU by investigating customer perceptions of the content and usability of services such as access to flight schedule and route information, special offers, booking, payment, and check-in over a mobile phone.

Regrettably, digital fraud is becoming increasingly widespread, and customers' feelings of insecurity or discomfort may outweigh the benefits they expect to gain by embracing new applications (Kaapu, 2005; Kindberg, Sellen & Geelhoed, 2004; Walker et al., 2002). Therefore, we also investigate customer perceptions of the *comfort and security* of mobile service usage.

Customer mobile *feedback* (m-feedback) is a key component of an e-CRM strategy (Cho, Im, Hiltz, & Fjermestad, 2002). It is important that companies have effective channels for customer feedback and procedures to resolve complaints, recover customers, and reduce switching (Fornell & Wernerfelt, 1987; Johnston & Mehra, 2002). Technological interfaces are important channels for customer complaints and quick service recoveries (Bitner, Brown, & Meuter, 2000). Such services are an important feature of customer relationship programs (Winer, 2001), and mobile CRM could provide one feedback channel. M-feedback can be used for suggesting ideas for service improvements, as well as for giving compliments or voicing complaints. The mobile channel could offer quick resolutions to problems, taking full advantage of mobility.

Benefits to the Firm

Offering mobile CRM applications to customers should have positive consequences for the firm. One such consequence is increased customer loyalty (Fjermestad & Romano, 2003). Another important consequence is the positive effects that it may have on the image of a brand and the company (Helenius & Liljander 2005; Lam & Chan, 2003; Nysveen, Pedersen, Thorbjørnsen, & Berthon, 2005). A CRM strategy must lead to a more distinct brand and to higher customer loyalty (Crosby & Johnson, 2001). Contacts with

customers through different channels can add to or detract from their feelings of loyalty towards the company (Shankar, Smith, & Rangaswamy, 2002). Therefore, the brand assets (cf. Aaker, 1996; Aaker & Joachimsthaler, 2000) *image and loyalty* were assessed in our study.

METHOD

To investigate the appeal of a new CRM technology solution to customers, that is, mobile services offered through a MIDlet application, a mail survey was constructed. Members of a Nordic airline's frequent flyer loyalty program were sampled for the study. The airline is one of the world's oldest operating airlines, with a turnover of 1,698 million euros in 2004. Among airlines, it has been at the forefront of electronic service development, and frequent flyers are offered self-services on the Internet (e.g., check-in), at the airport (e.g., electronic gates), and when on the move (e.g., an SMS(Short Message Service)-based mobile check-in service). The airline's frequent flyers are always among the first to be offered new technology services. Previously published data on these customers show that their technology readiness is comparatively high (Liljander, Gillberg, Gummerus, & van Riel, 2006). Thus they form an attractive segment for mobile CRM. To maintain its technology advantage, the airline is planning to offer new mobile services to its loyalty program customers. The services are designed with MIDlet applications. Our study was conducted to investigate the appeal of the proposed services to loyal customers.

Questionnaire Design

Background data were gathered on gender, age, loyalty program level (here called: bronze, silver, gold, platinum), and customers' current use of the airline's electronic services. To check for the suitability of their mobile devices for MIDlet applications, and their readiness to use such services, customers were asked how often they use mobile phone e-mail, Internet browsing, and WAP(wireless application protocol) services, and whether their mobile phone supports Java applications. A "do not know" alternative was offered for the last question, since many consumers are unaware of all available features in their mobile phone.

SCU was measured by asking the respondents to imagine that the airline offered a mobile phone-based application that made it possible to look up schedules, check frequent flyer information, book flights, and perform check-in at the airport through their mobile phone. The following scale and items were used (7-point Likert scales, strongly disagree-strongly agree):

"I would have significant use for":

- mobile phone-based flight schedule and route information (SCU1)
- frequent flyer-information and special offers (SCU2)
- flight booking and payment (SCU3)
- check-in services (SCU4)

An alternative to the formulation of this question would have been to use a perceived usefulness scale from the technology acceptance model (TAM) as a basis for our study (cf., Davis, 1989; Featherman & Pavlou, 2003; Venkatesh & Davis, 2000; Venkatesh, Morris, Davis, & Davis, 2003). However, our formulation of the SCU statements was chosen because it corresponded best with the local language. SCU taps into similar issues as the perceived usefulness component of TAM, in that it asks customers to evaluate if the services would be of use to them. Another reason for not using TAM for our research was that, since the services do not yet exist, customers would have been unable to evaluate their ease of use, which is an important TAM component.

Four statements relating to *comfortable and secure use* of mobile services were included (7-point Likert scales). It should be noted that "comfortable use" in the local language includes connotations of "free of effort," which has been considered important for technology acceptance (Davis, 1989, p. 320).

- I would feel comfortable booking my flight through a mobile phone service (Comfort1)
- I would feel secure booking my flight through a mobile phone service (Secure1)
- I would feel comfortable paying for a flight through a mobile phone service (Comfort2)
- I would feel secure paying for a flight through a mobile phone service (Secure2)

Mobile feedback was evaluated on the following items (7-point Likert scales):

- I would feel comfortable giving feedback through a mobile phone service (MFB1)
- I believe that the airline would handle mobile feedback in the same way as conventional feedback (MFB2)
- Mobile feedback would make it easier for me to contact the airline (MFB3)
- Mobile feedback could help the airline better solve my problems (MFB4)
- I would use the mobile feedback service regularly (MFB5)
- I would give mobile feedback in instances I otherwise would not (MFB6)

As timeliness is one of the key features of mobile feedback, customers were also asked:

- How quickly they believed that they would send mobile feedback (MFBSend)
- How quickly they expected to receive an answer (MFBReceive)

The response alternatives to MFBsend were: a) Immediately after a negative incident, b) Some time later when I sit down, c) Within the same time span as I would give regular feedback, and d) I do not think I would give mobile feedback at all.

Improved benefits to the firm, in the form of increased brand assets, were measured with the following statements:

- Mobile services would make the airline more desirable as an airline carrier (Image1)
- Mobile services would improve my picture of the airline as an airline carrier (Image2)
- Mobile services would distinguish the airline from other airlines (Image3)
- Mobile services are associated with a modern and technologically up-to-date company (Image4)
- Mobile services could be a key factor that keeps me from changing to another airline (Loyalty)

In addition, customers' willingness to pay for new services was asked for regarding mobile feedback (WillPayFeedback) and flight booking (WillPayFlight). They were also reminded of the fact that the price of phoning the call centre was 1.64 euros per call. The alternatives given for both questions were: a) nothing, b) the price of an SMS message, c) 2€, d) 5€, e) 10€ or more.

Customers' intentions to use the services (Adopt), were captured with one question: If this application were available, I believe I would: a) Begin using it instantly, b) Wait until I hear from other people who have used it, c) Wait until it becomes the standard way of using the air carrier's services, d) Probably never use it. Since the service did not yet exist, only adoption intentions could be measured. This is a common problem in many technology adoption studies (e.g., Anckar & D'Incau, 2002; Featherman & Pavlou, 2003; Plouffe, Vandenbosch, & Hulland, 2001). However, since we have collected infor-

mation on customers' adoption of other mobile services, we had data on their actual mobile service adoption. These data were used to explore differences between more experienced and less experienced mobile service customers. Research on technology adoption covers descriptions of adopter characteristics (e.g., Okazaki, 2006) but to a lesser extent differences between perceptions of technological applications in different adopter groups. For example, Anckar and D'Incau (2002) found significant differences in intentions to use mobile services between adopters and non-adopters of the Internet. Thus, we expected the experienced mobile service customers to evaluate the proposed services more positively than the less experienced customers.

Sample

Stratified sampling was used to include customers from all frequent flyer levels, representing a variety of customer loyalty to the company. Since there are fewer customers on the higher levels, a normal probability sampling procedure would have yielded a disproportionately high number of bronze members, many of whom fly infrequently and thus would not be the prime beneficiaries of the proposed services.

The survey was posted in an official airline-branded envelope, together with an introductory letter and a prepaid return envelope, to 262 frequent flyers, including 70 Bronze, 70 Silver, 70 Gold members, and all the Platinum members (52). The total response rate was 42%, yielding 104 completed questionnaires. In addition to the completed responses, nine were returned uncompleted. One questionnaire was discarded as incomplete, two were returned blank because the respondents were not proficient in the local language, and six envelopes were returned because of change of address.

The response rates for frequent flyer levels were: Bronze (32.9%), Silver (40%), Gold (45.7%), and Platinum (40.4%). There may be several reasons for a higher response rate among the more frequent flyers among loyalty card members. One reason could be that people who travel often are more likely to have sophisticated phones, with which they can access e-mail while being away from work. Another plausible reason is that customers who have reached a higher level within the loyalty program feel a greater attachment to the airline and thus are more inclined to respond to the survey.

Answers to the background questions revealed that 78.8% of all respondents were male, which is representative of the total sample that received the survey. Male customers are overrepresented on all loyalty program levels, except the Bronze level. The age distribution among survey participants was 18-25 years (1.9%), 26-35 (13.5%), 36-50 (46.2%), 51-65 (37.5%), and 66+ (1%). These figures correspond with previous studies of the firm's frequent flyers and suggest that the age distribution is representative of the airline's loyalty program clientele.

RESULTS

Customer Readiness to use Mobile Services

When new services and technologies emerge, customer adoption is often slower than expected by companies (Gilbert & Han, 2005). For example, customer adoption of self-service check-in automats at airports has been slow, as has been the adoption of electronic check-in (Liljander et al., 2006). However, the customers who responded to the present survey appear to be at the forefront of mobile service adoption. More than half of the respondents (53.8%) used the mobile Internet daily, weekly, or monthly, whereas only 26% had never used it, or had only tried it (20.2%). There was no relationship between the loyalty program level and the use of mobile Internet services (Chi-Square=5.049, p= 0.168).

In addition, Chi-square tests showed that there was no relationship between gender and mobile Internet adoption (p=0.258), but that there was a relationship between adoption and age (p=0.025). Not surprisingly, but contrary to insignificant findings in other mobile service contexts (Mort & Drennan, 2005), older customers (51-65, 66+) had adopted sophisticated mobile services to a lesser extent than younger customers.

Customers are not necessarily aware of what applications they use to access services, and thus they may possess Java-supporting phones without being aware of this. Among the respondents only 43.3% were confident that their phone supports Java, 22.1% said that it did not, and 33.7% did not know. Thus a fairly large percentage of loyalty program customers have the necessary equipment to access and receive new services, but the majority showed the need to either update their phones or receive help in recognizing and using inherent mobile features. The results are presented in Table 1.

Next, the attractiveness of the proposed services, as well as their impact on image and loyalty will be presented. The respondent data were divided into two groups, those who used the mobile Internet daily, weekly, or monthly (mobile Internet adopters) and those who never used it, or who had only tried it (mobile Internet non-adopters). As previously mentioned, the first group was expected to evaluate the services more highly than the second group.

Mobile Service Evaluation

Table 2 presents the mean result for customer evaluations of SCU; comfort and security; m-feedback; and improvement of brand assets. The results for the total sample show a neutral attitude towards the proposed mobile services, with means close to the middle value of the scale (4). T-tests were performed to investigate differences in means between adopters and non-adopters of the mobile Internet. Since mobile Internet adopters were expected to exhibit higher scores than non-adopters, one-tailed *t* tests are reported. As expected, customers who already use more sophisticated mobile services found the offered services significantly more attractive in terms of SCU, comfort, and security.

Of particular interest from a CRM perspective is the finding that frequent flyers evaluated frequent flyer information (SCU2) as the least interesting service. This result requires further investigation within the company to reveal the reasons for it. One reason may be that customers cannot imagine what kind of information could be communicated on the small screen, and what the benefits would be. Pairwise *t* tests revealed that the mean for customers' perceived use of check-in mobile services (SCU4) was significantly higher ($p<0.01$, two-tailed) than the means of other proposed services. One explanation is that check-in via various technological devices is becoming increasingly familiar to airline customers. Thus, familiarity with performing these services by using other technologies may have a positive effect on consumer interest in performing them also with their mobile phone. In addition, paired-samples *t* tests showed that customers felt significantly ($p<0.01$, two-tailed) more comfortable and secure booking (Comfort1 and Secure1) than paying for flights (Comfort2 and Secure2) with their mobile phone. This was the case in all customer groups (complete sample, adopters and non-adopters).

M-feedback would be a novel service, offering customers the possibility of immediate feedback to the company through a device that they always carry with them. Even though customers believed that mobile feedback would be handled in the same way as other feedback (MFB2 $\underline{M} = 5.40$), they expressed only a lukewarm interest in the service. Means of MFB1 and MFB3-6 ranged from 3.46 to 4.35 for non-adopters, and from 3.98 to 4.88 for adopters. Only the difference in the means of MFB1 and MFB5 was significant between adopters and non-adopters, showing that adopters would be more comfortable using

Table 1. Mobile Internet use and awareness of JAVA support

					Percentages
Use of mobile internet	Bronze	Silver	Gold	Platinum	Total
	N=23	N=28	N=32	N=21	N=104
Daily	17.4	28.6	46.9	47.6	35.6
Weekly	17.4	10.7	6.3	4.8	9.6
Monthly	4.3	10.7	15.6	0	8.7
Have tried a couple of times	17.4	25.0	18.8	19.0	20.2
Have never used	43.5	25.0	12.5	28.6	26.0

	Gender		Age				
Use of mobile internet	M	F	18-25	26-35	36-50	51-65	66+
	N=82	N=22	N=2	N=14	N=48	N=39	N=1
Daily	37.8	27.3	0	42.9	43.8	25.6	0
Weekly	8.5	13.6	0	0	16,7	5.1	0
Monthly	9.8	4.5	0	28.6	4.2	7.7	0
Have tried a couple of times	22.0	13.6	50.0	14.3	25.0	15.4	0
Have never used	22.0	40.9	50.0	14.3	10.4	46.2	100.0

Awareness of JAVA support in respondents' personal mobile phone	Total
Mobile phone has JAVA support	43.3
Mobile phone has no JAVA support	22.1
Do not know	33.7

Table 2. Item means for mobile Internet adopters, non-adopters, and the total sample

Components (7-point scales)	Mobile Internet Adopters N=48	Mobile Internet Non-Adopters N=56	*t* test p-value[1]	Total N=104	SD
Service content and usability					
Mobile phone-based flight schedule and route information SCU(1)	4.80	3.81	0.004	4.35	1.945
Frequent flyer-information and special offers SCU(2)	4.18	3.23	0.002	3.74	1.712
Flight booking and payment SCU(3)	4.68	3.77	0.008	4.26	1.926
Check-in services SCU(4)	5.84	4.71	0.000	5.32	1.541
Comfort and security					
I would feel comfortable booking my flight through a mobile phone service Comfort(1)	4.85	3.73	0.001	4.33	1.839
I would feel secure booking my flight through a mobile phone service Secure(1)	5.24	4.17	0.000	4.74	1.754
I would feel comfortable paying for a flight through a mobile phone service Comfort (2)	4.25	3.29	0.006	3.81	1.986
I would feel secure paying for a flight through a mobile phone service Secure(2)	4.40	3.64	0.014	4.05	1.793
Mobile feedback (MFB)					
I would feel comfortable giving feedback through a mobile phone service MFB(1)	4.71	3.98	0.026	4.37	1.927
I believe that the airline would handle mobile feedback in the same way as conventional feedback MFB(2)	5.41	5.40	0.430	5.40	1.523
Mobile feedback would make it easier for me to contact the airline MFB(3)	4.88	4.35	0.072	4.63	1.790
Mobile feedback could help the airline better solve my problems MFB(4)	3.98	3.75	0.258	3.88	1.810
I would use the mobile feedback service regularly MFB(5)	4.14	3.46	0.024	3.83	1.765
I would give mobile feedback in instances I otherwise would not MFB(6)	4.67	4.35	0.203	4.52	1.887
Brand assets					
Mobile services would make the airline more desirable as an airline carrier Image(1)	4.52	3.92	0.033	4.24	1.726
Mobile services would improve my picture of the airline as an airline carrier Image(2)	4.80	4.23	0.041	4.54	1.683
Mobile services would distinguish the airline from other airlines Image(3)	4.86	4.35	0.061	4.63	1.656
Mobile services are associated with a modern and technologically up-to-date company Image(4)	5.45	4.96	0.060	5.22	1.595
Mobile services could be a key factor that keeps me from changing to another airline Loyalty	3.34	2.44	0.005	2.92	1.810

[1] *t tests between adopters and non-adopters, one-tailed significance reported*

the mobile phone for feedback (MFB1) and that they would use it more regularly (MFB5). However, the low means overall (adopters $\underline{M}$= 4.14, non-adopters $\underline{M}$ = 3.46) for MFB5 suggests that most clients would hesitate in making mobile feedback their primary communication channel with the company.

Regarding m-CRM benefits to the firm in the form of improved brand assets, there were no significant differences between adopters and non-adopters (Table 2). According to the mean values, offering mobile services might improve the image only slightly. In particular Image4 (adopters $\underline{M}$ = 5.45 and non-adopters $\underline{M}$ = 4.96) showed that the airline with mobile services would be perceived as a modern and technologically up-to-date company. However, customers' responses to loyalty ($\underline{M}$ = 3.34 and $\underline{M}$ = 2.44) demonstrated that mobile services would probably not be a key factor in keeping customers from switching airlines. This mean score is the lowest in comparison with all other statements. Thus, the conclusion must be that customers do not expect the mobile services to be a bonding factor in their relationship with the company. They might be perceived as nice additions to existing services, but not as a relationship strengthening factor.

Speed of Mobile Feedback, Willingness to Pay and Intentions to Use

One of the key features of mobile feedback is its potential speed both in sending and in receiving feedback (MFBSend and MFBReceive). Only 49% said that they would send feedback immediately when they had experienced a problem, while the rest would do it later; 42.2% expected to get an answer immediately, or within 2 hours, while the rest expected to get it in one day or later. The results are presented in Table 3. Since quick handling of mobile feedback would require extra resources and thus added costs, customers were asked if they would be willing to pay for the mobile feedback service. Not surprisingly, the majority of customers were prepared to pay either nothing (31.4%), or the price of an SMS (54.9). Only a small percentage of customers (13.7%) were prepared to pay 2€ or more for the service. Similar results were obtained for customers' willingness to pay for flight booking services (WillPayFlight). Only 9.7% were prepared to pay 2€ or more for the services, while the others were prepared to pay nothing (38.8%) or the price of an SMS (51.5%). This result is in line with earlier findings on customer willingness to pay for mobile services (Jarvenpaa, Lang, Takeda, & Tuunanen, 2003)

Customers were also asked how soon they believed that they would start using these services if they were offered (StartUse, Table 3). The answers revealed that 35% would begin to use them immediately, while 48.6% would wait until more people had adopted the service, and 16.5% said that they would probably never use them.

DISCUSSION

A key finding of the study is that customers do not yet seem to be ready to fully embrace mobile services as part of an airline's relationship marketing program. Their attitudes towards the proposed services can be described as "wait and see" and "let others use it first." This is a typical consumer response to many innovations, and it does not in itself mean that they would not adopt any of the services, if they were available. Resistance to innovations is an instinctive response in many consumers, which is due to functional and psychological barriers (Ram & Sheth, 1989). So far, consumers have not embraced mobile commerce to the extent that was predicted at the beginning of this century (Anckar & D'Incau, 2002; Nordman & Liljander, 2004). However, consumers have expressed a higher interest in utility than in entertainment services (Anckar & D'Incau, 2002), which seems promising also for m-CRM programs. Our study showed that customers were most interested in utility mobile

Table 3. Quickness of m-feedback, willingness to pay and intention to use the services

Percentages			
MFBSend	Total	MFBReceive	Total
Immediately	49.0	Immediately	21.6
Sometime later	16.7	Within couple of hours	20.6
Same time frame as conventional feedback	22.5	The same day	24.5
Not at all	11.8	In due time	33.3

WillPay Feedback	Total
2 € or more	13.7
The price of SMS	54.9
Nothing	31.4

WillPay Flight	Total
2 € or more	9.7
The price of SMS	51.5
Nothing	38.8

StartUse	Total
Immediately	35.0
Wait until more people have adopted the services	48.6
Probably never	16.5

services that they were likely to have used previously on other technological interfaces (e.g., check-in services).

When dividing the data into two groups, adopters and non-adopters of mobile Internet, we found that the adopters had a more positive attitude than non-adopters towards many of the services. This supports the results of Anckar and D'Incau (2002), where adopters of the fixed Internet expressed a higher interest in mobile services compared to non-adopters. The mobile Internet adopters in our study were younger than non-adopters, indicating that there is a new generation of customers who are more positively tuned into this new channel. However, since all customers expressed a low interest in receiving frequent flyer information through their mobile phone, its use in CRM will have to be carefully considered. The study also revealed that customers are not prepared to pay additional costs for being able to use the mobile channel, whenever and wherever required. Customers expect the same feeless services through the mobile channel as they have become used to on the wired Internet. They are also not prepared to pay for quicker service, but probably see this as a normal service improvement in a competitive environment. For example, immediate feedback attracted customers to some extent but not enough to be paid for. However, although the new service would require additional investments from the companies, they should welcome customers' complaints as part of a defensive marketing strategy (Fornell & Wernerfelt, 1987).

Further, customers did not feel that the new services would have a strong positive effect on the company's brand assets in terms of improved image and retention. One reason might be that customers view mobile services as a hygiene

factor and not as a motivation factor. Thus in the same way as customers expect all companies to have an online presence, they expect them to offer mobile services. Customers might not use them regularly, but they expect them to be available when needed. Moreover, business customers probably know by experience that successful services are easily copied by competitors and that readily available applications do not offer unique and stable competitive advantages to a company. Further, customers may be afraid of their phones being cluttered with unwanted messages and may prefer companies to communicate with them in a less obtrusive way.

Since CRM aims to increase customer retention, the findings of the survey indicate that at present the suggested m-services to frequent flyers would not achieve this aim. The new means of getting flight information or buying flight tickets do not seem to be sufficiently attractive to enhance customer loyalty.

LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

The study has several limitations, which have to be taken into account when interpreting the results. In addition to the obvious limitations of studying a small sample of a single company, and the bias that comes from self-selection among those who received the survey, other limitations need to be mentioned. An important limitation is that customers had to imagine the proposed services and could not experience them first hand. It is possible that they would have had a more positive attitude if they had been able to try the service on a high quality mobile device. Multiple items in the questionnaire on customers' current mobile service use and on their loyalty to different service channels would have provided valuable information that would have helped in explaining the results. The study was conducted in close cooperation with the company, which put severe limitations on the constructs that were used for the study. Thus, future studies should include more of the well-established concepts in the consumer adoption literature. For example, future studies should include more information on customer innovation characteristics and behavior, which would make it possible to categorize customers into more specific adopter segments. Despite the limitations of TAM in studying customer interface usage of technology (Nysveen, Pedersen, & Thorbjørnsen, 2005) to explain the adoption of mobile CRM, measures from the consumer innovativeness and/or technology acceptance literature could also be used.

Our study should be seen as exploratory, in providing some initial findings on customer perceptions of mobile CRM services. More studies are obviously needed, in other companies, on other services, and on complete customer relationship programs. Since consumer innovativeness research has concentrated on tangible products (for a review, see Roehrich, 2004), it would be fruitful to apply this line of research on services and technologically novel products, and in particular on a combination of m- and e-services. Further, it would be of interest to study customers' reasons for their choice of channel to contact a company and to receive communications from it. Research on bank services has shown that customers use different channels for different purposes (Patrício, Fisk, & Falcão e Cunha, 2003), but there is also evidence that the new generation of customers make little difference between channels (Lindstrom, 2003). Finally, our study could be extended to examine specific use contexts that may influence the usability of mobile services (Kim, Kim, & Lee, 2005).

Managerial Implications

MIDlet technologies offer companies the opportunity to develop new, specialized services; bringing benefits, and thereby added value, to customer relationships. In the hype and speed of techno-

logical development, it is easy for companies to be fascinated by technological developments that may seem to improve both current services and brand image, but which attract little interest when they are first introduced on the market. Customers' habits change slowly. Although mobile banking has enjoyed a remarkable success throughout Europe, it is in many ways a unique context (Riivari, 2005). In other contexts, such as travel services (Wang & Cheung, 2004), neither the market, nor the devices seem to be ready for the complexity of mobile travel services. Therefore, companies that consider developing wireless services as part of their CRM strategy should first thoroughly investigate its potential in relation to costs. Our study showed that most customers expect companies to offer new CRM mobile services free of charge, as part of customer relationship maintenance costs. Companies need to carefully consider what charges can be claimed for services that are intended to add value to customer relationships. Further, companies need to educate customers in the use and benefits of mobile services and provide incentives to encourage trial.

In addition, when developing mobile services, it is important that the logic of using the service strongly resembles that which the customers have grown used to through other channels, or through other service providers. This is a huge challenge, since different channels differ considerably in how the service is presented to customers, and different applications result in different service logics and scripts. To give an example from airlines, customers already have had to learn different logics for checking in on the Internet and through an automat at the airport. In addition, the Internet check-in services and automats of different airlines have different interfaces and work in different ways. Thus, it is understandable if customers are unwilling to learn yet a third way to check in through their mobile phone. These types of problems have to be minimized through service development that gives the customers' perspective first priority.

From a relationship marketing perspective, it is important that customers are provided with a choice of how to interact with the company. Relationships are not enhanced by forcing customers to interact with certain channels. Therefore, we adopt a different standpoint from Winer (2001, p. 89), who suggests that "[the] essence of the information technology revolution and, in particular, the World Wide Web is the opportunity afforded companies to choose how they interact with their customers." Instead, we suggest that the new channels afford *customers* an opportunity to choose how to interact with the company, and that strong customer relationships can be built only through voluntary use of new technologies. When designing strategies, all channels need to be considered from a customer relationship perspective, designing the services of each channel so that it maximizes its benefits to customers.

Concluding Remarks

Our study on mobile CRM contributes to the literature on mobile services by being one of the first empirical investigations of customer attitudes towards loyalty program services provided through a mobile device. Although the study showed that loyalty program customers have little interest in mobile CRM services, it can be concluded that mobile CRM to some extent enhances the brand image of a company, which over time may have a positive effect also on customer retention. In addition, offering mobile services will demonstrate that the company is at the forefront of service technology development. This will attract early adopters with a strong interest in new technologies, whose expertise can be used, for example, by involving them in the service development process. Thus it is clear that the mobile channel should be included in companies' future CRM strategies, but also that more research is needed on the benefits of mobile CRM to both customers and companies.

REFERENCES

Aaker, D. A. (1996). *Building strong brands.* New York: The Free Press.

Aaker, D. A., & Joachimsthaler, E. (2000). *Brand leadership.* New York: The Free Press.

Adjari, J. (2001). *Java 2 mobile information device profile (MIDP).* Retrieved July 22, 2003, from http://www.tml.hut.fi/Studies/Tik-111.590/2001s/papers/jafar_ajdari.pdf

Akhgar, B., Siddiqi, J., Foster, M., Siddiqi, H., & Akhgar, A. (2002). Applying customer relationship management (CRM) in the mobile commerce market. *International Conference on Mobile Computing, Sponsored by EU (IST),* Greece.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. *Journal of Information Technology Theory and Application*, *4*(1), 43-64.

Balasubramanian, S., Peterson, R. A., & Jarvenpaa, S. L. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science*, *30*(4), 348-361.

Bitner, M. J., Brown, S. W., & Meuter, M. L. (2000). Technology infusion in service encounters. *Journal of the Academy of Marketing Science*, *28*(1), 138-149.

Chae, M., Kim, J., Kim, H., & Ryu, H. (2002). Information quality for mobile internet services: A theoretical model with empirical validation. *Electronic Markets, 12*(1), 38-46.

Cho, Y., Im, I., Hiltz, R., & Fjermestad, J. (2002). The effects of post-purchase evaluation factors on online vs. offline customer complaining behavior: Implications for customer loyalty. *Advances in Consumer Research, 29*(1), 318-327.

Crosby, L. A., & Johnson, S. L. (2001). Technology: Friend or foe to customer relationships? *Marketing Management, 10*(4), 10-11.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340.

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-services adoption: A perceived risk facets perspective. *International Journal of Human-Computer Studies, 59,* 451-474.

Feinberg, R., & Kadam, R. (2002). E-CRM Web service attributes as determinants of customer satisfaction with retail Web sites. *International Journal of Service Industry Management, 13*(5), 432-451.

Fjermestad, J., & Romano, N. C., Jr. (2003). Electronic customer relationship management. Revisiting the general principles of usability and resistance—An integrative implementation framework. *Business Process Management Journal, 9*(5), 572-591.

Fornell, C., & Wernerfelt, B. (1987, November). Defensive marketing strategy by customer complaint management: A theoretical analysis. *Journal of Marketing Research, 24,* 337-346.

Gilbert, A. L., & Han, H. (2005). Understanding mobile data services adoption: Demography, attitudes or needs? *Technological Forecasting & Social Change, 72,* 327-337.

Grönroos, C. (2000). *Service management and marketing—A customer relationship management approach.* New York: John Wiley & Sons, Ltd.

Heinonen, K. (2004). Reconceptualizing customer perceived value—The value of time and place. *Managing Service Quality, 14*(2/3), 205-215.

Heinonen, K. (2006). Temporal and spatial e-service value. *International Journal of Service Industry Management, 17*(4), 380-400.

Helenius, J., & Liljander, V. (2005). Developing brand assets with wireless devices. In I. Clarke III & T. B. Flatherty (Eds.), *Advances in electronic marketing* (pp. 176-192). Hershey PA: Idea Group.

Jarvenpaa, S. L., Lang, K. R., Takeda, Y., & Tuunanen, V. K. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41-44.

Johnston, R., & Mehra, S. (2002). Best-practice complaint management. *Academy of Management Journal, 16*(4), 145-154.

Jukic, N., Sharma, A., Jukic, B., & Parameswaran, M. (2002, May 19-22). M-commerce: Analysis of impact on marketing orientation. *Information Resources Management Association International Conference*, Seattle, WA.

Kaapu, T. (2005). The concept of information privacy in e-commerce: A phenomenographical analysis of consumers' views. *Conference Paper IRIS'28,* Kristiansand, Norway.

Kannan, P. K., Chang, A.-M., & Whinston, A. B. (2001) Wireless commerce: Marketing issues and possibilities. In *Proceedings of the 34th Hawaii International Conference on System Sciences*, Hawaii.

Kim, H., Kim, J., & Lee, Y. (2005). An empirical study of use context in the mobile internet, focusing on the usability of information architecture. *Information Systems Frontier, 7*(2), 175-186.

Kindberg, T., Sellen, A., & Geelhoed, E. (2004, July 7). *Security and trust in mobile interactions: A study of users' perceptions and reasoning.* Consumer Applications and Systems Laboratory, HP Laboratories Bristol, HPL-2004-113. Retrieved September 29, 2005, from http://www.hpl.hp.com/techreports/2004/HPL-2004-113.pdf

Lam, J., & Chan, S. S. (2003). *Exploring CRM implementation on the internet and mobile channels.* Chicago: Seminar, DePaul University, School of Computer Science, Telecommunication and Information Systems.

Liljander, V., Gillberg, F., Gummerus J., & van Riel, A. (2006). Technology readiness and the evaluation and adoption of self-service technologies. *Journal of Retailing and Consumer Services, 13*(3), 177-191.

Lin, H.-H. & Wang, Y.-S. (2006). An examination of the determinants of customer loyalty in mobile commerce contexts. *Information and Management, 43*(3), 271-282.

Lindstrom, M. (2003). *Brand child. Remarkable insights into the minds of today's global kids and their relationship with brands.* London: Kogan Page.

Money, A., Tromp, D., & Wegner, T. (1988). The qualification of decision support benefits within the context of value analysis. *MIS Quarterly, 12*(2), 223-236.

Mort, G. S., & Drennan, J. (2005). Marketing m-services: Establishing a usage benefit typology related to mobile user characteristics. *Database Marketing & Customer Strategy Management, 12*(4), 327-341.

Nordman, J., & Liljander, V. (2004). MSQ-model—An exploratory study of the determinants of mobile service quality. In S. Krishnamurthy (Ed.), *Contemporary research in e-marketing* (Vol. 1, pp. 93-129). Hershey, PA: Idea Group Publishing.

Nysveen, H., Pedersen, P. E., & Thorbjørnsen, H. (2005). Intentions to use mobile services: Antecedents and cross-service comparisons. *Journal of the Academy of Marketing Science, 33*(3), 330-246.

Nysveen, H., Pedersen, P. E., Thorbjørnsen, H., & Berthon, P. (2005). Mobilizing the brand. The effects of mobile services on brand relationships and the main channel use. *Journal of Service Research, 7*(3), 257-276.

Okazaki, S. (2005). New perspective on m-commerce research. *Journal of Electronic Commerce Research, 6*(3), 160-164.

Okazaki, S. (2006). What do we know about mobile Internet adopter? A cluster analysis. *Information & Management, 43,* 127-141.

Patrício, L., Fisk, R. P., & Falcão e Cunha, J. (2003). Improving satisfaction with bank service offerings: Measuring the contribution of each delivery channel. *Managing Service Quality, 13*(6), 471-482.

Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. *Journal of Marketing, 69*(4), 167-176.

Plouffe, C. R., Vandenbosch, M., & Hulland, J. (2001). Intermediating technologies and multigroup adoption: A comparison of consumer and merchant adoption intentions toward a new electronic payment system. *The Journal of Product Innovation Management, 18*(2), 65-81.

Pura, M. (2005). Linking perceived value and loyalty in location-based mobile services. *Managing Service Quality, 15*(6), 509-538.

Ram, S., & Sheth, J. N. (1989). Consumer resistance to innovations: The marketing problem and its solutions. *The Journal of Consumer Marketing, 6*(2), 5-14.

Riivari, J. (2005). Mobile banking: A powerful new marketing and CRM tool for financial services companies all over Europe. *Journal of Financial Services Marketing, 10*(1), 11-20.

Roehrich, G. (2004). Consumer innovativeness concepts and measurements. *Journal of Business Research, 57,* 671-677.

Shah, J. R., & Murtaza, M. B. (2005). Effective customer relationship management through Web services. *Journal of Computer Information Systems, 46*(1), 98-109.

Shankar, V., Smith, A. K., & Rangaswamy, A. (2002). Customer satisfaction and loyalty in online and offline environments. *International Journal of Research in Marketing, 20*(2), 153-175.

Sugai, P. (2005). Mapping the mind of the mobile consumer across borders—An application of the Zaltman metaphor elicitation technique. *International Marketing Review, 22*(6), 641-657.

Turban, E., King, D., Lee, J., Warkentin, M., & Chung, H. M. (2002). *Electronic commerce, a managerial perspective.* London: Prentice Hall.

Varshney, U., & Vetter, R. (2001). A framework for the emerging m-commerce applications. In *Proceedings of the 34th Hawaii International Conference on System Sciences,* Hawaii.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186-204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Verhoef, P. C. (2003). Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing, 67*(4), 30-45.

Walker, R. H., Craig-Lees, M., Hecker, R., & Francis, H. (2002). Technology-enabled service delivery. An investigation of reasons affecting customer adoption and rejection. *International Journal of Service Industry Management, 13*(1), 91-106.

Wang, S., & Cheung, W. (2004). E-business adoption by travel agencies: Prime candidates for mobile e-commerce. *International Journal of Electronic Commerce, 8*(3), 43-63.

Winer, R. S. (2001). A framework for customer relationship management. *California Management Review, 43*(4), 89-105.

Zablah, A. R., Bellenger, D. N., & Johnston, W. J. (2004). An evaluation of divergent perspectives on customer relationship management: Towards a common understanding of an emerging phenomenon. *Industrial Marketing Management, 33*, 475-489.

This work was previously published in International Journal of E-Business Research, Vol. 3, Issue 2, edited by I. Lee, pp. 24-40, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 5.15
Contractual Obligations Between Mobile Service Providers and Users

Robert Willis
Lakehead University, Canada

Alexander Serenko
Lakehead University, Canada

Ofir Turel
McMaster University, Canada

INTRODUCTION

The purpose of this chapter is to discuss the effect of contractual obligations between users and providers of mobile services on customer loyalty. One of the unique characteristics of mobile commerce that distinguishes it from most other goods and services is the employment of long-term contractual obligations that users have to accept to utilize the service. In terms of over-the-counter products, sold in one-time individual transactions in well-established markets, a strong body of knowledge exists that suggests that businesses may enhance loyalty through the improvement of quality and customer satisfaction levels. With respect to mobile commerce, however, this viewpoint may not necessarily hold true given the contractual nature of business-customer relationships.

In the case of mobile computing, it is suggested that loyalty consists of two independent yet correlated constructs that are influenced by different factors: repurchase likelihood and price tolerance. Repurchase likelihood is defined as a customer's positive attitude towards a particular service provider that increases the likelihood of purchasing additional services or repurchasing the same services in the future (e.g., after the contract expires). For example, when people decide to purchase a new mobile phone, they are free to choose any provider they want. In other words, repurchase likelihood is not affected by contractual obligations. In contrast, price tolerance corresponds to a probability of staying with a current provider when it increases or a competitor decreases service charges. In this situation, individuals have to break the existing contractual

obligations. Currently, there is empirical evidence to suggest that the discussion above holds true in terms of mobile computing. However, there are few well-documented works that explore this argument in depth. This chapter attempts to fill that void.

This chapter will present implications for both scholarship and practice. In terms of academia, it is believed that researchers conducting empirical investigations on customer loyalty with mobile services should be aware of the two independent dimensions of the business-customer relationship and utilize appropriate research instruments to ensure the unidimensionality of each construct. With regards to practice, it is suggested that managers and marketers be aware of the differences between repurchase likelihood and price tolerance, understand their antecedents, and predict the consequences of manipulating each one. It is noted that overall loyalty is not the only multidimensional constuct in mobile commerce. Recently, it was emperically demonstrated that perceived value of short messaging services is a second-order construct that consists of several independent yet correlated dimensions (Turel et al., 2007).

Theoretical separation of the overall loyalty construct into two dimensions has been already empirically demonstrated in three independent mobile commerce investigations. First, Turel and Serenko (2006) applied the American customer satisfaction model (ACSM) to study mobile services in North America. By utilizing the original instrument developed by Fornell, Johnson, Anderson, Cha, and Bryant (1996), they discovered a low reliability of the overall satisfaction construct, and found that the correlation between two items representing price tolerance and one item reflecting repurchase likelihood was only 0.21 ($p<0.01$, N=204). Second, Turel et al. (2006) adapted the ACSM to study the consequences of customer satisfaction with mobile services in four countries (Canada, Finland, Israel, and Singapore), and reported that the correlation between price tolerance and repurchase likelihood was 0.20 ($p<0.01$, N=736). Third, Yol, Serenko, and Turel (2006) analyzed the ACSM with respect to mobile services in the U.S. and again found the same correlation to be 0.45 ($p<0.01$, N=1,253). All these correlations fall into the small-to-medium range, and two of them are beyond the lowest cut-off value of 0.35 for item-to-total correlation (Nunnally & Bernstein, 1994). The statistical significance of these correlations is explained by large sample sizes. Therefore, it is impossible to design a single unidimensional construct in mobile commerce research consisting of both price tolerance and repurchase likelihood. In all of these studies, most users had long-term contractual obligations with their respective mobile service provider that confirms the validity of the aforementioned conceptual discussion.

To better understand the customer loyalty concept in light of contractual obligations, this chapter briefly describes the American customer satisfaction model (ACSM), and then discusses the concepts of price tolerance and repurchase likelihood. Finally, it presents a summary which outlines implications for research and practice.

THE AMERICAN CUSTOMER SATISFACTION MODEL

The mobile telephony market continues to be one of the fastest growing service sector markets, creating a fiercely competitive industry environment (Kim & Yoon, 2005). As has happened in other, subscription-based mobile service industries, the nature of this competition has changed from the acquisition of new customers to the retention of existing customers and the luring away of competitors' customers. This last strategy is known in the industry as *outbound churn* or, more simply, as *churn*. Given the increasing penetration of mobile computing devices and the maturation of the market, avoiding churn and maximizing customer loyalty has become a primary concern

for wireless providers. The first step in minimizing churn in a company's customer base is to understand its root causes.

The determinants of churn may be estimated by the adapted American Customer Satisfaction Model (see Figure 1). The original model suggests that satisfaction affects overall customer loyalty, where loyalty is a unidimensional construct that consists of price tolerance (i.e., the probability of staying with the current provider if it increases prices or if a competitor decreases prices) and repurchase likelihood (i.e., the probability of purchasing the same service again). At the same time, several recent works suggest that these loyalty dimensions are distinct yet correlated because of the contractual nature of the customer-service provider relationship.

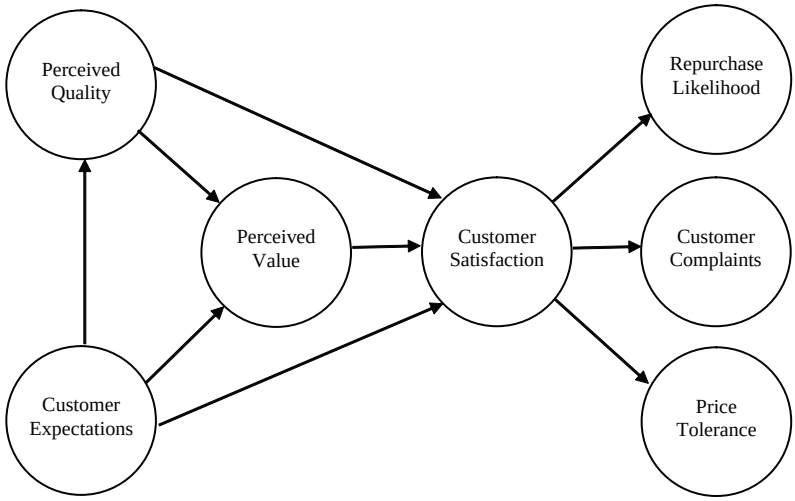
Customer loyalty is one of the major constructs in marketing, and a large part of a marketing manager's effort is aimed at creating and maintaining loyalty among an organization's customer base. The significance of loyalty comes from the positive impact it has on the operations of the company in terms of customer retention, repurchase, long-term customer relationships, and company profits (Caruana, 2004). In other words, loyalty is a primary factor in reducing churn.

The notion of switching costs affecting loyalty has been recognized and researched by several professional and academic disciplines, including marketing, economics, and strategy. "Switching costs are generally defined as costs that deter customers from switching to a competitor's product or service" (Caruana, 2004, p. 256). For managers and researchers, it is important to understand the concepts of switching costs and customer loyalty, and to clearly identify both their dimensions and their interaction.

PRICE TOLERANCE

Switching costs are generally defined as one-time costs facing the consumer/buyer of switching from one supplier to another (Porter, 1980; Burnham, Frels, & Mahajan, 2003). Several researchers have identified various attributes or types of switching costs (e.g., Thibault & Kelly, 1959; Klemperer, 1987; Guiltnan, 1989; Burnham et al., 2003; Hu & Hwang, 2006); however, for the purposes of this chapter, switching costs are broadly catego-

Figure 1. The American Customer Satisfaction Model (adapted from Fornell et al., 1996)



rized as three types: transaction, learning, and contractual. Transaction costs are costs incurred when a consumer begins a relationship with a provider and includes the costs associated with ending that relationship or terminating an existing relationship. Learning costs are associated with the effort required by the consumer to achieve the same level of knowledge and comfort acquired using a particular supplier's product, but which may not be transferable to similar/same products of other suppliers. Additionally, the notion of learning costs incorporates the implicit switching costs associated with decision biases, risk aversion, and market knowledge/familiarity. Learning what one's options are, what the relative competitive position of all suppliers are, and other such knowledge involves learning costs that will be differentially valued by individuals. In the case of mobile services, the switching costs are created by a service provider that requires customers to sign a long-term contract. If a customer wants to switch to another provider, he or she will have to pay a penalty to the current provider. As such, contractual costs are those costs that are directly provider induced in order to penalize churn and which are intended to prevent poaching of customers by other suppliers. With respect to the American customer satisfaction model, switching costs directly affect price tolerance. The ACSM survey instrument presents two questions: (1) by how much their current provider should increase its current prices in order for them to switch to a competitor, and (2) by how much a competitor should reduce its prices in order for them to switch. Peoples' answers to these questionnaire items are greatly affected by the direct switching costs they incur, such as a penalty.

Consistent with this proposition, Weiss and Anderson (1992) found that switching costs are a major consideration when consumers are making a churn decision, and that these costs (barriers) tend to reduce customers' churn behavior. These findings were further supported by research done by Jones, Mothersbaugh, and Beatty (2000). Burnham et al. (2003) suggested that switching costs are negatively correlated with a customer's intention to churn: the higher the costs, the lower the intention to switch. As Hu and Hwang (2006) point out, "the industry remains in a state of dynamic competition" (p. 75) and providers continue implementing flexible offerings that are aimed at reducing consumers' churn behavior. Shapiro and Varian (1999) found that *perceived* switching costs—which incorporate all of the explicit costs as well as the implicit costs discussed above—act as barriers to churn behavior. They suggest consumers will weigh the benefits of switching against the actual and psychological costs when considering churning.

Overall, the discussion above demonstrates that the concept of switching barriers has its own unique dimensions. In terms of the American customer satisfaction model applied in the context of mobile services, it is believed that two items pertaining to the customer switching behavior (conceptualized as price tolerance in the model) reflect a unique latent variable entitled *price tolerance*.

REPURCHASE LIKELIHOOD

The notion of overall customer loyalty has changed in both breadth and depth over the years in which it has been studied by academics and practitioners alike. The breadth of its definition is demonstrated by the multiplicity of areas that are examined, such as brand, product, vendor, or service loyalty. Initial research was primarily focused on brand loyalty, and mostly examined the behavioral aspects of the construct. In this view, Newman and Werbel (1973) defined customer loyalty as the repurchase of a brand that only considered that brand and which involved no brand-related information seeking.

Day (1969) was one of the first researchers to highlight the role of a positive attitude in the construct of loyalty. Following this line of rea-

soning, which incorporated both the behavioral and attitudinal conceptions of loyalty, operationalization of the construct of customer loyalty involved combining the aspects of purchasing a particular brand together with an affective attitudinal measure, whether that measure used a single scale or multi-scale items. With regards to the American customer satisfaction model, the discussion above relates to the unique dimension of loyalty as *repurchase likelihood,* or the probability of buying new services from the current provider when these purchases are not affected by prior contractual obligations, for example, when a contract has expired.

PRICE TOLERANCE AND REPURCHASE LIKELIHOOD

The literature—and intuition—suggests that higher switching costs are positively related to price tolerance—that is, that higher switching costs compel customers to remain loyal. Fornell et al. (1996) were among the first to include switching costs by adding them to the construct of customer satisfaction in the reflection of customer loyalty. In the ACSM, all items (i.e., two pertaining to price tolerance and one relating to repurchase likelihood) were believed to reflect overall loyalty. A number of subsequent studies demonstrated the unidimensionality of this construct. However, in the context of mobile services when high switching costs exist, unidimensionality does not apply. As such, it is suggested that, based on the theoretical rationale as well as empirical studies cited earlier, loyalty should be analyzed along two distinct dimensions: price tolerance and repurchase likelihood.

In terms of prior empirical research, Jones and Sasser (1995) included switching costs as one factor or competitiveness: since high switching costs discourage churning, they reduce the incentive for firms to compete. Bateson and Hoffman (1999) similarly suggest that customer satisfaction and switching costs are the primary influencers of loyalty. More recent studies have shown that switching costs have a direct and strong influence on the re-purchase decision (customer loyalty) in all markets, for example France (Lee, Lee, & Feick, 2001), Korea (Kim & Yoon, 2005), Australia (Caruana, 2004), Taiwan (Hu & Hwang, 2006), and Turkey (Aydin, Özer, & Arasil, 2005).

Jones et al.'s (2000) study examined the role of switching costs (barriers) in customer retention for services. They found that although core-service satisfaction was a primary issue in retention, switching factors in the form of interpersonal relationships, direct and indirect costs, and the perceived benefits of potential alternatives were also important. As such, these factors represented different unique dimensions of the overall loyalty concept. This supports the notion, outlined above, that loyalty of mobile service users must be considered as multidimensional and not simply as direct, contractual costs.

IMPACTS FOR MANAGERS AND RESEARCHERS

The findings of the many studies in the area show support for the intuitive link between higher switching costs and greater levels of customer loyalty (or at least, retention). More importantly, they also provide a greater understanding of the interaction between switching costs and loyalty, and refine the model that has, to date, served as a guide to management of mobile phone companies.

Management of mobile phone companies must understand the complexity and multidimensionality of the concepts of switching costs that directly influence price tolerance and repurchase likelihood that is not affected by contractual obligations. They must also understand that switching costs affect customer loyalty not solely through the contractual cost component of switching costs, but also through the learning and transac-

tion cost components. A customer's, or potential customer's, belief that he or she will end up with a 'bad deal' financially in switching to a new provider—and that assessment will include all of the implicit as well as explicit costs—is the most important issue in the churn decision. This highlights the point that managing customer relationships, so that they remain positive, acts to keep the customer attached, whether this is a result of satisfaction outweighing perceived benefits or simply of customer inertia (Burnham et al, 2003; Caruana, 2004). It also highlights the need for poaching strategies to emphasize not only the financial benefits, but the relational benefits as well (Hu & Hwang, 2006). It should be noted that existing studies point out that one of the primary issues affecting the learning cost component has been the lack of time to undertake a complete comparison of the many offerings in the market. Additionally, providers have tended in the past to couch their offerings in terms that vary widely from their competitors', thus introducing a level of uncertainty and confusion in the minds of the analyzing consumer (Hu & Hwang, 2006). These factors are becoming less and less viable as consumers turn to the Internet for their purchasing information and guidance, and as consumers demand—and get—a certain level of standardization in the offerings of providers in the market, whether that standardization comes from the providers themselves or from organizations that perform such analyses and offer them to the consuming (Internet- or magazine-based) consumer. Additionally, the increasing homogeneity of pricing strategies and service packages will lead to a lessening of the impact of explicit (transaction and contractual) switching costs on the churn decision (Hu & Hwang, 2006). Thus, management needs to concentrate on customer relationships. Swartz (2000) quotes two senior executives in the industry:

If service is poor, then customers will pay any cancellation fees to get rid of the service and choose another provider….

You have to look at your reasons for churn…You can't use a contract to make up for poor service. If your service is poor, you can lock them in for a year…but they're gone the minute month 13 rolls around.

More research needs to be done on the notion of overall loyalty as a multidimensional construct. Are there positive barriers, such as interpersonal relationships, as well as negative? What relative influence on customer satisfaction do core and non-core services have? How sensitive are costs as barriers? Research into whether or not there are services that are perceived as having low barriers as opposed to services that are perceived as having high barriers within the market offerings would help refine our understanding of the role of various costs.

REFERENCES

Aydin, S., Özer, G., & Arasil, Ö. (2005). Customer loyalty and the effect of switching costs as a modifier variable: A case in the Turkish mobile phone market. *Marketing Intelligence and Planning, 23*(1), 89-103.

Bateson, J. E. G., & Hoffman, K. D. (1999). *Managing services marketing, text and readings* (4th ed.). Fort Worth, TX: Dryden Press.

Burnham, T. A., Frels, J. K., & Mahajan, V. (2003). Consumer switching costs: A typology, antecedents and consequences. *Journal of the Academy of Marketing Science, 31*(2), 109-126.

Caruana, A. (2004). The impact of switching costs on customer loyalty: A study among corporate customers of mobile telephony. *Journal of Targeting, Measurement and Analysis for Marketing, 12*(3), 256-268.

Day, G. S. (1969). A two dimensional concept of brand loyalty. *Journal of Advertising Research, 9*(3), 29-36.

Fornell, C. (1992). A national consumer satisfaction barometer: The Swedish experience. *Journal of Marketing, 56*(1), 6-21.

Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American Customer Satisfaction Index: Nature, purpose, and findings. *Journal of Marketing, 60*(4), 7-18.

Guiltnan, J. P. (1989). A classification of switching costs with implications for relationship marketing. In T. L. Childers & R. P. Bagozzi (Eds.), *Proceedings of the Winter Educators' Conference: Marketing Theory and Practice* (pp. 216-220), Chicago.

Hu, A.W.-L., & Hwang, I.-S. (2006). Measuring the effects of consumer switching costs on switching intention in Taiwan mobile telecommunications services. *Journal of American Academy of Business, 9*(1), 75-85.

Jones, M. A., Mothersbaugh, D. L., & Beatty, S. E. (2000). Switching barriers and repurchase intentions in services. *Journal of Retailing, 76*(2), 259-274.

Jones, T. O, & Sasser, W. E. (1995). Why satisfied customers defect. *Harvard Business Review, 73*(1), 88-99.

Kim, H.-S., & Yoon, C.-H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy, 28*, 751-756.

Klemperer, P. (1987). Markets with consumer switching costs. *The Quarterly Journal of Economics, 102*, 375-394.

Lee, J., Lee, J., & Feick, L. (2001). The impact of switching costs on customer satisfaction-loyalty link: Mobile phone service in France. *Journal of Services Marketing, 15*(1), 35-48.

Newman, J. W., & Werbel, R. A. (1973). Multivariate analysis of brand loyalty for major household appliances. *Journal of Marketing Research, 10*(4), 404-409.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Oliver, R. L. (1996). *Satisfaction: A behavioral perspective on the consumer*. New York: McGraw-Hill.

Porter, M. E. (2003). *Competitive strategy: Techniques for analyzing industries and competitors*. New York: MacMillan.

Serenko, A., Turel, O., & Yol, S. (2006). Moderating roles of user demographics in the American customer satisfaction model within the context of mobile services. *Journal of Information Technology Management, 17*(4): in-press.

Swartz, N. (2000). Reconsidering contracts. *Wireless Review, 17*(4), 48-52.

Thibault, J. W., & Kelley, H. H. (1959). *The social psychology of groups*. New York: John Wiley & Sons.

Turel, O., & Serenko, A. (2006). Satisfaction with mobile services in Canada: An empirical investigation. *Telecommunications Policy, 30*(5-6), 314-331.

Turel, O., Serenko, A., & Bontis, N. (2007). User acceptance of wireless short messaging services: Deconstructing perceived value. *Information & Management, 44*(1), 63-73.

Turel, O., Serenko, A., Detlor, B., Collan, M., Nam, I., & Puhakainen, J. (2006). Investigating the determinants of satisfaction and usage of mobile IT services in four countries. *Journal of Global Information Technology Management, 9*(4), 6-27.

Weiss, A.M., & Anderson, E. (1992). Converting from independent to employee sales forces: The role of perceived switching costs. *Journal of Marketing Research, 29*(1), 101-115.

KEY TERMS

American Customer Satisfaction Model: The original model suggests that satisfaction affects the overall customer loyalty, where loyalty is a unidimensional construct that consists of price tolerance (i.e., the probability of staying with the current provider if it increases prices or if a competitor decreases prices) and repurchase likelihood (i.e., the probability of purchasing the same service again). If the customer's expectations of product quality, service quality, and price are exceeded, a firm will achieve high levels of customer satisfaction and will create *customer delight*. If the customer's expectations are not met, customer dissatisfaction will result. And the lower the satisfaction level, the more likely the customer is to stop buying from the firm.

Churn: This refers to the notion that a company will, over any given period of time, lose existing customers and gain new customers. Churn is, currently, mostly created by the luring away of competitors' customers.

Customer Loyalty: The notion that a customer will continue to use a particular brand or product; the behavior customers exhibit when they make frequent repeat purchases of a brand or product.

Price Tolerance: The extent to which price is an important criterion in the customer's decision-making process; thus a price-sensitive customer is likely to notice a price rise and switch to a cheaper brand or supplier.

Repurchase Likelihood: The probability of buying new services from the current provider when these purchases are not affected by prior contractual obligations, for example, when a contract has expired.

Switching Cost: One-time cost facing the consumer/buyer of switching from one supplier to another. Switching costs are composed of transaction costs (costs incurred when a consumer begins a relationship with a provider, and includes the costs associated with ending that relationship or terminating an existing relationship), learning costs (costs associated with the effort required by the consumer to achieve the same level of knowledge and comfort acquired using a particular supplier's product, but which may not be transferable to the same/similar products of other suppliers), and contractual costs (costs that are directly provider induced in order to penalize churn and which are intended to prevent poaching of customers by other suppliers).

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 143-148, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.16
Accessibility of Mobile Applications

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

The increasing affordability of devices, advantages associated with a device always being handy while not being dependent on its location, and being able to tap into a wealth of information/services has brought a new paradigm to mobile users. Indeed, the *mobile Web* promises the vision of universality: access (virtually) anywhere, at any time, on any device, and to *anybody*.

However, with these vistas comes the realization that the users of the mobile applications and their context vary in many different ways: personal preferences, cognitive/neurological and physiological ability, age, cultural background, and variations in computing environment (device, platform, user agent) deployed. These pose a challenge to the ubiquity of mobile applications and could present obstacles to their proliferation.

This chapter is organized as follows. We first provide the motivation and background necessary for later discussion. This is followed by introduction of a framework within which accessibility of mobile applications can be systematically addressed and thereby improved. This framework is based on the notions from semiotics and quality engineering, and aims to be practical. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

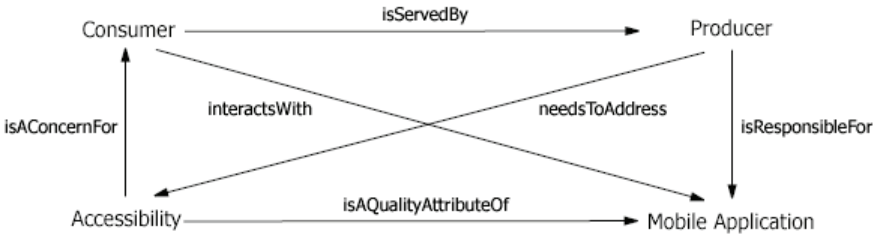
BACKGROUND

The issue of accessibility is not new. However, the mobile Web with its potential flexibility on both the client-side and the server-side presents new challenges towards it.

Figure 1 illustrates the dynamics within which the issue of accessibility of a mobile application arises.

We define a *mobile application* as a domain-specific application that provides services and

Figure 1. The interrelationships between a consumer, a producer, accessibility, and a mobile application



means for interactivity in the mobile Web. For example, education, entertainment, or news syndication are some of the possible domains. The issue of accessibility is intimately related to the user and user context that includes client-side computing environment. To that regard, we define *accessibility* in context of a mobile application as access to the mobile Web by everyone, regardless of their human or environment properties. A *consumer* (user) is a person that uses a mobile application. A *producer* (provider) is a person or an organization that creates a mobile application.

The Consumer Perspective of Mobile Accessibility

The accessibility concerns of a consumer are of two types, namely human and environment properties, which we now discuss briefly.

Human Properties

Human properties are issues relating to the differences in properties among people. One major class of these properties is related to a person's ability, and often the degree of absence of such properties is termed as a disability. We will use the term "disability" and "impairment" synonymously.

The statistics vary, but according to estimates of the United Nations, about 10% of the world's population is considered disabled. The number of people with some form of disability that do have access to the Internet is in the millions.

There are several types of disabilities that a producer of a mobile application needs to be concerned with. These can include visual (e.g., low visual acuity, blindness, color blindness), neurological (e.g., epilepsy), auditory (e.g., low hearing functionality, deafness), speech (e.g., difficulties in speaking), physical (e.g., problems using an input device), cognitive (e.g., difficulties of comprehending complex texts and complex structures), cultural/regional (e.g., differences in the use of idioms, metaphors leading to linguistic problems).

Environment Properties

Environment properties are issues relating to different situations in which people find themselves, either temporarily or permanently. These situations could be related to their connectivity, the location they are in, or the device/platform/user agent they are using. For example, a user using a computer in a vehicle shares many of the issues that some people have permanently due to a disability in hand motorics. Or, for example, a user may be accessing the *same* information using a personal digital assistant (PDA) or a cellular phone.

The Producer Perspective of Mobile Accessibility

The motivation for accessibility for a business is to reach as many users as possible and in doing so reduce concerns over customer alienation.

It is the producer of the mobile application that needs to adjust to the user context (and address the issue of accessibility), not the other way around. It is not reasonable for a producer to expect that the consumer environment will be conducive to *anything* that is delivered to him/her. In certain cases, when a consumer has a certain disability, such adaptation is not even possible.

If the success of a mobile application is measured by the access to its services, then improving accessibility is critical for the producers. Still, any steps that are taken by a producer related to a mobile application have associated costs and trade-offs, and the same applies to improvements towards accessibility.

Initiatives for Improving Accessibility in Mobile Contexts

There are currently only a few efforts in systematically addressing accessibility issues pertaining to mobile applications.

There are guidelines available for addressing accessibility (Chisholm, Vanderheiden, & Jacobs, 1999; Ahonen, 2003) in general and language-specific techniques (Chisholm et al., 2000) in particular.

ADDRESSING THE ACCESSIBILITY OF MOBILE APPLICATIONS

To systematically address the accessibility of mobile applications, we take the following steps:

1. View accessibility as a qualitative aspect and address it indirectly via quantitative means.
2. Select a theoretical basis for communication of information (semiotics), and place accessibility in its setting.
3. Address semiotic quality in a practical manner.

Based on this, we propose a framework for accessibility of mobile applications (see Table 1). The external attributes (denoted by E) are extrinsic to the mobile application and are directly the consumer's concern, while internal attributes (denoted by I) are intrinsic to the mobile application and are directly the producer's concern. Since not all attributes corresponding to a semiotic level are on the same echelon, the different tiers are denoted by "Tn."

We now describe each component of the framework in detail.

Semiotic Levels

The first column of Table 1 addresses semiotic levels. Semiotics (Stamper, 1992) is concerned with the use of symbols to convey knowledge.

From a semiotics perspective, a representation such as a mobile resource can be viewed on six interrelated levels: physical, empirical, syntactic, semantic, pragmatic, and social, each depending on the previous one in that order. The physical and empirical levels are concerned with the physical representation of signs in hardware and communication properties of signs, and are not of direct concern here. The syntactic level is responsible for the formal or structural relations between signs. The semantic level is responsible for the relationship of signs to what they stand for. The pragmatic level is responsible for the relation of signs to interpreters. The social level is responsible for the manifestation of social interaction with respect to signs, and is not of direct concern here.

We note that none of the layers in Table 1 is sufficient in itself for addressing accessibility and intimately depends on other layers. For ex-

Table 1. A semiotic framework for accessibility of mobile applications

Semiotic Level	Quality Attributes	Means for Accessibility Assurance and Evaluation	Decision Support
Pragmatic	**Accessibility** [T4;E]	• Training in Primary and Secondary Notation • "Expert" Knowledge (Principles, Guidelines, Patterns) • Inspections • Testing • Metrics • Tools	Feasibility
	Comprehensibility, Interoperability, Performance, Readability, Reliability, Robustness [T3;E]		
Semantic	Completeness and Validity [T2;I]		
Syntactic	Correctness (Primary Notation) and Style (Secondary Notation) [T1;I]		

ample, it is readily possible to create a document in XHTML Basic, a markup language for small information appliances such as mobile devices, that is syntactically correct but is semantically non-valid. This, for instance, would be the case when the elements are (mis)used to create certain user-agent-specific presentation effects. Now, even if a mobile resource is syntactically and semantically acceptable, it could be rendered in such a way that it is unreadable (and therefore violates an attribute at the pragmatic level). For example, this could be the case by the use of very small fonts for some text, or the colors chosen for background and text foreground being so close that the characters are hard to discern.

Quality Attributes

The second column of Table 1 draws the relationship between semiotic levels and corresponding quality attributes. We contend that the quality attributes we mention are necessary but make no claim of their sufficiency.

The internal quality attributes for syntactic and semantic levels are inspired by Lindland, Sindre, and Sølvberg (1994). At the semantic level, we are only concerned with the conformance of the mobile application to the domain it represents (that is, semantic correctness or completeness) and at the syntactic level the interest is in conformance with, with respect to the languages used to produce the mobile application (that is, syntactic correctness).

Accessibility belongs to the pragmatic level and depends on the layers beneath it. It in turn depends upon the other external quality attributes, namely comprehensibility, interoperability, performance, readability, reliability, and robustness, which are also at the pragmatic level. Since these are perceived as necessary conditions, violations of any of these lead to a deterioration of accessibility.

Means for Accessibility Assurance and Evaluation

The third column of Table 1 lists the direct and indirect (and not necessarily mutually exclusive) means for assuring and evaluating accessibility:

- **Training in Primary and Secondary Notation:** The knowledge of the *primary notation* of all technologies (languages) is necessary for guaranteeing conformance to tier T1. The Cognitive Dimensions of Notations (CDs) (Green, 1989) are a generic framework for describing the utility of information artifacts by taking the system environment and the user characteristics into consideration. Our main interest here is in the CD of *secondary notation*. This CD is about appropriate use (that is, *style*) of primary notation in order to assist in interpreting semantics. It uses the notions of *redundant recoding* and *escape*

from formalism along with spatial layout and perceptual cues to clarify information or to give hints to the stakeholder, both of which aid the tiers T2 and T3. Redundant Recoding is the ability to express information in a representation in more than one way, each of which simplifies different cognitive tasks. It can be introduced in a textual mobile resource by making effective use of orthography, typography, and white space. Escape from Formalism is the ability to intersperse natural language text with formalism. Mobile resources could be complemented via natural language annotations (metadata) to make the intent or decision rationale of the author explicit, or to aid understanding of stakeholders that do not have the necessary technical knowledge. Incidentally, many of the language-specific techniques for accessibility (Chisholm et al., 2000) are in agreement with this CD.

- **"Expert" Body of Knowledge:** The three types of knowledge that we are interested are principles, guidelines, and patterns. Following the basic principles (Ghezzi, Jazayeri, & Mandrioli, 2003; Bertini, Catarci, Kimani, & Dix, 2005) underlying a mobile application enables a provider to improve quality attributes related to tiers T1-T3 of the framework. However, principles tend to be abstract in nature which can lead to multiple interpretations in their use and not mandate conformance. The guidelines encourage the use of conventions and good practice, and could serve as a checklist with respect to which an application could be heuristically or otherwise evaluated. The guidelines available for addressing accessibility (Chisholm et al., 1999; Ahonen, 2003) when tailored to mobile contexts can be used as means for both assurance and evaluation of accessibility of mobile applications. However, guidelines tend to be more useful for those with an expert knowledge than for a novice to whom they may seem rather general to be of much practical use. The problems in using tools to automatically check for accessibility have been outlined in Abascal, Arrue, Fajardo, Garay, and Tomás (2004). Patterns (Alexander, 1979) are reusable entities of knowledge and experience aggregated by experts over years of "best practices" in solving recurring problems in a domain including mobile applications (Roth, 2002; Van Duyne, Landay, & Hong, 2003). They are relatively more structured compared to guidelines and, if represented adequately, provide better opportunities for sharing and reuse. There is, however, an associated cost of learning and adapting patterns to new contexts.
- **Inspections:** Inspections (Wiegers, 2002) are a rigorous form of auditing based upon peer review that can address quality concerns for tiers T1, T2, and most of T3, and help improve the accessibility of mobile applications. Inspections could, for example, use the guidelines and decide what information is and is not considered "comprehensible" by consumers at-large, or whether the choice of labels in a navigation system enhances or reduces readability. Still, inspections, being a means of static verification, cannot completely assess interoperability, performance, reliability, or robustness. Furthermore, inspections do involve an initial cost overhead from training each participant in the structured review process, and the logistics of checklists, forms, and reports.
- **Testing:** Some form of testing is usually an integral part of most development models of mobile applications (Nguyen, Johnson, & Hackett, 2003). However, due to its very nature, testing addresses quality concerns only at of some of the tiers (T1, subset of T2, subset of T3). Interoperability, performance, reliability, and robustness would intimately depend on testing. Unlike inspections, tool

support is imperative for testing. Therefore, testing *complements* but does not replace inspections.

- **Metrics:** In a resource-constrained environment of mobile devices, efficient use of time and space is critical. Metrics (Fenton & Pfleeger, 1997) provide a quantitative means for making qualitative judgments about quality concerns at technical levels. There is currently limited support for metrics for mobile applications in general and for their accessibility (Arrue, Vigo, & Abascal, 2005) in particular. Any dedicated effort of deploying metrics for accessibility measurement would inevitably require tool support, which at present is lacking.
- **Tools:** Tools that have help improve quality concerns at all tiers. For example, tools can help report violations of accessibility guidelines, or find non-conformance to markup or scripting language syntax. However, at times tools cannot address some of the stylistic issues (such as an "optimal" distance between two text fragments that will improve readability) or semantic issues (like semantic correctness of a resource included in a mobile application). Therefore, the use of tools as means for automatic accessibility evaluation should be kept in perspective.

Decision Support

A systematic approach to a mobile application must take a variety of constraints into account: organizational constraints (personnel, infrastructure, schedule, budget, and so on) and forces (market value, competitors, and so on).

A producer would need to, for example, take into consideration the cost of an authoring tool vs. the accessibility support it provides; since complete accessibility testing is virtually impossible, determine a stopping criteria that can be attained within the time constraints before the application is delivered; and so on.

Indeed, the last column of Table 1 acknowledges that with respect to any assurance and/or evaluation, and includes feasibility as an all-encompassing consideration on the layers to make the framework practical. There are well-known techniques such as analytical hierarchy process (AHP) and quality function deployment (QFD) for carrying out feasibility analysis, and further discussion of this aspect is beyond the scope of this chapter.

FUTURE TRENDS

Much of the development of mobile applications is carried out on the desktop. The tools in the form of software development toolkits (SDK) and simulators such as Nokia Mobile Internet Toolkit, Openwave Phone Simulator, and NetFront Mobile Content Viewer assist in that regard. However, explicit support for accessibility in these tools is currently lacking.

The techniques for accessibility for mobile technologies such as XHTML Basic/XHTML Mobile Profile (markup of information) and CSS Mobile Profile (presentation of information) would be of interest. This is especially an imperative considering that the widely used traditional representation languages such as Compact HTML (cHTML), an initiative of the NTT DoCoMo, and the Wireless Markup Language (WML), an initiative of the Open Mobile Alliance (OMA), have evolved towards XHTML Basic or its extensions such as XHTML Mobile Profile.

Identification of appropriate CDs, and an evaluation of the aforementioned languages for presentation or representation of information in a mobile context with respect to them, would also be of interest.

As mobile applications increase in size and complexity, a systematic approach to developing them arises. Indeed, accessibility needs to be a part of the *entire* lifecycle of a mobile application—that is, in the typical workflows of planning, model-

ing, requirements, design, implementation, and verification and validation. To that regard, integrating accessibility into "lightweight" process methodologies such as Extreme Programming (XP) (Beck & Andres, 2005) that is adapted for a systematic development of small-to-medium scale mobile applications would be useful. A similar argument can be made for the "heavyweight" case, for example, by instantiating the Unified Process (UP) (Jacobson, Booch, & Rumbaugh, 1999) for medium-to-large scale mobile applications.

Finally, a natural extension of the issue of accessibility is to the next generation of mobile applications, namely mobile applications on the semantic Web (Hendler, Lassila, & Berners-Lee, 2001). The mobile applications for the Semantic Web present unique accessibility issues such as inadequacy of current searching techniques (Church, Smyth, & Keane, 2006) and a promising avenue for potential research.

CONCLUSION

This chapter takes the view that accessibility is not only a technical concern, it is also a social right. In that context, the issues of credibility and legality are particularly relevant as both are at a higher echelon (social level) than accessibility within the semiotic framework.

Credibility is considered to be synonymous to (and therefore interchangeable with) believability (Hovland, Janis, & Kelley, 1953). Indeed, improvement of accessibility is necessary for a demonstration of *expertise*, which is one of the dimensions (Fogg, 2003) of establishment of credibility of the producer with the consumer.

Accessibility is now a legal requirement for public information systems of governments in Canada, the U.S., Australia, and the European Union. The producers need to be aware of the possibility that, as mobile access becomes pervasive in society, the legal extent could expand to mobile applications.

As is well known in engineering contexts, preventative measures such as addressing the problem *early* are often better than curative measures at late stages when they may just be prohibitively expensive or simply infeasible. If accessibility is to be considered as a first-class concern by the producer, it needs to be more than just an afterthought; it needs to be integral to mobile Web engineering.

REFERENCES

Abascal, J., Arrue, M., Fajardo, I., Garay, N., & Tomás, J. (2004). The use of guidelines to automatically verify Web accessibility. *Universal Access in the Information Society, 3*(1), 71-79.

Ahonen, M. (2003, September 19). Accessibility challenges with mobile lifelong learning tools and related collaboration. *Proceedings of the Workshop on Ubiquitous and Mobile Computing for Educational Communities: Enriching and Enlarging Community Spaces (UMOCEC 2003)*, Amsterdam, The Netherlands.

Alexander, C. (1979). *The timeless way of building.* Oxford, UK: Oxford University Press.

Arrue, M., Vigo, M., & Abascal, J. (2005, July 26). Quantitative metrics for Web accessibility evaluation. *Proceedings of the 1st Workshop on Web Measurement and Metrics (WMM05)*, Sydney, Australia.

Beck, K., & Andres, C. (2005). *Extreme programming explained: Embrace change* (2nd ed.). Boston: Addison-Wesley.

Bertini, E., Catarci, T., Kimani, S., & Dix, A. (2005). A review of standard usability principles in the context of mobile computing. *Studies in Communication Sciences, 1*(5), 111-126.

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *Web content accessibility guidelines 1.0.*

W3C Recommendation, World Wide Web Consortium (W3C).

Chisholm, W., Vanderheiden, G., & Jacobs, I. (2000). *Techniques for Web content accessibility guidelines 1.0.* W3C Note, World Wide Web Consortium (W3C).

Church, K., Smyth, B., & Keane, M.T. (2006, May 22). Evaluating interfaces for intelligent mobile search. *Proceedings of the International Cross-Disciplinary Workshop on Web Accessibility 2006 (W4A2006),* Edinburgh, Scotland.

Fogg, B.J. (2003). *Persuasive technology: Using computers to change what we think and do.* San Francisco: Morgan Kaufmann.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The unified software development process.* Boston: Addison-Wesley.

Ghezzi, C., Jazayeri, M., & Mandrioli, D. (2003). *Fundamentals of software engineering* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Fenton, N. E., & Pfleeger, S. L. (1997). *Software metrics: A rigorous & practical approach.* International Thomson Computer Press.

Green, T. R. G. (1989). Cognitive dimensions of notations. In V. A. Sutcliffe & L. Macaulay (Ed.), *People and computers* (pp. 443-360). Cambridge: Cambridge University Press.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The semantic Web. *Scientific American, 284*(5), 34-43.

Hovland, C. I., Janis, I. L., & Kelley, J. J. (1953). *Communication and persuasion.* New Haven, CT: Yale University Press.

Lindland, O. I., Sindre, G., & Sølvberg, A. (1994). Understanding quality in conceptual modeling. *IEEE Software, 11*(2), 42-49.

Nguyen, H. Q., Johnson, R., & Hackett, M. (2003). *Testing applications on the Web: Test planning for mobile and Internet-based systems* (2nd ed.). New York: John Wiley & Sons.

Paavilainen, J. (2002). *Mobile business strategies: Understanding the technologies and opportunities.* Boston: Addison-Wesley.

Van Duyne, D. K., Landay, J., & Hong, J. I. (2003). *The design of sites: Patterns, principles, and processes for crafting a customer-centered Web experience.* Boston: Addison-Wesley.

KEY TERMS

Cognitive Dimensions of Notations: A generic framework for describing the utility of information artifacts by taking the system environment and the user characteristics into consideration.

Delivery Context: A set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user, and other aspects of the context into which a resource is to be delivered.

Mobile Accessibility: Access to the Web by everyone, regardless of their human or environment properties.

Mobile Resource: A mobile network data object that can be identified by a URI. Such a resource may be available in multiple representations.

Mobile Web Engineering: A discipline concerned with the establishment and use of sound scientific, engineering, and management principles, and disciplined and systematic approaches to the successful development, deployment, and maintenance of high-quality mobile Web applications.

Quality: The totality of features and characteristics of a product or a service that bear on its ability to satisfy stated or implied needs.

Semantic Web: An extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

Semiotics: The field of study of signs and their representations.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 9-14, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.17
Propagating the Ideal:
The Mobile Communication Paradox

Imar de Vries
Utrecht University, The Netherlands

ABSTRACT

In this chapter, visions of mobile communication are explored by focussing on idealised concepts surrounding wireless technology. By examining sources on the development, marketing, and use of wireless technology, I contextualise these visions within earlier accounts of ideal communication found in media history and isolate the regularities that are part of these accounts. On close examination, a paradox reveals itself in these regularities, one that can be described as resulting from an uneasiness in the human communication psyche: an unfulfilled desire for divine togetherness that clashes with individual communication needs. While the exact nature of this paradox—innate and hardwired into our brains, or culturally fostered—remains unknown, however, I assert that the paradox will continue to fuel idealised ideas about future communication technology. I conclude with the observation that not all use of mobile technology can immediately be interpreted as transcendental, and that built-in locational awareness balances the mobile communication act.

INTRODUCTION

In October 2003, two British climbers were caught in a blizzard on a Swiss mountain. Rachel Kelsey and her partner Jeremy Colenso, both experienced climbers, were forced to stop behind a large rock at 3000 meters up and wait for the weather to clear. They soon realised that their chances of finding the abseil points in heavy snow were very slim, which meant they were stuck. They texted five friends, one of whom received the message in London at 5 a.m. and immediately notified the rescue services in Geneva. After having to wait another 36 hours because the conditions were too severe for the rescue team to pick them up, the two climbers were finally rescued (Allison, 2003).

The idea that Earth is becoming entirely networked is not new,[1] but the characteristics of mobile communication media have—just as with the first wireless revolution in the beginning of the 20th century[2]—fiercely fuelled the Western notion that through better communication technology all problems of communication will—finally—be solved (Peters, 1999). The "anywhere, anytime, anyhow, anyone" slogan, subliminally attached to every mobile apparatus, opens up a vision of a universally accessible communication space, in which the exchange of information comes to stand for the single most important condition of human progress. More than at any other time in history, this human progress is thought to depend on technological progress.

Rescue stories as those described in the opening paragraph play their part in keeping the idea alive that improvement through technological progress can be measured. The conventional wisdom is that human lives are the single most valuable things we can think of, and if new technology can help save them, it must be treasured. Moreover, if new technology such as mobile telephony makes possible a way of life that is never forsaken of human contact—which therefore is taken as *safe* because there will always be someone who can help—this technology is surely poised to be seamlessly adapted to and integrated in our being (Katz, 2003). Through the remediation of older dominant forms of communication and entertainment technology, the mobile device (or personal digital assistant (PDA) or smart phone, as it is increasingly being called by mobile phone operators and providers) does seem to try to provide an ultimate extension of the natural balance of our sense organs (Levinson, 1997, 2004). Future visions of mobile communication strive for setting up globally accessible meeting points that cater bodiless but perfect interaction, and ultimately for opening up a communication space in which everyone is represented.

This is the inherently human dream of reaching an ideal state, which is cunningly exploited by advertisements, telecom operators, service providers, and the like. We know it is a dream, and we know that we are confronted by it day after day. It will probably haunt us for centuries to come. However, just as "our desire for each other [is] a poor substitute for the primary Eros—and therefore doomed to fail" (Campe, 2000), so are our telecommunication media substitutes for the primary *closeness*—and bound to fail (Vries, 2005). The end result of this is a tragic search for ideal communication through a continuous so-called improvement of communication technologies, a search that will never end.

This chapter will investigate the paradox of this eternal futile quest that we seem to keep embarking on, and will do so by looking at how mobile discourse is framed within quest-ending narratives. By analysing texts from influential scholars such as Pierre Lévy, Howard Rheingold, and Paul Levinson, we will get a grasp of how idealised ideas of the power of new communication technology have pervaded the mobile realm. From there, an attempt is made to single out the recurrent elements in those ideas, whose pervasiveness in our culture will then be examined. Finally, we will look at a few current trends in mobile cooperation techniques that potentially realise certain ideals of communication, albeit in a more pragmatic sense than a sublime one.

UNWIRING THE KNOWLEDGE SPACE

So far, it has mainly been cyberspace and its accompanying access points in the form of personal computers and laptops that are associated with potentially establishing the universally accessible communication realm. However, with the amount of mobile phones growing at an enormous pace,[3] the mobile device has with stunning speed become an essential tool to establish and maintain social networks, as well as managing all kinds of data flows. In this capacity, the device seems perfectly

poised to morph itself into the logical choice of medium when accessing the ever-expanding Über network, the Internet.[4] Wherever, whenever, whatever: downloading or uploading information on the move, sharing news events as they happen with your carefully filtered online friends, checking in on your favourite weblog while lying on the beach; it is already possible and will be even more so when the devices grow into always-on mode. It is at this point where Pierre Lévy's (1997) imaginative *collective intelligence*, located in what he calls the knowledge space, starts to come into its own on an immense scale.

Lévy describes the evolution of earthbound living as being immersed in a succession of four types of space, in which man's identity is determined by the tools and symbols predominantly available in that space (see Table 1). The knowledge space is the fourth—and final—space in which we have come to live, and can best be seen as an informational cloud, a "space of living-in-knowledge and collective thought" (Lévy, 1997, p. 140). An important premise for its existence, growth, and preservation is that people interact with the informational cloud by adding, changing, and retrieving data in whatever way possible.[5] It is to "unfold and grow to cover an increasingly vast and diverse world" (Lévy, 1997, pp. 111-112), ultimately creating a universally accessible information realm. Already, we can recognise this vision in descriptions of the multiple thrusts behind both the Internet and the mobile revolutions, such as those found in marketing publicity and open source movements' manifests alike.

Lévy's hierarchical description of the four levels of space invoke Borgmann's (1999) distinction between information *about* ("my shed can be found next to the willow tree"), *for* ("this is how you build a cathedral"), and *as* reality ("hi, I am Imar's avatar, shall we start exchanging data?"). Both Lévy and Borgmann show us historical shifts that expose a dematerialising transition of the dominant form of information. Although—as is conspicuously evident from the title of his book *Holding on to Reality*—Borgmann warns us for a Baudrillard-like potentially dangerous split between information about/for reality and information as reality, Lévy is not so much concerned about the danger of leaving reality behind, as he frames the knowledge space firmly within the other three spaces: "[It is n]ot exactly an earthly paradise, since the other spaces, with their limitations, will continue to exist. The intention of collective intellect is not to destroy the earth, or the territory, or the market economy" (Lévy, 1997, p. 141).

Paradise or not, Lévy cannot help but describe the knowledge space in terms of "a u-topia ... waiting to be born," "a cosmopolitan and borderless space," "an electronic storm," and "a sphere of artifice shot through with streaks of light and mutating signs" (Lévy, 1997, pp. 138-141), thereby mimicking the eccentric cyberpunk style of William Gibson's *Neuromancer*. There is undeniably a religious element visible in the way Lévy writes about the knowledge space, in which information is to be uncoupled from its static base. This dematerialising movement fits perfectly with the

Table 1. Succession of spaces according to Lévy (1997)

Space	**Identity**
Nomadic Space of Earth	totems, lineage
Territorial Space	territorial inscription
Commodity Space	position within the domains of production and consumption
Knowledge Space	skill, nomadic cooperation, continuous hybridization

transcendental nature of going wireless: liberating things by releasing them from their carriers (be it wires, paper, or the brain) promises more opportunities to interconnect those liberated entities, as they form free-floating nodes in a dynamic network. In the end, in its most radical form, the idea is that every node can be connected to all others, providing instant and perfect transferral of whatever form of data.

As asserted previously, although the knowledge space is self-regulated and its transcendental nature gives rise to the supposition that it might leave the other spaces behind, Lévy holds that it can not be entirely separated from the three preceding spaces. Moreover, in a circular movement—"a return of the earth to itself," as Lévy (1997, p. 141) calls it—the knowledge space connects back to the first space through the recurrence of the nomadic identity. Again, this is a characteristic that is typically found in the mobile device, as has been shown by scholars in recent literature (Gergen, 2003; Kopomaa, 2000; Meyrowitz, 2003). The multiple social roles we possess are called upon in increasingly diverse geographical and social environments when a mobile device is carried along: we can perform parental tasks while at work, we can keep in touch with friends while on vacation, and we can consume entertainment while sitting in classrooms. Slowly, urban design is responding to the diminishing need to build strict and fixed divisions between sites for work, leisure, and family, creating heterogeneous zones in which the individual's social status is defined by the type of communication he or she engages with. The use of mobile technology therefore does not entail a full-circle return to the nomadic in the sense that it forces *us* to change location in order to find more fertile ground, as was the case in Lévy's first earthly space, but it forces our *locations* to adapt to our dynamic modes of being.

The transcendental and nomadic nature of the knowledge space calls for an intricate investigation of the points where it meets other spaces, and of the materiality of these meeting points. Considering the ease with which the mobile device has found its place as the essential data tool, such meeting points, which according to Rheingold (2002) seem to call for a "marriage of bits and atoms" (p. 100) or for us to be able to "click on reality," (p. 95) are set to be facilitated by the smart phones of the future. Or, as we will see in the next section, this is how it is envisioned in idealised ideas of communication.

THE LURE OF THE IDEAL

Although he admits to being utopian, and has subsequently tried to capture the dynamics of the collective intelligence in a formal language in order to make it more visible and tangible, Lévy has been criticised for painting an exaggeratedly pretty picture, ignoring the tough reality of political, economic, social, and other factors that influence the way communication technology is developed, produced, distributed, and used. In the fourth chapter of their book *Times of the Technoculture: From the Information Society to the Virtual Life*, Robins and Webster (1999) accuse Lévy of "promot[ing] and legitim[ising] the prevailing corporate ideology of globalization," and hold that "there is a desperate need for a richer debate of knowledges in contemporary societies — in place of the shallow, progressivist marketing that attaches itself to the cyberculture slogan (and reflects the hegemony of corporate interests)" (Robins & Webster, 1999, pp. 225, 227). In the same chapter, the aforementioned Rheingold receives similar flak for his—supposedly uncritical—belief in the Internet as a means of restoring communities.

However, Lévy and Rheingold are influential writers and are certainly not alone in taking an optimistic and idealised view on the possible contributions new communication technology can make to finally bring people together in an intelligent collective—nor will they be the last. If the years between the launch of the world's first

graphic Internet browser in March 1993 and the crash of the dotcom boom in early 2000 marked the building up of the cyberspace hype, then the subsequent years can be characterised as having been labelled the new and improved mobile or wireless era: countless press releases, research papers, news articles, advertisements, books, radio shows, and television programmes have heralded mobile technology as the ideal solution to many communication problems. Two books I would like to bring to the fore in this respect are *Smart Mobs: The Next Social Revolution* by Howard Rheingold (2002) and *Cellphone* by Paul Levinson (2004), as their structures show interesting similarities with Lévy's (1997) approach—and with it, the same dangerous tendency to overestimate communication technology's power to fulfill longtime ideals of communication.

Comprised of a large series of anecdotal, interview, and travel journal material, *Smart Mobs* intends to uncover the characteristics of the "next social revolution," which is to be cranked up by the new mobile devices that "put the power of instant and ubiquitous communication — literally—within everyone's grasp" (Rheingold, 2002, back cover). Describing an impressive amount of trends, experiments, news reports, and commercial projects within the global realm of mobile telephony and computing, Rheingold shows how "technologies of cooperation" have an inherent tendency to group people together—and where there is a group, there are opportunities to learn, create, or topple over. The well-known (albeit somewhat overused) example of the protest demonstration in the Philippines in 2001, in which more than 1 million people were rallied by text messages to oppose Joseph Estrada's regime, is used by Rheingold as a key argument in describing a pivotal cultural and political moment: the power of mobile, ad hoc social networks is not to be underestimated; it can even influence politics on a momentous scale! To be fair, Rheingold's argument does not hinge upon this example alone; next to three other activist movements, he also mentions the squads of demonstrators that, thanks to mobile coordination, *won* the "Battle of Seattle" during a World Trade Organization meeting in 1999. These *movements*, however, have been minor in impact and longevity, and do not appeal to the imagination as much as the Philippine regime change does. It is therefore that *Smart Mobs* focuses mainly on events and projects that contain a clearly visible potential to change things; after all, what better way is there to show that the social impact of mobile technology is not only measurable, but can also be described in terms of setting in motion an unstoppable voyage towards a better future?

Other examples of what the consequences of ubiquitous mobile communication might be are equally carefully chosen for their provocative nature. Among the phenomena that await us, Rheingold (2002) names WiFi neighbourhoods; wearable computing that makes our environment aware of our presence and can react accordingly; RFID tags that provide contextual information on any object; and swarm intelligence that makes possible useful emergent behaviour. He does his best to convince us of the inherent potential of these things to fundamentally change the way we are living—and does so with an obligatory nod to the possibility that some of those changes might not be as pleasurable as we would like—but fails to go much further beyond stating the mantra *together is good*. The majority of Rheingold's examples, however tangible and useful they may be within their own context, are used to construct a vision of a futuristic world in which the possibility to connect things (people and machines) is most highly rated. To connect is to solve, to evolve, to come closer to the ideal of sublime togetherness.

Levinson's *Cellphone*[7] is another very good example of how opportunistic ideas found in much cyberculture literature have been transferred to the mobile realm. Not wasting any time, the book's subtitle, which is as subtle as it is provocative, already promises to tell us "[t]he story of the world's most mobile medium **and how it**

has transformed everything" (bold in original). Working from within his Darwinian approach to media evolution—only the fittest media persist in the human environment—Levinson holds that "the cellphone has survived a human test," and that the human need it satisfies is "as old as the human species — the need to talk and walk, to communicate and move, at the same time" (Levinson, 2004, p. 13). This need, which "even defines the human species" (Levinson, 2004, p. 13), is satisfied by the mobile device to such an extent that Levinson foresees the end of the digital divide; the rise of new and more honest forms of news gathering and dispersal; and the birth of a smart world.

The most important (and obvious) characteristic Levinson stresses is that the mobile device blurs the boundary between inside and outside, rendering it unnecessary to confine ourselves to brick and mortar rooms when we want to call someone or find information. The consequence of this blurring is that it will enable us to "do more of what we want to do, be it business or pleasure, pursuit of knowledge, details, companionship, love," and that it will make "every place in the world in which a human may choose to thread ... well-read, or 'intelligent'" (Levinson, 2004, pp. 60-61). Dubbing this intelligent world a "telepathic society"—accompanied by the obligatory but hollow disclaimers that "our progress ... will be tough going at times" (Levinson, 2004, pp. 60-61) and that the mobile device not only solves things but generates new problems of privacy as well—Levinson sides with previous visions of emerging all-encompassing intelligence that have proved to be vulnerable to easy critique, including the Noosphere of Teilhard de Chardin (1959), the morphic fields of Sheldrake (1989) and the global brain of Bloom (2000). As we will see in the next section, the recurrence of these ideas is not coincidental.

RESEARCHING REGULARITIES

Clearly, optimistic visions of new futures are often met with scepticism, but this does not stop them from reoccurring through time; especially when new information and communication media find the limelight. To understand why this "almost willful, historical amnesia," as Mosco (2004, p. 118) calls it, occurs, it is necessary to investigate the underlying regularities of such idealised claims, and to map the basic elements that make up those regular elements. By focussing not on a new medium itself—nor on what it is that makes it unique—but on the path that lies before that medium, we can get a detailed view of the moments in time that mark significant contributions to the medium's earlier discourse. This can best be achieved using the so-called media archaeology approach, which aims to prevent historical amnesia by "(re)placing [the histories of media technologies] into their cultural and discursive contexts" (Huhtamo, 1994). Doing so, the emphasis is shifted "[f]rom a predominantly chronological and positivistic ordering of things, centered on the artefact, ... into treating history as a multi-layered construct, a dynamic system of relationships" (Huhtamo, 1994). It is these relationships that can clarify the intricate ways in which idealised regularities in the dynamic communication media discourse may have changed face, but not their core.

Huhtamo proposes to call the regularities *topoi*, or topics, which he defines as "formulas, ranging from stylistic to allegorical, that make up the 'building blocks' of cultural traditions." He stresses that these topoi are dynamic themselves: "they are activated and de-activated in turn; new topoi are created along the way and old ones (at least seemingly) vanish" (Huhtamo, 1994). In other words, topoi are highly political and ideologically motivated. As an example of a topos found in media history, Huhtamo considers the recurrent "panicky reactions" of public being exposed to visual spectacles, and finds

these in illustrations of the Fantasmagorie shows at the end of the 18th century, in reports of the showing of the arriving train in the Lumière brother's *L'Arrivee d'un train a La Ciotat* (1895) and in the stereoscopic movie spectacle *Captain EO* in Disneyland. There is, of course, a danger of over-interpreting historical sources that may well have served another function than to give an accurate account of what actually happened, but this is exactly Huhtamo's point: "unrealized 'dream machines,' or discursive inventions (inventions that exist only as discourses), can be just as revealing as realized artefacts" (Huhtamo, 1994). The Lumière showing may well not have created any panic at all, but it still remains a poignant reference, a media myth that is repeatedly used in numerous books, articles, and essays in which the reception and impact of new media is discussed. Media archaeology tries to expose these dubious but persistent stories, to collect and dust off forgotten elements of a medium's history by looking at discursive connections, however weak those connections may be. By looking at the many levels on which the discursive construction of a communication technology presents itself, media archaeology bridges the revolutionary gaps that are often found in teleological historiographies of that technology.

This archaeological approach has been put to practice by several scholars in recent years,[8] and has so far been successful in revealing and critically analysing media topoi such as the desires for immediacy, presence, liveness, and simultaneity. The most powerful (or overarching) topos, however, is the gnostic longing to transcend earthly life by improving technology, and to create a Universal Brotherhood of Universal Man. This ultimate topos unites every imaginable description of fulfillment, perfection, pureness, and harmony, and can be found in accounts of every communication medium, in every stage of its development, production, distribution, and use. The dream to finally fulfill the ultimate topos through improvement of communication technology can be comprehensively traced through media history, as many scholars (Mattelart, 2000; Mosco, 2004; Peters, 1999) have already shown. As I have written elsewhere, "[w]ireless telegraphy was seen as 'the means to instantaneous free communication'; telephony seemed to promise banishment of distance, isolation and prejudice; radio would pave the way for contact with the dead and television would transform its viewers into eyewitnesses of everything that went on in the world" (Vries, 2005, p. 11). With every development, be it technological, political, economical, or social, the regularities in discursive accounts of older media have been passed on to newer versions, thereby changing form but not essence.

The argument here is that mobile technology fits into a long line of media in which a limited set of regularly used *modes of reflection* determines the discursive domain of media reception. By analysing the discursive construction of mobile technology and comparing it to that of previous communication media, we can get a grasp of the topoi that have flourished or been revived—be it essentially unchanged or in disguised form—and of those that have floundered or been abandoned. Some of the most interesting indicators of these topoi are to be found in rationalisation techniques people use when explaining why they buy mobile phones, or what they are mainly going to use them for. On the surface, these explanations mostly point to very pragmatic reasons. Field study has shown that common justifications for acquiring a mobile phone are business, safety, and security (Palen, Salzman, & Youngs, 2000). On a deeper psychological level, however, these pragmatic reasons can be tied to fears of solipsism, a desire to increase the amount and strength of communication channels in the social network, and a wish for greater control over one's overall connectivity and availability. Just as we have seen in Rheingold's *Smart Mobs*, a need for the potential to increase *togetherness* is expressed in the mobile discourse, reflecting the ultimate topos of ideal communication.

The hints of religious elements present in these uncovered communication ideals is not surprising; just as Ludwig Andreas von Feuerbach stated in the middle of the 19th century that God is the projection of the human essence onto an ideal, so is an ultimate communicative Being One a projection of a human essence onto communication ideals. The religious motifs continue to exist today: authors such as Erik Davis (1998) and David Noble (1997) have written elaborate accounts of how contemporary technological discourses are still undeniably intertwined with religious beliefs, despite the widely held notion that since the Enlightenment these categories have slowly but surely separated. Such is the case with the topos of ultimate togetherness: the fears and desires disseminated by that topos are exponents of a mixture of the autonomous behaviour of the liberated Cartesian subject on the one hand, and a dream of a bodiless sharing of minds, described by Peters (1999) as angelic communication, on the other. This is a deeply paradoxical mixture, however. Angelic communication shows all the hallmarks of a divine togetherness: with no physical borders and direct one-on-one mappings of minds, every entity will ultimately know and be the same. This loss of individuality collides with the search for more control over ones individual connectivity found in the modern subject's autonomous behaviour. Both angelic communication and complete autonomy are idealised opposite poles on the same scale, and will therefore remain forever out of reach.

THINKING THROUGH PARADOX

The crux of the communication paradox can be described as an uneasiness in the human communication psyche, born out of the tension between the desire for ideal communication and the knowledge of never being able to reach that goal. This is not to say that every individual always wants to strive for perfection. Moreover, reaching perfection may not be what would actually be beneficial for human kind, as many dystopian answers to utopian projects, proposals, and literature have shown; there is no room for individuals or deviations in a society that can only function perfectly if every citizen is synchronised in the grand scheme.[9] Still, the paradox holds, as even in dystopian visions the utopian looms; in the end, Armageddon, the ultimate dystopian event, does nothing more than to destroy old structures in order to lay the foundation for a new, perfect one. A similar argument can be made for a dominant part of the communication media discourse: New media strive for the abolishment of old media in order to provide improved togetherness (Bolter & Grusin, 1999).

As we have seen in the previous section, the successive observations that the development phase and subsequent promotion of communication media are almost always framed within idealised expectations, that these are always accompanied by dystopian rebuttals, and that this process of touting and dismissing keeps reoccurring through time, give rise to the assumption that there is a steady undercurrent present, a topos that can be described as an idea of ideal communication that drives humankind to keep searching despite guaranteed failure. The objection to this assumption might be that this process is merely a marketing mechanism, but such a mechanism can only work if it addresses a human longing, one that is sensitive to promises of solving the communication tension.[10] The question, then, is whether the paradoxical attitude towards communication technology is innate, or if it is just a temporary, culturally sustained concept of progress left over from the Enlightenment, which, at some time in the future, is to be replaced by another concept. If it is innate, we will not be able to escape it; if it is not, we might be able to understand how to change or manipulate the structures in which the paradox resides.

To ask the question of innateness is to enter the realm of epistemology, the study of how we

can know the world around us. Until the middle of the 18th century, this field had known two fairly opposed visions: the rationalist and the empiricist view. The rationalist Innate Concept thesis holds that there are some concepts that are already in our minds when we are born, as part of our rational nature. The notion that we can have a priori knowledge, that we have some innate awareness of things we know to be true that is not provided by experience, rests on the premise that the concepts used to construct that knowledge are also innate. Empiricists, however, argue that there are no innate concepts, and that experience alone accounts for the raw material we use to gain knowledge. The most well-known proponent of empiricism, John Locke, wrote that humans are born with a blank mind, a tabula rasa, which is *written onto* by experience. Knowledge, therefore, is not brought to consciousness by experience, but is provided by that experience itself.

This distinction largely disappeared toward the end of the 18th century when the two views were brought together by Emmanuel Kant, who divided reality into the phenomenal world (in which things are what they appear to us to be, and can empirically be known) and the noumenal world (in which things are what they are *in themselves*, and where rationalism rules). According to Kant's transcendental idealism, innate concepts do exist, but only in the noumenal world, where they remain empirically unknowable. Arguably, these innate concepts are philosophical in nature and therefore proof of their existence remains hard to formulate, but this does not mean *innateness* is always metaphysical. For instance, genetic theory, a late 20th century science, claims to provide empirical evidence for the existence of innate mechanisms in cognitive evolution: Human brains are not tabula rasa, but prestructured in specific ways so that they can learn things other organisms can not. While some elements of evolutionary psychology (EP) are highly controversial,[11] it is increasingly accepted that we all come wired with what Chomsky (1957) has called a Language Acquisition Device (LAD): Not only do we possess an innate capacity to learn, but also an innate set of universal language structures. This means that, independent of our social, cultural, or ethic environment, we already *know* how language works before we even speak it. It is on this level that we have to look for the communication paradox if we believe it to be innate: Are we in some way hard-wired to have a tendency to long for goals that are impossible to reach, to be fascinated by things that are and yet are not? Is there some sense of divine togetherness that we come programmed with, that is at some point in time to be fulfilled but keeps slipping away when we think we come close? The long history of trying to overcome distance and time through the use of media makes a strong argument for such a claim, especially when looking at the positivist discourse this search is usually framed in.

Seen this way, the topos of increased togetherness through idealised communication is but one manifestation of a central paradoxical tendency generated by our brains, albeit one of the most dominant. An imaginative account of how this paradoxical core pervades all aspects of life is found in Hofstadter's (1979/1999) *Gödel, Escher, Bach: An Eternal Golden Braid*. In the new preface in the 20th anniversary edition Hofstadter stresses the paradoxical motive for writing the book by stating that he had set out to "say how it is that animate beings can come out of inanimate matter" (Hofstadter, 1979/1999, p. xx). Introducing so-called strange loops, instances of self-reference that can often lead to paradoxical situations, Hofstadter shows that these loops can not only be found in math, perspective drawings, and music, but also—and this is his main argument—in the very essence of conscious existence itself. Without paradoxes, it seems, life as we know it could not exist. A similar argument is made by Seife (2000), who explores our uneasy relationship with zero and infinity in *Zero: The Biography of a Dangerous Idea*. Innocent as they might seem, in many situations in many times the notions of zero and

infinity have been difficult to grasp, use, and explain; to such an extent even that people have equated them with the work of God and ignored them as not allowed by God at the same time. It was through the use of zero and the infinite that Zeno could create his paradoxical race, in which Achilles never overtakes the tortoise, and it is zero and the infinite that plague contemporary physicists' current understanding of our universe. Opposite poles that invoke as well as fight the paradoxical will always be with us, because we are born out of a paradox, Seife concludes.

EP is a relatively young field, and as such has not yet found very stable ground. The argument that there is a universally active module in our brain that triggers—or is even responsible for—a life with paradoxes is therefore to be very cautiously approached. As asserted previously, it may well be that our paradoxical attitude towards communication is not the manifestation of an innate concept, but of a culturally constructed one. A helpful nongenetic argument for the paradoxical inclination is found in existentialist theories, especially in Heidegger's treatment of *Gelassenheit* (releasement) and Sartre's description of *mauvaise foi* (bad faith). Whereas the former concept deals with fully accepting one's Being-in-the-world as something that has no intrinsic goal or pregiven content, as something that can only receive its significance through the meaning one chooses to give to it, the latter is the result of *not* accepting the open-ended nature of our existence, of continuously asking "why"? and trying to find the answer outside of one's own will. Such a denial of things-as-they-are and things-as-they-happen actively feeds and sustains a two-pole system, in which paradoxes reside: There is no coincidence when everything happens for a reason, and there is no sense when everything is contingent. People with bad faith—and there are a lot, according to Sartre—often face and cannot accept the most fundamental paradox: Sometimes things are just what they are, even when they are not.

Now all these observations may seem a far cry from our day-to-day experience of using mobile phones, but whenever we transfer any information in any way we are positioned as a node in a communication network, one that exists foremost because we as humans seek contact. We hope and strive for this contact to be instantaneous, clear, under control, and ideal, even when we want to mislead or deceive the other person; if we manage to use the medium and channel in such a way that it serves our intent, the contact has been ideal for its purpose. The desire is for a technologically induced complete fulfillment, which is omnipresent in mobile discourse. There is never any certainty about having reached this ideal state, however, as we have seen. The communication paradox makes sure that something always gets in the way of pure experience.

THE RETURN OF LOCATION

In light of this knowledge, the best way we can act, as Peters (1999) also argues, is to embrace the impossibility of ideal communication and make do with what forms of communication we *can* realise. The transcendental nature of wireless technology may at times lure us into thinking we have come close and need just a little push in the right direction, but this would be like chasing a mirage. What then are the elements of more appropriate pragmatic approaches to using new communication technology, ones that defy the urge to hand out idealised promises? Some interesting trends in recent innovative wireless concepts show that the independency of locality, the characteristic that seemingly constitutes the *essence* of mobile telephony, can be turned on its head. Where the most pure form of communication is equated with a bodiless presence and is therefore situated in a nondescriptive *anywhere*, part of the current crop of wireless projects inject exactly this sense of locality into the mobile communicative act. The resulting location based services (LBS) are put to use in a variety of ways: backseat games

that merge road context with virtual content (Brunnberg & Juhlin, 2003), portable devices that support the tourist experience by supplying on the spot information (Brown & Chalmers, 2003), systems that provide virtual annotation of physical objects (Persson, Espinoza, Fagerberg, Sandin, & Cöster, 2002), and mobile phone applications that can *sense* the proximity of people on your buddy list (Smith, Consolvo, Lamarca, Hightower, Scott, Sohn, et al., 2005). Of course, all these projects in some way reflect a drive towards making things easier, quicker, better, or simply more enjoyable, and therefore do not completely escape paradoxical idealised thinking, but they do not ostentatiously try to transcend our present experience of communication by denying its inherent grounding in lived space and time.

Another area where mobile phones are undeniably making a difference without having to resort to metaphysical musings is in developing countries. By leapfrogging older communication technology—in most cases this concerns landlines that had been too expensive to be installed nationwide—mobile technology is used to quickly set up cheap networks, thereby facilitating measurable boosts to local economies and communities. The mobile networks do not instantly connect all parts of a country, but remain localised in existing urban or rural environments. This localisation is further strengthened by the fact that, less tempted to use the mobile device to mix different social locales into one heterogeneous zone, as is more the case in Western metropolitan areas, people in these developing countries tend to see the mobile more as a landline that happens to be wireless. If there would have been a landline the impact would have largely been the same, something communication theorist Jonathan Donner (2003) concurs with. He conducted several field studies in Rwanda, and found that the use of mobile phones by Rwandan entrepreneurs enhanced their ability to do business, but also to satisfy their emotional and intrinsic needs. This is mostly due to the mere presence of a communication channel, and not to the mobile's intrinsic essence. Again, the underlying idealised implication is that appointments, deals, and transactions can occur faster and more streamlined when people are increasingly brought together in whatever way, but in cases such as those in Rwanda the results of introducing wireless technology are clearly visible and do not remain mostly theoretical.

CONCLUSION

With the global proliferation of mobile communication devices, a reinvigorated sense of ubiquitous connection possibilities has emerged. Covering large parts of the Earth, a networked informational skin seems set to revolutionise our way of living. The key new paradigm that is stressed in this "mobilisation" of the world is the ability to tap into an all-encompassing knowledge space, thereby making information addition, retrieval, and communication virtually instantaneous. The fundamental driving force behind this endeavour can be ascribed to a desire for establishing connections to everyone or everything in whatever way possible, a bodiless omnipresence. The radical consequences of this—almost angelic—desire are affecting traditional modes of interaction such as dialogue and dissemination.

This dream of idealised communication is subconsciously stressed by the dominant image of wireless communication that is found in advertisements, press releases, books on social change, government policies, and the like. Promises that things will get better, fuel our impatience when contemporary technology fails to deliver. In other words, the desire for ideal communication itself is part of a paradoxical system found in all layers of our existence. The dream can never be realised, and will therefore continue to recur through time. Whether we will be able to change our attitude towards this strange loop depends on its nature: If it is hard-wired into our brains, we will have to live with the paradox forever. If it is not, who

knows, we might come to see mobile communication for exactly what it is, a specific but not definitive "Being" of communication.

REFERENCES

Allison, R. (2003, October 7). Climbers on Alpine ridge rescued by text message. *The Guardian*. Retrieved May 16, 2005, from http://www.guardian.co.uk/uk_news/story/0,3604,1057271,00.html

Analysys Press Office (2005, May 5). Mobile penetration in Western Europe is forecast to reach 100% by 2007, says Analysys. *Analysys*. Retrieved May 16, 2005, from http://www.analysys.com/Articles/StandardArticle. asp?iLeftArticle=1897

Bloom, H. K. (2000). *The global brain: The evolution of mass mind from the big bang to the 21st century*. New York: Wiley.

Bolter, J. D., & Grusin, R. A. (1999). *Remediation: Understanding new media*. Cambridge, MA: MIT Press.

Borgmann, A. (1999). *Holding on to reality: The nature of information at the turn of the millennium*. University of Chicago Press.

Brown, B., & Chalmers, M. (2003). Tourism and mobile technology. In K. Kuutti & E. H. Karsten (Eds.), *Proceedings of the 8th European Conference on Computer Supported Cooperative Work* (pp. 335-355). Dordrecht: Kluwer Academic Press.

Brunnberg, L., & Juhlin, O. (2003). *Movement and spatiality in a gaming situation: Boosting mobile computer games with the highway experience*. Interactive Institute. Retrieved May 16, 2005, from http://www.tii.se/mobility/Files/BSGFinal.pdf

Campe, C. (2000). *Spheres I: An introduction to Sloterdijk's book*. Goethe-Institut Boston. Retrieved May 16, 2005 from http://www.goethe.de/uk/bos/englisch/Programm/archiv/2000/enpcamp100.htm

Chomsky, N. (1957). *Syntactic structures*. The Hague, The Netherlands: Mouton.

Clark, T. (2004). *Mobile communications and the wireless Internet: The Japanese experience*. Receiver 11. Retrieved May 16, 2005 from http://www.receiver.vodafone.com/11/articles/pdf/11_02.pdf

Davis, E. (1998). *Techgnosis: Myth, magic, mysticism in the age of information*. New York: Harmony Books.

Day, R. E. (1999). The virtual game: Objects, groups, and games in the works of Pierre Lévy. *The Information Society, 15*(4).

Donner, J. (2003). What mobile phones mean to Rwandan entrepreneurs. In K. Nyíri (Ed.), *Mobile democracy: Essays on society, self and politics* (pp. 393-410). Vienna: Passagen.

Gergen, K. (2003). Self and community in the new floating worlds. In K. Nyíri (Ed.), *Mobile democracy: Essays on society, self and politics* (pp. 103-114). Vienna: Passagen.

Hofstadter, D. R. (1999). *Gödel, Escher, Bach: An eternal golden braid*. New York: Basic Books. (Original work published 1979)

Huhtamo, E. (1994). *From kaleidoscomaniac to cybernerd: Towards an archeology of the media*. De Balie Dossiers Media Archaeology. Retrieved May 16, 2005 from http://www.debalie.nl/dossierartikel.jsp ?dossierid=10123&articleid=10104

Katz, J. E. (2003). *Machines that become us: The social context of personal communication technology*. New Brunswick, NJ: Transaction Publishers.

Katz, J. E., & Aakhus, M. A. (2001). *Perpetual contact: Mobile communication, private talk, public performance*. Cambridge, UK; New York: Cambridge University Press.

Kopomaa, T. (2000). *The city in your pocket: Birth of the mobile information society.* Helsinki, The Netherlands: Gaudeamus.

Levinson, P. (1997). *The soft edge: A natural history and future of the information revolution*. London; New York: Routledge.

Levinson, P. (2004). *Cellphone*. New York: Palgrave Macmillan.

Lévy, P. (1997). *Collective intelligence: Mankind's emerging world in cyberspace.* New York: Plenum Publishing Corporation.

Malik, K. (1998, December) The Darwinian fallacy. *Prospect, 36*, 24-30.

Mattelart, A. (2000). *Networking the world, 1794-2000.* Minneapolis: University of Minnesota Press.

Medosch, A. (2004). Not just another wireless utopia. *RAM5.* Retrieved May 16, 2005, from http://www.rixc.lv/ram/en/public07.html

Mosco, V. (2004). *The digital sublime: Myth, power, and cyberspace*. Cambridge, MA: MIT Press.

Noble, D. F. (1997). *The religion of technology: The divinity of man and the spirit of invention.* New York: A.A. Knopf.

Palen, L., Salzman, M., & Youngs, E. (2000). Going wireless: Behavior and practice of new mobile phone users. In W. A. Kellogg & S. Whittaker (Eds.), *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (pp. 201-210). New York: ACM Press.

Persson, P., Espinoza, F., Fagerberg, P., Sandin, A. & Cöster, R. (2002). GeoNotes: A location-based information system for public spaces. In K. Höök, D. Benyon, & A. Munro (Eds.), *Designing information spaces: The social navigation approach* (pp. 151-173). London; New York: Springer.

Peters, J. D. (1999). *Speaking into the air: A history of the idea of communication*. University of Chicago Press.

Rheingold, H. (2002). *Smart mobs: The next social revolution.* New York: Perseus Publishing.

Robins, K., & Webster, F. (1999). *Times of the technoculture. From the information society to the virtual life.* London; New York: Routledge.

Seife, C. (2000). *Zero: The biography of a dangerous idea.* New York: Viking.

Sheldrake, R. (1989). *The presence of the past: Morphic resonance and the habits of nature*. New York: Vintage Books.

Smith, I., Consolvo, S., Lamarca, A., Hightower, J., Scott, J. Sohn, T., et al. (2005). Social disclosure of place: From location technology to communication practices. In H. W. Gellersen, R. Want, & A. Schmidt (Eds.), *Proceedings of the 3rd International Conference on Pervasive Computing* (pp. 134-141). London; New York: Springer.

Standage, T. (1998). *The Victorian Internet: The remarkable story of the telegraph and the nineteenth century's on-line pioneers*. New York: Walker and Co.

Teilhard de Chardin, P. (1959). *The phenomenon of man.* New York: Harper.

Vries, I. de (2005). Mobile telephony: Realising the dream of ideal communication? In L. Hamill & A. Lasen (Eds.), *Mobiles: Past, present and future*. London; New York: Springer.

ENDNOTES

1. See Standage (1998) for a comparison of the telegraph age with the rise of the Internet.
2. See Medosch (2004) for an account of how both wireless eras are very similar in the way the technology was received.

[3] Mobiles in Europe are predicted to exceed Europe's population in 2007 (Analysys Press Office, 2005).

[4] See Clark (2004) for an account of how "educational policy, peer pressure, and most importantly, soaring use of internet-enabled mobile handsets" drive young people in Japan to use mobile phones instead of computers when sending and receiving e-mail.

[5] A fitting current example of an implementation of such a cloud would be Wikipedia, which thrives on user input and moderation. Other methods of knowledge storage and retrieval such as Google and archive.org rely on algorithms and filters, which makes them more archival than dynamic modes of knowledge preservation.

[6] See http://www.aec.at/en/festival2003/wvx/FE_2003_PierreLevy_E.wvx for a Webcast of his lecture at the 2003 Ars Electronica conference, in which he presented the system of this formal language.

[7] Levinson prefers to call the device a *cell phone* instead of a *mobile phone*, because "[it] is not only mobile, but generative, creative." On top of that, it "travels, like organic cells do," and it "can imprison us in a cell of omni-accessibility" (Levinson, 2004, p. 11). I tend to use *mobile device*, as this category includes not only the mobile (or cell) phone, but also smart phones and PDAs.

[8] Huhtamo names Tom Gunning, Siegfried Zielinski, Carolyn Marvin, Avital Ronell, Susan J. Douglas, Lynn Spiegel, Cecelia Tichi, and William Boddy (Huhtamo, 1994).

[9] Eager to show that a collective intelligence does not mean a loss of individuality, Lévy acknowledges that it is important to ask, in Day's words, "how we can pass from a group mentality characterised by a modern notion of the mass (and with that, mass broadcasting) to a collective intelligence wherein persons may remain individual and singular" (Day, 1999, p. 266).

[10] Claims that support the idea of a universal disposition towards what mobile communication is supposed to be about can be found in Katz and Aakhus (2001).

[11] Malik (1998) criticises EP because it can be used to explain sexual and racial discrimination as "biologically meaningful." Because our genes have not been able to keep up with cultural evolution, the EP argument goes, we are "stone age men in a space age world," and therefore cannot help but to exhibit hunter-gatherer behaviour. Malik claims that this would completely deny the fact that culture has evolved out of natural selection too, and that we consciously make choices.

This work was previously published in Information Communication Technologies: Concepts, Methodologies, Tools, and Applications, edited by C. Van Slyke, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.18
Portals Supporting a Mobile Learning Environment

Paul Crowther
Sheffield Hallam University, UK

Martin Beer
Sheffield Hallam University, UK

INTRODUCTION

Mobile computing gives a learner the ability to engage in learning activities when and where they wish. This may be formal learning, where the learner is a student enrolled on a course in an institution, or informal learning, where they may be engaged in activities such as a visit to an art gallery. This entry emphasises the importance of portals to this learning environment, using the MOBIlearn project as an example.

The MOBIlearn project intends to develop software that supports the use of mobile devices (smartphones, PDAs, Tablet PCs, and laptops with wireless network connection) for various learning scenarios, including noninstitutional learning. (MOBIlearn, 2005)

The project has two primary objectives:

- Develop a methodology for creating mobile learning scenarios and producing learning objects to implement them.
- Develop the technology to deliver the learning objects to users via mobile computing devices such as personal digital assistants, smart phones and tablet computers.

The pedagogic aim of the system is to provide users with the ability to engage in formal, nonformal and informal learning in a personal collaborative virtual learning environment. To this end four scenarios were used as the basis of developing the requirements for the system. These were a formal university course and a related orientation activity, a nonformal health care scenario and an informal scenario based around museums and galleries.

The philosophy behind the MOBIlearn system is that it provides a set of interoperable services.

Services should be able to communicate asynchronously using unstable communication channels (MOBIlearn, 2005). The primary component of the system is the Main Portal component. Central to the Main Portal component was the Portal Service (PO_POS) that represents the single access point for the user to all the services provided by the MOBIlearn system. As well as the Portal Service there are six other services that make up the Main Portal component.

PORTALS AND MOBILE COMPUTING ENVIRONMENTS

The scenarios used to develop the MOBIlearn system are all examples of environments supporting knowledge transfer. Portals act as a repository and transfer tool for that knowledge. This concept of a portal as a knowledge repository and transfer tool has been studied within business domains (Fernandes, Raja, & Austin, 2005). It is also relevant in a learning environment. In MOBIlearn, the users have an online presence and can engage in collaboration that can range from formal to informal. They can access formal content, but also develop their own.

For example, in the MOBIlearn health care domain, one of the main objectives is the sharing of tacit knowledge. Users can discuss case studies, and alternative approaches to specific problems can be evaluated and documented. This is then used and extended in future case studies. In this environment, individual health workers can use the system to advanced their skills, and in a "live" incident, use it for reference and indeed call for backup.

The formal learning domain exemplified by the MBA (Master of Business Administration) expands on existing teaching portals to deliver course material and facilitate individual and collaborative learning. In this scenario, the novel aspect is customising delivery to a variety of mobile devices in use simultaneously in the same course. The system uses the learners profile to deliver an appropriate view of the material.

Both of these applications require a secure access to the portal. In the case of the MBA, there is a fee involved. In the health care scenario, there is an initial requirement that it be restricted to a specific institution. Also in the health care environment, a supervisor would take responsibility for maintaining content and moderating some of the collaborative activities. However, it was thought inappropriate for users who were not health care workers to have access. In both the MBA and health care environments there is a need for providing trusted interactions between learners and providers (Kambourakis, Kontoni, Rouskas, & Gritzalis, 2005).

In the museum domain, the majority of mobile users are engaged in informal learning. The traditional support tool in a museum or gallery is the audio guide. This provides more detailed information about an artefact an individual is interested in. The art gallery, TATE Modern, has introduced a PDA-based multimedia guide, but the devices were loaned by the museum and did not allow collaboration between learners (Proctor & Burton, 2003). MOBIlearn extends the application via portals to allow a variety of personal devices to be used and the ability of users to collaborate on topics of mutual interest.

PEDAGOGIC DESIGN IN A MOBILE LEARNING ENVIRONMENT

The pedagogic basis of the system is the learner who interacts with the mobile learning portal to access learning objects and participate in online activities. Each of the test scenarios has its own learning objects. However, all these learning objects need to be delivered in a flexible way to a variety of devices (Stone, 2003). For example, the interface characteristics of a tablet computer are

far different from that of a PDA. One challenge is therefore to deliver the correct interface to a learning object, or oblette, to the mobile device.

There are a variety of ways of delivering learning materials to devices with differing characteristics including reauthoring, transcoding and the functional-based object model (Kinshuk & Goh, 2003). Ideally, an open standard should be used to allow different content providers to make their material available on mobile devices. The approach taken in MOBIlearn is to use reauthoring where page descriptions are held as XML, which is compatible with the standard suggested by Loidl (2005).

The second feature of the environment is that it facilitates communities of learners. In the case of the museum scenarios, the learners are operating in an informal environment motivated by their own interests (Cook & Smith, 2004). The methodology gives them the ability to join a virtual community with interests like their own. The learner is under no obligation to formally join (or leave) the community, and can participate as much or little as they wish. This particular scenario has many features in common with the Virtual Museum of Canada (Soren, 2005), but is also designed to be used in a real museum (the Uffizi Gallery in Florence, Italy being a test site) to give a richer experience than the traditional audio guides.

The health care scenario on the other hand is a nonformal learning environment where a community of practice is being established. The system is designed to deliver training scenarios that can then be discussed and delivered. Learning has no start or end point, and new members can join (and leave) at any time; however, it may be a condition of employment that staff engage with this continuing development. This does contradict some of Ellis et al.'s (Ellis, Oldridge, & Vasconcelos, 2003) criteria for a community of practice; specifically, a voluntary and emergent group. However, if staff engage with the learning environment, a virtual community of practice could develop meeting other criteria including a mutual source of gain.

Finally, there is the MBA scenario, which is based in formal learning, where students use the system to access resources, undertake tasks, and discuss topics with fellow students and academics. There is immersion and presence in the online learning environment. This encourages students to build trust and teamwork (Beer, Slack, & Armitt, 2005). The environment is more constrained, and there is a specific enrolment and end point. Although it is theoretically possible to start and end a course at any time, this does not yet happen.

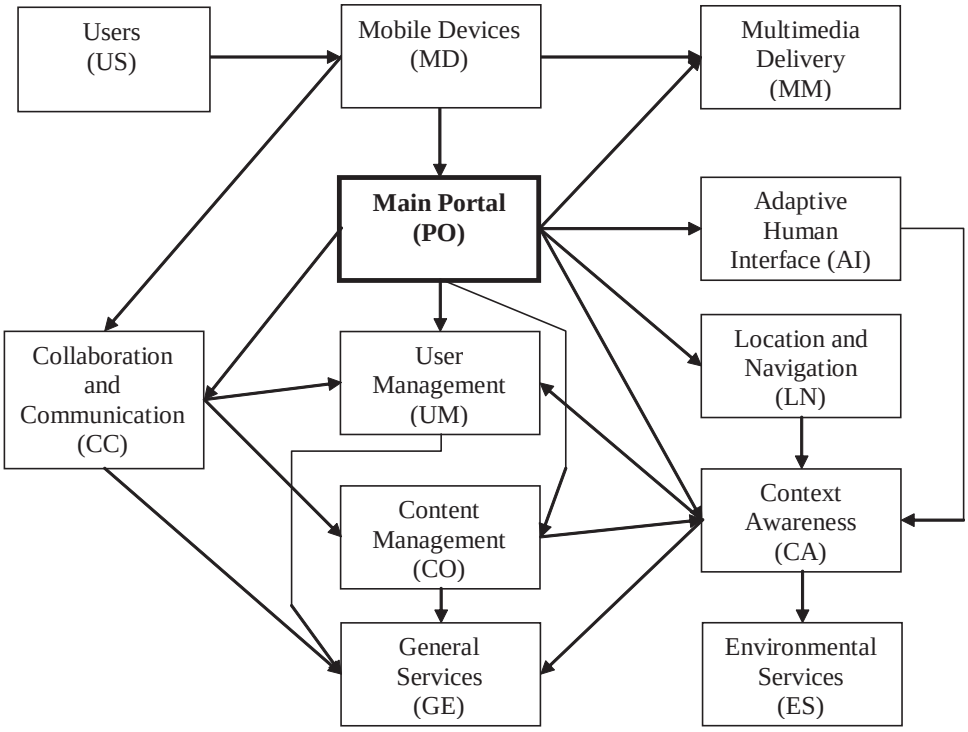
There is a framework common to all three scenarios. This includes the base content. In the case of the museums, this is the information about exhibitions and within that, information about specific exhibits. In the case of health care, there are a series of reference oblettes relating to various diseases and situations. For the MBA, there are the formal course materials. Also, there are the discussion areas, or forums, allowing collaborative learning and providing the foundations for a community of learning and practice to be built. All of these are facilitated through the MOBIlearn portal.

The MOBIlearn portal provides a tool to facilitate collaboration and teamwork. It expands on systems such as OTIS (Occupational Therapy Internet School) (Beer et al., 2005) to provide a framework that can be used in variety of learning situations.

A PORTAL DESIGN IN A MOBILE ENVIRONMENT

MOBIlearn is an example of a personal virtual environment (PVLE) (Xu, Wang, & Wang, 2005) consisting of domain level knowledge from the content provider (for example a museum or univer-

Figure 1. High level component diagram of the MOBIlearn architecture (p. 32 of MOBILlearn Documentation V 2.47)



sity) and a meta level model to allow the learners profile to be matched to the environment and the mobile device they are using.

Figure 1 shows the overall architecture of the MOBIlearn system. Users (US) are users of the system who interact with it using a variety of mobile devices (MD). These are the physical components of the system.

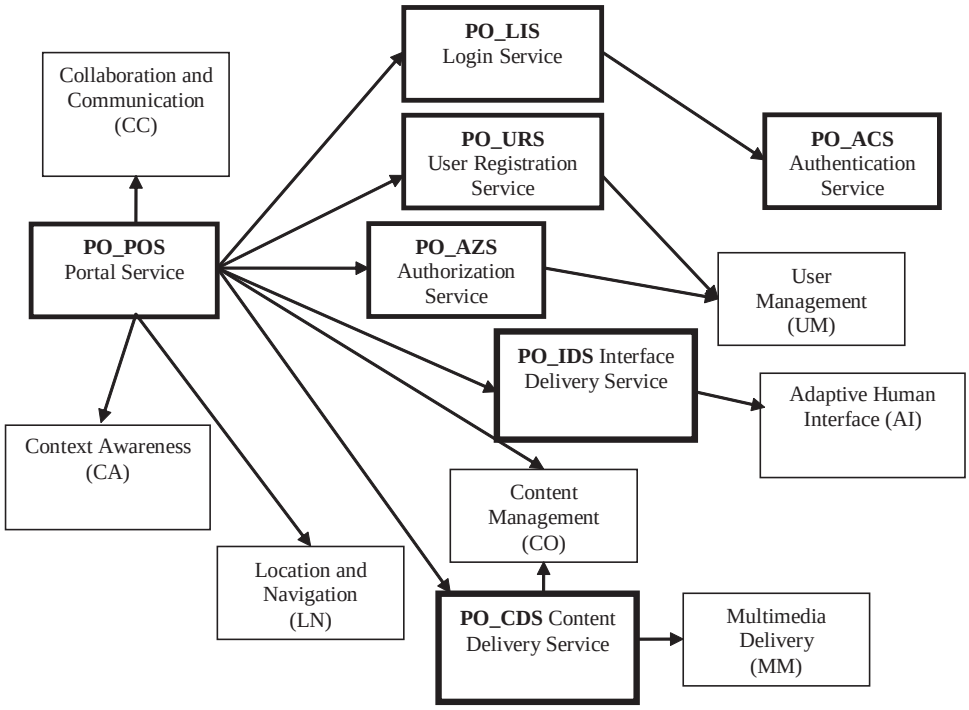
The main portal component is central to the software system and consists of seven services that are detailed in Figure 1, based on the descriptions in the MOBIlearn documentation (2005).

Portal Service (PO_POS)

This service represents the single access point for the user to all the services provided by the MOBIlearn system. It provides the main interface to the system and activates the logging in procedure. Once logged in, this service provides access to other services directly accessible to a user. All but one of the other portal services are called by this service.

A typical session would have a user interacting with the Portal Service. This would first request

Figure 2. Main portal component services (in bold) and their relationship to other components



the logging in procedure detail, which is handled by the Authentication Service (PO_ACS) and Authentication Service (PO_ACS). In the case of a new user, the User registration Service (PO_URS) would be called.

Once the user is logged in, the Authorisation service is called, which in turn uses the User management component of the system. The context of the user has now been established, and the appropriate interface can be delivered for the users device by the Interface Delivery Service (PO_IDS). Content can then be displayed using the Content Delivery Service (PO_CDS). Figure 2 shows the interaction of the Main Portal Components services with each other and the other components of MOBIlearn. The details of the other services are listed.

Login Service (PO_LIS)

This service manages data about users, user profiles, and services, so that authenticated users have access to resources they are authorized to use. The service provides a GUI for the input of user name and password, checks whether the user is authenticated, then allows entry to the system.

Authentication Service (PO_ACS)

The authentication service extends the log in service by verifying the authenticity of the user. It receives the user name and password from the Login Service, then checks if the user can be authenticated using information provided by the User Management component of MOBIlearn. It then returns an authenticated/not-authenticated message to Login Service.

User Registration Service (PO_URS)

If a new user wishes to use the system, they must first register. This service provides functionality for registering a new user. The data provided by the user is used as part of the user profile. The service provides a GUI with a form suitable for collecting user-related data, then activates the creation of a new user profile.

Authorization Service (PO_AZS)

This service is used to determine the level of access an authenticated user should have to resources. The service receives a user's identification data from the Portal Service and the user's profile data from the user management component. Using this information, the Authorization Service checks any requests for services, resources, and operations to see if the user is authorized. It returns an authorized/not-authorized message to the Portal Service.

Content Delivery Service (PO_CDS)

This service delivers the learning objects. It provides a framework for adapting the learning object to the specific context through the request of other correlated services. To do this, it receives identification data related to a selected learning object and retrieves it. The semantic priorities-based adaptation, multirendering-based adaptation are activated, followed by the rendering of the adapted learning object.

Interface Delivery Service (PO_IDS)

The adaptive human interface is delivered by this service. It provides a framework for adapting the Adaptive Interface to the specific context through the invocation of other correlated services. The service receives an XML description of content selected by the user, scenario name, and user identifier. The adaptive interface is then personalised, customised, and rendered on the users device.

CONCLUSION

MOBIlearn is an example of a portal-based mobile learning methodology and delivery system that can be used in a variety of learning situations ranging from formal university courses to informal communities with a common interest. Learners have an online presence and can engage in collaboration and teamwork. The delivery system is designed using a service-oriented structure, at the centre of which is a portal component. The portal is essential to deliver content and allow interaction that is customised to both the learners and their mobile devices.

ACKNOWLEDGMENTS

We acknowledge the EU for financial support through the MOBIlearn project (IST-2001-37440). The views expressed in this chapter are those of the authors, and may not represent the views of the EU

REFERENCES

Beer, M., Slack, F., & Armitt, G. (2005). Collaboration and teamwork: Immersion and presence

in an online learning environment. *Information Systems Frontiers*, *7*(1), 27- 35.

Cook, J., & Smith, M. (2004). Beyond formal learning: Informal community eLearning. *Computers and Education*, *43*, 35-37.

Ellis, D., Oldridge, R., & Vasconcelos, A. (2003). Community and virtual community. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 37, pp. 145-146).

Fernandes, K. J., Raja, V., & Austin, S. (2005). Portals as a knowledge repository and transfer tool: VIZCon case study. *Technovation*, *25*, 1281-1289.

Kambourakis, G., Kontoni, D. N., Rouskas, A., & Gritzalis, S. (in press). A PKI approach for deploying modern secure distributed e-learning and m-learning environments [electronic version]. *Computers and Education.*

Kinshuk & Goh, T. (2003). Mobile adaptation with multiple representation approach as educational pedagogy. *Proceedings of Wirtschaftsinformatik 2003—Medien—Markte - Mobilitat* (pp. 747-763). Heidelberg, Germany.

Loidl, S. (in press). Towards pervasive learning: WeLearn.Mobile. A CPS package viewer for handhelds [electronic version]. *Journal of Network and Computer Applications.*

MOBIlearn. (2005). *The MOBIlearn software documentation* V 2.47. Retrieved September 8 2005 from http://bscw.uni-koblenz.de/bscw/bscw.cgi

Proctor, N., & Burton, J. (2003). Tate modern multimedia tour pilots 2002-2003. *Proceedings of MLEARN 2003: Learning with Mobile Devices* (pp. 127-130). London.

Soren, B. J. (2005). Best practices in creating quality online experiences for museum users. *Museum Management and Curatorship, 20,* 131-148.

Stone, A. (2003). Designing scalable, effective m-learning for multiple technologies. *Proceedings of MLEARN 2003: Learning with Mobile Devices* (pp. 145-153). London.

Xu, D., Wang, H., & Wang, M. (2005). A conceptual model of personalised virtual learning environments. *Expert Systems with Applications, 29,* 525-534.

KEY TERMS

Community of Practice (CoP): A flexible group informally bound by common interests.

Formal Learning: Learning in a structured and controlled environment with fixed, specified learning objectives.

Informal Learning: Learning motivated by personal interest with no specific learning objective and structured by the individual or by an independent informal group.

Learning Portal: A portal that provides a point of access to a virtual learning environment.

MOBIlearn: A system that provides both a methodology and a technology to deliver flexible learning in a mobile environment.

Nonformal Learning: Learning in a formal environment but with no formal learning objectives.

Pedagogy: The activities of educating or instructing or teaching; activities that impart knowledge or skill.

Service-Oriented System: A set of interoperable services, which have been developed independently, that interact to provide the learning environment.

This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 826-830, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.19
Secure Collaborative Learning Practices and Mobile Technology

Hannakaisa Isomäki
University of Jyväskylä, Finland

Kirsi Päykkönen
University of Lapland, Finland

Hanna Räisänen
University of Lapland, Finland

INTRODUCTION

During the past few years, mobile technologies have become common in everyday life. Almost everyone carry some kind of mobile technological equipment with him or her, for example a personal digital assistant (PDA), a mobile phone, a multimedia player, such as an iPod, or a laptop computer. The use of these equipments is not limited only to workplaces, schools or homes. Particularly useful information and communication technologies (ICTs) are in educational settings. Especially wireless networks and laptop computers may promote many useful practices of collaborative learning (Cutshall, Changchit, & Elwood, 2006; Jones, Holmfeld, & Lindström, 2006).

On the one hand, the use of mobile ICTs may also diminish the fluency of studying. With technology both restricting and enabling different ways of action, even small changes in technology may lead to substantial changes in the way it is used in educational settings (Waycott & Kukulska-Hulme, 2003). The use of mobile devices and wireless networks in studying may even reduce communality, social contacts, and collaboration between students instead of increasing and supporting them (Kreijns & Kirschner, 2004). These kinds of deficiencies can restrain users from making good use of otherwise advantageous technology-supported interaction environments. On the other hand, if mobile technology is utilized successfully, it can engender students' feelings of

belonging to a safe virtual community, which helps to construct shared knowledge when members of the community collaborate and apply information and experiences received from others.

In order to successfully implement mobile ICTs for computer supported collaborative learning (CSCL) it is important to obtain information how students take into use mobile technologies in their studying and learning. In particular, different features of knowledge sharing and social usability in the virtual learning environment along with issues of data security within the wireless network become crucial with respect to CSCL that is supported by mobile technology.

This chapter explores the role of mobility and social usability features in a CSCL environment on a wireless campus. In our analysis we found features that either support or diminish the fluency of CSCL.

BACKGROUND

Long research tradition substantiates the benefits of computer supported collaborative methods for learning. The central tenet of CSCL is that a student is part of a studying and learning community that uses ICTs as a mediating tool for social interactions. Koschmann (1996) states the key to successful learning is to support interaction and sharing of experiences by means of technology. Through interaction, students also share distributed cognition (Hutchins, 1995), which means that it is beneficial for collaborative knowledge construction if the members of a community have their own special knowledge. Interpersonal knowledge can only be achieved through the social construction of it and learning can not be separated from its social context (Jones et al., 2006). Computer-supported collaborative learning is successful when students are active, maintain dialogical culture, share convergent goals, and complete tasks together (Dillenbourg, Baker, Blaye, & O'Malley, 1996).

Nowadays, it is essential that students can fluently take mobile technology into use in studying and learning in order to take advantage of the benefits of mobility in CSCL. Mobility, or the movability of devices used in studying and learning, such as laptop computers and wireless network (Luff & Heath, 1998), may benefit CSCL in several ways. Primarily the advantages have been in supporting flexible interaction and continuity between learning contexts. Mobility can also create adaptability and promote accessibility in studying. Finally, mobility can support managing time and learning (Hoppe, Joiner, Milrad, & Sharples, 2003; Roschelle, 2003).

Most importantly, in order to facilitate collaboration and knowledge construction it is of utmost importance that students can easily join mobile learning community, interact with each other, and thus reach a level of critical thinking, mutual understanding and deep learning (Stahl, 2004). For this reason, on a mobile campus students need to fluently interact through laptops in a wireless local area network (WLAN). However, technology supported social interaction does not emerge automatically and the features of the ICTs-based studying environment may even impede it. Therefore, the usability features of the studying environment have to be considered carefully. More precisely, the usability of the studying environment should support social interaction (Kreijns & Kirschner, 2004).

In previous studies, sociability and usability have been considered as two separate concepts: sociability is concerned with social interactions in the online community whereas usability is more focused on the human-computer interface (Souza & Preece, 2004). As a combination of these two viewpoints, social usability is concerned with those features of technology that influence the user's social interaction. We examine social usability in the context of collaborative learning through mobile technology, where it is seen as a prerequisite for taking technology into use and

for being able to participate in a virtual learning community.

Further prerequisites for successful CSCL, such as supporting interaction and sharing of experiences, are met when the students form virtual communities: groups of users that communicate via computer and share common interests, aims and resources (Lazar & Preece, 2002). If the users feel secure and confident about belonging to an online community, they want to participate and share their knowledge (Haythorntwaite, 2002). Therefore, in virtual learning communities social usability refers to the features of technology that facilitate students to take technology fluently into use and to join the community. Moreover, the sense of security emerging from the technology's features promotes social usability by promoting trustworthy interactions (cf. Johnston, Eloff & Labuschagne, 2003). Deficiencies in these factors of usability can restrain students from making good use of otherwise advantageous technology (Girgensohn & Lee, 2002).

MAIN FOCUS OF THE ARTICLE

Although a virtual learning community that works through the net is a social entity, the role of the mediating technology is significant. This paper describes initial empirical results from an ongoing five-year longitudinal study that aims at finding out how the implementation of laptop computers and a wireless network shape the collaborative practices of studying at the University of Lapland, Finland. In the fall of 2004, 582 (85 percent) new students of the university were provided with a laptop computer. At the same time, a wireless local area network (WLAN) was implemented using access points following the standards WLAN 802.11 A/B/G. Support for the standards is built into the same actual devices. The access points do not contain any configuration or information about the network but are connected to switches that guide the operation of the access points. Presently, the network consists of 70 access points, which cover the whole campus area. The wireless network is supposed to be predominantly used with laptops. The computer programs installed in the laptops include ordinary Open Office applications, the Windows operating system, web browsers, instant messaging, e-mail, computer conferencing, and data security applications, such as firewalls and virus detection. (Isomäki, Mattila, Kokkonen, & Päykkönen, 2004.)

In this chapter we examine collaborative mobile learning practices, where social usability is seen as a prerequisite for taking the above-mentioned technology into use, and for being able to participate in a virtual learning community. Our main aspect is that mobility brings about new features in computer-supported collaborative learning, and that social usability should be investigated as a prerequisite for the emergence of successful virtual learning communities on a wireless campus. The main research questions in this study are:

1. What features of mobility promote computer-supported collaborative learning on a wireless campus?
2. How does social usability appear when the students take laptop computers and the wireless LAN into use for computer-supported collaborative learning?

Research Method

After selecting informants by using theoretical sampling, we interviewed during summer 2005 the selected twenty students by using a qualitative interviewing method (Clemmensen, 2004; Kvale, 1996). Before the actual interview questions were presented, informed consent was acquired from the interviewees by asking permission to tape the interview and to write notes with laptop computer. The main points of maintaining confidentiality

and the interviewees' anonymity were also discussed. Every interview included the following themes: (1) experiences of taking the laptop and wireless network into use, (2) data security and usability, and (3) issues related to studying and learning. The interviewer wrote notes during the interview on a laptop computer and presented her own interpretations to the interviewee as the discussion proceeded, and the interviewee corrected the re-searcher's interpretations if needed. As we had interviewed twenty students, it seemed new thoughts or themes no longer appeared. Further, a data driven analysis was carried out (Strauss & Corbin, 1997). The interview data was analyzed in three steps: open, axial, and selective coding. Open coding led to the identification of both the learning technology's mobility features as means to support CSCL and related issues of social usability.

Findings

Wireless networks and laptop computers may promote many useful practices of collaborative learning, but they also may diminish the fluency of studying. According to our analysis, mobility

Table 1. Mobility and social usability features in CSCL

Evidential data examples	Mobility and social usability features in CSCL
"I can write for example in a bus, a hallway or when I'm waiting for something."	Mobility enables flexible management of time and diverse studying settings
"My studying methods have become more versatile."	New studying methods
"[…] we started having all studying material and lecture material as PowerPoint files [...]"	Receiving and seeking studying material through mobile network services flexibly
"If I didn't have friends who know more about computers than I, I would have needed more training in using my laptop."	Receiving or giving peer support instantly through mobile devices
"I have used my laptop in sending email, using Messenger and video conferencing as well as calling Internet phone calls on Skype."	Instant, fast-paced interaction
"Taking the laptop into use was easy, because everything was installed and ready to use."	Taking the laptop into use and using the basic functions is easy
"Software is nowadays so good, that they advise the user during working."	Being able to solve problems by oneself
"[...] friends came to help and advise me."	Peer support in problematic situations
"Touchpad is really annoying; I'd prefer a normal mouse."	Difficulties with laptops: learning to use the touchpad instead of a regular mouse takes time; insufficient duration of batteries; carrying a heavy computer
"During a four-hour lecture the battery of the laptop will last for only about the first half."	
"I was surprised at how heavy the laptop was [...]"	
"Connecting the computer to the wireless network was next to impossible to do."	Taking the WLAN into use is difficult and the wireless network works unreliably
"I trust the functioning of the laptop only when I don't use Bluetooth or WLAN."	Fear of lack of confidentiality in storing information and communicating with mobile technology

appears to promote CSCL in terms of flexible time management, diverse studying settings and versatile new studying methods. These seemed to promote studying motivation by enabling virtual presence in learning situations. Moreover, receiving and seeking studying material through mobile network services flexibly regarding time and place was regarded as one essential benefit of mobile learning. Mobility was also considered to enable instant, fast-paced peer interactions, which were seen as indispensable for overcoming problems in the use of mobile technology. Overall, receiving or giving peer support instantly through mobile devices was largely the most common reason for social interactions.

The most important social usability features emphasize the ease of use of mobile devices and software for instant messaging, earlier experience on computers, and peer support. Features that seem to diminish the fluency of mobile CSCL concentrate on usability problems of the laptop computers and WLAN. The usability difficulties concerned the features of laptops that were different from traditional PCs. For example, learning to use the touchpad instead of an ordinary mouse was mentioned as problematic. Hardware appeared as difficult in use due to insufficient duration of batteries, and the heavy weight of the 'portable' computers. Also the means to connect to WLAN were regarded troublesome. Further, some students were afraid of inadequate data security and privacy protection. The features of mobility and social usability in CSCL that emerged in our analysis are presented in table 1.

FUTURE TRENDS

Mobility brings new features to studying and learning. It can entail new working practices, but if not successfully implemented, it can also rather hinder than promote collaboration. One important advantage of using laptop computers and WLAN is that they make studying more adaptable. Compared to the results of previous studies of CSCL, mobility creates more possibilities for the student to decide when, where and how to study. Students embrace the fact that they have several studying environments within reach and can use new versatile methods in studying. For example, flexibility regarding time and place may previously have meant that a student was able to log on to a network-based studying environment from different localities, whereas now mobility enables logging in, for instance, while sitting on a pier. The traditional campus area is no longer a fixed infrastructure as students set up studying environments where it suits them best. The possibility to receive or seek studying materials through mobile network services further promotes adaptability regarding time and place. Having the material available in an electronic form also means that if a student is unable to attend a lecture, the absence does not automatically mean lagging behind or dropping out.

Another essential feature that promotes collaborative learning is a possibility to instant interaction through mobile devices: receiving or giving peer support, and taking part in instant, fast-paced mobile interaction. Mobility enabled by wireless technology makes it possible to use network-based communication systems, such as e-mail and instant messaging, in places where it has not been possible before. Being a part of a learning community does not necessarily mean being physically present at the university. Students carry the memberships, and the communities they are a part of, with them as they move between places. Virtual presence enables not only continuity between different learning contexts but also an expansion within the time-space scale of learning situations.

Social usability in terms of successful technology-supported interactions puts stress on basic computer skills and especially being able to manage various functionalities of software. Deficiencies in the social usability of technology forges students into face to face interactions

in order to obtain peer support in overcoming problems in technology use. Instant messaging systems, since they are easy to swiftly take into use, are another way for students to give and receive peer support.

Difficulties with the laptops' hardware characteristics, such as the short duration of batteries and poor possibilities to connect to WLAN, may appear as crucial usability deficiencies. Product development is needed to improve the usability of wireless hardware devices. Affordable, lightweight laptops with more durable batteries would add students' eagerness to use their computers in campus area.

Trust in data security and confidential communication in WLAN appear as a critical factor in supporting versatility of mobile computing in CSCL, and is also an important issue in systems design that aims to support social usability. In order to join and study in virtual learning communities, students must be able to trust the mediating technologies. This is a challenge for mobile technologies, since the means of ensuring data security and privacy are essential student demands in mobile CSCL. Without this fundamental prerequisite students cannot evaluate the trustworthiness of the virtual community and, for example, cannot make assumptions whether it is safe to share knowledge with other students. Especially the user interfaces should, on the one hand, convey the information security properties to users and, on the other hand, users should be able to observe and control the system's information security status.

CONCLUSION

In this chapter we have discussed the features that the movability of technological devices and social usability bring to mobile CSCL. In our study we interviewed 20 students, who study on a wireless campus with a laptop computer. According to the results, laptop computers and wireless network can improve the adaptability of studying if certain prerequisites are met. The most important prerequisites are good skills in using mobile devices, software that support instant interactions, light laptops with powerful batteries, easy ways to connect to WLAN, and satisfactory data security and privacy protection. In addition, Universities should organize pedagogical practices to support virtual presence.

The social usability of mobile technology is a critical part of CSCL. The easiness of taking laptops into use and managing basic functions enables students to join virtual learning community. Also the students' ability to overcome technological problems either by themselves, with the guidance of the software's built-in instructions, or with peer support, is a significant factor in successful mobile studying.

These findings serve teachers, tutors and designers who develop computer-supported collaborative learning especially in mobile environments. When the demands for data security, privacy protection and social usability of the mobile ICTs are met, students can fluently mobile CSCL.

REFERENCES

Clemmensen, T. (2004). Four approaches to user modelling —A qualitative research interview study of HCI professionals' practice. *Interacting with Computers, 16*(4), 799–829.

Cutshall, R., Changchit, C., & Elwood, S. (2006). Campus laptops: What logistical and technological factors are perceived critical? *Educational technology & society, 9*(3), 112–121.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.),

Learning in humans and machine: Towards an interdisciplinary science (pp. 189–211). Oxford: Elsevier.

Girgensohn, A., & Lee, A. (2002). Making web sites be places for social interaction. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work* (pp. 136–145). NY: ACM.

Isomäki, H., Mattila, H., Kokkonen, K., & Päykkönen, K. (2004). Uudet opiskelukäytännöt ja mobiili teknologia Lapin yliopistossa. [New learning practices and wireless technology in University of Lapland.] In M. Lehtonen, H. Ruokamo, R. Rajala, H. Jaakkola, J. Multisilta, & J. Viteli (Eds.), *Lapin tietoyhteiskuntaseminaari tutkijatapaamisen abstraktit 2004* [*Abstracts of the Lapland Information Society Seminar Researcher Workshop 2004*] (p. 17). Rovaniemi: University of Lapland.

Haythornewaite, C. (2002). Building social networks via computer networks. In K. Renninger & W. Shumar (Eds.), *Building virtual communities: Learning and change in cyberspace*. New York, NY: Cambridge University.

Hoppe, H.U., Joiner, R., Milrad, M., & Sharples, M. (2003). Guest editorial: Wireless and mobile technologies in education. *Journal of Computer Assisted Learning, 19*(3), 255–259.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Jones, C., Dirckinck Holmfeld, L., & Lindström, B. (2006). A relational, indirect, meso-level approach to CSCL design in the next decade. *Computer Supported Collaborative Learning, 1*(1), 35–56.

Johnston, J., Eloff, J.H.P., & Labuschagne, L. (2003). Security and human computer interfaces. *Computers & Security, 22*(8), 675–683.

Koschmann, T. (1996). Paradigm shifts and instructional technology. In T. Koschmann (Ed.), *CSCL: Theory and practice of an emerging paradigm* (pp. 1–23). Mahwah, NJ: Lawrence Erlbaum Associates.

Kreijns, K., & Kirschner, P. (2004). Designing sociable CSCL environments: Applying interaction design principles. In J.W. Strijbos, P.A. Kirschner, & R.L. Martens (Eds.), *What we know about CSCL* (pp. 221–243). Boston: Kluwer Academic.

Kvale, S. (1996). *InterViews. An introduction to qualitative research interviewing.* Sage: Thousand Oaks.

Lazar, J., & Preece, J. (2002). Social considerations in online communities: Usability, sociability, and success factors. In H. van Oostendorp (Ed.), *Cognition in the Digital World* (pp. 127–151). Mahwah, NJ: Lawrence Erlbaum Associates.

Luff, P., & Heath, C. (1998). Mobility in collaboration. In *Proceedings of CSCW'98* (pp. 305–314). New York: ACM.

Roschelle, J. (2003). Keynote paper: Unlocking the learning value of wireless mobile devices. *Journal of Computer Assisted Learning, 19*(3), 260–272.

Souza, C.S., & Preece, J. (2004). A framework for analyzing and understanding online communities. *Interacting with Computers 16*(3), 579–610.

Stahl, G. (2004). Building collaborative knowing: Elements of a social theory of CSCL. In P. A. Kirschner, R.L. Martens, J.-W. Strijbos (Eds.), *What we know about CSCL and implementing it in higher education* (pp.53–85). Boston, MA: Kluwer Academic.

Strauss, A., & Corbin, J. (Eds.). (1997). *Grounded theory in practice*. Thousand Oaks, CA: Sage.

Waycott, J. & Kukulska-Hulme, A. (2003). Students' experiences with PDAs for reading course

materials. *Personal and Ubiquitous Computing, 7*(6), 30–43.

KEY TERMS

Computer-Supported Collaborative Learning: Studying and learning in knowledge building communities that have convergent goals and whose interaction is supported by computers and networks.

Data Security: Technical means of ensuring that data is kept safe from corruption and that access to it can be suitably controlled. Data security helps to protect personal data.

Education on Virtual Organization: Teaching, studying and learning in network-based environments.

Mobile Technology: The use of communication infrastructures, protocols, and portable devices like laptop computers, personal digital assistants (PDAs) and mobile phones, that enable users to communicate, study or work flexibly in terms of time and place.

Social Usability: A part of usability that is concerned with those features of technology that influence users' social interaction.

Trust: Human trust means that the someone is willing to put him or herself in a position of vulnerability to or risk from another party. Technical trust means that the application does what it is supposed to do and not what it is not supposed to do.

Virtual Community: Group of users that often are widely separated geographically, communicate via computer and share common interests, aims and resources.

This work was previously published in Encyclopedia of Networked and Virtual Organizations, edited by G. Putnik and M. Cunha, pp. 1407-1412, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.20
Gender Difference in the Motivations of Mobile Internet Usage

Shintaro Okazaki
Autonomous University of Madrid, Spain

INTRODUCTION

The rapid pace of adoption of Web-enabled mobile handsets in worldwide markets has become an increasingly important issue for information systems professionals. A recent survey indicates that the number of global mobile Internet adopters is expected to reach nearly 600 million by 2008 (Ipsos-Insight, 2004; Probe Group, 2004), while the number of Internet-connected mobile phones will exceed the number of Internet-connected PCs by 2005 (*The Economist*, 2001). Such drastic convergence of the Internet and the mobile handset has been led by Asian and Scandinavian countries, where penetration has been especially meteoric. For example, roughly 70 million people in Japan, or 55% of the population, have signed up for mobile Internet access, in comparison to 12% in the United States (Faiola, 2004; Greenspan, 2003). Consequently, mobile phones or *Keitai* have been converted into devices for surfing the Internet, and by 2004 monthly mobile spending per consumer exceeded 35 euro.

Much of this success can be traced back to 1999, when NTT DoCoMo introduced the "i-mode" service. i-mode is a mobile service offering continuous Internet access based on packet-switching technology (Barnes & Huff, 2003). Through an i-mode handset, users can access a main micro-browser, which offers such typical services as e-mail, data search, instant messaging, Internet, and "i-menu." The "i-menu" acts as a mobile portal that leads to approximately 4,100 official and 50,000 unofficial sites (NTT DoCoMo 2003). Many such mobile portal sites can thus be considered as a pull-type advertising platform, where consumers can satisfy diverse information needs.

Several researchers have attempted to conceptualize the success of i-mode in comparison to WAP (Baldi & Thaung 2002) and in the light of the technology acceptance model (TAM) (Barnes & Huff 2003). Okazaki (2004) examined factors influencing consumer adoption of the i-mode pull-type advertising platform. However, there is a dearth of empirical research in this area, and

especially in developing a model that captures the specific dimensions of mobile Internet adoption. In this respect, this study aims to propose a measurement scale of consumer perceptions of mobile portal sites.

The present study adopts, as its principal framework, the attitudinal model suggested by Dabholkar (1994). This includes "ease of use," "fun," and "performance" as important determinants of attitude. These are often referred to as "ease of use," "usefulness," and "enjoyment" in, for example, the TAM proposed by Davis (1986; Davis, Bagozzi, & Warshaw, 1989, 1992). The relevant literature suggests that dimensions similar to "ease of use" and "fun" are important antecedents of new technology adoption. For example, Shih (2004) and Szymanski and Hise (2000) found "perceived ease of use" and "convenient," respectively, as important antecedents of online behavior. Likewise, Moon and Kim (2001) found "perceived playfulness" to be a factor influencing WWW usage behavior, similar to the "fun" dimension. However, unlike earlier studies of m-commerce adoption, this study drops the third dimension of the TAM, "usefulness," in favor of "performance," because the former is appropriate only for tangible products, but not relevant for technology-based services (Dabholkar & Bagozzi, 2002). In contrast, "performance" represents a dimension that encompasses the reliability and accuracy of the technology-based service, as perceived by the consumer (Dabholkar, 1994). These three dimensions capture customer perceptions, which would initiate the attitude-intention-behavior causal chain (Davis, 1986).

BACKGROUND

Prior Theories on Technology Adoption

The technology acceptance model has been used to explain online user behavior (Featherman & Pavlous, 2002; Moon & Kim, 2001). Originally, TAM was based on Ajzen and Fishbein's (1980) theory of reasoned action (TRA), which is concerned with the determinants of consciously intended behaviors. TRA has been described as one of the most widely studied models in social psychology (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975). According to TRA, "a person's performance of a specified behavior is determined by his or her behavioral intention (BI) to perform the behavior, and BI is jointly determined by the person's attitude (A) and subjective norm (SN) concerning the behavior in question (Figure 1), with relative weights typically estimated by regression: BI = A + SN" (Davis et al., 1989). Here, BI refers to the degree of strength of one's intention to perform a specified behavior, while A is defined as an evaluative effect regarding performing the target behavior. SN is meant to be "the person's perception that most people who are important to him think he should or should not perform the behavior in question" (Fishbein & Ajzen, 1975).

TAM extends TRA with attempts to explain the antecedents of computer-usage behavior. TAM comprises five fundamental salient beliefs: perceived ease of use, perceived usefulness, attitudes toward use, intention to use, and actual use. Perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance," while perceived ease of use is "the degree to which a person believes that using a particular system would be free of effort" (Davis et al., 1989). Although they are not the only variables of interest in explaining user behavior, perceived ease of use and perceived usefulness have been proven empirically to be key determinants of behavior in a wide range of academic disciplines, such as the learning process of a computer language, evaluation of information reports, and adoption of alternative communication technologies, among others. However, TAM excludes the "influence of social and personal control factors on behavior"

Figure 1. Theory of reasoned action (TRA)

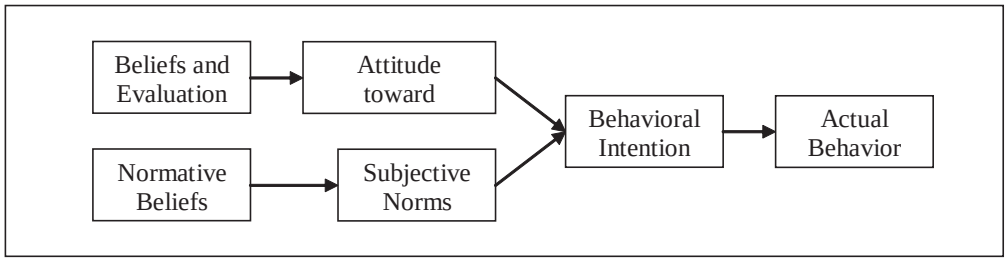
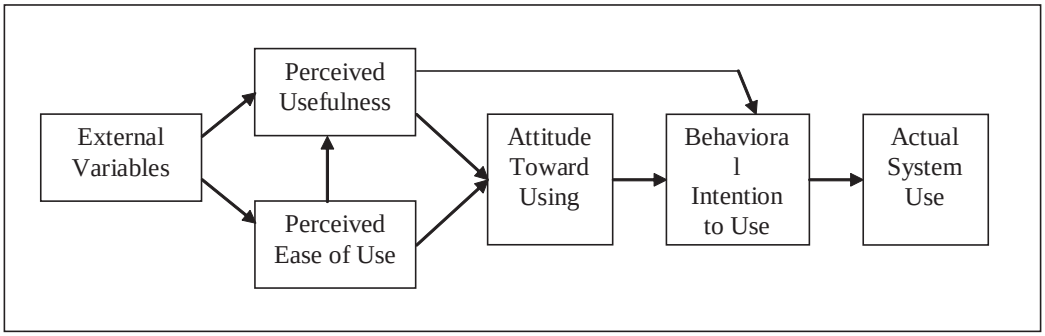


Figure 2. Technology acceptance model (TAM)



(Taylor & Todd, 1995). Consequently, Ajzen and Fishbein (1980) proposed another extension of TRA, the theory of planned behavior (TPB), to account for those conditions in which individuals may not have complete control over their own behavior (Taylor & Todd, 1995).

A key objective of TAM is "to assess the value of IT to an organization and to understand the determinants of that value" (Taylor & Todd, 1995). Hence, much IT research has aimed to enhance companies' effective IT resource management. However, TAM has been expanded to emerging new forms of IT, such as the wired as well as the wireless Internet.

In a pioneering study, Moon and Kim (2001) conducted empirical research on extending TAM to the World Wide Web context. They constructed an extension of TAM based on an individual's intrinsic motivation theory, and found that "perceived playfulness" had a positive effect on individuals' attitudes toward using the WWW and their behavioral intentions to use the WWW.

Furthermore, TAM has been applied to explain the adoption of telecommunication technology, such as telework (Hun, Ku, & Chang, 2003), mobile devices (Kwon & Chidambaram, 2000), and m-commerce services (Pedersen & Ling, 2003). Generally, these studies suggest certain modifications of the original TAM in order to include social influence and behavioral control variables (Pedersen & Ling, 2003). In the following section, we explore the possible extension of TAM to m-commerce adoption.

Gender Differences in Mobile Internet Adoption

Gender has frequently been used as part of the social and the cultural meanings associated with developing marketing strategy via advertising messages, and in market segmentation strategy in particular, because it is easily: (1) identifiable, (2) accessible, (3) measurable, (4) responsive to marketing mix, (5) sufficiently large, and (6) profitable (Darley & Smith, 1995). However, although there has been much research on new technology adoption, little attention has been paid to gender differences in electronic communication. Yang and Lester (2005) argue that "research on gender and CMC has consistently demonstrated that gender inequalities define professional and scholarly electronic communication and that men are over-represented in electronic communities". This is considered a serious lacuna, since evidence has been found of important gender differences in human communication, including advertising (Wolin, 1999).

Our literature review found only one study that examined gender differences in online purchasing behavior. Yang and Lester (2005) conducted a series of studies on purchasing textbooks online at universities, and found that female students at an urban university tended to demonstrate fewer computer/Internet skills than male students, and that their level of skill was a more consistent predictor of purchasing textbooks online: the higher their level of skill, the more likely female students were to buy books online, and the effect of level of skill was greater for female than for male students.

To date, no gender studies of mobile Internet adoption have been reported. However, following Yang and Lester (2005), we may assume that, in learning and accessing wireless Internet with mobile handsets, female users may be less skillful than their male counterparts. For example, in terms of TAM, females may perceive more negatively ease of use, which is one of the essential determinants of attitude toward new technology adoption.

PROPOSED MODEL OF CONSUMER MOTIVATIONS

Although the specific motivations to use *wired* and *wireless* Internet must differ between individuals, the *overall* motivations of online information search may be similar for the two media. Thus, we adopted three primary motivations from prior research on wired Internet adoption: (1) performance, (2) ease of use, and (3) fun. First, Shih (2004) empirically examined online purchasing behavior, and found perceived usefulness to be the major determinant of behavioral intentions to use the Internet, while perceived ease of use is a secondary determinant. We adopt these concepts as performance and ease of use. It has been pointed out that the term *performance* is preferred to *usefulness*, in the case of intangible technology adoption. Second, Moon and Kim (2001) introduced an additional determinant of attitude formation, perceived playfulness or fun, to capture WWW usage behavior. Hence, we propose these three constructs as the principal drivers or motivations of enhanced mobile Internet usage. These constructs are essentially in line with Davis et al.'s (1989) TAM, which has frequently been used to explain and predict user adoption of a new information technology. Hence,

our aim in this study is to examine whether there are any important differences between male and female mobile Internet users in terms of these constructs.

SURVEY METHOD

The survey was conducted via an online questionnaire that was made available in a popular commercial Web site in Japan. There were no restrictions on access, and the survey was open to the public audience. The questionnaire consists of a variety of questions, on general demographics, media usage, habits and spending, motives to use mobile Internet site, and so forth. As an incentive to complete the questionnaire, respondents were given an e-coupon as a reward for their participation. In total, 1,637 responses were obtained.

We assigned four adjectives for each of the three constructs: detailed, reliable, educational, and updated for performance; interesting, appealing, helpful, and killing time for fun; and easy, free, intelligible, and practical for ease of use. In order to identify the importance of each item, we used a dichotomous measure, asking whether respondents perceived a given adjective as describing his or her own perception of the mobile Internet site. For example, if they accessed a mobile Internet site because it seemed "reliable," they marked the answer "yes." In order to conduct statistical analysis, these dichotomous variables were converted into fictitious variables by assigning "1" to "yes" and "0" to "no."

RESULTS

With regard to the demographic composition by gender, the distribution of age and marital status differ little across gender; important differences can be observed in education and occupation. The proportion of people with bachelor or higher degrees is much greater in males than in females. On the other hand, females dominate junior college graduates. With regard to occupation, administrative, managerial, and professional workers are primarily male. A similar tendency can be seen in self-employed and skilled labor, although the magnitude is much less. There are more female workers in services.

To examine the dimensionality of the variables, we first conducted an exploratory factor analysis (EFA) with a principal component method. Although dichotomous variables are not ideal in EFA, fictitious variables are considered acceptable in this usage (Hair, Anderson, Tatham, & Black, 1998). The Varimax rotation was used, while a scree plot was carefully examined. Only variables with eigenvalue greater than 1 were retained. After several attempts using trial and error, we determined a three-factor solution to be the best, in which 12 proposed items were converged. However, as Table 1 shows, some of the items were classified into different constructs. Because of the exploratory nature of the study, we deemed this convergence to be reasonable and acceptable for the subsequent analysis.

Next, a logistic regression was performed with gender as a dependent variable and the importance (existence or absence) of adjective items as independent variables. It was possible to use binary data for both dependent and independent variables, because logistic regression does not require the normality assumption, as multiple regression does (Hair et al., 1998). However, because multicollinearity can seriously distort the results, a diagnostic was carried out via VIF and Tolerance values. Both values for each independent variable ranged from .80 to 1.23, showing no serious presence of multicollinearity.

The results of logistic regression are shown in Table 2. As clearly shown, ease of use plays an important role in separating male and female mobile Internet users. Chi-square tests reveal significant differences between male and female users in terms of easy, killing time, and free. Interestingly, female users are likely to perceive

Table 1. Rotated component matrix

		Component 1	Component 2	Component 3
Performance	Detailed	.695		
	Updated	.639		
	Intelligible	.619		
	Reliable	.449		
Ease of use	Easy		.644	
	Killing time		.589	
	Interesting		.536	
	Free		.440	
	Educational		.410	
Fun	Appealing			.642
	Helpful			.629
	Practical			.437
Total variance		21.1	30.8	40.0
Eigenvalue		2.53	1.17	1.08

Table 2. Logistic regression results

Theoretical constructs	Variables	Mobile site		Internet		Satellite TV		Newspaper		WOM	
Performance	Detailed	-.117		-.005		.209		-.070		.005	
	Updated	.070		-.053		-.117		.202	*	-.107	
	Intelligible	.117		.073		.036		-.025		.042	
	Reliable	-.832	**	-.419	**	-.206		-.022		-.057	
Ease of use	Easy	.253	**	.294	**	.241		.134		.297	***
	Killing time	.311	**	.131		.255	*	-.285	**	-.345	**
	Interesting	.126		-.023		-.210		.098		.206	*
	Free	-.490	*	-.531	***	-.712	**	-.218		-.413	***
	Educational	-.042		-.021		.043		.137		-.318	**
Fun	Appealing	-.566		-.101		.021		-.226		.017	
	Helpful	-.019		.133		-.505		-.138		.518	**
	Practical	.103		.197		-.364	*	.189		.835	***

mobile Internet sites as an easy medium for killing time significantly more than their male counterparts. The opposite is true for free: male users essentially appreciate a mobile Internet site as a free information source. With regard to reliability in performance, male users tend to perceive this item more strongly than female users. Finally, logistic regression was also performed for different media, such as (wired) Internet, satellite TV, newspapers, and word of mouth (WOM). Despite the dangers of oversimplification, it seems that a mobile Internet site exhibits the combined effects of a wired Internet and word of mouth.

IMPLICATIONS

Our findings clearly show that there are important differences between male and female mobile users in terms of motivations to access mobile Internet sites. Specifically, female users are more prone to access a mobile Internet site for spare-time leisure and ease of use, while male users do so for free information. It should be noted that both genders perceive a mobile Internet site as a reliable, updated, and intelligible information source. In comparison with other media, a mobile Internet site is considered to be a combination of Internet and word of mouth. This makes sense because a mobile device is essentially and uniquely characterized as a personalized telecommunication medium, which is accessible only via a mobile telephone.

Given that Japanese mobile Internet services focus on information and entertainment (Okazaki, 2004), the findings of this study may provide useful implications for IT managers. That is, female users are more likely to appreciate a mobile Internet site for its entertainment value, while male users may seek more pragmatic results or outcomes, that is, reliable daily information. For example, typical male white-collar workers may seek daily financial news and replace newspapers with a mobile device as an information source. On the other hand, female users are attracted by more enjoyable usage, which provides an easy escape from daily routine. In this regard, IT managers should be aware of the importance of tailoring the content of mobile Internet according to gender-specific needs and wants.

Limitations

A few limitations should be recognized to make our findings more objective. First, our study examined only 12 adjective items with three proposed constructs. Future research should include a larger variety of items that may be related to consumers' perceptions associated with mobile Internet sites. Second, this study did not specify a type of "mobile Internet site." That is, our findings should be interpreted at most as general evaluations of mobile Internet adoption. Given a rapid proliferation of mobile Internet services, future research should specify the type of mobile Internet site and its benefits as a unit of analysis. Finally, the binary nature of data means that the construct reliability and validity were not established. Any future study should use a semantic differential scale, instead of a dichotomous scale, as a measure, and much effort should be made to improve the reliability indices.

REFERENCES

Ajzen, I. (1985). From intentions to actions: A theory of planned behaviour. In *Action control: From cognition to behaviour* (pp. 11-39). New York: Springer-Verlag.

Darley, W., & Smith, R. (1995). Gender differences in information processing strategies: An empirical test of the selectivity model in advertising response. *Journal of Advertising, 24*(1), 41-56.

Davis, F., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Durlacher. (1999, November). *Mobile commerce report*. Retrieved from http://www.durlacher.com/fr-research-reps.htm

Featherman, M., & Pavlos, P. (2002). Predicting e-services adoption: A perceived risk facets perspective. *Proceedings of AMCIS 2002*, Dallas, TX.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, behaviour: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Hair, J. Jr., Anderson, R., Tatham, R., & Black, W. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Höflich, J., & Rössler, P. (2001). Mobile schriftliche Kommunikation oder: E-mail fürdas handy. *Medien & Kommunikationswissenschaft, 49,* 437-461. Cited by Pedersen & Ling (2003).

Hun, S., Ku, C., & Chang, C. (2003). Critical factors of WAP services adoption: An empirical study. *Electronic Commerce Research and Applications, 2*(1), 42-60.

Juniperresearch.com. (2004). Mobile data sales predicted to bolster operator revenues. *New Media Age*, (October 21), 6.

Kleijnen, M., Wetzels, M., & Ruyter, K. (2004). Consumer acceptance of wireless finance. *Journal of Financial Service Marketing, 8*(3), 206-217.

Lin, C. (1996). Looking back: The contribution of Blumler and Katz's 'Uses of mass communication' to communication research. *Journal of Broadcasting & Electronic Media, 40*(4), 574-581.

Moon, J., & Kim, Y. (2001). Extending the TAM for a World-Wide-Web context. *Information & Management, 38*, 217-230.

NTT DoCoMo. (2003, October 30). *I-mode subscribers surpass 40 million.* Retrieved January 2004 from http://www.nttdocomo.com/

Okazaki, S. (2004). How do Japanese consumers perceive wireless advertising? A multivariate analysis. *International Journal of Advertising, 23*(4), 429-454.

Pagani, M. (2004). Determinants of adoption of third generation mobile multimedia services. *Journal of Interactive Marketing, 18*(3), 46-59.

Pedersen, P., & Ling, R. (2003). Modifying adoption research for mobile Internet service adoption: Cross-disciplinary interactions. *Proceedings of the 36th IEEE Hawaii International Conference on System Sciences 2003*, (pp. 90-91).

Sadeh, N. (2002). *M-commerce: Technologies, services, and business models.* New York: John Wiley & Sons.

Shih, H. (2004). Extended technology acceptance model of Internet utilization behaviour. *Information & Management, 41,* 719-729.

Taylor, S., & Todd, P. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research, 6*(2), 144-176.

Wolin, L. (2003). Gender issues in advertising—An oversight synthesis of research: 1970-2002. *Journal of Advertising Research, 43,* 111-129.

Yang, B., & Lester, D. (2005). Sex differences in purchasing textbooks online. *Computers in Human Behaviour, 21,* 147-152.

KEY TERMS

i-mode: A broad range of Internet services for a monthly fee of approximately 3 Euro, including e-mail, transaction services (e.g., banking, trading, shopping, ticket reservations, etc.), infotainment

services (e.g., news, weather, sports, games, music download, karaoke, etc.), and directory services (e.g., telephone directory, restaurant guide, city information, etc.), which offers more than 3,000 official sites accessible through the i-mode menu.

Mobile Commerce (M-Commerce): Any transaction with a monetary value that is conducted via a mobile telecommunications network. In a broader sense, it can be defined as the emerging set of applications and services people can access from their Internet-enabled mobile devices.

Mobile Portal: Typically, m-commerce takes place in a strategic platform called a "mobile portal," where third-generation (3G) mobile communication systems offer a high degree of commonality of worldwide roaming capability, support for a wide range of Internet and multimedia applications and services, and data rates in excess of 144 kbps. Examples of such mobile portals take many forms: NTT DoCoMo's i-mode portal, Nordea's WAP Solo portal, Webraska's SmartZone platform, among others. So far, Japan's i-mode portal has been asserted to be "the most successful and most comprehensive example of m-commerce today."

Technology Acceptance Model (TAM): Extends TRA with attempts to explain the antecedents of computer-usage behavior. TAM comprises five fundamental salient beliefs: perceived ease of use, perceived usefulness, attitudes toward use, intention to use, and actual use.

Theory of Reasoned Action (TRA): This model explains that a person's performance of a specified behavior is determined by his or her behavioral intention (BI) to perform the behavior, and BI is jointly determined by the person's attitude (A) and subjective norm (SN) concerning the behavior in question, with relative weights typically estimated by regression: BI = A + SN.

Uses and Gratifications Theory: A theory derived from media communication studies that focuses on individual users' needs or motivations of a particular medium. According to a recent study of mobile phone users, seven gratifications were identified: fashion/status, affection/sociability, relaxation, mobility, immediate access, instrumentality, and reassurance.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 296-301, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.21
Hand Measurements and Gender Effect on Mobile Phone Messaging Satisfaction:
A Study Based on Keypad Design Factors

Vimala Balakrishnan
Multimedia University, Malaysia

P. H. P. Yeow
Multimedia University, Malaysia

ABSTRACT

A total of 110 participants were interviewed to investigate the effect of hand measurements and gender on mobile phone messaging satisfaction. Physical measurements of hand-size and thumbs were recorded. This study focused on mobile phone keypad design factors; namely, key size, shape, texture, space between keys, layout and keypad simplicity. Females were found to be more satisfied with the key size and space between keys, whereas males are more satisfied with key shape. Users with smaller hands and thumbs were found to be more satisfied with key size and space between keys compared to those with larger hands and thumbs. One of the recommended improvements was to have larger keys with more space between them. Results obtained can be used by mobile phone designers to design customized mobile phones, for example, mobile phones that suit users with larger hands and thumbs, especially males.

INTRODUCTION

Mobile phone messaging is a popular service that allows users to communicate nonverbally, expressing themselves via combinations of alphanumerical characters with a maximum of 160 characters per single message. These mes-

sages, colloquially called SMS (Short Message Service) have recorded tremendous success in most of countries, including Asian countries like Singapore, Philippines and Malaysia. Ericsson reported that SMS has been the biggest mobile data service thus far in Malaysia (Wong & Pang, 2005). SMS growth is being driven by inexpensive, convenient, interpersonal communication, as well as by applications in business and games. Moreover, it is a fast medium of communication, as a message can be delivered to the recipient within a matter of seconds.

The popularity of SMS has heightened the interest in mobile phone research. A lot of studies have been done on the adoption of mobile phone and SMS in certain countries (Faulkner & Culwin, 2005; Hőflich & Rőssler, 2002; Ling, 2005). Social and psychological effects of SMS messaging were also studied to examine the underlying motivations of using SMS (Reid & Reid, 2004). Some researchers have done usability studies of mobile phones (Balakrishnan, Yeow, & Ngo, 2005; Soriano, Raikundalia, & Szajman, 2005) and some have compared the performance of the text entry methods (Friedman, Mukherji, Roeum, & Ruchir, 2001; James & Reischel, 2001). Although numerous studies have been conducted related to SMS, very few were related to SMS users' subjective satisfaction (Han, Kim, Yun, Hong, & Kim, 2004; Yun, Han, Hong, & Kim, 2003).

BACKGROUND

Mobile Phone Keypad Design

Mobile phones still have a keypad designed for dialing numbers, which makes text messaging difficult. The standard ISO mobile phone has only 12 keys ("0"-"9," "#" and "*") to input the entire alphabet, punctuations and numerical characters. Each physical key is therefore overloaded with three or four alphabetical characters: for example, the digit "9" is overloaded with "W," "X," "Y" and "Z." Consequently, this requires the users to make multiple key presses to make any intended selections. The most popular forms of text input on a standard 12-key mobile phone are either multitap or predictive text entry. In the multitap system, one or multiple key presses need to be made to make certain selections. For example, the digit "2" is pressed once for "a," twice for "b" and thrice for "c." As an example, "life" is entered as 555-444-333-33. On the other hand, predictive text entry uses linguistic knowledge and allows the user to choose from possible combinations of characters, shown from the most frequent words to the least frequent words (James & Reischel, 2001).

Studies related to keypad designs are numerous; however, most attempt to tackle keypad design problems by focusing on the text input mechanism (Mackenzie, 2002; Silfverberg, Mackenzie, & Korhonen, 2000; Wigdor & Balakrishnan, 2004). The Fastap keypad was designed by placing 52 independent keys onto an area the same size as the standard ISO keypad. Although it offers an increased performance over an ISO keypad, it remains to be seen how mobile phone users will assess the trade-off between the increased performance of advanced input technologies and their additional cost (Cockburn & Siresena, 2003). Tiny mobile phone key sizes were also identified as one of the problems related to mobile phones by several studies (Kurniawan, Mahmud, & Nugroho, 2006; Maragoudakis, Tseios, Fakotakis, & Avouris, 2002; Soriano et al., 2005). A study conducted with a group of elderly people revealed that keys that are placed too close to one another cause problems while handling a mobile phone (Ornella & Stephanie, 2006).

Thus far, no studies have been conducted to study the effect of varying hand-sizes and thumbs on mobile phone messaging satisfaction; hence, this study aims to investigate and evaluate the influence of hand-size, thumb size and gender of mobile phone users on their SMS messaging satisfaction, focusing only on the keypad design factors.

DESIGN OF STUDY

User satisfaction in using SMS based on the standard ISO mobile phone keypad design was identified as the dependent variable, whereas keypad design factor(s) was the independent variable. The list of keypad design factors are shown in Table 1. These factors were obtained from studies conducted to identify mobile phone design features that are critical to users' satisfaction (Han et al., 2004; Ling, Hwang, & Salvendy, 2007; Yun et al., 2003). Other studies that have reported on issues related to some of these features are: problems related to tiny keys (Kurniawan et al., 2006; Soriano et al., 2005), problems related to key size and space between them (Balakrishnan et al., 2005; Ornella & Stephanie, 2006). None of these studies, however, took hand measurements and gender into consideration.

Materials and Methods

Participants

SMS is hugely popular among the young (Ling, 2001; Reid & Reid, 2004); thus a total of 110 youth between 17-25 years old (mean = 21.5 years, SD = 1.64) were recruited. The interviews were conducted in two states, Melaka and Perak. Both these states have a high mobile phone penetration rate per 100 inhabitants. Moreover, both the states have numerous higher education institutions such as colleges and universities, and therefore have a high youth population. This resulted in the majority of the participants (76.3%, 84/110) to be students. Twenty-five (22.7%) participants were working young adults and only one was unemployed. All the participants have used SMS before, with an average of 3.8 years of experience and SD = 1.19. All the participants composed messages single-handedly using their thumbs. 80.9% (89/110) of the participants used multitap for text entry, 11.8% (13/110) used both multitap and predictive text entry interchangeably and only 7.3% (8/110) used predictive text entry. The participants responded to the interview questions based on their own mobile phones; therefore, the data collected captured the users' real feelings toward their own mobile phones. Some of the mobile phones used in the study were Nokia (6100, 3310, 8250, etc.), Motorola (C300, C261, etc.), and Samsung (X430, E700, etc.), among others. All the mobile phones had a 4 x 3 keypad layout and supported predictive text entry.

Table 1. Mobile phone keypad design factors

Keypad design factors	Explanation
Size	Size of the keys for messaging and navigating
Simplicity	The overall simplicity of the keypad design
Space	Existing space between the keys
Shape	Shape of the keys (square, rectangle, oval etc.)
Layout	Arrangement of keys (4 x 3 etc., QWERTY)
Texture	Tactual satisfaction related to key texture/material (E.g. soft, hard, coarse, etc.)

Hand Measurements

Users with large hands may find it difficult to message using a small mobile phone, whereas small hand-sized users may find it difficult to message via a large mobile phone. In both cases, users will be dissatisfied as they will not be able to hold the mobile phones and message comfortably. In the present study, hand-size is measured based on the hand breadth measurement, taken at the distal ends of the metacarpal bones (the joints of index finger to the little finger). Users with large thumbs may find it cumbersome keying in messages via the tiny keys; hence, thumb circumference was measured at the narrowest point of the thumb. All the measurements were taken using measurement tape, twice, and the average is noted to maintain the consistency. Table 2 shows the summary statistics for these anatomical measurements based on gender. None of the measurements were significantly different between the dominant and the opposite hands; thus, only the dominant hand measurements are displayed in the table.

Three hand-size groups (small, medium and large) were defined based on the hand breadth sizes used in You, Kumar, Young, Veluswamy, & Malzahn (2005), that is, for males, <8.8 cm is small, 8.8–9.2 cm is medium and >9.2 cm is large; for females, <7.3 cm is small, 7.3–7.7 cm is medium and >7.7 cm is large. The number of participants for each hand-size groups and gender is: males (14 small, 18 medium and 23 large) and females (14 small, 25 medium and 16 large). Participants were not grouped based on their thumb circumference measurements.

DESIGN OF THE INTERVIEW QUESTIONNAIRE

A structured questionnaire was designed based on Sinclair's (1995) guidelines, tested on five participants and revised before finalizing it. The questionnaire was developed in English and had two major sections. The first section, Section A, is to obtain the demographic profiles of the participants and the mobile phone characteristics. It consists of 22 questions addressing issues like age, dominant hands, finger(s) used when composing messages, years of experience in using SMS, frequency of using abbreviations, emoticons, average time spent on SMS daily, mobile phone brand, model and many others. The participant's hand measurements were also measured and recorded in this section. Section B is for the participants to rate their satisfaction/dissatisfaction levels of SMS usage based on the keypad design factors using Likert's five-point scale, whereby 1 means "Strongly Dissatisfied;" 2 means "Dissatisfied;" 3 means "Neutral;" 4 means "Satisfied" and 5 means "Strongly Satisfied." Some of the questions asked are as follows:

Table 2. Hand-size statistics based on genders

Measurements	Male (N=55) Mean ± SD (Min–Max)	Female (N=55) Mean ± SD (Min–Max)
Hand breadth (cm)	9.0 ± 0.5 (8.0 – 9.4)	7.3 ± 0.4 (6.0 – 8.2)
Thumb circumference (cm)	5.8 ± 0.75 (4.5 – 7.8)	5.4 ± 0.58 (4.5 – 7.2)

Box A.

a.The size of the keys used for messaging.	1 2 3 4 5
b. The ease in which you can compose a message based on the keypad design.	1 2 3 4 5
c. The amount of space available between the keys.	1 2 3 4 5
d. The shape of the keys used to SMS.	1 2 3 4 5
e. The way the keys are arranged on your mobile phone.	1 2 3 4 5
f. The tactile feedback felt when key presses are made on your mobile phone.	1 2 3 4 5

How would you rate your satisfaction level for the criteria in Box A?

Interviews

Face-to-face interviews were conducted using the above questionnaire on a one-to-one basis, beginning with the participants' background information, which includes their age, gender, years of experience in sending SMS, the finger(s) used in composing SMS and so forth. The interviewers then recorded the hand-size measurements. Mobile phone characteristics like brand, model and support of predictive text entry were also recorded. All the questions were read out to the participants. The structured questionnaire interviews enabled the capturing of both some quantitative and qualitative data. The qualitative data were gathered by encouraging the participants to give comments, opinions and suggestions. All verbal comments were recorded by the interviewers. Each interview session lasted for about 30 minutes. Two interviewers were involved in the interviewing activities and this consumed approximately 6 to 8 weeks in total. Both interviewers were knowledgeable of mobile phone features and SMS application so that they could easily interact with the participants during the interview sessions.

RESULTS

The data collected were analyzed using Statistical Package for the Social Sciences (SPSS) version 13.0. Descriptive statistics, Analysis of variance (ANOVA), Tukey Post-Hoc and Pearson correlations were used to analyze the collected data. All results are considered significant at $p < 0.05$ level.

Table 3 shows that the majority of the males (52.7%) and females (69.1%) spent more than 5 minutes daily on SMS. The overall statistics show that females spent more time messaging than the males. Twenty-five (45.5%) of the males sent between three to five messages daily compared to 38 (69.1%) of the females who sent more than five messages daily. The females also composed longer messages than males. Thirty-two females (58.2%) composed messages that were 75 to 160 characters in length, whereas the majority of males (60.0%) composed messages that were 25 to 75 characters in length.

Table 4 shows that the effect of gender is significant for mobile phone key size, space between keys and key shape. Females were found to be more satisfied with the key size (mean = 4.04) and space between keys (mean = 4.02) than males (mean = 3.36 and 3.49, respectively).

Table 3. Profile summary on SMS pattern based on gender (Highest frequency in each category)*

Profile	Categories	Male (N = 55)		Female (N = 55)	
		Frequency	Percentage	Frequency	Percentage
Average time SMS daily (min)	1 – 3	7	12.7	6	10.9
	3 – 5	19	34.6	11	20.0
	> 5	29*	52.7*	38*	69.1*
Average SMS sent daily	1 – 3	18	32.7	4	7.3
	3 – 5	25*	45.5*	13	23.6
	> 5	12	21.8	38*	69.1*
Average length of SMS sent	< 25	10	18.2	4	7.3
	25 – 75	33*	60.0*	19	34.5
	75 – 160	12	21.8	32*	58.2*

Table 4. ANOVA test for keypad design factors satisfaction, based on gender and hand-size (Significant at $p < 0.05$)*

Keypad design factors	Gender F-ratio (p-value)	Hand-size F-ratio (p-value)
Size	13.35 (< 0.001*)	12.96 (< 0.001*)
Simplicity	0.25 (0.622)	0.218 (0.804)
Space	7.33 (0.008*)	13.89 (< 0.001*)
Shape	4.33 (0.04*)	0.56 (0.573)
Layout	2.37 (0.127)	1.26 (0.292)
Texture	0.00 (1.000)	0.27 (0.974)

*Table 5. Pearson correlations between thumb circumference and keypad design factors (*Only significant results are shown, $p < 0.01$)*

Keypad design factors	*p*-value	Pearson Coefficient
Size	0.001	-0.309
Space	0.004	-0.412

However, the males were more satisfied with the key shape (3.87) than females (3.53).

The effect of hand-size is significant for key size and space between keys. Tukey post-hoc analysis revealed that small hand-sized participants are more satisfied with the key size than participants with medium hand-size ($p = 0.005$) and large hand-size ($p < 0.001$). They are also more satisfied with the space between keys than medium hand-sized ($p = 0.008$) and large hand-sized participants ($p < 0.001$).

Table 5 shows significant correlations between thumb circumference and users' satisfaction toward key size ($p = 0.001$, $r = -0.309$), and with users' satisfaction toward space between keys ($p = 0.004$, $r = -0.0272$).

DISCUSSION

Females spent more time messaging and sent more messages daily than males. This phenomenon accords with findings reported by other studies whereby a higher adoption of SMS by young females was found by several studies worldwide. For example, in Norway, Ling (2005) reported that women are more enthusiastic in using SMS than males, based on an analysis of SMS corpus. Their data showed that only 36% of males reported messaging daily compared to more than 40% of females. Females being more frequent users of SMS than males can be attributed to psychological reasons. Lohan (1997) stated that males are more "task-oriented" in the use of telephones, whereas females are more "person-oriented." Similarly, Skog (2002) observed that females valued social functionality of the mobile phone higher than males, who, on the other hand, stressed more on technical functionality and noninteractive uses like gaming.

The present study also revealed that females tend to write longer text messages as compared to the males, with the majority of females (58.2%, 32/110) composing messages that were 75 to 160 characters in length, whereas the majority of males (60.0%, 33/110) composed messages that were 25 to 75 characters in length. A similar finding was reported by Kasesniemi and Rautiainen (2002) who observed that Finnish young females tend to write longer messages than males, often using up all the 160 characters of an SMS, filled with references and social gossip, while the males often wrote messages of 40 to 50 characters with plain language. Females being more verbose in their SMS than males are consonant with conclusions other researchers have reached about gender

differences in face-to-face spoken language (Clark & Schaeffer, 1981; Treichler & Kramarae, 1983) and written communication (Cheshire, 2002). Women were also found to often talk on landline telephones longer than men (Moyal 1989) and to get involved in gossips more than men (Mante & Piris, 2002; Potts, 2004).

Judging from the above results, SMS seems to be a gendered practice with more females using the application. Analyses of the use of SMS also revealed females wrote lengthier messages than males, indicating females are more comfortable in using the mobile phone keypads to SMS. Perhaps this is because of the designs of mobile phone keypads that are awkward for larger hands and fingers, and thus makes messaging more difficult for males. As shown in Table 4, gender and hand-size were found to significantly affect users' satisfaction toward key size and space between keys. Females are more satisfied compared to males with the key size and space between them. Having smaller hands and thumbs makes it easier for the females to key in messages via the tiny and closely placed keys. This tallies with the effect of hand-size as well, whereby small hand-sized participants were found to be more satisfied with key size and space between keys than participants with medium and large hand-size. Miniaturization seems to be the trend toward designing the mobile phones nowadays. As the mobile phone size shrinks, it causes the key size and space between keys to decrease as well. Tiny keys are one of the major problems among mobile phone users with larger hands and thumbs. This is due to the difficulty in making multiple key presses error-free. The majority of the males (67.2%; 37/55) reported that messaging becomes cumbersome, as they tend to make more errors while composing messages as wrong key presses are made frequently, especially when it is done within a rapidly changing physical environment such as when moving. Frequently having to correct their errors hinders these large hand- and thumb-sized users from adopting SMS at certain times, as making a phone call is faster. This finding is consistent with Soriano et al. (2005), who reported that four out of five male participants in their study claimed that the size of the keys became an issue when messaging especially among those with larger fingers. Small key sizes and limited space between keys were also reported as one of the mobile phone usability problems in other studies; however, none took hand measurements into consideration (Axup, Viller, & Bidwell, 2005; Balakrishnan et al., 2005; Ornella & Stephanie, 2006). Moreover, participants with larger hands and thumbs tend to be more careful when making key presses to avoid making unwanted errors, and this increases the time spent on composing a message. Due to this, participants tend to make phone calls that are faster, instead of making slow key presses to message.

Gender was also found to be significantly affecting users' satisfaction toward key shape, with the males being more satisfied than females. Thirty-one females reported that keys that are rectangular or square in shape provide better satisfaction when key presses are made, especially when the keys are "raised" as a better feedback is provided while messaging. In other words, it provides a better sense of knowing that a key press has been made, hence reducing the unintended key presses and enables "eyes-free" messaging among the expert users. One male participant in this study responded that *"the look, shape or color of the keys is not important as long as I get to key in the message fast…."* Psychological reasons could be attributed as to why females emphasize more on aesthetic values than males. A similar finding was reported by Yun et al. (2003), whereby female participants identified mobile phones' body color, button shape and brightness of color as some of the features that affect their satisfaction compared to males, who feel that clearness of menu item and softness of bell sound are more important.

Thumb circumference was found to significantly correlate negatively with key size.

This confirms that as users' thumb increases in size, their satisfaction decreases toward key size. Large thumbed users find it difficult to make multiple key presses on the tiny keys. This is further aggravated by the limited or no space between keys. These users tend to accidentally hit the wrong keys when entering messages. Having to correct the errors cause frustrations among these users, hence decreasing their satisfaction with respect to key size. This finding is consistent with Soriano et al. (2005), who reported that four out of five male participants in their study claimed that the size of the keys became an issue when messaging, especially among those with larger fingers; however, the researchers did not take any finger measurements in their study. Small key sizes were also reported as one of the mobile phone usability problems by Axup et al. (2005) and Ornella and Stephanie (2006); however, none took hand measurements into consideration. Anderson (2005) reported that any tool that involves a struggle to be used earns a "D" or worse for usability. A common criticism is that mobile phones have become too small, causing aim and accuracy to suffer when adult hands finger child-sized buttons.

Thumb circumference also significantly correlates with satisfaction toward space between keys. The negative correlation indicates that as users' thumb circumferences increase, their satisfaction toward space between keys decrease. Miniaturization of the mobile phones also causes the keys to be placed closely together, hence limiting the space between them. Large thumbed users find messaging a tedious task due to the close placement of the keys, which is further aggravated by the tiny keys. Thirty-two participants commented that they tend to hit the neighbouring keys accidentally while messaging, especially when it is done in a hurry or while in motion (e.g., walking and talking). It can be a frustrating task, as they have to waste their time correcting the errors instead of messaging efficiently. Moreover, they also mentioned that they need to constantly focus on the screen to make sure they have pressed the correct key, hence eliminating the possibility of "eyes-free" input among the large thumbed users. Frequently having to correct their errors hinders these users from adopting SMS at times or to use it only when it is deemed necessary, for example, to send simple and short messages, especially single line messages. Ornella and Stephanie (2006) also found limited spaces between keys to be a problem among the elderly mobile phone users (60–80 years old); however, no hand measurements were taken into consideration.

CONCLUSION AND RECOMMENDATION

The effect of hand measurements and gender on mobile phone messaging satisfaction was studied. These moderating variables were tested against mobile phone keypad design factors. The following results were drawn:

- Females are more actively involved in messaging as compared to males. Apart from psychological reasons, poor keypad design can also be attributed to males using SMS application less frequently than females.
- Females are more satisfied with key size and space between keys than males.
- Males are more satisfied with the key shape than females, indicating that females emphasize more on aesthetic values.
- Small hand-sized users are more satisfied with the key size and space between keys than medium and large hand-sized users.
- The increase of thumb circumference decreases users' satisfaction toward key size and space between keys as messaging becomes tedious due to accidentally hitting the wrong keys.

Varying hand measurements were found to affect mobile phone users' messaging satisfaction, with differences noted between genders as well. Customized mobile phones have been designed to suit the elder people (Croasmun, 2005) and also kids (Budnick, 2005) to improve usability and increase satisfactions. With this in mind, the results from this study can be used to design customized mobile phones that suit specific users, such as users with larger hands and thumbs or even males and females. Some of the recommendations are as follows:

- *Enlarge keys or increase the space between keys*: Mobile phone manufacturers and designers should look into the possibility of enlarging the keys or increasing the space between the keys. This may result in a larger mobile phone and a change in the keypad layout; however, this will be eventually accepted as users' messaging speed will be increased. This will be advantageous not only for the larger hand-sized males, but also to those users who are capable of messaging without looking at the keypad as the larger keys and increased space between keys will make it easier for them to find and press the desired keys. This eventually results in an increased messaging speed and also on users messaging satisfaction.
- *Provide better tactile feedback*: The female participants were found to be less satisfied with the key shape than males, with preferences for rectangle or square shaped keys that are "raised" in nature. With this knowledge, mobile phone manufacturers should look into the possibility of designing keypads with "raised" keys that are rectangle or square in shape. The process of appropriating technology to a specific group, especially the female, exists in the Japanese, Korean, and Chinese markets where mobile phones are specially designed to appeal to female users. For example, Japan is known for its culture of kawaii or "cute culture" that has been extended to include mobile phone as the latest female fashion item, using flashy colours and cute characters as decorations (Hjorth, 2003).

It is believed that the above recommendations will not only benefit the specific groups of users (e.g., large hands and thumbs) but the overall SMS population as well.

REFERENCES

Anderson, J. (2005). *Cell phone design given a failing grade for usability.* Retrieved May 9, 2008, from http://www.ergoweb.com/news/detail.cfm?print=on&id=1177

Axup, J., Viller, S., & Bidwell, N. (2005). Usability of a mobile, group communication prototype while rendezvousing. In *Proceedings of the CTS'05 International Symposium on Collaborative Technologies and Systems-Special Session on Mobile Collaborative Work*, St. Louis, USA.

Balakrishnan, V., Yeow, P.H.P., & Ngo, D.C.L. (2005). An investigation on the ergonomic problems of using mobile phones to send SMS. In P.D. Bust & P.T. McCabe (Eds.), *Contemporary ergonomics* (pp. 195-199). UK: Taylor & Francis.

Budnick, P. (2005). *A cell phone designed just for the kids.* Retrieved May 9, 2008, from http://www.ergoweb.com/news/detail.cfm?id =1092

Cheshire, J. (2002). Sex and gender in variationist research. *Handbook of language variation and change*. Oxford: Blackwell.

Clark, H., & Schaeffer, E.W. (1981). Contributing to discourse. *Cognitive Science, 13*, 259-295.

Cockburn, A., & Siresena, A. (2003). Evaluating mobile text entry with the Fastap Keypad.

In *Proceedings of People and Computers XVII: British Computer Society Conference on Human Computer Interaction*, England, (Vol. 2).

Croasmun, J. (2005). *A simple cell phone*? Retrieved May 9, 2008, from http://www.ergoweb.com/news/detail.cfm?id=1084

Faulkner, X., & Culwin, F. (2005). When fingers do the talking: A study of text messaging. *Interacting with Computers, 17,* 167-185.

Friedman, Z., Mukherji, S., Roeum, G.K., & Ruchir, R. (2001). Data input into mobile phones: T9 or keypad?. *Student online HCI research experiments*. Retrieved April 2, 2007, from http://www.otal.umd.edu./SHORE2001/mobilePhone/index.html

Han, S.H., Kim, K.J., Yun, M.H., Hong, S.W., & Kim, J. (2004). Identifying mobile phone design features critical to user satisfaction. *Human Factors and Ergonomics in Manufacturing, 14*(1), 15-29.

Hjorth, L. (2003). Cell phone use in social settings: Preliminary results from a study in the United States and France. In N. Gottlieb & M. McLelland (Eds.), *Japanese cybercultures* (pp. 50-59). New York: Routledge.

Hőflich, J.R., & Rőssler, P. (2002). More than just a telephone. The mobile phone and use of Short Message Service (SMS) by German adolescents: Results of a pilot study. *Journal of Studies on Youth*, (57), 79-99.

James, C.L., & Reischel, K.M. (2001). Text input for mobile devices: Comparing model prediction to actual performance. In *Proceedings of CHI 2001*, (Vol. 3, No. 1, pp. 365-371).

Kasesniemi, E.L., & Rautiainen, P. (2002). Mobile culture of children in Finland. In J.E. Katz & M. Aakhus (Eds.), *Perpetual contact: Mobile communication, private talk, public performance* (pp. 170-819). England: Cambridge University Press.

Kurniawan, S., Mahmud, M., & Nugroho, Y. (2006). A study of the use of mobile phones by older persons. In *Proceedings of CHI 2006*, Canada.

Ling, R. (2001). We release them little by little: Maturation and gender identity as seen in the use of mobile telephony. *Personal and Ubiquitous Computing, 5*, 123-136. Springer-Verlag.

Ling, R. (2005). The socio-linguistic of SMS: An analysis of SMS use by a random sample of Norwegians. In R. Ling & P. Pedersen (Eds.), *Mobile communications: Renegotiation of the social sphere* (pp. 335-349). London: Springer-Verlag.

Ling, C., Hwang, W., & Salvendy, G. (2007). A survey of what customers want in a cell phone design. *Behaviour and Information Technology, 26*(2), 149-163.

Lohan, E.M. (1997). *Men, masculinity and the domestic telephone. A theoretical framework for studying gender and technology.* Retrieved May 9, 2008, from http://www.dcu.ie/communications/iegis/Marial2.htm

Mackenzie, S.I. (2002). Mobile text entry using three keys. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction: NordiCHI 2002*, (pp. 27-34). New York: ACM.

Mante, E.A., & Piris, D. (2002). SMS use by young people in the Netherlands. *Journal of Studies on Youth*, (57), 47-58.

Maragoudakis, M., Tselios, N.K., Fakotakis, N., & Avouris, N.M. (2002). Improving SMS usability using bayesian networks. In I.P. Vlahavas & C.D. Spyropoulos (Eds.), *Methods and applications of artificial* intelligence (pp. 179-190). Berlin: Springer-Verlag.

Moyal, A. (1989). The feminine culture of the telephone: People patterns and policy. *Prometheus, 7*, 5-31.

Ornella, P., & Stephanie, B. (2006). Universal designs for mobile phones: A case study. In *Proceedings of CHI 2006* (Work in Progress). Quebec, Canada.

Potts, G. (2004). *College students and cell phone use: Gender variation.* Retrieved May 9, 2008, from http://personalwebs.oakland.edu/$gapotts/rht160.pdf

Reid, F.J.M., & Reid, D.J. (2004). Text appeal: The psychology of SMS messaging and its implications for the design of mobile phone interfaces. *Campus-Wide Information Systems, 21*(5), 196-200.

Silfverberg, M., Mackenzie, I.S., & Korhonen, P. (2000). Predicting text entry speed on mobile phones. *CHI 2000, 2*(1), 9-16.

Sinclair, A.M. (1995). Participantive assessment. In J.R. Wilson & E.N. Corlett (Eds.), *Evaluation of human work—a practical ergonomics methodology* (pp. 69-100). London: Taylor & Francis.

Skog, B. (2002). Mobiles and the Norwegian teen: Identity, gender and class. In J. E. Katz & M. Aakhus (Eds.), *Perpetual contact: Mobile communications, private talk, public performance* (pp. 255-273). Cambridge: Cambridge University Press.

Soriano, C., Raikundalia, G.K., & Szajman, J. (2005). A usability study of short message service on middle-aged users. In *Proceedings of OZCHI 2005*, Canberra, Australia.

Treichler, P.A., & Kramarae, C. (1983). Women's talk in the ivory tower. *Communication Quarterly, 3*, 118-132.

Wigdor, D., & Balakrishnan, R. (2004). A comparison of consecutive and concurrent input text entry techniques for mobile phones. In *Proceedings of CHI 2004*, (Vol. 6, No. 1, pp. 81-88).

Wong, C.C., & Pang, L.H. (2005). Correlations between factors affecting the diffusion of mobile entertainment in Malaysia. In *Proceedings of ICEC 2005*, Xi'an, China.

You, H., Kumar, A., Young, R., Veluswamy, P., & Malzahn, D.E. (2005). An ergonomic evaluation of manual Cleco Plier designs: Effect of rubber grip, spring recoil and work surface angle. *Applied Ergonomics, 36*, 575-583.

Yun, M.H., Han, S.H., Hong, S.W., & Kim, J. (2003). Incorporating user satisfaction into the look-and-feel of mobile phone design. *Ergonomics, 46*(13/14), 1423-1434.

This work was previously published in the International Journal of Technology and Human Interaction, edited by B. Stahl, Volume 4, Issue 4, pp. 54-67, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 5.22
User Acceptance of Mobile Services

Eija Kaasinen
VTT Technical Research Centre of Finland, Finland

ABSTRACT

Personal mobile devices are increasingly being used as platforms for interactive services. User acceptance of mobile services is not just based on usability but includes also other interrelated issues. Ease of use is important, but the services should also provide clear value to the user and they should be trustworthy and easy to adopt. These user acceptance factors form the core of the Technology Acceptance Model for Mobile Services introduced in this chapter. The model has been set up based on field trials of several mobile services with altogether more than 200 test users. The model can be used as a design and evaluation framework when designing new mobile services.

INTRODUCTION

Research on mobile services has thus far mainly concentrated on the usability of alternative user interface implementations. Small mobile devices pose significant usability challenges and the usability of the services is still worth studying. However, more attention should be paid to user acceptance of the planned services. The reason for many commercial failures can be traced back to the wrongly assessed value of the services to the users (Kaasinen, 2005b).

User evaluations of mobile services often have to be taken into the field as the service would not function properly otherwise, or it would not make sense to evaluate it in laboratory conditions. This would be the case, for instance, with GPS systems and route guidance systems. In long-term field trials with users, it is possible to gather feedback on the adoption of the service in the users' everyday lives. Such studies gather usage data beyond mere usability and pre-defined test tasks (Figure 1). Field trials help in studying which features the users start using, how they use them and how often, and which factors affect user acceptance of the service.

Business and marketing research already have approaches whereby new technology is studied on a wider scale. The Technology Acceptance Model

Figure 1. Taking user evaluations from the laboratory to the field makes it possible to evaluate user acceptance on new services

by Davis (1989) defines a framework to study user acceptance of a new technology based on perceived utility and perceived ease of use. Each user perceives the characteristics of the technology in his or her own way, based for instance on his or her personal characteristics, his or her attitudes, his or her previous experiences and his or her social environment. The Technology Acceptance Model has been evolved and applied widely, but mainly in the context of introducing ready-made products rather than in designing new technologies.

In this chapter an extension to the Technology Acceptance Model will be introduced. The model is based on a series of field trials and other evaluation activities with different mobile Internet and personal navigation services and over 200 test users (Kaasinen, 2005b). The Technology Acceptance Model for Mobile Services constitutes a framework for the design and evaluation of mobile services.

BACKGROUND

Technology acceptance models aim at studying how individual perceptions affect the intentions to use information technology as well as actual usage (Figure 2).

In 1989, Fred Davis presented the initial technology acceptance model (TAM) to explain the determinants of user acceptance of a wide range of end-user computing technologies (Davis 1989). The model is based on the Theory of Reasoned Action by Ajzen and Fishbein (1980). TAM points out that perceived ease of use and perceived usefulness affect the intention to use. Davis (1989) defines perceived ease of use as "*the degree to which a person believes that using a particular system would be free from effort"* and perceived usefulness as "*the degree to which a person believes that using a particular system*

Figure 2. The basic concept underlying technology acceptance models (Venkatesh et al., 2003)

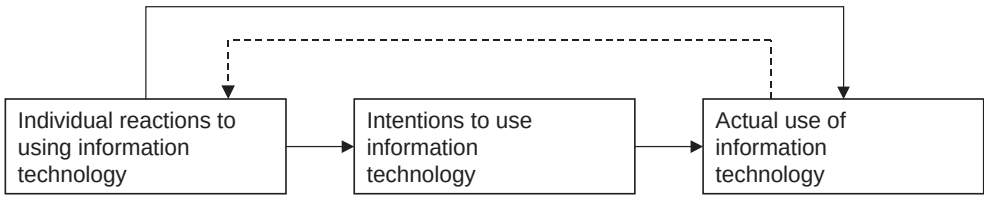
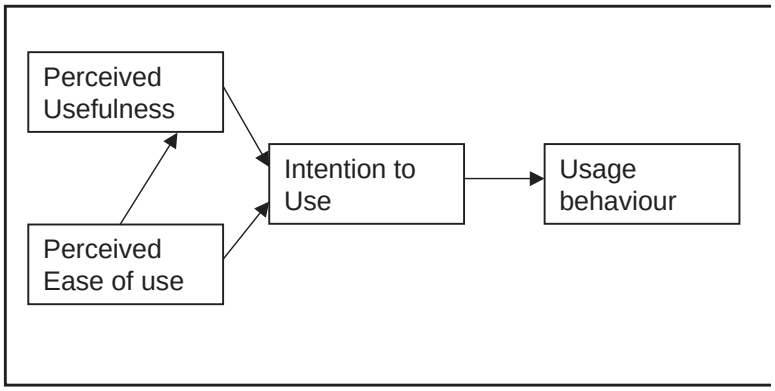


Figure 3. Technology acceptance model (Davis, 1989)



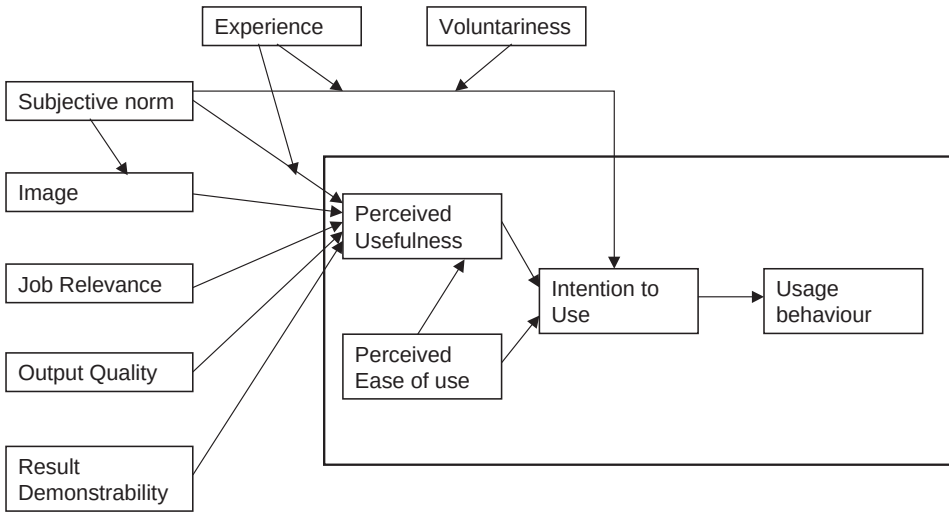
would enhance his or her job performance." Perceived ease of use also affects the perceived usefulness (Figure 3). The intention to use affects the real usage behavior. TAM was designed to study information systems at work to predict if the users will actually take a certain system into use in their jobs. The model provides a tool to study the impact of external variables on internal beliefs, attitudes and intentions.

TAM deals with perceptions; it is not based on observing real usage but on users reporting their conceptions. The instruments used in connection with TAM are surveys, where the questions are constructed in such a way that they reflect the different aspects of TAM. The survey questions related to usefulness can be, for instance: "Using this system improves the quality of the work I do" or "Using this system saves my time." The survey questions related to ease of use can be, for instance: "The system often behaves in unexpected ways" or "It is easy for me to remember how to perform tasks using this system."

TAM has been tested and extended by many researchers, including Davis himself. Venkatesh and Davis (2000) have enhanced the model to TAM2 (Figure 4), which provides a detailed account of the key forces underlying judgments of perceived usefulness, explaining up to 60 percent of the variance in this driver of usage intentions. TAM2 showed that both social influence processes (subjective norm, voluntariness and image) and cognitive instrumental processes (job relevance, output quality, result demonstrability, and perceived ease of use) significantly influenced user acceptance.

Mathieson, Peacock and Chin, (2001) have extended TAM by analyzing the influence of perceived user resources. They claim that there may be many situations in which an individual wants to use an information system, but is prevented by lack of time, money, expertise and so on (Mathieson et al., 2001) classify resource-related attributes into four categories: user attributes, support from others, system attributes and general control-related

Figure 4. Enhanced technology acceptance model (TAM2) by Venkatesh and Davis (2000)



attributes that concern an individual's overall beliefs about his or her control over system use. In their extended model, external variables affect perceived resources that further affect perceived ease of use and the intention to use.

TAM was originally developed for studying technology at work, but it has often been used to study user acceptance of Internet services as well (Barnes & Huff, 2003; Chen, Gillenson & Sherell, 2004; Gefen, 2000; Gefen & Devine, 2001; Gefen, Karahanna & Straub, 2003). Gefen et al. (2003) have studied TAM in connection with e-commerce. They have extended TAM for this application area and propose that trust should be included in the research model to predict the purchase intentions of online customers.

The Technology Acceptance Model constitutes a solid framework to identify issues that may affect user acceptance of technical solutions. Davis and Venkatesh (2004) proved that the model can be enhanced from the original purpose of studying user acceptance of existing products to study planned product concepts, for example, in the form of mock-ups. This indicates that TAM could also be used in connection with technology development projects and processes to assess the usefulness of proposed solutions.

APPLICABILITY OF EARLIER APPROACHES FOR MOBILE SERVICES

The focus of traditional usability studies is on specified users performing specified tasks in specified contexts of use (ISO13407, 1999). In field trials the users can use prototype services as part of their everyday life. The research framework can then be enhanced to identify the actual tasks that users want to perform and the actual contexts of use. Technology acceptance models provide a framework for such studies.

Mobile services targeted at consumers have several specific characteristics that may mean that their user acceptance cannot be studied using the same models as with information systems in the workplace. When dealing with consumer services, individuals make voluntary adoption decisions and thus the acceptance includes assessing the benefits provided compared with either competing solutions or the non-acquisition of the service in question. As pointed out by Funk (2004), mobile services are disruptive technology that may find their innovation adopters elsewhere than expected, as highlighted by the experiences with the Japanese i-mode. Focusing too early on only limited user groups may miss possible early adopters. With the Japanese i-mode, other services were boosted through e-mail and personal home pages (Funk, 2004). This suggests that the focus of user acceptance studies of mobile services should be extended to interrelated innovations, as proposed by Rogers (1995).

Perceived usefulness included in TAM may not indicate an adequate purchase intention in a market situation. Product value has been proposed as a wider design target both in software engineering and HCI approaches. A value-centered software engineering approach was proposed by Boehm (2003) to define more clearly what the design process is targeted at, and identifying the values that different stakeholders—including end-users—expect of the product. Although not using the actual term "value," Norman (1998) emphasizes the importance of identifying big phenomena related to user needs and communicating them early on to the design. Cockton (2004b) points out that in value-centered HCI existing HCI research components, design guidance, quality in use and fit to context need to be reshaped to subordinate them to the delivery of product value to end-users and other stakeholders.

Mobile services are increasingly handling personal information of the user, for instance due to the personalization and context-awareness of the services. The functionalities of the increasingly complex systems are not always easy for the users to comprehend. Context-aware services

may include uncertainty factors that the users should be able to assess. Mobile service networks are getting quite complex and the users may not know with whom they are transacting. Technical infrastructures as well as the rapidly developed services are prone to errors. All these issues raise trust as a user acceptance factor, similar to TAM applied in e-commerce (Chen et al., 2004; Gefen et al., 2003). Trust has been proposed as an additional acceptance criterion for mobile services by Kindberg, Stellen and Geelhoed, (2004) and Barnes and Huff (2003). Trust has also been included in studies of personalization in mobile services (Billsus et al., 2002) and studies of context-aware services (Antifakos, Schwaninger & Schiele, 2004).

Ease of adoption is included in the studies by Sarker and Wells (2003) and Barnes and Huff (2003). Sarker and Wells (2003) propose a totally new acceptance model that is based on user adoption. Barnes and Huff (2003) cover adoption in their model within the wider themes of compatibility and trialability. *Perceived user resources* in the extension of TAM by Mathieson et al. (2001) and *Facilitating conditions*, in the Unified Theory of Acceptance and Use (Venkatesh et al., 2003) also include elements related to ease of adoption.

In the following, the technology acceptance model for mobile services (Kaasinen, 2005b) is described in detail. The model aims at taking into account the aforementioned special characteristics of mobile consumer services, and previous studies on user acceptance described in this chapter. The model can be utilized when designing new services and assessing them to ensure that key user acceptance factors are considered in the design.

TECHNOLOGY ACCEPTANCE MODEL FOR MOBILE SERVICES (TAMM)

The technology acceptance model for mobile services (TAMM) was constituted based on a series of field trials and other user evaluation activities involving over 200 users. The studies were carried out as parts of technology development projects in 1999-2002 by project usability teams comprising altogether 13 researchers from VTT and three researchers from other research organizations. The focus of the studies was in particular on mobile Internet services and location-based services targeted at consumers (Kaasinen, 2005b). Mobile Internet studies were carried out in connection with the development of mobile browsers and the first WAP (wireless application protocol) services for mobile phones. In addition to commercial services, the test users could access many Web services because our project developed a Web-WAP conversion proxy server. Based on identified user needs, our research team also developed specific WAP services, for instance, for group communication. The services were evaluated in long-term field trials with users. The studies of location-based services were carried out within a horizontal usability support project, part of the Personal Navigation (NAVI) research and development program in Finland. The aim of the program was to facilitate co-operation between different actors who were developing personal navigation products and services. Our research group supported individual projects in usability and ethical issues and, beyond this, identified general guidelines for acceptable personal navigation services. We studied user attitudes and preliminary acceptance by evaluating different service scenarios in focus groups. In addition we evaluated some of the first commercial location-based services and carried out user evaluation activities in co-operation with the NAVI projects that were developing location-based services. Table 1 gives an overview of the user evaluation activities that the technology acceptance model for mobile services is based on.

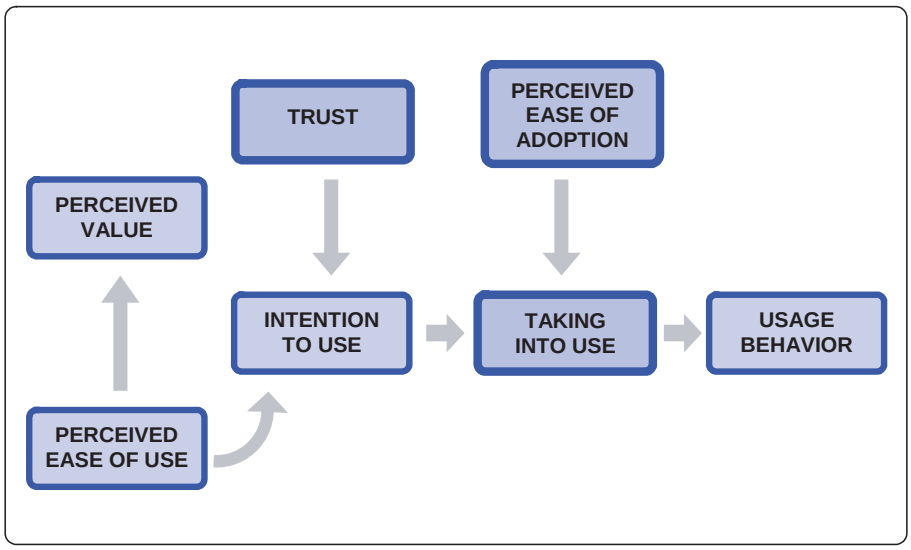
The original technology acceptance model was chosen as the starting point for the new model because it provided a framework for connecting field study findings of ease of use and usefulness.

Table 1. The user evaluation activities that the technology acceptance model for mobile services was based on.

Service, application or device	Research methods	Users	Original results published in
WAP services	Laboratory evaluation with phone simulator	6	Kaasinen et al., 2000
WAP-converted Web services	Laboratory evaluation with phone simulator	4	
WAP services WAP-converted Web services	Field trial 2 months	40	Kaasinen et al., 2001
	Interviews with service providers	25	
WAP services WAP-converted Web services Web/WAP Message board for group communication	Field trial 2 months	40	
	Interviews with service providers	11	
Scenarios of personal navigation services	Group interviews	55	Kaasinen, 2003
Benefon GPS phone and services	Field evaluation	6	
Sonera Pointer location-aware WAP services	Laboratory evaluation	5	
Garmin GPS device	Field evaluation	5	
Magellan GPS device	Field evaluation	5	
Location-aware SMS services	Field evaluation	6	
Weather and road conditions by SMS	Field trial, 1 month	10	Kaasinen, 2005a
Location-aware integrated service directory	Field trial, 3 weeks	7	
Mobile topographic maps	Field evaluation	6	
Mobile 3D maps	Laboratory evaluation	6	
	Field evaluation	4	
Scenarios of context-aware consumer services	Interviews in anticipated contexts of use	28	

The user acceptance framework is especially suitable for field trials where the focus is to study how different users start using the mobile services in their everyday lives and which features make the services acceptable in actual usage. As not all the field study findings could be fit to the original TAM model, it was necessary to update the model according to the repeated field study findings and themes identified in related research. The new model extends the original core model by Davis (1989) by identifying two new perceived product characteristics that affect the intention to use, that

Figure 5. Technology acceptance model for mobile services (Kaasinen, 2005b) as an extension and modification of TAM by Davis (1989)



is trust and ease of adoption, and by redefining the theme of usefulness as value to the user.

The framework (Figure 5) suggests that perceived ease of use, perceived value, and trust affect the intention to use a mobile service. To get from an intention to use to real usage, the user has to take the service into use. This transition is affected by the perceived ease of adoption. Perceived value, perceived ease of use, trust and perceived ease of adoption need to be studied in order to assess user acceptance of mobile services.

The technology acceptance model for mobile services (Kaasinen, 2005b) constitutes a framework that helps designers of mobile services to identify key issues that should be focused on in the design to ensure user acceptance. Thus the motivation of the model is different than the motivation of the original TAM, which was built to explain user acceptance and underlying forces for existing technical solutions.

Perceived ease of use was included in the original TAM and it is also included in the TAMM model. Davis (1989) defined perceived ease of use as "*the degree to which a person believes that using a particular system would be free from effort.*" At first, perceived ease of use is based on external factors such as the user's attitude towards technology in general, experiences of using similar services and information from other people. In actual use and sustained use, perceived ease of use is increasingly affected by the user's own experiences of using the system in different contexts of use.

In the case of mobile services that are used on small devices such as mobile phones or PDAs, the limitations of the device have a major influence on perceived ease of use. The limitations include the small screen, small and limited keyboard, the absence or limited functionality of pointing devices, limited amount of memory, limited

battery power, and slow connections. As new devices and mobile networks are being introduced to the market, these limitations have somewhat diminished but still mobile networks are slower than fixed ones and the requirements for ease of carrying and holding the device do not allow very large screens or large keyboards. Designing mobile services for ease of use is to a large extent about coping with the limitations of the device. In addition, the design should adapt to the variety of client devices and available networks and other infrastructures.

The ease of use of mobile services has been studied quite a lot and different usability guidelines are available. It is a pleasure to note that many of the usability problems identified in early mobile Internet studies have already been corrected in current mobile devices, browsers and services. However, location-aware services pose even more challenges for ease of use. Location-aware services are not just mobile in the sense that they can be easily carried around but, typically, they are used while the user is moving. These kinds of usage situations require extreme ease of use. Personalization and context-awareness are expected to improve ease of use, but they may also introduce new usability problems, for example in the form of personalization dialogues.

Perceived value replaces perceived usefulness in the TAMM model because in our field trials with consumers it became evident that in the consumer market, perceived usefulness may not indicate adequate motivation to acquire the mobile service. As the focus group studies by Järvenpää et al. (2003) point out, consumers may lack a compelling motivation to adopt new mobile services unless those services create new choices where mobility really matters and manage to affect people's lives positively. In a value-neutral setting each requirement is treated as equally important in the design (Boehm, 2003). This easily leads to featurism—the product becomes a collection of useful features but as a whole it may not provide enough value to the user. Value not only includes rational utility but also defines the key features of the product that are appreciated by the users and other stakeholders, that is the main reasons why the users are interested in the new product. As Roto (2006) points out, costs of using the service also affect the perceived value as user expectations tend to be higher for more expensive products. Values are made explicit by the identification of objectives, which are statements about what the user wants to achieve. Fundamental objectives are directly related to the user's current problem or situation at hand, whereas means objectives help to achieve the fundamental objectives (Nah et al., 2005).

Defining the targeted values and concentrating on them in design and evaluation helps to focus the design on the most essential issues. This is in line with the concept of value-centered software engineering proposed by Boehm (2003) and value-centered HCI proposed by Cockton (2004a, b). Focusing on perceived value in user acceptance studies supports the wider scope of value-centered design, where user value can be studied in parallel with business value and strategic value as proposed by Henderson (2005).

Trust is added as a new element of user acceptance in the TAMM model. The original TAM (Davis, 1989) was defined for information systems at work, and in those usage environments the end-users could rely on the information and services provided and the ways their personal data was used. When assessing user acceptance of e-commerce applications, Gefen et al. (2003) proposed to enhance TAM with trust in the service provider, as in their studies trust-related issues turned out to have a considerable effect on user acceptance. In our studies with mobile Internet, consumers were using mobile services that were provided to them via complex mobile service networks. In this environment trust in the service providers turned out to be an issue. As location-based services collect and use more and more information about the usage environment and the user, ethical issues arise. Especially ensuring the

privacy of the user was a common concern of our test users. As the users get increasingly dependent on mobile services, reliability of the technology and conveying information about reliability to the user becomes more important.

In the technology acceptance model for Mobile Services, trust is defined according to Fogg and Tseng (1999). Trust is an indicator of a positive belief about the perceived reliability of, dependability of, and confidence in a person, object or process. User trust in mobile services includes perceived reliability of the technology and the service provider, reliance on the service in planned usage situations, and the user's confidence that he or she can keep the service under control and that the service will not misuse his or her personal data.

Perceived ease of adoption is related to taking the services into use. In the original TAM settings with information systems at work, this certainly was not an issue as users typically got their applications ready installed. In our field trials it turned out that a major obstacle in adopting commercial mobile services was the users' unawareness of available services, as well as problems anticipated in taking services into use (Kaasinen, 2005b). Furthermore, as usage needs were typically quite occasional, people often did not have enough motivation to find out about these issues. And finally, configuration and personalization seemed to require almost overwhelming efforts (Kaasinen, 2005b). Introducing the services to users would definitely require more attention in service design (Kaasinen et al., 2002).

As mobile services are typically used occasionally and some services may be available only locally in certain usage environments, ease of taking the services into use becomes even more important. The user should easily get information about available services and should be able to install and start to use the services easily. Finally, he or she should be able to get rid of unnecessary services.

Compared with the original TAM (Davis, 1989), the technology acceptance model for mobile services includes an additional phase between the intention to use and the actual usage behavior. Taking a service into use may constitute a major gap that may hinder the transfer from usage intention to actual usage (Kaasinen, 2005b). Perceived ease of adoption is added to the model at the stage when the user's attention shifts from intention to use to actually taking the service into use.

The characteristics of the user and his or her social environment affect how the user perceives the service. These issues are not included in the core TAMM model that aims to identify key characteristics of mobile services that generally affect user acceptance of mobile services. Further research is needed to fit previous TAM enhancements such as TAM2 (Venkatesh & Davis, 2000) and UTAUT (Venkatesh et al., 2003) to the model to identify external factors such as characteristics of the users and their social environment that affect the user acceptance factors in the model.

In the following section the technology acceptance model for mobile services is analyzed further and design implications are presented for each user acceptance factor, based on the synthesized results of the original case studies (Kaasinen, 2005b). The technology acceptance model for mobile services, together with the design implications, communicates previous user acceptance findings to the design of future mobile services.

DESIGN IMPLICATIONS

The technology acceptance model for mobile services defines four main user acceptance factors: perceived value, perceived ease of use, trust and perceived ease of adoption. How these factors should be taken into account in the design of individual mobile services depends on the service in question. However, there are many attributes of the acceptance factors that repeat from one

service to another. These attributes form a set of design implications that can be used in the design of mobile services. The design implications can additionally be used in designing user acceptance evaluations to define the issues to be studied in the evaluation. In the following, design implications for each user acceptance factor are presented by combining results from the original studies (Kaasinen et al., 2000; Kaasinen et al., 2001; Kaasinen, 2003; Kaasinen, 2005a; Kaasinen 2005b) and results from related research.

Because of the quality of the case study material, the design principles cover best mobile information services targeted at consumers. For other kinds of services, the technology acceptance model for mobile services as well as the design implications can certainly be used as a starting point but they may need to be revised.

Perceived Value

Values define the key features of the services that are appreciated by the users and other stakeholders, that is the main reasons why the users are interested in the new services. Defining the targeted values helps in focusing the design on the most essential issues. Value is also related to the costs of using the service, and for commercial products the relationship of these two attributes should be studied, as proposed by Roto (2006). The following list gives some ideas about where in our studies the value was found.

Successful Service Content is Comprehensive, Topical, and Familiar

In the early days of mobile Internet, service providers often thought that small devices would require just a small amount of contents. Our studies showed that mobile users need access to all relevant information, as deep as they are ready to go, but the information has to be structured in such a way that the user can choose to get the information in small portions. Users appreciate comprehensive services in terms of geographic coverage, breadth (number of services included) and depth (enough information in each individual service).

Topical information is likely such that the mobile service is the best way to keep up to date with what is going on. In our field trials examples of successful topical content included weather forecasts, traffic information, news topics and event information. Topical travel information, for instance, does not just give timetables but informs about delays and traffic jams and recommends alternative routes.

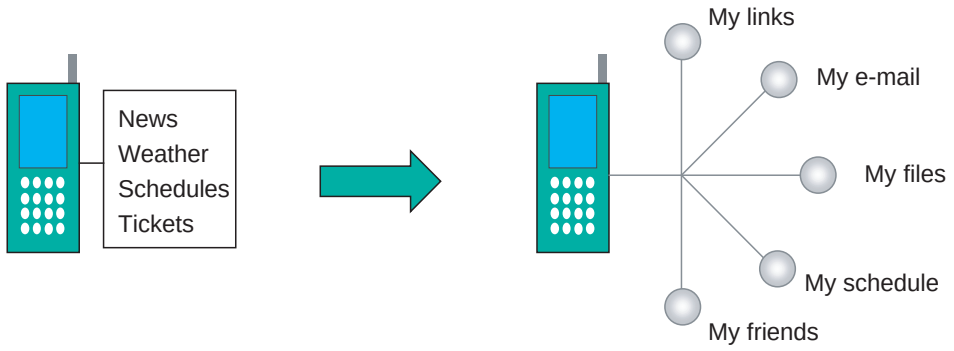
The user may perceive the service as being familiar because it resembles other mobile services that he or she has been using or because it resembles the same service or brand in a different environment such as Web, TV or newspaper. For instance, in our mobile Internet studies teletext services converted from the Web were well accepted because of their familiarity. Familiarity was also related to the provider of the service, as test users pointed out that they preferred using, for instance, news services from a familiar and trusted service provider.

The Service Should Provide Personal and User-Generated Content

Personalization is not just about selecting services and contents within services but also about making the user's own personal items available, as illustrated in Figure 6 by the setup of the personal mobile Internet pages in our trials. Also with location-based services the users appreciated the possibilities to complement, for example map data with their own information such as important places, favorite routes, and self-written notes.

In the mobile Internet trials many users were keen to use services such as discussion groups where they could contribute as content providers. Letting the mobile users contribute to content cre-

Figure 6. The shift from common to personal increases the appeal of the services



ation could enhance many services. Such content may enrich the service, bring in additional users and encourage a sense of community among users. For instance, information generated by users at a particular location may be of interest to the next visitors. With the growing trend of social media services, the role of users' own content generation is expected to become increasingly important. Mobile users have key roles in many social media services as they can contribute by bringing in topical information from the field, such as mobile video of important occasions.

The Users Appreciate Seamless Service Entities Rather than Separate Services

In the mobile Internet trials it turned out that usage needs for many individual services were quite occasional, even if the users would have assessed the services as being very useful in those occasional situations. The value of mobile Internet to the user was based on the wide selection of services rather than any individual service.

The studies with location-based services pointed out the need for seamless service entities, whereby the user is supported throughout the whole usage situation, for example while looking for nearby services, getting information on the services, contacting the services, and getting route guidance to find those services. The usage may even extend from one terminal device to another.

The Services Need to Provide Utility, Communication or Fun

In addition to personally selected content, interactive services also take mobile services to a more personal level, providing the users with new ways of communicating and participating. A mobile phone is basically a communication device and thus it is no wonder that services that enhanced or enriched communication were well accepted in our field trials.

Location-awareness can provide the users with services that are really intended for mobile use, not just secondary access points to Web services. Examples of such services include traffic information, weather forecasts, route guidance, travel information, event information and help services in emergency situations. Those services turned out

to be popular as location-awareness made them both easier to use and more personal.

Perceived Ease of Use

Many ease of use attributes are already well known but as mobile services are getting increasingly complex and enhanced with new characteristics such as personalization and context-awareness, new usability challenges are raised. Key design principles that in our trials turned out to affect the perceived ease of use of mobile services are described in the following.

Clear Overview of the Service Entity

The most common usability problem with both mobile Internet services and location-based services was that when accessing the services, the users did not know what to expect from the service. The users would need a clear and intelligible overview of the whole range of available information, services and functions. The first impression may encourage and motivate the user or frighten him/her away. Enough design efforts and user evaluations should be invested in designing the main structure and the front page of the service. There are already efficient solutions available such as Minimap introduced by Roto (2006). Minimap gives the overview by showing a miniaturized version of the original Web page layout on the mobile device, and Roto's (2006) studies showed that this approach clearly improved the usability of Web browsing with mobile phones.

The information and functions that the user will most probably need should be the easiest to access. By proceeding further, the user should be able to access any information available within the service. Occasional usage typical of mobile services emphasizes the need for a clear overview of available services, including information on how the service should be used, where the content comes from, how often it is updated, and how comprehensive it is.

Fluent Navigation on a Small Screen

The mobile Internet trials showed that a single scrollable page (Figure 7) is good for browsing through information, whereas separate pages are better for navigation. The users need ways to browse quickly through less interesting information: for instance, an adaptive scroll speed and an illustrative scroll bar are useful.

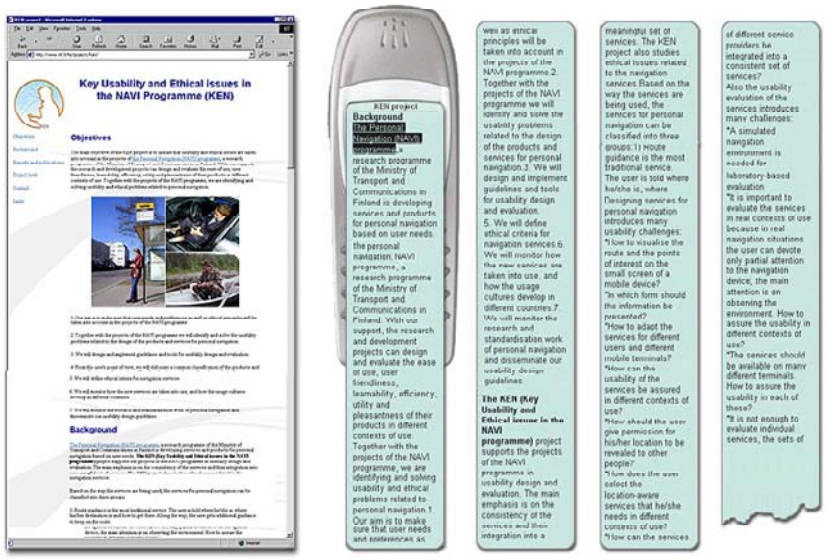
The user needs clear feedback on which service and where in it he or she currently is. In our evaluations, this was facilitated by descriptive and consistent link/page header pairs for back, forward, exit, home and other safe heavens within the service.

The usability of the sites can be further improved by making the structure adaptive according to each user. A novice user may want to get instructions first, whereas more experienced users may want to go straight into the service. For frequent users, the structure could be adaptive so that the most recently or most often used items are easily available.

Smooth User Interaction with the Service

User interface restrictions of mobile devices and the implementation of the user interface elements may hinder smooth user interaction. In our trials text input was often a major effort for the users, especially if the usage took place while moving. Still the users needed and wanted to give input to the services. In the mobile Internet evaluations ready-made selection lists turned out to be useful when the user was getting acquainted with the service, whereas experienced users often preferred text input. Preferably, both alternatives should be available. Text input should be predicted and interpreted to suggest corrections to possible misspellings. Location-awareness as such could also

Figure 7. On a small screen, there is a lot to scroll, even when accessing a simple Web page



be utilized in text input, as suggested by Ancona et al. (2001). For instance, nearby streets or the most popular search terms at a certain location could be suggested to the user. Even though user input may be tedious, it should not be avoided in the services. However, the user should not be obliged to input information that is not absolutely necessary.

Personally Relevant Services and Information without Expending Effort on Personalization Set-up

Our trials repeatedly showed that users were not willing to do much regarding personalization, although they would have appreciated the personalized solution. Personalization should be voluntary, and strongly supported in the beginning.

Users could be provided with ready-made service package alternatives, as we did in our mobile Internet trials, or they could be guided through personalization services. New service offerings could be sent automatically based on user profiles if the user accepts that. The user should be able to see and refine the personalization with his or her mobile device on the fly, even though the personalization could also be done with a desktop PC. New approaches such as group profiles, profiles shared by several services and learning profiles may ease personalization.

Easy Access to Situationally Relevant Information and Services

Mobile contexts vary a lot and may even change in the middle of a usage session. Our trials with location-based services showed that in services targeted at a limited area, such as travel guides, service catalogues and event guides, the context of use can be predicted quite well according to

user location and time. This gives possibilities for different context-aware features in the services, easing their use and giving the users personalized access to the services.

Location-awareness can be utilized to provide the users with local services such as tourist guides, event information and shopping guides. Context-awareness can be complemented with personalization to adapt to user preferences that in different contexts may vary from one individual to another. This may, however, indicate lots of personalization efforts.

Facilitating Momentary Usage Sessions on the Move

On the move the users can devote only part of their attention to using the service while their main attention is on their main task of moving. In our trials with location-based services, on-the-move use was typically non-continuous. A user could, for instance, activate a route guidance service and start using it but occasionally he or she had to put the device aside and do something else. Later on he or she returned to the service. For these kinds of usage sessions task resumability should be supported both in the terminal device and in the services. Pousman et al. (2004) point out that resumability can be supported, for example by atomic interaction sessions, by appropriate timeouts on unfinished operations, and by a stateless interaction model. The users should be able to use the services both on and offline.

Design for Device and Network Variety

One of the main challenges in designing mobile services is the growing variety of mobile devices, networks and other infrastructures. The Design for All approach (EDeAN, 2007) with regard to mobile services requires taking into account all kinds of devices, not just the most advanced ones. In our development work on mobile Internet services we found that a good starting point is a simple service, suitable for any device. The usability and the attractiveness of the service can then be improved by utilizing the unique features of each device in separate implementations. Our experiences from mobile Internet trials show that in mobile environments there may be needs for adaptive search services that would not only look for particular content, but also take into account the current client device. The search results could be prioritized according to how suitable the content is for the device and network that the user is currently using.

Trust

In the TAMM model, user trust in mobile services is quite a wide concept that includes perceived reliability of both the technology and the information and functions provided, reliance on the service in planned usage situations, and the user's confidence that he or she can keep the service under control and that the service will not misuse his or her personal data. The design principles that in our evaluations turned out to affect user trust in mobile services are described in the following.

The User should be able to Rely on the Service in Intended Contexts of Use

In our user trials errors with mobile services were often difficult to cope with for the users as they did not know whether the problems were in the mobile device, in the network or in the services. Repeated malfunctions that the user could not understand or solve were a major source of bad usage experiences and often made the user stop using the service in question. To avoid these kinds of situations, the user should get easy-to-understand information to help him or her to understand and recover from the error situation. User errors should be prevented by all means, for example by trying to interpret,

correct or complete user input. In the event of the user losing the connection to the service, it should be assured that no harm will be done.

With location-based services the users often would have liked to get feedback on the power still available and estimates of the sufficiency of batteries with different combinations of add-on devices and functions (Kaasinen, 2003). A user on the move may need to make decisions regarding which combination of functions he or she can afford to keep on in order to avoid exhausting the battery power totally.

Evaluations of personal navigation scenarios and prototype services revealed that users may get quite dependent on mobile services such as navigation services. That is why the users should be made aware of the possible risks of using the product and they should be provided with information about the reliability of the service so that they can assess whether they can rely on the service in the planned usage situations.

Measurement without Estimated Accuracy is of no Use

The accuracy of the location information was often questioned in our trials. In addition to location, future mobile services will be using and providing the user with increasing amounts of different measurement data (Kaasinen et al., 2006). Accuracy requirements for the data need to be considered in the design. The accuracy should be sufficient for the kinds of tasks for which the user will be using the service. The users should get feedback on the freshness of the data and its accuracy, especially if these vary according to the usage situation. Both actual reliability and perceived reliability need to be ensured in the design as these may be only loosely mapped, as found out by Kindberg et al. (2004).

Context-aware systems have several error possibilities: the system may offer the user wrong things either because it predicted the context wrongly or because it predicted the context correctly but predicted the user's needs in that context wrongly. Displaying uncertainty to the user may improve the acceptability of the services by making them more intelligible, as pointed out by Antifakos et al. (2004).

The Privacy of the User must be Protected Even if the User would not require it

User data should be protected even if—like in some of our trials—the users themselves would be trusting enough not to require it. The user should be provided with easy mechanisms for giving permission to use the data for a predefined purpose. Histories of user data should not be stored purposelessly and without user consent. When location data is conveyed to others, it is worth considering whether they will need the exact location coordinates or a more descriptive but less intrusive description. It should also be considered whether it is necessary to connect personal data to the user identity.

The legislation in most countries requires the user's permission before he or she can be located. Also social regulation can create rules and norms for different situations in which location-aware services are used (Ackerman et al., 2001). In practice, trade-offs between privacy protection and effortless use need to be resolved.

In future services, it can be expected that in addition to user location, a lot of other personal data may be collected. This may include health-related measurements, shopping behavior; services used and so on (Kaasinen et al., 2006). The same principles as with location are to great extent valid also with this data.

The User Needs to Feel and Really be in Control

The more complicated the mobile services and the service networks behind them get, the less possibilities the user has to understand what is

happening in the service. The services need to be somewhat seamless to ensure effortless use. On the other hand, some issues need to be clearly differentiated so as to ensure that the user understands what is going on. Seamless services may hide details from users when aiming to provide ease of use. This may prevent the user from understanding what is happening "behind the scenes" (Höök, 2004).

Based on the findings of the trials with location-aware services, the main user requirement is that the user needs to feel and really be in control. For instance, the users more easily accepted context-aware behavior of the services if they could understand the reason for the behavior. To be able to be in control, the user needs to understand enough about the system's capabilities and rules of reasoning. The user needs to get feedback on what is going on and why, even if it is unnecessary to understand all the details. As automated functions may take control away from the user, the user should be able to control the degree of automation and intrusiveness. The user should be able to override the recommendations of the system, as suggested by Cheverst et al. (2000).

Similar to the findings by Cheverst et al. (2002), also in our trials the users tended to accept push services because of the effortless use. However, as the amount of push features grows, the attitude of the users may soon change. That is why the user should be able to fine-tune or cancel the push feature easily—ideally as he or she receives a push message.

Perceived Ease of Adoption

As mobile services will increasingly be available from different sources and in complex service networks, it becomes important to ensure that the users get reliable information about available services and the necessary guidance when taking the services into use. Based on user feedback in our trials, key design principles regarding ease of adoption of mobile services are described in the following.

Real Values of the Services Need to be Emphasized in Marketing

Users often have a poor understanding of mobile devices and services (Kolari et al., 2002). The users may have misconceptions about the services behind acronyms or different technologies. In our trials the users were often unaware of the features and services available on their personal phones.

As a part of our research work a Trade Description Model (Kaasinen et al., 2002) was set up to help consumers to compare different products and, on the other hand, to help service providers to describe their products in a consistent way (Table 2). Although the model was designed for personal navigation services, it is general enough to be adopted for the description of other mobile services as well. The trade description model can also be used as a checklist of issues to be covered when writing "Getting started" manuals.

Disposable Services for Occasional Needs

IBM has issued guidelines on how to design out-of-box experiences that are productive and satisfying for users (IBM, 2005). Ideally, the services should be installed on the user device at the point of sale, and the user should at the same time get personal usage guidance, but presumably this will be possible with only a few services.

In our trials with location-based services, the users often said that they wanted to have the services easily available when a spontaneous need for a certain service arose. Context-aware services pose additional challenges for taking new services into use. The services may be available only locally or in certain contexts. The user should be able to identify, understand and take into use these services easily while on the move.

Table 2. Trade description model for personal navigation products and services (Kaasinen et al., 2002)

Classification	*Trade description*
User	**Is this product/service suitable for me?** • Targeted specially at a certain user group • Targeted only at a certain group • Accessibility for disabled users
User goal	**What can I do with this product / service?** • Locate myself • Be located by other people • Locate other people • Track my property • Get route guidance • Find and use nearby services • Get help in emergency situations • Have fun
Environment	**Where can/cannot I use this product/service?**
Equipment	**What do I need to know about the technology?** • What kind of technology do I need to be able to use the service? • How compatible is this product/service with other products/services? • How accurate is the positioning? • To what extent can I rely on this product?
Service characteristics	**What specific features does this service include, what is the added value of this product compared with competing products or current ways to act?**

As the selection of available services grows, it will also become increasingly important to get rid of unnecessary services easily.

The Service has to Support Existing and Evolving Usage Cultures

Personal mobile devices should be designed to be both intuitive for first-time use and efficient in long-term use (Kiljander, 2004). This is true also with mobile services, which should be designed for gradual learning. New services shape the usage, but the usage should also shape the services (Norros et al., 2003). Existing and evolving usage cultures should be studied in parallel with the technology development to identify and support natural usage patterns. The design should fit in with the social, technical and environmental contexts of use, and it should support existing usage cultures. Ideally, the technology should provide the users with possibilities that they can utilize in their own way, rather than forcing certain usage

models fixed in the design (Norros et al., 2003). Although the users will benefit from clear usage guidance, they should also be encouraged to discover and innovate their own ways to utilize new services.

FUTURE TRENDS

The current technology acceptance model for mobile services (TAMM) is based on studies with mobile Internet services and location-based information services targeted for consumer use. The identified user acceptance factors can be utilized in designing these kinds of services, but they can also be applied when designing other kinds of mobile services. In future visions, mobile devices are increasingly interacting with their environment and are transforming into tools with which the user can orient in and interact with the environment. As the user moves from one environment to another, the available services will change accordingly (Kaasinen et al., 2006). These kinds of services will require extreme ease of adoption, and, as the services will increasingly deal with personal data, the user's trust in the services will become an even more important user acceptance factor.

Further studies will be needed to study the mutual relations of the four user acceptance factors. As with the original TAM, the model can be enhanced by studying key forces underlying the judgments of perceived value, perceived ease of use, trust and perceived ease of adoption.

The technology acceptance model for mobile services was set up by analyzing and combining the results of several individual evaluation activities of different mobile services. When developing future mobile technologies and infrastructures, human-centered design can be expanded similarly. By synthesizing and generalizing the results of parallel research activities, key user acceptance factors and design implications for future service development can be identified.

The technology acceptance model for mobile services seems to have potential as a framework for ubiquitous computing applications as well. The model has already been successfully applied in connection with a project that aims to develop a mobile platform for ubiquitous computing applications that utilize wireless connections to sensors and tags (Kaasinen et al., 2006).

CONCLUSION

In this chapter, the technology acceptance model for mobile services has been introduced. According to the model, user acceptance of mobile services is built on three factors: perceived value of the service, perceived ease of use, and trust. A fourth user acceptance factor: perceived ease of adoption is required to get the users from intention-to-use to actual usage. Based on the technology acceptance model for mobile services, design implications for each user acceptance factor have been proposed.

Instead of implementing collections of useful features, the design of mobile services should be focused on key values provided to the user. The value of mobile services can be built on utility, communication or fun. Successful service content is comprehensive, topical and familiar, and it includes personal and user-generated content. The users appreciate seamless service entities rather than separate services. Ease of use requires a clear overview of the service entity, fluent navigation on a small display, and smooth user interaction with the service. The users should get personally and relevant services and information without needing to expend effort on personalization. The services should be designed to be adaptive to a wide variety of devices and networks. As the services increasingly support individual users in their daily tasks and increasingly deal with personal data, user trust in the services is becoming more and more important. The user should be able to assess whether he or she can rely on the service

in the intended contexts of use. The user needs to feel and really be in control, and the privacy of the user must be protected.

Occasional usage and momentary usage sessions on the move are typical of mobile services. In addition, services are increasingly available only locally or in certain contexts of use. This indicates the need for disposable services: services that are easy to find, take into use, use and get rid of when no longer needed. The user needs realistic information about the actual values of the services, so that he or she can realize how to utilize the service in his or her everyday life and discover new usage possibilities.

The technology acceptance model for mobile services provides a tool to communicate key user acceptance factors and their implications to the design. The model can be used in all design and evaluation activities throughout the design process, but it is especially useful in identifying issues that should be examined in field studies.

REFERENCES

Ackerman, M., Darrel, T., & Weitzner, D. J. (2001). Privacy in context. *Human-Computer Interaction, 16,* 167–176.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behaviour.* Prentice Hall.

Ancona, M., Locati, S., & Romagnoli, A. (2001). Context and location aware textual data input. *SAC 2001* (pp. 425-428). Las Vegas: ACM.

Antifakos, S., Schwaninger, A., & Schiele, B. (2004). Evaluating the effects of displaying uncertainty in context-aware applications. In Davies, N., Mynatt, E., & Siio, I. (Eds.), in *Proceedings of Ubicomp 2004: Ubiquitous Computing 6th International Conference* (pp. 54-69). Springer-Verlag.

Barnes, S. J., & Huff, S. L. (2003). Rising Sun: imode and the wireless internet. *Communications of the ACM, 46*(11), 79–84.

Billsus, D., Brunk, C. A., Evans, C., Gladish, B., & Pazzani, M. (2002). Adaptive interfaces for ubiquitous Web access. *Communications of the ACM*, *45*(5), 34–38.

Boehm, B. (2003). Value-based software engineering. *Software Engineering Notes, 28*(2), 1–12.

Chen, L., Gillenson, M. L., & Sherell, D. (2004). Consumer acceptance of virtual stores: A theoretical model and critical success factors for virtual stores. *ACM SIGMIS Database archive*, *35*(2), 8–31.

Cheverst, K., Davies, N., Mitchell, K., Friday, A., & Efstratiou, C. (2000). Developing a context-aware electronic tourist guide: some issues and experiences. *CHI 2000 Conference Proceedings* (pp. 17–24). ACM.

Cheverst, K., Mitchell, K., & Davies, N. (2002). Exploring context-aware information push. *Personal and Ubiquitous Computing, 6*, 276–281.

Cockton, G. (2004a). From quality in use to value in the world. In *Proceedings of CHI2004* (pp. 1287-1290). ACM.

Cockton, G. (2004b). Value-centred HCI. In *Proceedings of the Third Nordic Conference on Human-Computer Interaction* (pp. 149–160).

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quartely, 13,* 319–339.

Davis, F. D., & Venkatesh, V. (2004). Toward preprototype user acceptance testing of new information systems: implications for software project management. *IEEE Transactions on Engineering Management, 51*(1).

EDeAN. (2007). European design for all e-accessibility network. *Homepage*. Retrieved February 13, 2007, from www.e-accessibility.org

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of CHI 99 Conference* (pp. 80–87).

Funk, J. L. (2004). *Mobile disruption. The technologies and applications driving the mobile Internet.* Wiley-Interscience.

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega: International Journal of Management Science, 28,* 725–737.

Gefen, D., & Devine, P. (2001). Customer loyalty to an online store: The meaning of online service quality. *Proceedings of the 22nd International Conference on Information Systems* (pp. 613–617).

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Inexperience and experience with online stores: The importance of TAM and Trust. *IEEE Transactions on Engineering Management, 50*(3), 307–321.

Henderson, A. (2005). Design: The innovation pipeline: Design collaborations between design and development. *ACM Interactions, 12*(1), 24–29.

Höök, K. (2004). Active co-construction of meaningful experiences: but what is the designer's role? *Proceedings of the Third Nordic Conference on Human-Computer Interaction* (pp. 1-2). ACM Press.

IBM. (2005). Out-of-box experience. Retrieved January 10, 2005, from www-3.ibm.com/ibm/easy/eou_ext.nsf/publish/577

ISO 13407. (1999). *Human-centred design processes for interactive systems.* International standard. International Standardization Organization. Geneve.

Järvenpää, S. L., Lang, K. R., Takeda, Y., & Tuunanen V. K. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41–44.

Kaasinen, E. (2003). User needs for location-aware mobile services. *Personal and Ubiquitous Computing, 6,* 70–79.

Kaasinen, E. (2005a). User acceptance of location-aware mobile guides based on seven field studies. *Behaviour & Information Technology,24*(1), 37–49.

Kaasinen, E. (2005b). *User acceptance of mobile services—Value, ease of use, trust and ease of adoption.* Doctoral dissertation. VTT Publications 566. Espoo: VTT Information Technology.

Kaasinen, E., Aaltonen, M., Kolari, J., Melakoski, S., & Laakko, T. (2000). Two approaches to bringing internet services to WAP devices. *Computer Networks, 33,* 231–246.

Kaasinen, E., Ermolov, V., Niemelä, M., Tuomisto, T., & Välkkynen, P. (2006). *Identifying user requirements for a mobile terminal centric ubiquitous computing architecture.* FUMCA 2006: System Support for Future Mobile Computing Applications. Workshop at Ubicomp 2006.

Kaasinen, E., Ikonen, V., Ahonen, A., Anttila, V., Kulju, M., Luoma, J., & Södergård, R. (2002). *Products and services for personal navigation—Classification from the user's point of view.* Publications of the NAVI programme. Retrieved January 4, 2005, from www.vtt.fi/virtual/navi

Kaasinen, E., Kasesniemi, E.-L, Kolari J., Suihkonen, R., & Laakko, T. (2001). Mobile-transparent access to web services—Acceptance of users and service providers. In *Proceedings of International Symposium on Human Factors in Telecommunication.* Bergen, Norway

Kiljander, H. (2004). *Evolution and usability of mobile phone interaction styles.* Doctoral thesis. Helsinki University of Technology. Publications in Telecommunications Software and Multimedia. TML-A8. Espoo: Otamedia.

Kindberg, T., Sellen, A., & Geelhoed, E. (2004). Security and trust in mobile interactions—A study of users' perceptions and reasoning. In Davies, N., Mynatt, E., & Siio, I. (Eds.), *Proceedings of Ubicomp 2004: Ubiquitous Computing 6th International Conference* (pp. 196-213). Springer-Verlag.

Kolari J., Laakko T., Kaasinen E., Aaltonen M., Hiltunen T., Kasesniemi, E.-L., Kulji, M., & Suihkonen, R. (2002). *Net in pocket? Personal mobile access to web services.* VTT Publications 464. Espoo: Technical Research Centre of Finland.

Mathieson, K., Peacock, E., & Chin, W. W. (2001). Extending the technology acceptance model: The influence of perceived user resources. *The DATA BASE for Advances in Information Systems, 32*(3), 86–112.

Nah, F. F.-H., Siau, K., & Sheng, H. (2005). The value of mobile applications: A utility company study. *Communications of the ACM, 48*(2), 85–90.

Norman, D. A. (1998). *The invisible computer.* Cambridge, MA: MIT Press.

Norros, L., Kaasinen, E., Plomp, J., & Rämä, P. (2003). *Human-technology interaction. Research and design. VTT Roadmap.* VTT Research Notes 2220. Espoo: Technical Research Centre of Finland.

Pousman, Z., Iachello, G., Fithian, R., Moghazy, J., & Stasko, J. (2004). Design iterations for a location-aware event planner. *Personal and Ubiquitous Computing, 8*, 117–125.

Rogers, E. M. (1995). *The diffusion of innovations.*(Fourth Ed.). New York: Free Press.

Roto, V. (2006). *Web browsing on mobile phones—Characteristics of user experience.* Doctoral dissertation. Espoo: Helsinki University of Technology.

Sarker, S., & Wells, J. D. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM, 46*(12), 35–40.

Venkatesh, V., & Davis, F. D. (2000). Theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478.

KEY TERMS

Ease of Adoption (TAMM): Perceived ease of identifying, understanding and taking into use new products.

Innovation Diffusion: User adoption of different innovations in target populations

Location-Aware Service: A special case of location-based service: a mobile service that adapts according to the location.

Location-Based Service: A mobile service that utilizes location data.

Perceived Ease of Use (TAM and TAMM): The degree to which a person believes that using a particular system would be free from effort (Davis, 1989).

Perceived Usefulness (TAM): The degree to which a person believes that using a particular system would enhance his or her performance in a certain task (Modified from Davis, 1989).

Technology Acceptance: User's intention to use and continue using a certain information technology product (Davis, 1989).

Technology Acceptance Model (TAM): Technology acceptance models aim at studying

how individual perceptions affect the intentions to use information technology as well as the actual usage. The Technology Acceptance Model was originally defined by Davis (1989), but it has subsequently been modified and augmented by other researchers.

Technology Acceptance Model for Mobile Services (TAMM): Extension of the original Technology Acceptance Model to take into account the specific characteristics of mobile services (Kaasinen, 2005b)

Trust (TAMM): An indicator of a positive belief about the perceived reliability of, dependability of, and confidence in a product (modified from Fogg & Tseng, 1999).

Value (TAMM): The key features of the product that are appreciated by the users and other stakeholders, i.e. the main reasons why the users are interested in the new product (Kaasinen, 2005b).

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 102-121, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.23
User–Centered Mobile Computing

Dean Mohamedally
City University London, UK

Panayiotis Zaphiris
City University London, UK

Helen Petrie
City University London, UK

INTRODUCTION

Mobile computing and wireless communications continue to change the way in which we perceive our lifestyles and habits. Through an extensive literature review of state-of-the-art human-computer interaction issues in mobile computing (Mobile HCI), we examine recent pertinent case studies that attempt to provide practical mobile capabilities to users. We thus contribute to the reader a primer to the philosophy of developing mobile systems for user centred design.

User centred design elicits the needs and requirements of end users. Its purpose in mobile systems is to enable useful computing and communicating experiences for diverse types of users, anywhere at anytime and on demand. We shall therefore illustrate to the reader some of the key constraints of mobile devices such as limited visuals, contextual awareness and mobility itself, and more importantly how they can be overcome through innovative design and development.

INFORMATION VISUALISATION

One of the most fundamental objectives in the miniaturisation of computer technologies is to present a platform from which users can maintain usable levels of interaction with their data from wherever they are. Information visualisation has come a long way since the days of two-colour text-only format screens. Yet constraints determined by factors of physical engineering

feasibility, such as screen quality and resolution, battery longevity and network capabilities, give us a particularly popular arena for exploration in mobile HCI research.

Such constraints are being addressed in a number of novel interfaces, such as Electronic Ink-based screens (E-Ink, 2004) that have a high resolution similar to that of natural paper, have low power requirements, and will eventually be capable of being rolled up for storage. TabletPCs (Microsoft, 2003) have the capability to use ultra-low powered pressure-sensitive pen input devices to elicit the amount of pressure incurred and also capture motion gestures on a visual interactive user interface, creating a sense of depth perception visualisation. They also enable very distinctive handwriting recognition without the need for learning preset handwriting letter shapes.

The Rapid Serial Visual Presentation (RSVP) concept (Bruijin, Spence, & Chong, 2001; Goldstein, Oqvist, Bayat-M, Ljungstrand, & Bjork, 2001) is one of many research investigations into methods of presentation of information on a small screen (Jones, Buchanan, & Thimbleby, 2002). By the beginning of 2001, over 88 million WAP (Wireless Access Protocol) hits on mobile phone screens were made in the UK alone (WAP Forum, 2003), and therefore significant effort has been undertaken on design factors of WAP site browsing globally.

LOCATION AND GEOGRAPHIC AWARENESS

Much of the research in geographic and location aware mobile systems correlate with mixing user requirements in information visualisation with geographic sensors. For example, audio user interfaces for Global Positioning Satellites (GPS) receivers have typically been designed to meet the needs of visually handicapped users by giving audible signals as they travel past real-world coordinates that have been pre-designated in their system.

The proliferation of GPS-based location services has seen an impressive array of uses for location awareness in mobile deployment activities. They are now becoming an embedded part of upcoming generations of mobile phones and smart personal devices in mass consumer products. Examples include experimental tourist guides (Cheverst, Mitchell, & Davies, 2001; Davies, Mitchell, Cheverst, & Blair, 1998), navigation systems for disabled and elderly people (Holland & Morse, 2001; Petrie, Furner, & Strothotte, 1998), and also in location aware collaborative systems (Rist, 1999).

Close range location awareness technologies include Bluetooth (1998) based and RFID—Radio Frequency Identification (2003) based devices with which broadcast points can beam radio data to compatible handheld receivers. The results of current research in this domain is leading to opportunities in a variety of user scenarios that are made aware of your unique presence, for example, walking into a personally aware room will turn on the lights at your chosen settings, or location enabled advertising billboards will be able to read your public profile and communicate suitable electronic media to you wirelessly.

CONTEXTUAL AWARENESS

An issue in HCI research is investigating models and scenarios for maintaining consistency between the user's understanding of their environment, the understanding of the environment reported by the system and the actual state of the environment. Context-aware systems have to react not only to the user's input but also input (i.e., context) from the user's environment (Brown, Bovey, & Chen, 1997; Schilit, Adams, & Want, 1994). This offers opportunities for helping people to accomplish their goals effectively by understanding the value of information.

A realisation in this domain involves a specifically mobile systems orientated question—do we push the contextual information into the mobile system as they move within monitored zones or pull it on demand at their request? One of the challenges in context awareness is discovering computing services and resources available in the user's current environment, utilising discovery protocols such as Jini (2001) and SDP by Czerwinski, Zhao, Hodes, Joseph, and Katz (1999).

SENSORY-AIDED MOBILE COMPUTING

Mobile HCI has also changed the nature of computing for the demographics of users that have sensory disabilities, or alternatively require sensory enhancement. A low visibility prototype with supplemented tactile cues is presented in Sokoler, Nelson, and Pedersen (2002) with the TactGuide prototype. This is operated by subtle tactile inspection and designed to complement the use of our visual, auditory and kinesthetic senses in the process of way finding. It was found to successfully supplement existing way finding abilities. A mobile system that lets a blind person use a common laser pointer as a replacement of the cane in demonstrated by Fontana, Fuiello, Gobbi, Murino, Rocchesso, Sartor, and Panuccio (2002), who presented an electronic travel aid device that enables blind individuals to "see the world with their ears." A wearable prototype was constructed using low-cost hardware with the ability to detect the light spot produced by the laser pointer. It would then compute its angular position and depth, and generate a correspondent sound providing the auditory cues for the perception of the position and distance of the pointed surface. Another wearable system for blind users to aid in navigation is presented by Petrie et al. (1998), which projects a simple visual image in tactile form on the back or stomach.

Aside from aiding those disabilities it should be noted that sensory enhancement is an area for particular growth in Mobile HCI. Mobile systems that can augment the senses such as vision with heat and electrical sensors and sonic receivers are all ideas that can be investigated with undoubtedly a wide arena of scenarios.

COLLABORATIVE SYSTEMS

Mobile systems in general are becoming refined as instruments for co-operative wireless computing communications in various forms of collaborative HCI. One successfully enhanced scenario is presented with air traffic controllers by Buisson and Jestin (2001). They constructed an effective solution for a distributed interaction prototype that would assist desk-based systems with the collaboration of a mobile operations manager in a fast paced and safety critical environment. The ability to work in mobile teams simulating CSCW (Computer Supported Collaborative Work) models is an important consideration for time critical and location dependent processes.

Collaboration is also an extensive area in augmented reality systems like that found in Nigay, Salambier, Marchand, Renevier, and Pasqualetti (2002). Augmented reality systems overlay information visualisation on top of physical views of the real world. Here it addresses the combination of the physical and digital worlds seamlessly in the context of a mobile collaborative activity. Mobile collaboration also takes place in VNC (Virtual Network Computing) based solutions for mobile devices such as PalmVNC for Palm PDAs (Minenko, 1998). VNC technology allows several users to view a desktop remotely and may allow them access to basic interactivity. This has a significant scope for future research as a remote-access collaboration method.

HOME AND DEVICE CONTROLLERS

Much of the research in the area of mobile human-computer interaction has focused on the user interfaces to the mobile devices themselves such as their input methods and displays. We can envisage how devices can fit into user-centred domains of information and control space.

Mobile device controllers for the home are still currently a relatively new and underdeveloped area for consumer interests beyond the usual remote controls (Weiser, 1991). We find now that protocols are being developed by companies and peer driven international committees, such as those that will enable the future wave of mobile and non-mobile devices to co-operate together on much more ubiquitous levels and facilitate growth in this area. Examples of these include Bluetooth short range wireless networking, Jini (2001) embedded devices networking in everyday home consumer equipment like light switches and kettles, and HAVi (2000) home media connectivity networking for audio visual equipment, to name a few.

An example of modelling future simulations of device controllers is found in Huttenrauch and Norman (2001), where a device is simulating the control of household robots. A popular consumer orientated protocol being corporately developed is the X10 (2000) protocol, which can control and relay electrical hardware information to other protocols such as e-mail and Internet connectivity. This combined with mobile scenarios can give rise to a host of ubiquitous and ambient intelligent hardware in physical locations. For example, one may wish to turn on their house lights, open the car garage and start the kettle boiling remotely with a single text message home to an X10 driven mobile interfacing system.

SOCIOLOGICAL VIEWS

Sociological aspects of mobile HCI are changing the way we live our everyday lives beyond our desktop computers at work and at home and there are new paradigms forming in these areas. For example, social communication by SMS text messaging using innovative short hand letter sequences rather than usual language syntax, and the development of assistive technologies such as Tegic Corp's patented T9 (1997) text prediction algorithm.

SMS text messaging and mobile chat messaging has changed our dimension of communications, despite its weak and sometimes unreliable connectivity. It is an asynchronous channel of communication that operates upon a principle of "store and forward": the sender sends a message when his or her device has a connection, and the message is forwarded to the recipient when the recipient's device has a connection. Given its asynchronous nature, we find that it is less obtrusive than real-time communications to utilise and respond to, as users may reply at their discretion. This contrasts with traditional voice telephony over mobiles—a synchronous channel of communication, which requires both mobile devices to establish a connection simultaneously.

The future of text messaging has been described as the advent of picture and video messaging and streaming, and it is already becoming apparent that its usage patterns are changing our cultural perspectives. In some countries it is already banned for religious, security, and privacy reasons as any unsuspecting person or entity can be the subject of a discrete imaging mobile device. Text messaging in combination with instant messaging and blogging (an Internet-viewable personal diary) techniques will present interesting dimensions to our social patterns.

MOBILE-BASED LEARNING SYSTEMS

M-learning systems that utilise mobile technologies and models of ubiquity are an area for growth in mobile HCI, though popular in their own right. Primarily considered to be a classical model of knowledge presentation in mobile and wireless classroom scenarios, the blackboard model has had numerous developments to enhance the capabilities of electronic learning, such as Chang and Sheu's (2002) ad hoc classroom system which enables students to migrate their daily activites to PDAs for digital recording of all of their events and contributions.

Learner-Centred Design (LCD) is an approach to building software that supports students as they engage in unfamiliar activities and focuses on enabling them to learn about a new area. LCD has been successfully used to support students using desktop computers for a variety of learning activities, and in Luchini, Quintana, and Soloway (2002), LCD is extended to the design of educational software for handheld computers. Here they presented a case study of ArtemisExpress, a tool that supports learners using handheld computers for online research. The results from this demonstrate that while user-centred design methods typically focus on software to support the work of expert computer users, LCD techniques in mobility focuses on directly providing learners with the educational support needed to learn about the content, tasks, and activities of the new domain they are exploring at their own pace and in their own environments.

NAVIGATION AND READABILITY

Voice recognition and synthesis has come an impressive way in Computer Science. Motorola's Mya Voice Browser as described by Chesta (2002) uses Automatic Speech Recognition (ASR) to understand and process human speech, capture requested information from voice-enabled Web sites, and then deliver the information via pre-recorded speech or Text-To-Speech (TTS) synthesis software that "reads" the relevant data to the user. The issues of internationalisation with such a system are covered in their research.

Researchers have to remain critical of the choices of navigation design, as found in Chesta (2002). An important requirement in HCI evaluations is to derive the usability and functional accuracy of a designed system in its domain. As Chittaro and Cin (2001) found in their results, the WAP/WML protocol navigation capabilities needs to be reviewed for user performance, in particular they observed the WAP methods for navigating links, list of links, action screens and selection screens.

GRAPHICS ENGINEERING

Computer graphics engineering associates closely with HCI, especially where constraint user interfaces are concerned. In general, researchers have been trying to find the most perceptually accurate and aesthetically pleasing representations for allowing humans to access and visualise computational information as responsively as possible. Several distinct disadvantages of current day mobile systems is demonstrated by the lack of screen space available and the hardware demands of the limited processing and power usage capabilities available to generate fast computer graphics. As technologies and industry standards develop however, some of these constraints will be removed altogether.

Constraint visual computing experiences are pushing mobile user interface requirements into constructing new and more powerful miniature hardware and software for the support of video and

real-time 2D/3D acceleration. This can be seen in the impressive work by the Khronos group (2002) to construct an open platform for mobile graphics software developers. NVidia Corp's mobile embedded graphics processors (2004) augment this with new research opportunities by creating platforms for embedded real-time 3D computer graphics processing on mobile devices.

CONCLUSION

We have presented a review of current state of the art issues in user centred mobile systems that researchers have been involved in, and explored some of their solutions. In this youthful field, research that has published applications and techniques from the period 1997 – 2004 are going on to influence research directions in the field of mobile HCI and have been described in this section as several key categories of computer science.

The future of mobile HCI research holds great promise in the culmination of the user centred design issues as noted, which will lead into the field of ubiquitous computing. This will stem from trends in digital capture, processing and presentation of real-time and real-world data that is embedded in our environment. For users of mobile systems, it gives hope to the development of systems that will one day provide tools that can react, adapt and assist our dynamic lifestyles and enhance both our naturally individual and collaborative ways of life.

REFERENCES

Bluetooth official membership site (1998). Retrieved on March 14, 2004, from *http://www.bluetooth.org*

Brown, P.J., Bovey, J.D., & Chen, X. (1997). Context-aware applications: From the laboratory to the market place. *IEEE Personal Communications, 4*(5), 58-64.

Bruijin, O.D., Spence, R., & Chong, M.Y. (2001). RSVP browser: Web browsing on small screen devices. *ACM Mobile HCI Workshop 2001,* Lille, France.

Buisson, M. & Jestin, Y. (2001). Design issues in distributed interaction supporting tools: Mobile devices in an ATC working position. *ACM Mobile HCI Workshop 2001,* Lille, France.

Chang, C.Y. & Sheu, J.P. (2002). Design and implementation of Ad Hoc classroom and eSchoolbag systems for ubiquitous learning. *IEEE WMTE 2002*.

Chesta, C. (2002). Globalization of voice-enabled Internet access solutions. *ACM Mobile HCI Conference 2002,* Pisa, Italy.

Cheverst, K., Mitchell, K., & Davies, N. (2001). Investigating context-aware information push vs. information pull to tourists. *ACM Mobile HCI Workshop 2001,* Lille, France.

Chittaro, L. & Cin, P.D. (2001). Evaluating interface design choices on WAP phones: Single choice list selection and navigation among cards. *ACM Mobile HCI Workshop 2001,* Lille, France.

Czerwinski, S., Zhao, B., Hodes, T., Joseph, A., & Katz, R. (1999). An architecture for a secure service discovery service. *Proceedings of MobiCom '99,* Seattle, Washington, August.

Davies, N, Mitchell, K., Cheverst, K., & Blair, G. (1998). Developing a context sensitive tourist guide. *ACM Mobile HCI Workshop 1998*, Glasgow, UK.

E-Ink (2004). Retrieved on March 14, 2004, from *http://www.eink.com*

Fontana, F., Fuiello, A., Gobbi, M., Murino, V, Rocchesso, D, Sartor, L., & Panuccio, A. (2002). A cross-modal electronic travel aid device. *ACM Mobile HCI Conference 2002,* Pisa, Italy.

Gershon, N., Card, S., & Eich, S.G. (1997). Information visualization. *Chi '97 Tutorial Notes*, Atlanta, Georgia, March 22-27.

Goldstein, M., Oqvist, G., Bayat-M, M., Ljungstrand, P., & Bjork, S.(2001). Enhancing the reading experience: Using adaptive and sonified rsvp for reading on small displays. *ACM Mobile HCI Workshop 2001*, Lille, France.

HAVi Consortium. Retrieved on March 14, 2004, from *http://www.havi.org*

Holland, S. & Morse, D.R. (2001). Audio GPS: Spatial audio in a minimal attention interface. *ACM Mobile HCI Workshop 2001*, Lille, France.

Huttenrauch, H. & Norman, M. (2001). Pocket-CERO - Mobile interfaces for service robots. *ACM Mobile HCI Workshop 2001*, Lille, France.

Jini - The community resource for Jini technology (2001). Retrieved on March 14, 2004, from *http://www.jini.org/*

Jones, M., Buchanan, G., & Thimbleby, H. (2002). Sorting out searching on small screen devices. *ACM Mobile HCI Conference 2002*, Pisa, Italy.

Khronos Working Group. (2004). Retrieved on March 14, 2004, from *http://www.khronos.org*

Luchini, K., Quintana, C., & Soloway, E. (2002). ArtemisExpress: A case study in designing handheld interfaces for an online digital library. *ACM Mobile HCI Conference 2002*, Pisa, Italy.

Minenko, V. (1998). PalmVNC. Retrieved on March 14, 2004, from *http://www.wind-networks.de/PalmVNC/*

Nigay, L., Salambier, P., Marchand, T., Renevier, P., & Pasqualetti, L. (2002). Mobile and collaborative augmented reality: A scenario based design approach. *ACM Mobile HCI.*

NVidia Corp. (2004). Retrieved on March 14, 2004, from *http://www.nvidia.com*

Petrie, H., Furner, S., & Strothotte, T. (1998). Design lifecycles and wearable computers for users with disabilities, *ACM Mobile HCI Workshop 1998*, Glasgow, UK.

RFID Organisation (2003). Retrieved on March 14, 2004, from *http://www.rfid.org*

Rist, T. (1999). Using mobile communication devices to access virtual meeting places. *ACM Mobile HCI Workshop 1999*, Edinburgh, UK.

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *Proceedings of the Workshop on Mobile Computing Systems and Applications*, Santa Cruz, California.

Sokoler, T., Nelson, L., & Pedersen, E.R. (2002). Low-resolution supplementary tactile cues for navigational assistance. *ACM Mobile HCI Conference 2002*, Pisa, Italy.

TabletPC (2003). Microsoft Corp. Retrieved on March 14, 2004, from *http://www.microsoft.com/tabletpc*

Tegic Corp T9 text prediction system. (2004). Retrieved on March 14, 2004, from *http://www.tegic.com*

WAP Forum press release (2004). Retrieved on March 14, 2004, from *http://www.wapforum.org/*

Weiser, M. (1991). The computer for the 21st century. *Scientific American,* September, 09-91.

Weiss, S. (2002). *Handheld usability.* UK: Wiley Publishing.

X10 electrical-to-computer-software hardware protocol and controller equipment (2004). Retrieved on March 14, 2004, from *www.x10.com*

KEY TERMS

CSCW: Computer-supported collaborative work, a discipline of computer science dedicated to the use of computer tools to allow groups of participants to work together in the resolution of a problem domain.

Information Visualisation: A process of transforming information into a visual form enabling the user to observe information (Gershon, Card, & Eich, 1997).

M-Learning: The use of mobile devices as tools in the computer science discipline of electronic learning.

Mobile HCI: The Human Computer Interaction (HCI) aspects of the design, evaluation and application of techniques and approaches for all mobile computing devices and services. (ACM Mobile HCI, 2001).

PDA: Personal Digital Assistant; a handheld computing device that may contain network facilities but generally is used for personalised software purposes beyond a standard organiser.

RFID: Radio Frequency Identification; a technology that uses radio frequency waves to communicate data between a moveable item with a small digital tag and a reader to identify, track, or locate that item.

Ubiquitous Computing: The evolution of mobile HCI whereby user centred principles of hardware and software development embed the nature of mobile computing into the background of everyday life.

WAP: Wireless Access Protocol; a protocol for implementing advanced telecommunications services for accessing Internet pages from mobile devices.

WML: Wireless Markup Language; based on the XML language it has been derived to create a user interface and content specification for WAP-enabled devices. (WAP Forum 2003).

This work was previously published in Encyclopedia of Multimedia Technology and Networking, edited by M. Pagani, pp. 1021-1026, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.24
User Experience of Camera Phones in Social Contexts

Hanna Stelmaszewska
Middlesex University, UK

Bob Fields
Middlesex University, UK

Ann Blandford
University College London, UK

ABSTRACT

This chapter reports on a qualitative study into people's use of camera phones for social interaction in co-present settings. The study examined people's behaviour and positive experiences (e.g., fun, enjoyment, or excitement) when camera phones were used in different spaces (public and private). It was found that camera phones influence social practices. Three distinct practices were observed: sharing a moment now, sharing a moment later, and using photos to initiate social interaction with strangers. The knowledge obtained through the study will offer a conceptual contribution that deepens our understanding of how this emerging and evolving technology is coming to be accommodated into the leisure-related practices of its users.

INTRODUCTION

What do we know about photography? Photography has been a part of our life for a long time. We document family celebrations, important events in our lives and those of our family and friends; we take pictures when visiting museums or if we want to illustrate everyday items and people in a funny way and when we want to create stories (Mäkelä et al., 2000). It seems that photography and photos bring either smiles when reminiscing about something pleasant or tears when emotions take over. They preserve memories, capture feelings, and provide a means to communicate with others. One of the most common and enjoyable experiences is to share photos with others through story telling (Balanovic et al., 2000; Chalfen, 1987). Photos can be shared using technology

and then they can be used as means for interaction with others.

Recent technological developments not only support new ways of working but also provide new mechanisms for social interaction. Mobile phones and camera phones, in particular, are examples of such technology. In the past decade, mobile phones have allowed profound changes to take place in people's behavior and practices in relation to communication (Ling, 2004), from being extensively used as a medium of verbal and text communication to one that uses pictures to facilitate people's social life. Mobile phones with integrated camera and video features have changed forever the way people communicate and interact, and have shaped both their individual and their social lives (Ito, 2005; Kato, 2005; Kindberg et al., 2005a, 2005b;Okabe, 2004; Scifo, 2004).

Although there is a vast body of literature focussing on the use of camera phones (Kindberg, et al., 2005a, 2005b; Okabe, 2004; Scifo, 2004) the issues relating to how camera phones are used to mediate social interaction between co-located users have been neglected. In this chapter, we report on the study of the collaborative use of camera phones by co-located users in various spaces.

BACKGROUND TO THE RESEARCH

In recent years, there has been substantial interest in digital photography, with a particular interest on how the digital medium facilitates sharing of images (Balanovic et al., 2000; Frohlich et al., 2002, Van House et al., 2005). Studies of sharing digital photographs include the use of Web-based systems, mobile applications, and multimedia messaging. Most of the studies focus on personal applications for sharing images remotely (Kato, 2005; Kindberg et al., 2005a; Van House et al., 2005) work on sharing images in co-present settings is in its infancy.

The issues of what people capture on mobile phones and what they do with these images were extensively investigated by Kindberg et al (2005a). They proposed a six-part taxonomy to describe the intentions behind the use of camera phone images. Intentions were grouped along two dimensions. The first intention defines whether people captured the images for affective (e.g., sentimental) or functional reasons. The second one defines social or individual intensions.

Others, such as Licoppe & Heurtin (2001) and Taylor and Harper (2003), focused on teenagers using their phones for social practices. The latter claim teenagers' practices are similar to 'gift-giving' rituals, which shape the way teenagers understand and use their mobile phones. The 'gift-giving' practices included sharing certain text messages, call-credits and even the mobile phones themselves. All these practices establish and cement allegiances and sustain rivalries (Taylor & Harper, 2003).

A field study conducted by Kato (2005) explored how the use of mobile phones/camera phones changes people's daily activities in Japan. He argues that the new ways of pervasive photo taking through camera phones allows people to document their lives on a daily basis, which can be preserved and shared as a life of a local community.

A different approach to studying mobile phone users was taken by Okabe (2004). He studied practices of Japanese camera phone users, which included personal archiving, intimate sharing, and peer-to-peer news sharing. Okabe (ibid) argues that capturing and sharing visual information cannot be understood without also understanding the social relationships and contexts within which those activities take place. Scifo (2004) provides similar views on this matter, arguing that taking photographs on camera phones and using MMS communication allows users (particularly youngsters) to identify themselves within social groups, and will intensify communication within that community.

The relevance of social relations to the uses of photographs was also identified by Van House

et al. (2005). They discovered five distinct social uses of personal photos. These are: creating and maintaining social relationships, constructing personal and group memory, self-expression, and self-presentation and functional communication with self and others.

Photos could also be used for social discourse. For example, a mobile picture system (MobShare) developed by Sarvas et al. (2005) supports that by transferring photos from the phone to different devices. These include transfers (1) to another phone over the network (e.g., MMS), (2) to a PC, (3) to a network server over the network, and (4) to a printer using a cable connection or Bluetooth.

Many methods have been used to study people's uses of mobile phones, including diaries, interviews and field studies (Kato, 2005; Kindberg et al., 2005a, 2005b; Okabe, 2004; Sarvas et al., 2005). The approach employed by Sarvas et al. (2005) involved asking people to fill out a diary including all activities their performed using their camera phones. This was followed by a set of interviews focusing on photographic habits and social networking involving photography. The same methods were employed by Okabe (2004) when investigating social practices, situations and relations of the use of camera phone.

Kato (2005) applied a fieldwork study to observed and record the practices of camera phone users encouraging them not only to take pictures but also to collect and store them as visual field notes on a specially designated web site. When conducting an in-depth study of camera phone use Kindberg et al. (2005a, 2005b) applied a set of interviews asking the subject to show images that were not private from their camera phones and talk about them.

Taking inspiration from such research, semi-structured in-depth interviews and field observational studies were employed in the study reported here, which will be discussed later in this chapter.

METHODOLOGY

This study is specifically concerned with peoples' experiences when using camera phones for social interaction in a co-present setting (i.e., when participants are present at the same location at the same time). The chapter builds on an earlier more general study into peoples' experience and emotions using personal technologies such as PDAs, digital cameras and mobile phones (Stelmaszewska et al., 2005).

Because we wanted to obtain the insights of the ways people use, their camera phones as a medium for social practices we adopted Kindberg et al.'s (2005a) method of asking participants about circumstances and reasons for taking these images and their life cycle. A series of observational field studies was conducted to develop a better understanding of peoples' practices using camera phones. The use of dual methods strengthened the results obtained and provided a means of triangulation between the interviews and observations to confirm that the reported practices really did occur when the observations took place. In addition, field observations of the phenomena provided richer insights into the circumstances and contexts in which practices described in interviews actually take place.

Five students were interviewed including two PhD students, two undergraduates, and one college student, all aged between 18 and 27; all participants had been camera phone users for at least a year. Each interview took between 25 and 45 minutes and was recorded and later transcribed. The participants were asked to describe how and for what reasons they used their camera phones. The participants were also asked to show a few of the images (pictures or video) stored on their phones and encouraged to discuss where the images were taken, in what circumstances, by whom and for what reason. Also of interest was whether pictures were taken by the participant

or received from another person, the means of storage and transfer employed (e.g., infrared, Bluetooth, MMS, e-mail), how long these pictures were stored, and whether they were shared with others, or retained for a private use.

The data from the field studies was gathered in a variety of public spaces, including pubs, restaurants, leisure and entertainment places, museums, and public transport (tube and buses). The first author spent around 35 hours in public spaces observing camera phone usage. In this time, 18 individual instances of individuals and groups interacting with photos on cameras were observed and noted.

As the data gathered from interviews and field observations was of a qualitative nature, data collection and analysis was carried out iteratively. This allows for 'theoretical sampling' on the basis of concepts and themes that emerge from the analysis and allows concepts to be explored and hypotheses to be tested as they are developed from the data (Strauss & Corbin, 1998, p.46). Data from both studies was transcribed and then analyzed by first, coding it using qualitative methods to identify emerging themes, and then the themes were merged to extract the high level concepts that gave the outline of the use and practices of camera phones.

SITUATED USE OF CAMERA PHONES

The field observation study revealed many instances of people being engaged in social interaction using camera phones in different co-present settings. The in-depth interviews provided extended information to support these phenomena. The data shows the relationships between space and place as well as the photo/video sharing practices, which will be discussed in the following sections.

The concepts of place and space have been researched by many like Casey (1997), Ciolfi (2004), Dourish (2001), and Salovaara et al. (2005) just to name a few. Casey (1997) discusses this phenomena as 'space refers to abstract geometrical extension and location' whereas 'place describes our experience of being in the world and investigating a physical location or setting with meaning, memories and feelings' (cited in Ciolfi, 2004, p.1). A similar view has been taken by Dourish (2001) who gives an example of a space like a shopping street being a different kind of a place depending on the time of a day. According to Salovaara et al. (2005) the 'concepts of space and place are mutually dependent and co-occur in the context' (p. 1).

Camera Phone Use in Different Spaces

Camera phones have become a part of our lives. People carry them to work, to social events, to leisure activities, even when going shopping. Every time we use camera phones, we experience something. The experience, however, does not exist in a vacuum, but rather in a dynamic relationship with other people, places and objects (Mulder & Steen, 2005). What we experience and how camera phones are used is also determined by place and space, which will be explored in the consecutive sections.

Public Space

It appeared in the data that people use their camera phones differently depending on where they are. It was observed that when using public spaces like a tube or a bus people tend to use their camera phones for individual purposes; that includes reading and answering text messages, playing games, viewing and sorting out images, playing music or ring tones, or examining different functions on their camera phones. Interview data indicated that people do these things to overcome the feeling of boredom or simply to 'kill time' while waiting for a bus, as one of the participants (Steve) commented:

I listen to the radio … when I'm on the tube, when walking around or waiting for a bus and I don't have anything to amuse me. To amuse me, I use the calendar and the diary quite a bit. Otherwise I'd forget everyone's birthday.

Similarly, another participant (Luisa), on using camera phone on a bus, commented:

…the setting itself is boring not much inspiration to take pictures and things … you have to be with someone to do it.

It was reported in the literature that some public spaces are regulated by different means: signage, announcements and by more informal peer-base regulations (Ito, 2003, 2004; Okabe & Ito, 2005). The former claims that these regulations are mostly exercised in public transport. Posters and signage exhort passengers from putting their feet on the seats or not smoking. The study by Okabe & Ito (2005) reported that people use email rather then voice calls when on trains and subways following 'sharing the same public space' regulations. Although, this kind of behavior was observed amongst Japanese youth population similar findings were reported by Klamer et al. (2000) who conducted a European survey investigating if the mobile phones used in public spaces disturb people.

A different kind of behavior was observed in museums (Science Museum and Natural History Museum in London). Camera phones were rarely used and only for individual purposes: receiving calls or messages, making phone calls, or texting. People treat museums as places to go on outings with friends and family, which they plan for and therefore they take a digital camera with them to capture something specific that they would like to keep as a reminder. In this case, the quality of pictures is of high importance. The comments of Maria confirm this:

…I like to take pictures of a nice scenery or … er… flowers or trees or just a really nice views or things… then I use my digital camera because of the quality of the picture.

Figure 1. A girl sitting with her family and taking picture of the artist playing

Figure 2. People taking photos of the pantomime artist

Other public spaces like pubs, restaurants, clubs, places of entertainment and leisure provide a different social context for camera phone activities, which is in line with our previous research reported elsewhere (Stelmaszewska et al., 2005, 2006). The data illustrates that people more often engage themselves in social interaction using camera phones during gatherings with friends and family, when going out with friends or during trips or excursions with friends (see Figures 1 and 2). Most of the participants claimed that the important issue for using camera phones is to be with other people. It is people who create experiences that people enjoy, as Adam noted:

When you have other people around you then you have a different kind of experience.…you are more likely to do silly things. So then you take pictures and when you view them you can laugh and have fun. When you are on your own … no, you don't do these things. You need to have people around you to have fun.

Private Space

A similar behavior was reported when groups of participants use their camera phones in private spaces (e.g., homes or cars); that is people took pictures or videos of friends, members of family or even themselves behaving funny or silly and then shared them with others co-present or they viewed pictures and videos taken previously. The comment from Adam supports this view:

...so what we did was just running through clips and passing them from one group of people to another … [laughing] this was funny… I like to take pictures of funny situations and when my friends are drunk they do funny things so we go back and try to remember what happen and we always have a good laugh. Sometimes we like to compare who managed to take the most funny shots … it is really funny seeing people doing crazy things.

Since the camera screens are small and do not support easy and clear viewing for a group of people when sharing pictures in the home environment, people often made use of external display technology, such as TV or computer. This issue will be explored further in the next section, *'Sharing a moment later.'*

As discussed in this section peoples' use of camera phones changes in relation to the space they are in; private vs. public. It was found that people's practices when using camera phones differs in different spaces. The next section will discuss this phenomenon in more detail.

Social Uses of Camera Phones

Camera phones have been used for individual as well as group purposes. Consistent with other studies (Kindberg et al., 2005a, 2005b) we found that people take photos for individual purposes that include creating memories and evocations of special events, trips, holidays, or beautiful landscapes. A common practice is to share images with friends and family, in a way that is deeply embedded in social interaction (Stelmaszewska et al., 2005, 2006). Sharing digital photos is often done remotely via email or by posting them on the web (Counts & Fellheimer, 2004; Stelmaszewska et al., 2005). Despite the growing popularity of using web-based applications and services (e.g., Flickr, YouTube, or Mobido) that allow their users to share photos there were no accounts reported using these services by the participants involved in this study.

However, we observed other practices that occur in co-present social contexts. These include 'sharing a moment now,' 'sharing a moment later,' or using photos to initiate social interaction with strangers.

'Sharing a Moment Now'

This study shows a different way people share photos taken on a camera phone that appears to

be less about evoking or recreating an event or scene after the fact, and more about augmenting that event as it happens. It was observed that people take a 'spur of the moment' photo or video and share it with people who are present at the same location at the same time. People reported having fun when taking photos or videos of their friends behaving funnily and then viewing them collectively at the location. This kind of behavior seems to motivate and shape social interaction, as Adam reported:

…she was happy and funny (referring to a friend) … far too engaged with dancing to notice what was happening around her … and I just thought that I'll just take that picture. … there were few of us friends so then I showed them and then other friends were taking more pictures of her dancing and we were waiting for her to realize what was going on … we were all taking pictures of her … we shared all the pictures and picked out the funniest ones. It was so funny because she couldn't believe that we did that and she didn't even notice it.

Whereas Lucy said:

When I'm out with my friends then I'll definitely use it (referring to a camera phone). … Sometimes I take pictures of my friends and then we'll sit down and go through them selecting the best once.

Data shows that photos were used for functional purposes as well, which is consistent with the findings of other research (e.g., Kindberg et al., 2005a; Van House et al., 2005). It was observed that when on a trip, people took a picture of a map displayed by a leader and then pursued his instructions using a display on their camera phones. This kind of activity allowed every person within the group to see clearly the map and use it for further reference.

Another common practice observed and reported by participants was to transfer photos between phones using the Bluetooth technology so that everybody concerned could store and use them when needed. The following observed episode is a typical example:

Episode 1: Pub, evening

Ten people are sitting at the table (three females and 7 males). Jim takes the camera phone out of his pocket and plays with it.

Jim: 'I have something really cool to show you.' He does something with his phone. After a while Jim said: 'OK, I've got it.' He plays the video and passes his phone over to a neighbor, Roy.

Jim: 'Just press the button.' Roy plays the video and moves the phone towards another male, Paul. Another male, Martin moves from his seat and stands behind Roy and Paul watching the video clip.

Martin: 'I want this clip. Can you Bluetooth it?'

Jim: 'Yeah' Jim takes his phone back from Roy and sets up the Bluetooth. Martin does the same on his phone. After a short while Jim transfers the clip over to Martin's phone.

However, it appeared that some people found it difficult to use it and either abandon the transfer or asked for help. When discussing issues related to managing pictures on the phone Maria said:

I Bluetooth them … I can do it now but I had to ask my friend to show me how to do it so I'm OK now.

'Sharing a Moment Later'

When people who you want to share photos with are around, it creates opportunities for social interaction to take place so that people can

enjoy the moment of sharing pictures together. What happens when they are not around? Other studies reported this kind of practice; that is to view the photos when the occasion arises, and not immediately after they have been taken. For example, Okabe (2004) described situations where people show their friends the photos from their archives (photo gallery) on occasions that they get together.

A co-present social interaction was reported to be associated with participants' experience when viewing pictures or videos stored on individual's phones but taken previously (not at the time of gathering). The intentions behind it were reported to include sharing memories of special events, reporting on events to those who were absent at the time of events, or creating and sharing a documentary of a friendship or family life as Maria remarked:

with the cam_phone I can capture the moment … and being able to view them later will bring all the memories and the fact that those pictures can be shared … so people can have fun.

People were more inclined to use photos for storytelling, which is in line with (Balanovic et al., 2000; Kindberg et al., 2005a) and, as suggested by Fox (2001) and Vincent & Harper (2003), mobile phones have been used to maintain personal relationships between friends and family. Since camera phones are becoming a part of our everyday lives, it is not surprising that the same behavior was observed in the context of camera phone use when photos or videos were shared during social gatherings.

However, given that phone screens were claimed to be very small it was common amongst participants to use other media like computer or TV to display photos in order to improve their visibility and enhance the experience of people participating. Adam reported:

I transferred them onto my computer … I'm quite organized with my pictures so I categorize them and put them in kind of albums and sometimes when I'm with friends we like to go through pictures and have fun.

Maria commented:

…sometimes what we do is we Bluetooth to transfer our pictures to one of our computers and then have a slide show so everybody can see it …you see the phone screens are very small and if we all want to have fun we need to see those pictures simultaneously. With camera phones we can't see it clearly if there are more then two or three people looking. It's just not enough space …

Sharing photos at co-present settings proved to be a way of social interaction that brings fun and joy to people's lives. The remarks of an interviewee, Steven, appear to confirm this point:

I'll show them (referring to family) what I managed to capture and then we have a good laugh.

Supporting the view, Lucy commented:

…you take pictures and when you view them you can laugh and have fun.

Ito & Okabe (2003, p.6) claim that: "Mobile phones … define new technosocial situations and new boundaries of identity and place … create new kinds of bounded places." We argue that camera phones go beyond that. When people view pictures together and tell the story behind them, they are transported to the place and space where those pictures were taken. Pictures conjure memories, feelings, and emotions and evoke sensations associated with the events that were photographed. Lee, another study participant, remarked when showing pictures from a group trip:

... The first dive was really s…. it was sooo cold, remember, … and we didn't see much... The vis was absolutely s…. yeah and then we had to get warmer ha, ha, ha …

Comments from other participants suggest the same:

Adam: … *when you are having a good time you don't always know what's happening around you. … I don't always know what everybody is doing so I miss a lot of stuff but when we view all the pictures taken during a particular party or we go for a short trip together … so only then you really can see what happened. We really like doing that.*

Maria: … *you can not only see the pictures but there are always some stories behind every picture. … so later when you show the pictures everybody gets involve and just add a story to it and that's great. I like it. And others who were not there can feel like they were there err...kind of.*

Social Interaction with Strangers

Studies reported by Weilenmann and Larson (2002) explored the collaborative nature of mobile phones use in local social interaction amongst teenagers. They suggest that mobile phones are often shared in different forms including: minimal form of sharing (SMS messages), taking turns (several people handling a phone), borrowing and lending of phones, and sharing with unknown others. The latter involves the phones being handled by teenagers who are unacquainted until one of them makes the initial contact. Weilenamm and Larson (2002) describe practices of teenagers (boys giving girls their mobile phone) to enter their phone numbers. This kind of social interaction is similar to the one that emerged from our studies.

Social interaction can coalesce around different media, from text and graphics, to interactive games (Stelmaszewska et al., 2005, 2006). Such interactions often occur between friends or family members sharing the same technology (i.e., computer, digital camera or mobile/camera phone). However, a striking finding was that camera phones were used as a new channel and medium for initiating social interaction with strangers. It was reported that people take photos of others (whom they like) in order to show their interest, introduce themselves, or simply start a new social relationship.

The comment from Luisa supports this claim:

I was at the Harvester, a restaurant/pub thing, …and there was a small window with glass between it looking like a fake door and the guys were looking through that doing (mimicking facial expressions) and then I saw one holding his camera phone against one of the window things and there was a picture of me going (shows facial expression) and I didn't know that they were taking it … I didn't really mind. It's a good humor… it was kind of friendly, sort of vague flirting without talking … just taking pictures.

So does another comment by Maria:

We were in the bar … having fun and there was this guy dancing [laughing] kind of a very funny dance … almost like an American Indian kind of dance … and one of the girls from our group took a photo of him because she liked him and she was showing it to us so instead of looking at him we could see his picture … and when he saw her taking pictures of him he did the same to her… the whole situation was funny … at least we had fun watching them two taking pictures of each other instead of talking …

This kind of behavior typically occurred in public spaces such as pubs, bars, or clubs where people usually gather for social events, and interaction with others is a part of the entertainment. In our

study, the focus was on social interaction that took place through and around digital photos. Such interaction is not always appreciated by those involved. Some participants felt offended and annoyed with those taking photos without obtaining agreement. For example, Lucy noted:

I don't know if I would be offended so much. I think it depends what for … sometimes you get photographers going like around pubs and clubs … and I never said yes to the photo. The other night when I was there with my friend and this group of guys we met before errr … this guy said: 'Oh yeah, let's get a picture' but we went like: 'no, we really don't want to.' And they had one done anyway and this kind of annoyed me a bit because … it's fair they wanted the picture of us but we didn't really want to be in it. … I think it depends how much choice you are given as whether or not you want your photo taken.

It appeared that pictures are not the only phone-related way people try to 'chat up' others. Phone features like Bluetooth can be used to connect to strangers and initiate communication. This kind of behavior was observed in public places (pubs, restaurants, bars). The practice was to switch on the Bluetooth and ask others (whoever is picked up by the Bluetooth) to activate the connection. However, this kind of interaction often raised some suspicions, as people did not know who wants to 'chat up' to them. Here is an extract from one of the participants expressing his concerns:

… someone wants me to activate the connection … but what do I do … I don't want any 'Boss' [the name of the Bluetooth connection] connecting to my phone. What if they do something to my phone?

The fact that people do not see the 'talker' and they do not have the full control of who they interact with seems to be a barrier to engaging in interaction with a stranger.

It seems that communication takes place not only through technology but also alongside it, a finding that is consistent with our earlier studies (Stelmaszewska et al., 2005). Moreover, Van House et al. (2005) argue that technology (e.g., online photo blogs) is used to create new social relationships. Although this study is at an early stage and further evidence is required, we suggest that camera phones provide new channels and foci for social interaction within co-present settings.

Barriers to Sharing

Although camera phones appear to be a new medium for social interaction that is enjoyable and fun, they are not without problems that limit the extent to which they are used. The data illustrates that people experience different kinds of trouble that hinder their experience or make it impossible for sharing to happen.

Firstly, the lack of compatibility between different camera phones stops people from sending photos. Several participants reported not using MMS features because it was difficult to use. In addition, people often know (not always) that those who they want to send pictures to will not be able to retrieve them as was commented by Luisa:

… none of mine friends really do this … you have to have the same phone or something to be able to send it and for them not to just say: 'message not being able to deliver or whatever.' Some people tried to send pictures on my phone but I never got them.

Secondly, for many camera phone users it is difficult to send pictures either via MMS or Bluetooth. People reported having difficulties to find the functions to do so or they could not set them up (in case of the Bluetooth—see comments in the section on 'Sharing a moment now').

Another barrier to sharing photos was the lack of a quick and easy way to find archived pictures. People spent time, sometimes a long

time, trying to find the pictures they wanted to share with their friends. This caused frustration and dissatisfaction as Jim said:

Where is it?!!! S… Hrrrrrrrrrr

Quick access to camera functionality and photo image features is an important issue in a context of sharing and it raised concerns amongst participants as Maria noted:

… one of my friends helped me to set it up so I can use it by pressing just a couple of buttons instead of going through menus and stuff. It was horrible. I missed so many great pictures because of that and I was very upset about it. … it's very important. I could have so many great pictures but couldn't find the camera function on my phone … it was very frustrating.

All these barriers affect not only experience of camera phone users but also their engagement in social interaction. So providing functionality that is transparent and supports users sharing activities is of a paramount importance when designing systems. It might also enhance the use of camera phones by creating pleasurable and fun experiences instead of satisfying only functional purposes.

DISCUSSION AND CONCLUSION

It seems that phone technology is moving from facilitating its original primary goal, supporting distance communication, to supporting new ways of social interaction that happens through sharing activities (photos and videos) as well as providing bridges between contexts. When people share photos or videos, they are transported from the context of a present space (pub, restaurant, or home) to the one that a specific photo or video clip conjures up.

In addition to providing resources for communication and interaction, camera phones have been used as a kind of archive of a personal life, a viewpoint on the world, or a collection of fragments and stories of everyday life. Okabe (2004) suggests that photos are often taken for purely personal consumption, whereas text messages are generally created with the intent to share with others. However, the findings from this study contradict Okabe's claim; people often take photos with the intention to share them with others, which is a more selective and intimate activity than sharing text.

When technologies are used in different places and spaces they become part of a specific environment and this often shapes the use of technology and experiences connected to it. As a consequence of this, technologies are often used in unexpected ways (Taylor & Harper, 2003). In the case of this study, these ways are 'sharing the moment now,' 'sharing the moment later' and using camera phones for 'social interaction with strangers.'

This chapter has described distinctive practices of camera phone users occurring in co-present settings, and how these practices change in relation to the place and space in which they were used. It has been argued that camera phones provide a new medium through which people can sustain and enrich they social interaction through taking and sharing photo images or videos. However, these activities are inseparable from social relations and context, which is in line with Okabe's (2004) and Scifo's (2004) findings. Moreover, we argue that this study provides a better understanding of how this emerging and evolving technology facilitates social interaction in the leisure-related practices of its users.

We agree with Rettie's (2005) view that mobile phone communication affects the role of space and we have shown that camera phones go beyond this: they bring people together, creating experiences through social interaction. No other technology has supported this to such an extent, and to so many

people. The multi-functionality of camera phones provides a different means of social interaction, which is unique to a place and space.

More generally, when designing camera phones that facilitate social interaction, understanding of emerging uses, practices and social activities is essential for the effective design of camera phones and related systems. Moreover, identifying problems within existing systems might be a good starting point for discussing user requirements, helping designers to develop systems that fulfill utilitarian as well as user experience needs.

Although the notions of 'sharing' might be a new phenomenon it is a manifestation and reflection of needs that relate to social identity (Scifo, 2004; Taylor & Harper, 2002) and are shaped by social context (Okabe, 2004; Stelmaszewska et al., 2005, 2006). This study is part of an ongoing effort to explore issues related to the use of camera phones for social interaction within co-present settings, and further studies will be required to investigate what affects such interaction, how camera phones' design, usability and context of use influence the nature of users' experience.

Furthermore, more work is needed to identify and understand problems when camera phones are used for social interaction, and how we can improve the design of camera phones so that they can evoke experiences such as pleasure, excitement, or fun.

ACKNOWLEDGMENT

We would like to thank all anonymous participants who took part in this study.

REFERENCES

Balanovic, M., Chu, L., & Wolff, G. J. (2000). Storytelling with digital photographs. In *Proceedings of CHI 2000: Conference on Human Factors in Computing Systems* (pp. 564-571). ACM Press.

Casey, E. S. (1997). *The fate of place: A philosophical history.* Berkeley: University of California Press.

Chalfen, R. (1987). *Snapshot version of life.* Bowling Green, OH: Bowling Green State University Press.

Ciolfi, L. (2004). *Digitally making places: An observational study of people's experiences of an interactive museum exhibition.* Paper presented at the Proceedings of the 2nd workshop on 'Space, Satiality and Technologies.' Edinburgh, UK.

Counts, S., & Fellheimer, E. (2004, April 24-29). Supporting social presence through lightweight photo sharing on and off the desktop. In *Proceedings of CHI 2004: Conference on Human Factors in Computing Systems* (pp. 599-606). Vienna: ACM Press.

Dourish, P. (2001). *Where the action is: The foundation of embodied interaction.* Cambridge, MA: MIT-Press.

Fox, K. (2001). Evolution, alienation and gossip: The role of mobile telecommunications in 21st century. *Social Issues Research Centre Report.* Retrieved from http://www.sirc.org

Frohlich, D., Kuchinsky, A., Pering, C., Don, A., & Ariss, S. (2002). Requirements for photoware. *Proceedings of the 2002 ACM conference on Computer Supported Cooperative Work* (pp. 166-175). ACM Press.

Ito, M. (2003). *A new set of social rules for a newly wireless society. Japan Media Review.* Retrieved July 28, 2006 from http://www.ojr.org/japan/wireless/1043770650.php

Ito, M. (2004). *Personal Portable Pedestrian: Lesson from Japanese Mobile Phone Use.* Paper

presented at The 24 International Conference on Mobile Communication Social Change. Seoul, Korea.

Ito, M., & Okabe, D. (2003, June 22-24). Mobile phones, Japanese youth, and the re-placement of social contact. In R. Ling (Ed.), *Front stage—back stage: Mobile communication and the renegotiation of the public sphere*. Grimstad, Norway.

Ito, M., & Okabe, D. (2005). Technosocial situations: Emergent structurings of mobile email use. In M. Ito, Okabe, D. & Matsuda, M. (Ed.), *Personal, portable, pedestrian: Mobile phones in japanese life*.

Kato, F. (2005, April 28-30). *Seeing the "seeing" of others: Conducting a field study with mobile phones/mobile cameras.* Paper presented at the T-Mobile conference, Budapest, Hungary.

Kindberg, T., Spasojevic, M., Fleck, R., & Sellen, A. (2005a). An in-depth study of camera phone use. *Pervasive Computing, 4*(2),42-50.

Kindberg, T., Spasojevic, M., Fleck, R., & Sellen, A. (2005b). I saw this and thought of you: Some social uses of camera phones. *In CHI '05 extended abstracts on Human factors in computing systems* (pp. 1545-1548). ACM Press.

Klamer, L., Haddon, L., & Ling, R. (2000). The qualitative analysis of ICTs and mobility, time stress and social networking. (No. P-903): EURESCOM.

Licoppe, C., & Heurtin, J.P. (2001). Managing one's availability to telephone communication through mobile phones: A French case study of the development dynamics of mobile phone use. *Personal and Ubiquitous Computing, 5*(2), 99-108.

Ling, R. (2004). *The mobile connection: The cell phone's impact on society.* Morgan Kaufmann.

Mäkelä, A., Giller, V., Tscheligi, M., & Sefelin, R. (2000). Joking, storytelling, artsharing, expressing affection: A field trial of how children and their social network communicate with digital images in leisure time. In *Proceedings of CHI 2000: Conference on Human Factors in Computing Systems* (pp. 548-555). ACM Press.

Mulder, I., & Steen, M. (2005, May 11). *Mixed emotions, mixed methods: Conceptualising experience of we-centric context-aware adaptive mobile services.* Paper presented at the Pervasive 2005. Presented at workshop on User Experience Design for Pervasive Computing, Munich, Germany.

Okabe, D. (2004, October 18-19). *Emergent Social practices, situations and relations through everyday camera phone use.* Paper presented at the International Conference on Mobile Communication, Seoul, Korea.

Okabe, D., & Ito, M. (2005). Ketai and public transportation. In M. Ito, Okabe, D., & Matsuda, M. (Eds.), *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life*. Cambridge: MIT.

Rettie, R. M. (2005). Presence and embodiment in mobile phone communication. *PsychNology Journal, 3*(1), 16-34.

Salovaara, A., Kurvinen, E., & Jacucci, G. (2005, September 12-16). *On space and place in mobile settings.* Paper presented at the Interact '05, Rome.

Sarvas, R., Oulasvirta, A., & Jacucci, G. (2005). Building social discourse around mobile photos—A systematic perspective. *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services* (pp. 31-38). ACM Press.

Scifo, B. (2004, June 10-12). *The domestication of camera-phone and MMS communication: The early experience of young Italians.* Paper presented at the T-Mobile Conference, Hungary.

Stelmaszewska, H., Fields. B., & Blandford, A. (2005, September 5-9). *Emotion and technology:*

An empirical study. Paper presented at the Emotions in HCI design workshop at HCI '05.

Stelmaszewska, H., Fields, B., & Blandford, A. (2006, September 11-15). Camera phone use in social context. In *proceedings of HCI 2006* (Vol. 2, pp. 88-92), Queen Mary, University of London, UK.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research. Techniques and procedures for developing grounded theory.* Newbury Park: Sage.

Taylor, A., & Harper, R. (2002, April 20-25). *Age-old practices in the 'new world': A study of gift-giving between teenage mobile phone users.* Paper presented at the CHI, Minneapolis, MN, USA.

Taylor, A., & Harper, R. (2003). The gift of the grab: A design oriented sociology of young people's use of mobiles. *Journal of Computer-Supported Cooperative Work, 12*(3), 267-296.

Van House, N., Davis, M., Ames, M., Finn, M., & Viswanathan, V. (2005, 2-7 April). The uses of personal networked digital imaging: An empirical study of cameraphone photos and sharing. *CHI '05 extended abstracts on human factors in computing systems* (pp.1853-1856). ACM Press.

Vincent, J., & Harper, R. (2003). *Social Shaping of UMTS: Preparing the 3G Customer.* University of Surrey, Digital World Research Centre. Retrieved July 26, 2006, from http://www.dwrc.surrey.ac.uk

Weilenamann, A., Larsson, C. (2002). Local use and sharing of mobile phones. In B. Brown, Green, N., and Harper, R. (Ed.), *Wireless world: Social and interactional aspects of the mobile age* (pp. 92-107): Springer.

KEY TERMS

Bluetooth: A wireless protocol that is used to connect compliant devices that are in close proximity with each other in order to transfer information between them. Bluetooth is commonly used with phones, hand-held computing devices, laptops, PCs, printers, digital cameras.

Camera Phone: A mobile phone with a camera built-in that allows the user to take pictures and share them instantly and automatically via integrated infrastructure provided by the network carrier. Camera phones can transfer pictures via Bluetooth, Infrared, or MMS messaging system.

Co-Present Interaction: Interaction that happens between two or more people that are physically present at the same time and location.

Digital Photo Sharing: An activity of two or more people, who share images by showing pictures to others. Sharing digital photos can occur at the co-present location or remotely. The former happens using different devices like camera phone screen, digital cameras, TV screen, or computer screen. The latter is often done via email or by posting them on the web.

Digital Photography: A type of photography where pictures are taken on digital cameras or camera phones. Images can be viewed, edited, stored, or shared with others using different means of communication medium such as email, Web-based applications and services, Bluetooth, Infra-red, MMS, computers or TV screens.

Field Observation Studies: A qualitative data collection method, which is used to observed naturally occurring behavior of people in their natural settings. The data can be gathered in a form of: film or video recording, still camera, audio type (to record spoken observation), or hand-written note taking.

Qualitative Data Analysis: A collection of methods for analyzing qualitative data, such as interviews or field notes. One example of such method is Grounded Theory, which is used to generate theory through the data gathering and analysis. Data is sorted to produce categories and themes of concepts emerging from the data.

Social Interaction: Interaction that happens between individuals typically mediated by, or in the presence of technological artifacts.

Theoretical Sampling: The process of data collection for generating theory where the researcher collects, codes and analyses data and makes decisions about what data to collect next. Researchers consciously select additional cases to be studied according to the potential for developing new insights or expanding and refining those already gained. Sampling decisions depend on analysis of data obtained, which relate to the developing theory.

Triangulation: The application and combination of at least two research methods or data gathering exercises to research the same phenomena in order to cross-checking one result against another, and increasing the reliability of the results.

User Experience: A term that is used to describe the overall experience and satisfaction of a user while using a product or system.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 55-68, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.25
Mobile Evaluations in a Lab Environment

Murray Crease
National Research Council of Canada, Canada

Robert Longworth
University of New Brunswick, Canada

ABSTRACT

The evaluation of mobile applications is increasingly taking into account the users of such applications' mobility (e.g., Mizobuchi, Chignell, & Newton, 2005; Mustonen , Olkkonen, & Hakkinen, 2004). While clearly an important factor, mobility on its own often does not require the user's visual focus to any great extent. Real-life users, however, are required to be aware of potential hazards while moving through their environment. This chapter outlines a simple classification for describing these distractions and two evaluations into the effect visual distractions have on the users of a mobile application. In both cases, the participants were required to monitor both their environment and the display of their mobile device. The results of both evaluations indicated that monitoring the environment has an effect on both task performance and the subjective workload experienced by the participants, indicating that such distractions should be considered when designing future evaluations.

INTRODUCTION

Mobile computing devices are becoming increasingly popular but the evaluation of such devices has not developed at the same rate. Many early evaluations were undertaken on desktop emulators, often because the real devices were not capable of supporting the applications being evaluated. As the availability and power of mobile devices grew, so too did their use in evaluations, but these evaluations were often still run in a static laboratory environment carefully devoid of distractions. Gradually, however, mobility has become an increasingly common component in the evaluation of mobile devices. Clearly, the fact that a user is likely to be mobile is the single greatest difference in context between users of mobile and

desktop devices. This mobility, however, leads to dynamic changes in users' context which may mean the users are not capable of solely focussing on the task at hand. Users may, for example, have to use their visual focus to navigate or they may be listening to a conversation while receiving audio feedback from their device.

This chapter describes two evaluations investigating how visual distractions affect the task performance of a mobile user. In the first evaluation participants were required to navigate a virtual 'maze' using different forms of navigational cues. While navigating through the maze, the participants were required to monitor projections on either side of them. In the second evaluation the participants were required to monitor the display of their wearable computer while moving through a lab and monitoring projections in front of them. The design of both evaluations highlighted the benefits in making the experimenter mobile while running the evaluation. The results showed that forcing the participants to monitor their environment had an impact on the results and should, therefore, be considered in the design of all mobile evaluations.

The remainder of this chapter is structured as follows. The second section gives an overview of the state of the art in mobile evaluations while the third section introduces a classification that can be used to describe the different forms of distraction that can affect mobile users. The fourth and fifth sections describe two experiments that evaluated the effect of visual distractions on users of mobile applications while the sixth section discusses the results of these evaluations in terms of their experimental design in general, and the distractions used in particular.

MOBILE EVALUATION

There are many examples of mobility having an effect on a user's task performance. Brewster, for example, showed that the amount of data entered using button presses was significantly reduced when comparing a seated, indoor user with a mobile, outdoor user (Brewster, 2002). It was also found that the subjective workload experienced by the participants was significantly increased. Brewster suggests that this is not surprising and goes on to say that further research is required to develop appropriate evaluation techniques for the evaluation of mobile devices in realistic situations. This section presents previous research that has incorporated mobility into the evaluation of mobile applications; and in particular, where it has been used in lab evaluations.

Mustonen et al. (2004) investigated the effect of walking on the legibility of mobile phone text. Four walking conditions—natural speed in a corridor, natural speed on a treadmill, fixed speed of 1.5 km/h on a treadmill, and fixed speed of 3 km/h on a treadmill—were compared to determine if the effect of mobility varied with speed. It was found that although mobility had an effect on legibility when reading normal text, there was no significant effect when parsing pseudo-text with a view to finding a text pattern. The overall workload of both tasks, as measured by NASA TLX ratings (Hart & Staveland, 1988), was significantly effected by mobility.

Mizobuchi et al. (2005) investigated the effect of walking on text input. Participants were required to enter English language sayings using one of four sizes of soft keyboard when either stationary or walking along a corridor. The size of the keyboard had a significant effect on the text input speed but walking had no significant effect on the speed. Furthermore, walking only had a significant effect on the number of errors when the participants were using the smallest keyboard. It is suggested that these results indicate that text input and walking can, in general, be viewed as separate tasks that have no effect on each other apart from a fixed cost to each task due to the presence of the other. This was indicated by a

reduction in walking speed when inputting text and a reduction in input speed when walking, although these effects were not significant. It could also be argued that the inputting and walking tasks were such that, other than when using the smallest keyboard, the participants had sufficient cognitive and visual capacity to successfully manage both but if the load was increased (e.g., when the keyboard was very small or if more complex navigation was required) then a noticeable effect may become apparent.

Crossan, Murray-Smith, Brewster, Kelly, and Musizza (2005) described a quantitative approach to measuring the effect of walking on usability. Using an accelerometer attached to the serial port of a PDA, it was possible to determine the users' gait while walking and consequently that the rhythm of walking affected the users' ability to select on-screen targets. Participants were required to tap targets that appeared on the screen at random intervals. This was done both when walking and when seated with, not surprisingly, far greater accuracy achieved when seated. In the walking condition it was found that there was a correlation between the phase of the participants' gait and both the accuracy and number of taps.

The examples of mobility in the evaluations presented thus far have been in controlled experimental scenarios where the participants are required to be mobile but are not required to monitor their surroundings as would be the case in a real-world scenario. Kjeldskov and Stage (2004) compared six techniques which could be used to increase the realism of such evaluations: sitting at a table; walking on a treadmill at a constant speed; walking on a treadmill at a variable speed; walking at a constant speed on a course that is constantly changing; walking at a variable speed on a course that is constantly changing; and in a field evaluation (walking in a pedestrian street). The five lab-based techniques related to the five possible combinations of motion (none, constant, and variable) and attention required navigating (yes or no). The different techniques were compared in terms of the number of usability problems found, the task performance, and the subjective workload experienced by the participants. Interestingly, the participants were best able to find usability problems when sitting at a table. It is suggested that this is because the participants are able to devote the most attention to the means by which problems were recorded—thinking aloud. Mobility had no significant impact on the task performance of the participants. It did, however, have an impact on the workload experienced by the participants. Simply being mobile, however, was not sufficient since as walking on a treadmill at a constant speed did not significantly increase the workload experienced. For this to happen, an additional cognitive load was required whether it be variable speed, variable course, a combination of the two, or being in a real-world situation such as a street. Although the overall workload was significantly increased when a variable course was employed, there was no significant increase in mental demand which would be expected due to the increased demand in following a variable course. It was hypothesised that this was due to the way the variable course was managed, with the participants required to follow an experimenter who followed a variable path. This enabled the participants to merely follow the experimenter with no real effort required to manage the navigation.

Kjeldskov and Stage also experienced difficulties in collecting data when running the experiments in the field. It was hard to video the participants as they moved through the streets and the realism of the setting was compromised as other pedestrians tended to avoid the participant and the experimenters. Goodman, Brewster, and Gray (2004) report that a further problem with field studies is the difficulty in controlling confounding variables. They suggest that this problem may be minimised by removing data where it varies too greatly from appropriate control

levels. This, however, can be both expensive and time consuming so Goodman et al. suggest that such results should be included as they are part of the real-world context in which the evaluated system is expected to operate.

Kjeldskov, Skov, Als, and Høegh (2004) investigated whether the added effort required to undertake a usability evaluation in the field is worth it in terms of the results such an evaluation produces. The investigation compared the effectiveness of two evaluations of a mobile Electronic Patient Record (EPR) system. Two forms of the evaluation were undertaken: one in a lab-based simulation of a hospital ward and the other in a real hospital ward. The comparison of the two forms of evaluation was based on the number of usability problems identified. Surprisingly, significantly more serious and cosmetic problems were discovered in the lab-based evaluation than in the field-based evaluation. Only one problem discovered in the field was not discovered in the lab and this problem was not even directly linked to the usability of the system but rather to the veracity of the data entered into the system and its storage in a database. Furthermore, it was found that running the experiment in the field posed challenges with respect to the collection of data. Participants in the lab, for example, were prepared to use a note taking facility to document problems they found whereas nurses operating in a real life context (not surprisingly) did not. Although the particular context of this study—a hospital ward where patient safety is the most critical factor—may have been a factor, the results do indicate that if the real-world context is taken into account when creating a lab environment, a lab evaluation may be at least as good as a field evaluation.

Duh, Tan, and Chen (2006) also undertook a comparison of lab and field evaluations. Two groups of participants undertook an evaluation of a mobile-phone based application in one of two settings: seated in a lab with the usage scenario textually described or in the field in the actual usage scenario. In both cases, the think aloud technique was used and the participants' interaction with the application was recorded. In contrast to Kjeldskov et al. (2004), significantly more critical errors were found by the participants in the field than by those in the lab. Although no definitive reason is given, there are several possibilities. The lab-based participants were seated during the evaluation so no attempt was made to mimic the real-life context of use. Also, the participants in the field expressed increased nervousness and stress which may have been an experimental artifact caused by the requirement to verbally describe everything they were doing in a public location.

This section has described some of the work that has been undertaken on making evaluations of mobile devices more realistic and, as a consequence, more effective. The majority of this work, however, has concentrated solely on mobility. Mobility on its own, however, does not require much of the participants' attention whereas in a real-world context, users are required to manage the consequences of mobility—a dynamically changing environment—which do require more attention. One way to solve this problem would be to undertake evaluations in the field but this brings its own set of problems such as difficulties in controlling the environment and capturing data. The remainder of this chapter describes two evaluations that attempted to produce a more realistic environment through the use of controlled, visual distractions.

DISTRACTIONS

As previously discussed, mobility alone is not sufficient when evaluating mobile applications. Real-life users of such applications are required to constantly monitor their environment while moving through it to avoid hazards such as other pedestrians or lamp-posts. Clearly, this monitoring will have an impact on both the attention and cog-

nitive load the user is able to devote to the mobile application. Furthermore, in a real-world context, users will be distracted by the sights, sounds, and smells of the environment regardless of whether these distractions pose a potential risk to the user. This section proposes a simple classification of the different forms of distraction that could be used in lab-based evaluation of mobile devices.

- **Passive distractions** distract users but require no active response. A real-life example of such a distraction is a billboard.
- **Active distractions** require the user to respond in some way. The required response will vary according to the distraction. A mobile phone ringing, for example, may require the user to answer it whereas the presence of a lamp-post will require the user to navigate around it.
- **Interfering distractions**—which may be passive or active—interfere with the user's interaction with the mobile device. The sound of a passing car or an ongoing conversation, for example, may limit a user's ability to correctly perceive audio feedback being presented by the mobile device. A distraction may be interfering in one context (e.g., the example given) but not in another (e.g., if the user was not relying on audio feedback).

Clearly, in this scheme, the classification of some distractions is subject to debate. A mobile phone ringing, for example, is only an active distraction if the user chooses to answer the call. These decisions, however, can be controlled in an experimental setting (e.g., by instructing users to answer all calls). In this way it is hoped that the effect of different types of distraction and different techniques for managing distractions (e.g., answer all calls immediately or ignore most calls) may be investigated. The remainder of this chapter describes two evaluations that investigate the use of active, visual distractions that interfere with the participants' ability to interact with the visual interface of a mobile application.

EVALUATION 1: COMPARISON OF AUDIO AND VISUAL NAVIGATIONAL CUES

A lab evaluation investigated whether simple, non-spatialised sounds could be effective at enabling users to navigate. Such sounds would have the advantage of not requiring headphones which may occlude environmental sounds of interest to the user.

Background

There are several systems which have used audio feedback to provide users with navigational information. MOBIC (Petrie, Johnson, Strothotte, Raab, Fritz, & Michel, 1996) is an example of a system designed to allow blind users to navigate. MOBIC used GPS information mapped to speech that provided users with assistance for macro-navigation—the navigation through the distant environment—which is typically done using visual cues such as church steeples. One of the main findings of the user analysis undertaken was that the users did not want to wear headphones as it was felt that this would block out useful environmental sounds which are especially important to visually-impaired users. More recently, the Personal Guidance System (Loomis, Marston, Golledge, & Klatzky, 2005) was evaluated using different forms of spatial display. As with MOBIC, the system mapped GPS information into direction and distance feedback which was presented to the users sonically.

The use of audio navigation cues is not limited to systems designed for visually impaired users. AudioGPS (Holland, Morse, & Gedenryd, 2002) was designed to allow sighted users to receive navigational information using spatialised non-speech audio presented using headphones. Non-speech

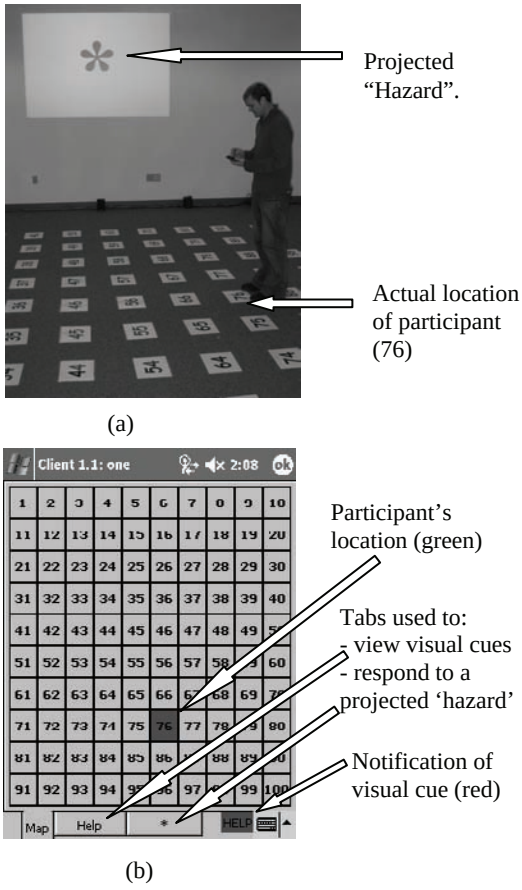
sounds were chosen to minimise any interference with the users' conversations. The system presented the users with two pieces of information: distance to their destination (or intermediate waypoint) and its direction. Initial trials indicated that the sounds were effective in allowing users to discern the direction of the destination but that the implementation of the system meant that it was slow to respond to a user's change in direction. The gpsTunes system (Strachan, Eslambolchilar, Murray-Smith, Hughes, & O'Modhrain, 2005) modified currently playing music to provide direction and distance information to the user. As the distance to the target decreased, the volume of the music was increased up to a user defined maximum. The stereo pan of the music indicated the correct direction of the target. When the target was reached, a pulsing sound was played over the music. As with AudioGPS, the sounds were presented using headphones. An initial field evaluation indicated that users were able to navigate successfully using the system but it is unclear whether the annoyance of modifying the music would prove to be annoying to users in the longer term.

Experimental Design

An experiment was designed to investigate two main questions: what affect do visual distractions have on users of mobile devices; and how effective are simple, non-spatialised sounds at providing navigational cues? The experimental task was in the form of a game with the participants required to navigate through a virtual 'maze' as quickly as possible. The participants could not see the maze but only the numbered grid in which the maze was located (Figure 1). As the participants moved through the maze they were provided with a cue indicating in what direction they were to move next (forwards, backwards, left, or right). When they had moved, the participants clicked on their new location in the client interface (Figure 1b) and the task proceeded until the participant reached the end of the maze.

Two forms of navigational cue representing the directions left, right, forward, and backward were used. Visual cues took the form of an arrow pointing in the appropriate direction. These cues were accessed by pressing the "Help" tab at the bottom of the screen (Figure 1b). The arrows were

Figure 1. Two views of the same maze: (a) as represented by a grid of cells on the floor through which the participants must navigate; and (b) as it is presented to the participants on the handheld device

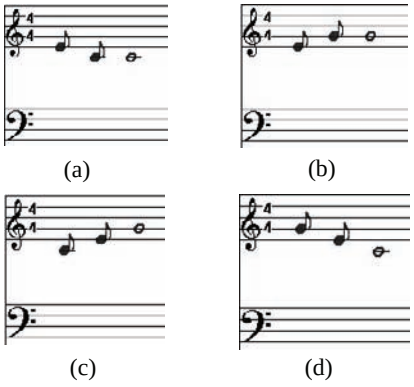


visible for 350ms before the display returned to the view of the map. The arrow could only be viewed once for each step. 350ms was chosen as a suitable length of time to display the visual cue as Öquist and Goldstein (2001) report that the average time required to fixate—or parse—visual information is ~330 ms.

Four earcons (Blattner, Sumikawa, & Greenberg, 1989) which all shared the same basic structure—two notes of duration 80ms followed by a third note of duration 480ms—were used for the audio cues. All the notes were played in a piano timbre. The directions were differentiated using the notes' pitch:

- Left—E3, followed by C3, followed by C3[1] (Figure 2a)
- Right—E3, followed by G3, followed by G3 (Figure 2b)
- Forward—C3, followed by E3, followed by G3 (Figure 2c)
- Back—G3, followed by E3, followed by C3 (Figure 2d)

Figure 2. The four sounds: (a) left; (b) right; (c) forward; and (d) backward



By using pitch as the only parameter by which the sounds can be distinguished, the possibility for encoding more complex navigational cues which could inform the user about a more detailed direction and/or the distance to be travelled is left open. By relying only on relative pitch (i.e., the way the pitch changes within the earcons) as opposed to relying on absolute pitch (i.e., the way the pitch changes between earcons), these sounds follow guidelines on the design of audio feedback (Lumsden, Brewster, Crease, & Gray, 2002). Because the earcons are not spatialised, the user was not required to wear headphones which may block out other sounds (Petrie et al., 1996).

To simulate a realistic mobile environment—where users are required to be aware of their surroundings—the participants were also required to monitor their surroundings and react to distractions accordingly. Projections were displayed at pseudo-random intervals on either side of the participants. Figure 1a, for example, shows a '*' being projected to the right of the participant. Two forms of projection were used: characters and images. In the characters condition, six characters were used to represent non-hazards—'u,' 'v,' 'w,' 'x,' 'y,' 'z;' a seventh—'*'—was used to represent a hazard. In the images condition, seven images replaced the seven characters of the characters condition. The non-hazard images were pictures of empty roads. The hazard image had a single moving car on a road. Participants were required to respond to the projection of a 'hazard' by pressing a tab at the bottom of the interface on the handheld (Figure 1b).

Experimental Procedure

The experiment consisted of three conditions run between three groups. The three conditions were: visual cues only; audio cues only; and both audio and visual cues. The three groups were: character distractions; image distractions; and no distractions. The third group was used as a control group to determine whether the participant monitoring

their environment had an effect on the evaluation results. Twenty-four participants were divided equally between the three groups. The order in which the audio and visual conditions were presented within each group was counterbalanced with the audio-visual condition always presented third. This design allowed the effectiveness of the two different forms of cues to be determined. The audio-visual condition allowed us to quantitatively determine whether the participants chose to use the audio or visual cues after experiencing both. This was measured by recording how often the participants clicked on the 'Help' tab to view the visual navigational cues.

Each condition consisted of a short training session where the participants were able to familiarise themselves with the navigational cues for that condition followed by a short training maze consisting of 16 steps. The participants then had to navigate a 40 step maze for the actual condition. Three full length and three training mazes were designed; each with an equal number of forward and backward steps (10 of each for the full length mazes and four of each for the training mazes). This was done to ensure the mazes were all of similar difficulty as pilot participants found it harder to move backwards than forwards. To eliminate the mazes as an experimental variable every participant was required to navigate the same mazes in the same order regardless of condition. Similarly, the same projection sequences were used in the same order for every participant.

Results

The different forms of visual distractions used had no significant effect on the participants' performance in the navigation task. The different forms of distraction did, however, effect the participants workload with a two factor ANOVA showing that distraction type significantly effected time pressure ($F_{2,68}$ =5.72, p<0.01), performance level achieved ($F_{2,68}$ =7.65, p<0.01), and the overall workload ($F_{2,68}$ =4.14, p=0.021) experienced by the participants as shown in Table 1.

Post hoc Tukey HSD tests showed that the use of image distractions significantly increased the workload compared to the no distractions group (time pressure p<0.01, performance level p<0.01, overall workload p<0.02). There were no significant differences in the subjective workload between the character distractions group and the other groups.

A two factor ANOVA showed that the different distraction types significantly affected the participants' ability to correctly detect projected hazards ($F_{2,68}$ =10.68, p<0.01). *Post hoc* Tukey HSD tests showed that the participants were significantly better at detecting hazards in the character group—67% correctly noticed—com-

Table 1. Average workload results across the different distraction groups

Workload / Distraction	Average Time Pressure	Average Performance Level Achieved	Average Overall Workload
Character	8.96	14.08	44.94
Image	11.9	11.87	51.19
None	7.29	16.67	37.56

pared to the image group—40% correctly noticed ($p<0.01$). This result was confirmed by a significant increase in the number of correct responses in the character group—on average 8.25 per condition—compared to the images group—on average 4.83 per condition ($p<0.01$).

When analysing the use of the audio cues the audio-visual condition was ignored due to the likely training effect caused by this condition always being third. Two analyses were undertaken, one across the two distraction groups and one across all three groups.

A two factor ANOVA test across the two groups with distractions, showed the average time taken to make a correct step was significantly reduced from 5.5 seconds in the visual condition to 3.9 seconds in the audio condition ($F_{31,1}=21.27$, $p<0.01$). The overall task time was not significantly reduced, however, as there were more navigational errors made in the audio condition (4.5) than in the visual condition (1.1). This, however, was not a significant difference ($F_{31,1}=2.99$, $p=0.09$).

When considering the subjective workload, a paired T-test showed that the physical workload experienced by the users in the audio condition was significantly reduced from 3.9 on average in the visual condition to 2.6 ($T_{15}=2.4$, $p<0.04$) (Figure 3).

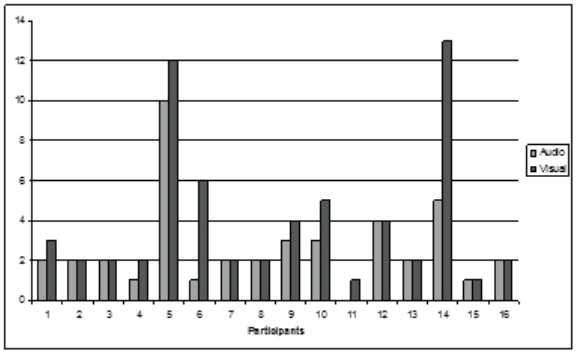
When similar comparisons were performed across all the groups the improvement in the time taken to take a correct step when using the audio feedback was maintained. When considering the subjective workload, however, significant increases in mental demand ($F_{1,47}=15.56$, $p<0.01$), effort expended ($F_{1,47}=5.52$, $p<0.03$), frustration ($F_{1,47}=5.01$, $p<0.04$), and overall workload ($F_{1,47}=15.56$, $p<0.01$) were found in the audio condition.

The audio-visual condition was subjectively analysed to see how often the participants chose to use the visual feedback when they had the option (Table 2).

Table 2. Percentage use of visual help in the audio-visual conditions across the three distraction groups

Distraction Type	% Use of Visual Help
Character	20.78%
Image	29.40%
None	40.84%

Figure 3. Average physical demand experienced by the participants who were monitoring distractions



Discussion

Overall, the audio cues were not wholly successful. Although the average time to make a complete step was significantly improved the overall task time did not reflect this. This was almost certainly due to the increased number of navigational errors made in the audio condition. Although not a significant increase, the cost of an error (each required up to 5 extra steps to be made) meant that any benefit in the time taken to make a step was lost. The improvement in the time to take a step can almost certainly be attributed to the fact that the audio cues were automatically presented to the participants whereas the visual cues had to be actively retrieved.

Anecdotally, however, the audio cues did seem to be preferred by many users with one in particular commenting "I found the visual to be the hardest even though I expected it to be easier." Other common comments among those who preferred the audio cues were that they allowed the participants to monitor their surroundings more easily, with the visual cues requiring too much of their visual attention. Conversely, many participants commented that they found the sounds hard to differentiate and therefore unsuitable for the task at hand. In these instances, the participants were most likely to fail to differentiate the sounds representing forward/right and/or back/left. This was especially true if the participant had moved several steps in the same direction before being required to change to the direction they found hard to differentiate. Observing the evaluations, it became obvious that many participants struggled with the audio cues initially but eventually reached an understanding of how the sounds were structured. Some participants, for example, navigated poorly in the training maze for the audio condition but performed flawlessly in the audio condition itself, indicating that despite having an opportunity to learn the sounds prior to the training maze, perhaps a longer training maze would have been appropriate.

Of more interest is the interaction between the distractions and the audio cues which was perhaps best highlighted by the number of times the visual cues were used in the audio-visual condition, Table 2. The visual help was only used 21% of the time when character distractions were used, 29% of the time when images were used, and 41% of the time when no distractions were used. In the latter case it is likely that, because the users were not required to monitor their surroundings, their visual focus could remain on the PDA and so the effort required to parse the sounds was not as necessary. The difference between the character and image distractions could be explained in terms of the participants' increased difficulty in determining what a hazard in the image condition was. The extra effort required to spot the hazards meaning the participants could spare less effort to parse the sounds.

It was also noticeable that, when comparing the overall workload in the audio condition across all three groups, it was significantly higher in the audio condition than in the visual condition. If only the two groups that had visual distractions were considered, then the only significant difference in workload was an increase in physical demand in the visual condition. This indicates that when the distractions were present, the audio meant the participants only had to turn left and right to monitor their surroundings while without the audio the participants were required to monitor the iPaq also. If no distractions were present, however, this extra physical demand was not present, meaning the extra effort required to parse the audio feedback became a significant factor.

The number of character hazards successfully detected was significantly higher than the number of image hazards. This was somewhat surprising given that in both instances the participants were looking for a single projection out of seven possibilities. In one case it was a '*' and in another it was the picture of a car on a road. The characters projected were chosen to be as similar as possible (u, v, w, x, y, z, and *)

as they largely all consisted of diagonal straight lines. It may have been that the '*' stood out due to the density of lines? The images were also all similar, containing a road disappearing into the horizon but perhaps the inclusion of the car was not as noticeable as the '*'.

EVALUATION 2: COMPARISON OF WEARABLE DISPLAYS

While mobile computing is typically associated with hand-held devices such as PDAs and cell phones, wearable computers are becoming increasingly common due to several advantages they possess: speed of access; hands free use; and privacy (Starner, 2003). Typically, wearable displays take the form of head or glasses-mounted displays (HMD or GMD) which utilise small displays located close to the eye enabling a high resolution image to be shown. Displays may be visible in one eye (monocular) or two (binocular). They may be transparent—enabling the user to see the surrounding environment through the display—or opaque. An alternative form of wearable display is a standard flat panel screen that is worn by the user—typically attached to a belt. The advantages of such displays include larger screen size and less occlusion of the environment. The main disadvantages are: the screens can be awkward to wear; reduced privacy; and users are required to actively move their visual focus away from the environment to the screen. This section describes a lab evaluation evaluating the effectiveness of two different forms of wearable display (a fold-down screen and a GMD) under different conditions.

Background

Sheedy and Bergstrom (2002) evaluated the effectiveness of different HMDs compared to different forms of monitor and hard copy for reading tasks. They found that the participants' performance was comparable across all displays. This result differed from previous research where the performance of participants using head-mounted displays was not comparable. It is suggested that these differences may have been caused—in part—by a lack of movement in the evaluation. Revels and Kancler (2000) undertook an evaluation of head-mounted displays that did incorporate mobility. Participants were required to perform three different tasks using a GMD while navigating one of two courses. All the tasks required the participants to press an appropriate button on a wrist mounted keypad based on a visual cue. Three forms of visual cue were used: graphical (a graphical representation of the keypad was shown with the appropriate key highlighted); numerical (the numerical label of the key to be pressed was shown); and textual (a textual description of the location of the key to be pressed was shown). The two courses were: clear (a straight course 100 feet long); and obstructed (participants were required to slalom through a series of different obstacles on a 100 foot long course). The results indicated the participants found the textual task the hardest, followed by the numerical and graphical tasks. The impact of having to avoid obstacles also significantly affected the participants' performance. Most interesting, however, was the fact that in the clear course condition the time taken to respond to a cue was significantly different between all three tasks whereas in the obstacle course condition only the graphical task was significantly different from the other tasks. This indicates the importance of incorporating a realistic environment in mobile evaluations in order to achieve the most accurate results.

Experimental Design

A lab evaluation was undertaken to determine the effectiveness of two common forms of wearable display: a glasses mounted display and a fold down screen. The experimental application ran on a Xybernaut MAV wearable computer running

Windows XP. The fold-down screen was attached to the waistcoat that held the wearable. The glasses-mounted display used was the MicroOptical SV-9 monocular display. In both cases, the display had a resolution of 640x480. The experimental task required the participants to walk between six locations in a lab while monitoring both the display of the wearable computer and their environment. As with the previous experiment, images were projected onto the walls of the lab with the participants required to acknowledge hazards. In this case, however, only images of the same form as those used in the image condition of the previous experiment were used. The participants were required to acknowledge the presence of a hazard simply by shouting "car," whereas in the previous experiment the participants were required to press a button on the experimental interface. This form of acknowledgement was chosen so as to remove a confounding variable—the participants' use of an unfamiliar interaction device: a handheld mouse. Acknowledging the hazards this way meant that once the participants had signalled they were ready they could move the cursor over the only active button on the experimental interface and press it when required without needing to move the mouse again. When a participant shouted "car" the experimenter recorded this acknowledgment using an experimenter application running on an iPaq. The experimental interface is shown in Figure 4. While moving between locations in the lab, the participants were also required to monitor the three characters displayed in the centre of the interface. When the three characters matched, the participants acknowledged this by pressing the "Match" button on the interface.

The layout of the lab is shown in Figure 5. The participants were required to walk from their current location to the location indicated in the top left corner of the experimental interface. The locations were physically indicated by squares approximately a meter across marked on the lab floor with the location number marked both in the centre of the square and behind the loca-

Figure 4. The experimental interface showing the next location to move to (top left), the characters to be matched (centre), and the "Match" button (centre left) which is pressed when the characters match. The "Ready" button (bottom left) is used only to signal the participants' readiness to start and is subsequently disabled. Both buttons flashed red when pressed to confirm that a press occurred.

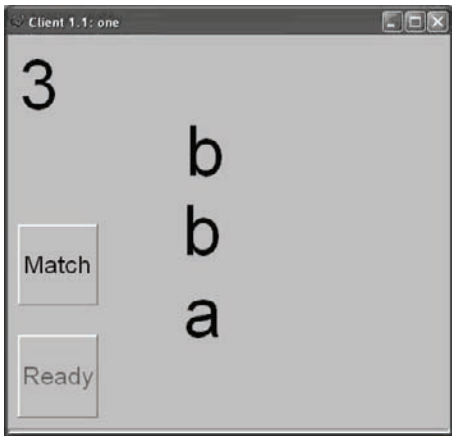
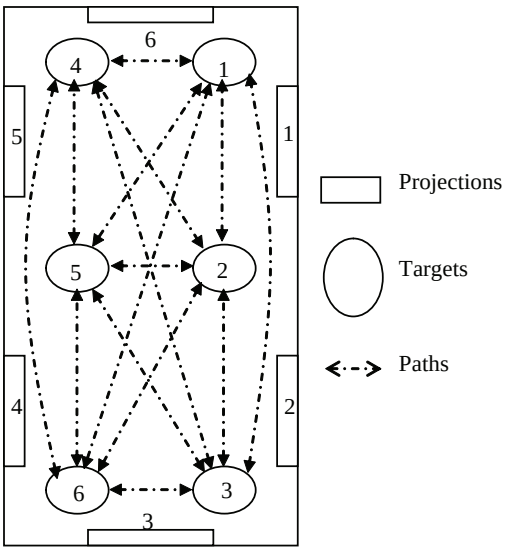


Figure 5. Lab layout showing the different locations participants were required to walk between and the location of the projections used to present potential hazards



tion on the wall. When a participant placed both feet in the square the next location was updated by the experimenter using the experimenter application. As with the acknowledgements of the hazards, this technique eliminated the need for the participant to move the cursor to press a different button—eliminating a confounding variable. At any given time the two projectors behind the location to which the participants were walking were used to display potential hazards. The definition of 'behind' depended not only on the destination but also the previous location. For example, if a participant was walking from location 5 to location 2, projectors 1 and 2 would be used. If, on the other hand, a participant was walking from location 3 to location 2, projectors 1 and 6 would be used.

Experimental Procedure

The experiment consisted of two conditions run between three groups. The conditions were display: glasses mounted (GMD) or fold-down screen (FDS). In the GMD condition, if the participant wore glasses they could chose to clip the display onto their own glasses or wear the glasses that came with the display over their own glasses. The FDS was not removed from the waistcoat containing the wearable but the screen was folded up so it was not visible. In the FDS condition, the participants were required to wear glasses (either their own or the non-prescription glasses that came with the GMD) but without the display attached to remove glasses as a confounding variable. The groups were hazard distractions: none

(No Distraction); single set of images (Single Distraction); or multiple sets of images (Multiple Distraction). The hazards were created by mapping one of seven characters to images. In the single distraction group, each character mapped to a single image. In the multiple distraction group, each character mapped to three images. In both cases only one character represents a hazard. A total of 18 participants performed the experimental task under both conditions; with the participants split equally across the three groups. The order in which the conditions were undertaken was counter-balanced within each group to eliminate any training effect.

Each condition consisted of a short training session where the participants were able to familiarize themselves with the task to be undertaken and the particular display being used. This training task required the participants to move to 18 different locations while monitoring their display and environment. After the training task, the experimental task—which consisted of moving between 60 locations in the lab—was undertaken. The paths the participants were required to follow were designed so that in each full condition the participant had to walk between each location in each direction exactly twice. To eliminate the paths as an experimental variable, every participant was required to navigate the same paths in the same order regardless of condition. Similarly, the same projection sequences were used in the same order for each participant. After completing the condition, the participants were required to complete a NASA TLX workload questionnaire (Hart & Staveland, 1988) to give an indication of the subjective workload experienced.

Results

The most surprising result was that a series of two-factor ANOVA tests showed no significant differences in either the subjective or quantitative results between the two displays. Total task time ($F_{1,30}$=0.00, p=0.992), percentage missed hazards ($F_{1,30}$=0.02, p=0.895), percentage missed matches ($F_{1,30}$=0.34, p=0.564), and overall workload ($F_{1,30}$=0.80, p=0.379) were all statistically similar.

Two-factor ANOVA tests did indicate, however, that the different distraction groups did have an effect on the results, with significant differences found in total task time ($F_{1,30}$=0.4.25, p=0.024), percentage missed matches ($F_{1,30}$=4.04, p=0.028), mental demand ($F_{1,30}$=5.18, p=0.012), physical demand ($F_{1,30}$=4.65, p-0.017), effort expended ($F_{1,30}$=3.89, p=0.032), performance level achieved ($F_{1,30}$=12.72, p<0.001), and overall workload experienced ($F_{1,30}$=6.60, p=0.004). *Post Hoc* Tukey HSD showed that, as one would expect, the requirement to monitor the environment had a negative effect on the results. The average total task time for the no distraction group (535.23 seconds) was significantly faster than for the single distraction group (618.85 seconds, p=0.0248) but not for the multiple distraction group (553.6 seconds, p=0.0938). The percentage of missed character matches was again significantly lower in the no distraction group (1.01%) compared to the single distraction group (7.61%, p=0.0403) but not the multiple distraction group (7.04%, p=0.0651). Similarly, the average mental demand experienced by the participants was significantly lower in the no distraction group (10.25) compared to the single distraction group (14.25, p=0.0121) but not the multiple image group (13.33, p=0.0616); and the average physical demand in the no distraction group (3.83) was significantly lower compared to the single distraction group (9, p=0.015) but not the multiple distraction group (7.33, p=0.124). The effort expended by the participants, however, was significantly lower in the no distraction group (8.33) when compared to the multiple distraction group (12, p=0.0418) but not the single distraction group (11.58, p=0.0778). The performance level the participants felt they achieved was significantly higher in the no distraction group (15.5) than both the single distraction (12.83, p=0.0369) and multiple distraction (10.33, p<0.001) groups. The

overall workload experienced by the participants was also significantly lower in the no distraction group (6.98) than in the single distraction (9.84, p=0.013) and multiple distraction (10.028, p=0.008) groups.

Discussion

The most surprising aspect of the results was that the participants performed equally as well with the fold-down screen as with the glasses-mounted display. At first glance, it is reasonable to think that the participants would perform better when using the GMD as both the characters to be matched and the hazards to be noticed would be in the same field of vision. This, however, was not the case with participants missing fewer hazards in the FDS condition (8.48%) than in the GMD condition (8.96%) although this difference was not significant. Participants did miss fewer character matches in the GMD condition (4.61%) than in the FDS condition (5.84%) but, again, this difference was not significant. Even the physical demand experienced by the participants showed no significant differences ($F_{1,30}$=1.05, p=0.315) between the GMD (6) and FDS (7.44) conditions. This was surprising given the need to constantly bend down and look at the fold-down screen. Furthermore, due to the way the FDS was mounted (essentially on a belt), the screen had a tendency to angle away from the participants, requiring them to hold it in place. These results may, however, be explained by the difficulty participants experienced focusing on the GMD. Although the GMD was in the same field of vision as the projected images, the participants still found it necessary to actively shift their focus from their surroundings to the screen. This may be explained in terms of binocular rivalry (or, simply put, the confusion caused by two eyes seeing different things) and the different depth of focus required by the two eyes as discussed by Laramee & Ware (2002).

It was not surprising, however, that requiring the participants to monitor their environment while moving around the lab had a detrimental effect on their performance (Table 3). What is interesting, however, is that there were no significant differences in task performance between the single and multiple distraction groups. The single distraction group, on average, took slightly longer to complete the task and missed slightly more character matches but neither of these differences was significant. The multiple distraction group, on the other hand, experienced a slightly higher overall workload but, again, this difference was not significant. What this would seem to imply is that the participants were required to devote their attention to the two main tasks—monitoring the characters on the display and monitoring the projections for hazards—and the participants

Table 3. Summary of the results highlighting the impact of monitoring the environment on the participant's task performance

	Avg. Task Time (Secs.)	Avg. % Of Missed Matches	Avg. Overall Workload
Single Distraction	618.85	7.61%	9.85
Multiple Distraction	553.61	7.04%	10.03
No Distractions	535.24	1.02%	6.99

devoted slightly different amounts of effort to each. The overall performance, however, was approximately equivalent.

DISCUSSION

This chapter has described two evaluations of mobile systems that attempted to increase the realism of the lab environment through the use of visual distractions. In the first experiment, the participants' use of the audio cues was noticeably lower in the audio-visual condition when there were no visual distractions. This was most likely due to the participants' visual focus not being required anywhere other than on the device's screen. When the participants' visual focus was required elsewhere (to monitor the projections for hazards), the participants were prepared to make the added cognitive effort to parse the sounds. In the second evaluation, the % of character matches missed by the participants was significantly greater when the participants were required to monitor the environment for hazards. What these results demonstrate is that when evaluating a mobile application it is important to consider more than just mobility when considering the context in which to undertake the evaluation.

While it is not possible to draw any conclusions regarding the realism of the environment created, it would appear that the use of pseudo-realistic distractions is preferable to more abstract distractions. In the first evaluation, participants were better able to discern the hazards when presented as characters as opposed to photographic images. While the use of different forms of abstract distraction may reduce this problem, the use of images would seem to be the simplest approach. Interestingly, the use of a single set of images (with one hazard image out of a total of seven) did not have any effect on the results when compared to the use of multiple sets of images (with 3 hazard images out of a total of 21). This implies that the participants were not memorizing the hazards, meaning that if the distractions are of an appropriate form, a relatively small set is appropriate. Future work is necessary to validate these hypotheses and to determine whether the form of distractions presented here provide results that are similar to a real-world scenario.

It is also worth mentioning the importance of experimenter mobility in the two evaluations presented. In both instances, a mobile application allowed the experimenter to interact with the experimental server. This eased the running of the evaluations by allowing the experimenter to move around the lab, assisting the participant when necessary. For example, in both cases, the experimenter could ensure that the participant was in the appropriate location at the start of the experiment and was ready to begin before starting the experiment. While this could have been achieved from the observation room off the lab, the mobility of the experimenter made this easier. It also enabled the participants to acknowledge the hazards in the second evaluation without having to shout to an unseen experimenter in the observation room.

CONCLUSION

This chapter has described two evaluations which demonstrated that the use of visual distractions can have a significant effect on the results of a usability evaluation of a mobile device. Participants' task performance was significantly affected by the requirement to monitor their environment for such distractions and the subjective workload experienced was also increased. Interestingly, there was some evidence that the introduction of such distractions changed the results of an evaluation, with the workload experienced when using audio navigational cues—compared to visual cues—being significantly higher when no distractions were present but being the same when distractions were used. This would indicate that—although further work is required in this field—the use of

visual distractions is something that should be considered when designing an evaluation of a mobile application.

ACKNOWLEDGMENT

We would like to thank all the experimental participants for their participation. The software was written in the EWE (www.ewesoft.com) and Java (www.java.com) programming languages. The sounds were designed using MIDI Studio V4.20 (www.sonicspot.com/midistudio/midistudio.html) and converted to wave files for playback on an iPaq using Midi2Wave Recorder V3.5 (www.midi2wav.com). This work was performed under NRC REB ethics approval numbers 2005-46 and 2005-47.

REFERENCES

Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *Human-Computer Interaction, 4*(1), 11-44.

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing, 6*(3), 188-205.

Crossan, A., Murray-Smith, R., Brewster, S., Kelly, J., & Musizza, B. (2005). Gait phase effects in mobile interaction. In *CHI '05 Extended Abstracts on Human factors in Computing Systems* (pp. 1312-1315). Portland, OR, USA: ACM Press.

Duh, H. B.-L., Tan, G. C. B., & Chen, V. H.-h. (2006). Usability evaluation for mobile device: a comparison of laboratory and field tests. In *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services (Mobile HCI 2006)* (pp. 181-186). Helsinki, Finland: ACM Press.

Goodman, J., Brewster, S. A., & Gray, P. D. (2004). Using field experiments to evaluate mobile guides. In *Proceedings of HCI in Mobile Guides (Workshop at MobileHCI 2004)*, Glasgow, Scotland.

Hart, S., & Staveland, L. (1988). Development of NASA_TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam, Holland.

Holland, S., Morse, D. R., & Gedenryd, H. (2002). AudioGPS: Spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous Computing, 6*(4), 253-259.

Kjeldskov, J., Skov, M. B., Als, B. S., & Høegh, R. T. (2004). Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In S. Brewster & M. Dunlop (Eds.), *Mobile human-computer interaction - Mobile HCI 2004* (Vol. 1, pp. 71-73). Glasgow, Scotland: Springer-Verlag.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human Computer Studies (IJHCS), 60*, 599-620.

Laramee, R. S., & Ware, C. (2002). Rivalry and interference with a head-mounted display. *ACM Transactions Computer-Human Interaction, 9*(3), 238-251.

Loomis, J. M., Marston, J. R., Golledge, R. G., & Klatzky, R. L. (2005). Personal guidance system for people with visual impairment: A comparison of spatial displays for route guidance. *Journal of Visual Impairment & Blindness, 99*(4), 219-232.

Lumsden, J., Brewster, S., Crease, M., & Gray, P. D. (2002). Guidelines for audio-enhancement of graphical user interface widgets. In F. Détienne, X. Faulkner & J. Finlay (Eds.), *Proceedings of HCI 2002* (Vol. II, pp. 6-9). London: Springer-Verlag.

Mizobuchi, S., Chignell, M., & Newton, D. (2005). Mobile text entry: relationship between walking speed and text input task difficulty. In *Proceedings of the 7th international Conference on Human Computer Interaction with Mobile Devices & Services* (pp. 122-128). Salzburg, Austria: ACM Press.

Mustonen, T., Olkkonen, M., & Hakkinen, J. (2004). Examining mobile phone text legibility while walking. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (pp. 1243-1246). Vienna: ACM Press.

Öquist, G., & Goldstein, M. (2001). Towards an improved readability on mobile devices: Evaluating adaptive rapid serial visual presentation. In F. Paternò (Ed.), *Proceedings of Mobile HCI 2002* (pp. 225-240). Pisa, Italy: Springer-Verlag, Berlin Heidelberg.

Petrie, H., Johnson, V., Strothotte, T., Raab, A., Fritz, S., & Michel, R. (1996). MOBIC: Designing a travel aid for blind and elderly people. *Journal of Navigation, 49*(1), 45-53.

Revels, A., & Kancler, D. (2000). Evaluation of mobile computing displays. In *Proceedings of SPIE, 2000* (Vol. 4021, pp. 33-44).

Sheedy, J., & Bergstrom, N. (2002). Performance and comfort on near-eye computer displays. *Optometry and Vision Science, 79*(5), 306-312.

Starner, T. (2003). The enigmatic display. *Pervasive Computing, 3*(1), 15-18.

Strachan, S., Eslambolchilar, P., Murray-Smith, R., Hughes, S., & O'Modhrain, S. (2005). GpsTunes: controlling navigation via audio feedback. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services* (pp. 275-278). Salzburg, Austria: ACM Press.

KEY TERMS

Active Distraction: A distraction that a user must respond to in some way. Examples of such distractions include hazards that a user must avoid (e.g., a lamp-post) or a mobile-phone that the user answers when it rings.

Earcon: Abstract, structured sounds used to provide information to a user. The musical qualities of the sound (e.g., rhythm, timbre, or pitch) can be varied to convey information to users. Earcons can be combined sequentially (compound earcons) or concurrently (parallel earcons).

Field Studies: The evaluation of a mobile application that takes place in the actual context of use. The advantage of such evaluations is that problems that only arise in the particular context will be detected. The disadvantages of such evaluations include difficulties in controlling the environment and capturing evaluation data.

Head-Mounted Display: A display that allows a user to view the visual output of a wearable computer at all times. Such displays may cover one eye (monocular) or two (binocular). They may be opaque (for immersive environments) or transparent (allowing the user to simultaneously view the surrounding environment). The displays may attach to a pair of glasses (glasses-mounted display) or be attached to a form of headwear such as a hat or band.

Interfering Distraction: A distraction that interferes with a user's ability to interact with their mobile device. Such distractions may be passive or active. Examples of such distractions include traffic noise interfering with a mobile device's audio feedback or oncoming pedestrians that limit a user's ability to monitor the visual display of a mobile device.

Lab Evaluations: The evaluation of a mobile application that takes place in a laboratory. The

advantages of such evaluations include ease of controlling the environment and data capture. Disadvantages include difficulties in creating an appropriately realistic evaluation setting.

Passive Distraction: A distraction that may "put-off" the user but can be ignored. Examples of such distractions include billboards mounted on the side of buildings or the sound of traffic on a distant road.

Wearable Computer: A mobile computing device that the user wears rather than carries. The processing unit may be worn in a pouch on a belt or in a bag. Various interaction devices may be secreted around the user's body including: a handheld input device such as a trackball, a wrist mounted keyboard, or a head-mounted display. The advantages of such devices over handheld mobile devices include: faster access to the device, increased privacy, and hands-free access to data.

ENDNOTE

[1] C3 is middle C (261.63Hz)

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 910-926, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.26
Mobile E-Work to Support Regional and Rural Communities

Sirkka Heinonen
VTT Building and Transport, Finland

INTRODUCTION

Telework, or e-work as it is now more frequently called in Europe, means working outside one's regular workplace, utilising sophisticated ICT. E-work is an alternative form of organising work, a "love child" of the information society. E-work manifests itself in numerous forms and modes. These various solutions emerge as an evolutionary process along with the technological developments, economic pressures, and changes in socio-cultural patterns such as new information-age lifestyles (e.g., Castells, 1996; Heinonen, 2000). E-work can be carried out at home, in a telework centre or at any other facility. It can also be done as a mobile mode on a train, bus or some other vehicle, as well as at airports, railways and bus stations-in other words on the move from one place to another. Such mobile e-work is primarily increasing, owing to technological and social developments. ICT has become smaller in size, more portable and more efficient.

MOBILE E-WORK AS A SOCIAL INNOVATION

Mobile e-work is, however, not only a result of technological breakthroughs and penetration of ICT equipment in society. It is essentially a social innovation where various goals coincide. First, it may ease the stress of working life when the long commuting hours can be used for reducing the work load. Second, it is an instrument for employers to recruit people from a wider geographical area. Last, but not least, the implications of mobile e-work on regional development and rural communities must be taken into careful consideration (Heinonen, 2001). Along with various obvious benefits that are to be expected from e-work, prejudices persist and obstacles are still abundant (e.g., Anderson et al., 1996). Mobile e-work as a social innovation primarily awaits a breakthrough of the trust culture in working milieus.

The data available on the numbers of e-workers is somewhat unreliable and incomparable.

This is largely because various surveys measure e-workers' numbers using different criteria or definitions. Mobile e-work is a recently new phenomenon in the field of e-work in general. Therefore, it is particularly difficult to get statistical data on relevant numbers of mobile e-workers. Some figures can be given, though. The number of teleworkers varies from country to country within Europe. Scandinavia and Finland have the highest proportion of teleworkers out of the total number of white-collar workers, as a result of low-cost technologies, legislative frameworks, and corporate culture. IDC Research has forecast that the number of teleworkers in Europe will increase to over 28.8 million by 2005, up from 10 million in 2000. According to IDC, the mobile workers are defined as those who spend at least 20% of their working hours away from home, their main place of work, or both. There will be over 20.1 million mobile workers in Europe by 2005, up from 6.2 million in 2000 (Jüptner, 2001).

Various models and practices on e-work were developed, tested and recommended for communities and regional authorities in a recently completed three-year research project on Eco-Managed Introduction of Telework, carried out at VTT Building and Transport (Heinonen et al., 2004; Heinonen, 2001). The perspectives chosen were an analysis of environmental impacts, as well as a scrutiny of socio-cultural implications from various e-work contexts (for environmental impacts see also Arnfalk, 2002). A case study was included to experiment with mobile e-work in the Regional Council of Häme, Finland, which will be presented further in this article in more detail.

Mobile e-work can be seen as a means to bridge up the gap between regions. The general processes of centralisation and urbanisation are shifting emphasis on metropolitan areas and a few other urban growth areas. Other regions continue to lose their educated young brainpower to cities, and struggle with economic hardships. By promoting e-work and especially mobile e-work, the regions could have more balance in a socio-economic sense. The skilled labour could remain living in rural regions or semi-urban communities if their employers permitted e-work as a way to organise their work and commuting. In a traditional e-work case, an employee e-works one or two days per week at home or at a nearby telework centre, while on other days he or she commutes to the main office. Mobile e-work adds relevant benefits to the traditional e-working. In mobile e-work, trips to and from work can be used for working and thus the working hours at office will be cut down correspondingly.

In regional development, legislative efforts to diminish the digital divide between cities and rural areas could include, for example, tax deductions to the companies that permit mobile e-working, as well as to the employees who regularly practice mobile e-work.

E-WORK AS A TOKEN OF MOBILE LIFESTYLE?

In a survey by the Helsinki Metropolitan Area Council (YTV), the results showed that a typical e-worker in the Helsinki Metropolitan area is a highly educated and well-off male employee, younger than an average (YTV, 2001; Heinonen et al., 2004). He lives in a detached house, drives a personal car to the office and has a longer distance from home to the job than on average. Does this imply that an e-worker is prone to more mobility when trips to work are reduced? Or is the diminished commuting a quality-of-life target for a person who is already accustomed or obliged to undertake much travelling? In this survey, 3.6% of all the respondents claimed to have teleworked on the day the questionnaire was administered. Of the respondents active in working life, more than 5% teleworked at least one day per week and 13% replied to have teleworked occasionally during the last six months.

Mobile e-work is understandably more natural to persons with a mobile or nomadic lifestyle.

They are already accustomed to embracing continuous change of place and perhaps more easily concentrating on working on the move than those persons who consume their energy on the act of moving from place A to place B.

MONITORING MOBILE E-WORKING CONDITIONS

Mobile e-work was launched and tested in the case of Regional Council of Häme in 2002. Two employees working in Hämeenlinna, the oldest inland town in Finland, were selected to participate in the experiment. Their one-way commuting times were 1 hour 15 minutes and 1 hour 45 minutes by train, respectively. For three months, it was monitored by questionnaires and detailed diaries how well a daily commuting trip on train was suitable for e-working by using a portable computer and mobile telephone. The employees signed special contracts of e-work where it was agreed to compensate their working time on the train by reducing the normal working hours. They were also asked to carefully write down any advantages, obstacles, and observations that might be relevant for the outcome of mobile e-work.

The seat reservation for ICT seats was considered very important. Such seats were equipped with ICT plugs for portables, and they were isolated from other passengers by a glass wall. Thus, peace for e-working was guaranteed unless the other person sitting in such a compartment was talkative or otherwise disturbing the working conditions.

The main benefits from this experiment on mobile e-work were the increased efficiency of working, the decreased sense of stress, the enhanced working motivation, and the improved quality of life. The employees had more time to their families, hobbies and leisure time. They did not have any pressure to move their homes nearer to their office (Heinonen et al., 2004). The main issues that still need more developing as regards e-working conditions were too small table space, and occasionally too weak field access for mobile telephones. The data security also needs more thorough attention. Even the best data security procedures are not sufficient if someone simply robs the e-worker's computer "on the road."

FUTURE CHALLENGES

Besides the promising potential of mobile e-work in support of the development of regional communities there are some hindrances, risks and threats involved in the process of promoting mobile e-work on a wider scale. Mainly, they are concerned with data security or rising costs for companies. The cost of mobile e-work can be a major obstacle to the penetration of mobile e-work in society. However, in many cases the employer has already provided the worker with a laptop, mobile phone and Internet connection. Then practically no extra costs arise from mobile e-work. On the other hand, it must be borne in mind that mobile e-work is often considered as a serious risk for data security. This may create some additional costs, primarily regarding new software.

Transport companies are beginning to realise the potential of more revenues from the increasing number of mobile e-working passengers. This is especially the case if people who normally commute using a personal car, transfer to commuting by train or by bus.

In Finland and in other countries as well, mobile e-work could be promoted as a two-fold instrument for supporting regional and rural communities. Firstly, the metropolitan areas become congested and the quality of air, for example, is deteriorating while the traffic is increasing. Therefore, rural communities seem more and more attractive as a living environment, if only they could provide work opportunities. Mobile e-workers could choose to live in rural communities and maintain their work at the metropolitan or

other greater urban area. Without the possibility of mobile e-work, such commuting trips would be too burdensome. Secondly, in regional policy, attempts are made to decentralise governmental offices and units from the metropolitan area to regions located farther away. The resistance of employers has so far been adamant in almost all such decentralisation actions that have been accomplished. Mobile e-work would satisfy the needs of those employees who prefer to stay at their housing location in the metropolitan area, and still keep - thanks to a mobile e-work opportunity - their job that moved to a farther away location. It is a challenge for public authorities to harness mobile e-work as an instrument to support decentralisation.

Mobile e-work should be seen as one of the options inside the complete toolbox for e-work solutions. It has turned out that the best performing solutions are often combinations of various e-work patterns or solutions that are adaptable to modifications. Such changes in e-work models should in each case follow up the different life situations of the employee or development stages of the employing organisation. The ICT equipment and infrastructure should not ultimately determine the applications of mobile e-working. The main impetus should always come from the motivation of the mobile e-worker and his or her needs to adopt mobile e-work as a factor of social welfare. Various types of equipment can be experimented with, and eventually a tailor-made solution can be found.

In the future, mobile e-work will more and more become a facilitator of differentiating choices of housing location. Consequently, rural communities will have a better chance to attract mobile e-workers as new residents. Mobile e-work could thus bridge the digital divide between cities, more distant regions and rural communities.

REFERENCES

Anderson, H.K. et al. (1996). Distance working-Motives and barriers: Experimenting with the impact of distance working on transportation. *Proceedings of the 3rd European Assembly on Telework and New Ways of Working*, Vienna, 4-6 November (pp. 221-233).

Arnfalk, P. (2002). *Virtual mobility and pollution prevention: The emerging role of ICT based communication in organisations and its impact on travel*. Doctoral dissertation, May 2002. Lund University, The International Institute for Industrial Environmental Economics. IIEEE Dissertations 2002(1).

Castells, M. (1996). The rise of the network society. *The Information Age: Economy, Society and Culture* (Vol. I). Oxford: Blackwell Publishers.

Heinonen, S. (2000). Finland in the information society. Collapsing myths and revealing paradoxes. *Finnish Architectural Review*, (1), 32-36.

Heinonen, S. (2001). Eco-managed eWork as a new urban and regional strategy. Paper at the *Conference Telework2001, Session on Community Developments*, September 13, 2001, Finlandia Hall.

Heinonen, S., Huhdanmäki, A., Niskanen, S., & Kuosa, T. (2004). Eco-managed telework. The Finnish Environment 701, Helsinki (in Finnish with an English abstract). Retrieved from *http://www.vtt.fi/rte/projects/yki4/etatyoeng.htm*

Jüptner, O. (2001). Teleworking on the increase in Europe. Retrieved from *http//www.e-gateway.net*

YTV. (2001). *The present state of mobility.* Publications of the Helsinki Metropolitan Area Council B, (10), 24. (In Finnish).

KEY TERMS

Digital Divide: Digital divide means unequal access and use of data, information and communication. More specifically, digital divide means unequal access to ICT infrastructure or lacking skills in using it.

E-Work: E-work is work carried out outside one's normal workplace, at home, telework centre or satellite office, by using ICT equipment and infrastructure. A previously common term for this concept is telework or telecommuting (used especially in the USA).

Mobile E-Work: Mobile e-work is a concept referring to e-working in a mobile mode. The term mobile e-work can be defined as "e-work being done while commuting" or "e-working while commuting." To be precise, the overall concept of mobile e-work covers e-working on other trips as well, not just while commuting to and from the regular workplace. For example, e-working can be done on a person's way to meet clients, to attend conferences, etc., whenever "on the move" outside the regular office. The most general case of mobile e-work refers to a commuter's e-working. To make matters more complicated, "mobile" e-work also refers to situations that are stationary - such as sitting on a bench in an airport and teleworking with one's laptop. Mobility here means that you have moved away from your office, you are on the move outside the office, even if not moving in any vehicle at the moment of e-working.

Mobile Work: Mobile work means work that is carried out while moving. The work of bus drivers and pilots, as well as other personnel in various vehicles is mobile work. Mobile work can also be conducted by passengers. Mobile work is the opposite of work carried out as fixed on a given location.

Nomadic Lifestyle: Nomadic lifestyle is a mobile lifestyle where a person has adopted a high degree of mobility in his or her life. The person moves a lot and frequently between various places of housing, work, errands, hobbies, recreation, and socialising.

This work was previously published in Encyclopedia of Developing Regional Communities with Information and Communication Technology, edited by S. Marshall, W. Taylor and X. Yu, pp. 497-500, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.27
Mobile Phone and Autonomy

Theptawee Chokvasin
Suranaree University of Technology, Thailand

ABSTRACT

This chapter is to offer a critical study of what the human living condition would be like in a new era of hi-tech mobilization, especially the condition of self-government or autonomy, and how, in the Thai perspective, the condition affects culture. Habermas' analysis of individuation through socialization and Heidegger's question concerning technology and being are used in the study, and it is revealed that we are now confronted with a new technological condition of positioned individuals in the universe of communication through mobile phones. This situation surely will be realized in a world highly mobilized by the phenomenon of connectedness. This means that we are concerning ourselves with our concrete individuality for our self-expression in that universe. I offer an interpretation that we would hold this kind of individuality to be valuable because of an effect from technological thinking. In addition, comparing this view on individuality with Buddhism, I found that the view offered here is not similar to the Buddhist concept of self as a construction. I offer an argument to show that these concepts are basically different for ethical reasons; while the Buddhist concept still preserves the nobility of the moral agent (Buddhism, after all, is a religion and needs to concern itself with morality), the concrete individuality discussed here is considered only as an instrumental value in a world of hi-tech mobilization.

PROBLEM AND SIGNIFICANCE

Does the formulation of autonomy come from inside or outside an individual? From the investigation in the theory of subjectivity comes Habermas' individuation through socialization; one can achieve greater autonomy when he or she is engaged in a process of social integration to become socialized individuals (Habermas, 1992). His approach is sketched out in an intersubjective understanding that emerges from communicative action when individuals enter a public sphere to share their volition or opinion in order to make a reasoned agreement that later becomes a so-called universal law (Habermas, 1990). It is interpreted that, while Kantian pure reason inside us is the source of autonomy (as interpreted in Guyer, 2003), Habermas' theory of communication

sheds light on the question by suggesting that an outer source, the process of socialization by communication, is the case.

If we accept for the sake of argument that Habermas' theory of communication is suitable to explain autonomy, one question still remains, particularly in our time of modernity: Can the same explanation be applied simultaneously to communication on mobile phones, especially the hi-tech ones? Habermas' theory primarily aims at our communication when we are face-to-face with those with whom we are communicating, but we did not see how a situation will be realized when the communication occurs between a distance, not a face-to-face one. The term *hi-tech mobile phones* that I used here means a kind of cellular phone that can be a credit card, Internet connection, e-mail port, voicemail junction, and so forth, according to the usage of Myerson (2001). We may imagine that it will be like a pocket personal computer. This concept of a hi-tech mobile phone somehow would be realized in the future. It sticks with its owner everywhere he or she goes, even in water (if waterproof and not easily broken). However, its dominant characteristic is that it is an important item of personal belongings. Its owner is the only one to hold and use it. If we were to routinely share the mobile phone with another person, there would be no difference at all between it and a public telephone or a house telephone. Therefore, a real mobile phone has a characteristic of being able to identify its owner in order for it to become the most efficient way to communicate with the one with whom we are trying to connect in such a way that is not possible when one receives (or sends) a call from (to) an unexpected person. Certainly, the communication on mobile phones is not a face-to-face one; we are not in a position of being body-to-body with him or her with whom we are talking. Sometime in the future, there might be a great development in mobile phone technology so we can see faces on the screens of mobile phones, but still, we normally do not consider this a face-to-face communication.

My topic, "Mobile Phone and Autonomy," may lead someone to think that the mobile phone itself becomes a thing that keeps us always in control. It sticks with us all the time; we have to use it in our daily lives, and we find it so indispensable that we will never reject it. Therefore, it is a channel through which another person can reach us directly and control us so that we behave according to the rules of social conduct. We may be afraid of being monitored by an online e-policeman through the channel of our mobile phone, and that feeling would prevent us from doing something illegal. In that kind of social management, everywhere we went surely would be known by the police, so if we did something against the rules, we could be abruptly caught, or we could be tracked down by the system in our mobile phone. Or we would be shocked by a dangerous flow of electricity caused by a police officer through the battery of the mobile phone to prevent us from escaping the scene. However, even though it seems that those situations might be possible in the near future or that someone might want to say that a side effect of using a personal mobile phone is a utopian society in which people dare not commit a crime, I do not have any intention in this chapter to talk about these surveillance roles of the mobile phone. As the highest status of moral development, autonomy or self rule of a moral agent is not explained as fear of being punished by the law. Autonomy is understood as a concept of self-expression as an agent who has his or her own freedom and intention to do according to his or her volition for his or her end for himself or herself. Moreover, the concept of autonomy always goes along with the concept of rationality rather than with emotion or any stimulus that does not stem from pure reason. I consider the concept of autonomy only in a dimension that involves rationality. It is possible that in the mobilization era, people will be aware of being

monitored and controlled, and they will have to conduct themselves strictly according to the law and social rules, but I think that this consequence is only the tip of the iceberg. There seems to be a hidden and more important phenomenon.

OBJECTIVES

In this chapter, I would like to accomplish two things. First, I would like to study how hi-tech mobile phones affect our self-governance (autonomy), depending largely on Habermas' and Heidegger's theories of communication. Second, I analyze whether the mentioned theories lead one to a conclusion that the mobile phone is strengthening one of the crucial aspects of autonomy: our self-responsibility or self-positioning in the universe (of communication).

WHAT WILL YOU BECOME WHEN YOU ARE IN COMMUNICATION? BECOME CONNECTED

Mobile phones make us closer. How should we understand the meaning of this sentence, especially its deep meaning lying at the heart of communicative acts on mobile phone? One may say that we are connected with one another, and we ourselves want to make that connection. Mobile phones are only tools that obliterate any distance between us and those with whom we want to communicate. Anyway, if we look at this phenomenon carefully and in a very different way, we may encounter something striking. Is it the case that the phenomenon of communication itself is using mobile phones as tools to tie our interlocutor and us in the line of communication? The communication cannot exist without us and the one with whom we are talking, and certainly without the tool functioning as a medium; therefore, it needs all of these factors in order to exist. This kind of technology really brings us closer.

Receiving a call from someone or, in other words, being connected can be considered a phenomenon that occurs. A question then arises: What is a characteristic of this phenomenon? Following Heidegger's (1977) phenomenological assumption that we should begin to answer from the first thing on our minds after a question is asked, perhaps the question could be answered that the most distinct characteristic of communication is connection. Only when at least two units of communication (the speaker and the receiver) are connected (i.e., related in some context of discourse) is the communication accomplished. If the communication analysis of Habermas is correct, it may be said that the relationship is the source of our own self-understanding as a unit in the accomplished communication. But Heidegger may have something more to say; that is, self-understanding is a phenomenon that already presumes some ontological status of the one who understands his or her own being in that sphere. This entity who understands his or her own being and can take being as an issue is what Heidegger calls Dasein ("being there"). To take being as an issue is a Heideggerian term that refers to Dasein's unique ability to conceive the *being of beings*; in other words, to have a preontological understanding of beings. Because this self-understanding has its source from communication, it means that the self becomes known as being there in that connection through communication. The self is not lying somewhere outside or beyond that connection in order for it to enter inside that sphere of connection with the other self. He or she learns of the being of himself or herself only when he or she is connected in the sphere of communication with his or her interlocutor (even when his or her interlocutor is himself or herself). I would call this phenomenon that occurs in communication *connectedness.* Connectedness means that that a person understands his or her own self or his or her own existence through the connection with the other person in the sphere of accomplished communication. Accomplished communication

means any talks about any thing that the speaker and his or her interlocutor have a consensus in understanding. Therefore, it may be considered that this meaning of communication is rather close to what Habermas called *meaning* and *communicative rationality* that constitute a mutual understanding between the speaker and the interlocutor (Habermas, 1979). Consequently, from a successful communication through which we are connected to the other person in conversation, the sphere of communication is where we infer our own existence from that person's response in a communicative action (we listen to him or her, or he or she hears us). Communication on telephone brings people near, even though there is a long distance between their physical bodies, and especially on mobile phones, we could say that there is always nearness between them.

What is this nearness, anyway? I would like to offer an investigation of the meaning by comparing communication on mobile phones with other forms of communication. Imagine a case in which we want to talk to a prime minister; it is possible that we have at least some chance of having a conversation with him or her directly. Nevertheless, the possibility is not so high, because we might be obstructed by the prime minister's staff. The next day, we write a letter of inquiry and send it by e-mail; however, it is possible that the answer we receive back is not directly from the prime minister but from an assistant. Even when we try to connect to the prime minister on a workplace telephone line and ask to talk to him or her directly, we could be obstructed again by an assistant, or an assistant might persuade us that we should leave our message with him or her or use him or her as a medium to talk to the prime minister (which means that the prime minister actually was sitting nearby but did not want to accept the line directly, so we actually receive the prime minister's answer but not directly). However, the situation will be totally different if we have a communication with the prime minister on a mobile phone. If our call is accepted, we certainly will be sure that the one who accepts the call is the prime minister (in the case that a personal mobile phone is allowed to be used only by its owner). If that call is outgoing from our own mobile phone, then the prime minister also will know who wants to talk to him or her. In any case, whatever answer the prime minister is telling us via the connection on a mobile phone is not an interesting characteristic to investigate here, because I find two more interesting characteristics to consider; namely, that the mobile phone is a source of constructing a sphere in which a real individual meets a real individual, and it is a hi-tech device to guarantee the connectedness between those individuals. In what follows, I will explain why these two characteristics are the two aspects that answer what nearness is.

Based on Habermas' critique of communicative rationality that can be found in a successful communication, I infer that there also may be some kind of rationality in a communication on a mobile phone. If this is the case, communication on a mobile phone will be a source of the smallest sphere in which rationality can be found. That is because this type of communication typically requires only one speaker and one hearer on the line. They are being connected; in other words, getting near each other. Being near, which occurs in that type of communication, can be understood clearly from our experience using a mobile phone. We experience more convenience reaching the one with whom we are trying to connect when using a mobile phone rather than when using nonmobile devices. This phenomenon of being near could be called the phenomenon in which two real individuals can talk to each other without being intruded by a third party and without a medium messenger. The situation in which we can talk directly to a prime minister, as mentioned previously, is of this type—a real individual encounters a real individual. So I am inclined to say that this situation always occurs when people are connected by their mobile phones, but the possibility would be lessened if the communication were not via a

mobile device. Mobile phones make human beings get closer as two real individuals in connection with each other. From here, I have some further questions: Why is this communication technology aimed at allowing this phenomenon to reveal itself? What is the value of this phenomenon? In other words, why is a real individual so necessary to develop from the hi-tech tool of communication? I will elaborate my argument further in the next section to show that, in the mobilization era, an instrumental value of the individuality is indispensable. Communication technology needs it for its own best efficiency.

Imagine two people—Mr. Black and Mr. Jones. Black has owed Jones a sum of money for a long period of time, and today, Jones needs his money back. Jones tries to find Black by visiting him at his company to have a face-to-face communication about the debt. It is possible that Black can see from afar that Jones is coming. Black goes straight to a doorway of a fire exit and hides himself behind the door, peeping through a little space between the door and the wall to see what Jones is going to do. Jones cannot see that his cunning debtor has run away, but could he be sure that Black did not want to see him and already escaped? It does not seem so. Could he be sure to believe in Black as before and console himself that this is because Black is not at work today? Not either. From this event, I conclude that between Black and Jones there is no connection. Moreover, there is no connectedness either, because Jones cannot be certain to infer any meaning or message from that disconnection.

Two hours later, Jones wonders whether Black is at his house, so he dials Black's house phone. His call is not received at all. Again, Jones cannot be certain to infer that this is because Black is not there. Black might guess correctly that the call was from Jones and did not want to receive it. Jones cannot be certain that Black tries to evade talking with him about the debt, because it could be the case that Black was then actually not at his house. From this event, I again conclude that there is no connection and connectedness, because Jones cannot be sure to infer any meaning from that disconnection.

However, would all these problems still happen in the case of communication on a mobile phone? If Jones uses his mobile phone to connect with Black's and finds that his call is not accepted at all, there is almost a zero probability that Jones would infer another meaning except that Black is trying to avoid meeting him or does not pay attention to Jones' needs. If a great development of mobile phone technology makes it so convenient that we could accept every call at any time and any place, except when we are dead, this would mean that when we live in this world, we are always in connectedness with other people in the sphere of communication. In the latter case between Black and Jones, even though there is no connection because of Black's not receiving the call, there already is connectedness. Jones can infer that Black pays no heed to him from the very act of Black's not receiving the call. There is connectedness without successful connection, but not without meaning, so we may conclude that this time there is an accomplished communication. From this investigation, I conclude that a mobile phone brings us near, because it always guarantees connectedness, and it is the most suitable channel for two individuals to have an immediately direct conversation. Even through the act of not receiving our calls, we can at least infer that our expected recipient is loath to do so. In a face-to-face communication or on a nonmobile phone, there is noticeably no guarantee of connectedness, but in mobile phone communication, there is a guarantee that none of the real individuals can ever escape from connectedness.

INDIVIDUATION, AUTONOMY, AND RATIONALITY

The next questions are: Why does this modern technology of communication need to control

an individual to be present in the sphere of communicative world? Why is there such a strong need for such exact identification of individuals? Communication on the mobile phone is aimed so much at the individual being that it seems to be itself a cause of that individuation. But could we then conclude that the mobile phone is a tool of individuation? Some might not say so, because the mobile phone is considered only as a tool of convenient communication that was invented by human beings. We have already individuated a person as an individual being long before the tool was invented, so it does not seem to make any sense to argue that individuals are individuated through mobile phones. Or could we say that the mobile phone is a tool of identification? Some might not agree either, because it would be more reasonable to explain that the mobile phone is only a tool that stores personal information, and to do so, we must identify in the first place which mobile phone is our own in order to keep our information in it. In sum, the process of individuation is not causally relevant to the invention of mobile phones, whether as cause or effect. It is only for the sake of our convenience that direct communication is done best via a mobile phone, and the process of individuation is only a necessary condition for that.

I find, however, that the previous summary is too big a leap to conclusion. The claim underlying that summary is that a communication device is only a tool for a more convenient orientation. There are some questions that those who argue along this line did not ask; for example, if a mobile phone is only a tool, then why is this technology aimed so much at the individuality of its owner, and why does it have a role of controlling its owner to be always within the sphere of connectedness, as shown in the previous analysis?

I would like to offer an argument to show that the role of technological thinking is to control the individual person as a positioned individual in the communicative sphere. This is all because of our need for efficiency in communication. I do not say that the communication technology is about to have an influence on us without our consent to it. This is not a matter of willing to be or not willing to be in that influence. It is a matter of necessity of the best communicative efficiency that we have to be individuated from each other in order to be in an exact position of receiving or sending a call. If we consider its efficiency as a desired end of communication, then mobile phone technology is here to help us reach that end. Everything that is involved in bringing about that efficiency is provided by a systematized technology of the mobile phone, and all of them are considered only in terms of resource preserving for the technology, or, in the Heideggerian term, a standing-reserve. A human being in the mobilization era would rather be revealed as a unit provided for technological efficiency. The best technology of communication reveals its own essence--Enframing, as Heidegger (1977) called it. If this is true, I am inclined to say that human beings are the resource of it; not the concept of a human being as a living person or an agent capable of intentional action, but a human being as revealed in its essence, which is its being there as a positioned individual. In other words, in the best communication technology, the position in a communicative sphere of human beings is resource-preserving for its own efficiency. To consider the mobile phone technology only as tool is not a way to see its underlying essence, because there is another aspect of it that is not merely a tool. Ironically, we ourselves are considered vital tools playing our exact individual roles in that world of the most efficient communication.

Again, the mobile phone always guarantees that the connectedness of meaning occurred in communication or, at least, some meaning that could be inferred from it. No, we are not influenced to receive any calls, especially those to which we are loath to do so. We are not influenced to receive any connection via our hi-tech mobile phone with a reason of efficiency in our communication with each other. But are we always in connectedness? In other words, are we always "there" to be in-

ferred some meaning from this communicative sphere? The answer is yes. Surely, in our common sense, we feel that we are not compelled to do the connection, but it is necessary for us to be in the connectedness.

From here, I am in a position to offer a critique of mobile phone technology and its impact on our autonomy or our self-directing. I have presented an argument purporting to show that the technology is really there to make our load much lighter, even in our nearest task of self-directing. Then what is the significance of this situation that has an impact on autonomy at all? As I have said, if, in the new age of modern technology, human beings are necessarily tied to the connectedness, then a question arises: Can we consider that human beings have freedom to choose their own lives as they prefer? If we are unavoidably tied to the sphere of communication in which everyone is in connectedness, how are we to speak of our freedom? Here, I will use a concept of autonomy from Habermas to clarify this problem. An interpretation of the Habermasian concept of autonomy by Warren (1995) is that Habermas considers autonomy a normative ideal with six implications, shown in Habermas' works. These are as follows:

- Self-identity (one can locate oneself in terms of biographical projections)
- Capacities of agencies and origination
- Capacity of having freedom (one can distance oneself from social context)
- Capacity for critical judgment (in Habermas, it means an individual's capacity to participate in communication with reason)
- Capacity of reciprocal recognitions of the identities of speaking subjects
- Some measure of responsibility (by giving reasons for behaviors to others)

For the sake of argument, I consider here that all of these implications are interwined with the concept of freedom of choosing the best lifestyle for the individual who lives that life of his or her own. Nevertheless, the freedom is developed from a communicative rationality when one participates in the social sphere to share his or her own judgment of the lifestyle, defend its validity claims in the argumentation, and, of course, be morally responsible for his or her own self conduct. But wait. Can this kind of freedom be realized in a sphere of hi-tech mobilization age? If human beings are about to comport themselves in a mobilization way of life that involves one-dimensional thinking of the efficiency of technology, then how can we have that freedom?

This question is explained in a book by Myerson (2001), *Heidegger, Habermas and the Mobile Phone*. Myerson suggests that Habermas' theory will not be so smooth when applied to the communication in the mobilization world. In that world, the idea of communication is changed dramatically. What is being communicated—message or meaning? asked Myerson. He says that communication on a mobile phone is transformed into a one-dimensional version of meaning, which is the message. Anyway, what Habermas needs for his theory to work well is that people should be in a suitable circumstance to provide reasons for their expressions; therefore the message version of communication is not a good one in Habermas' view (Myerson, 2001). The result is that a mobilization world is not a sphere in which the source of rationality can reside. However, rationality is one of the crucial aspects of autonomy. Then how can autonomy be developed, if a mobilization world is not a place for rationality at all? Unfortunately, Myerson has no answer for this, but it is one that I am trying to figure out. Is autonomy possible in that world? If it is, then what would it be like? Would autonomy in this case be in a similar formulation, as in Habermas or Kant? I am inclined to say no.

How can Habermas compromise the thesis that autonomy is developed from social participation with communicative rationality in a public sphere with my thesis that something, which can

be called self-positioning, is developed from connectedness in mobile phone communication? If the compromise does not work, Habermas may have to accept that his theory cannot explain some aspects of communication. In a public sphere, face-to-face communication can be represented fully in processes of engaging in reasoning or in a critical examination of self and others, and one can be recognized as an individual from the processes. But one can be recognized as an individual via mobile phone communication, for I showed that communication on the mobile phone is where a real individual meets a real individual and where there is no other but only those individuals connecting on the line. When the processes come to terms with the development of autonomy, must we limit the autonomy development only to the extent of face-to-face communications in a public sphere? Or do we have to say that we are seeing a new consequence from this hi-tech communication that has an impact on autonomy development? What we are seeing is that the mobile phone is bringing about a new phenomenon of sphere revealing, the smallest sphere in which communicative rationality can be developed. This sphere is where a real individual meets another real individual to make an intelligible communication on the line. However, what kind of rationality is there on that sphere? The individuals are able to have intelligible communication with each other on their mobile phones, as they do when they participate in a face-to-face communication. So must there be some kind of rationality in that sphere? I would venture to say yes.

We know that the mobile phone is the best way to communicate efficiently. This new technology gradually is necessitating itself within our modern age. People can have a conversation about politics, for example, face-to-face with one another with communicative rationality. They also can talk about the same political story via mobile phones with communicative rationality. But there surely is something more in this mobile phone talking. This is because we only can make the communication best via mobile phones to feel like being together when we really are some distance from each other. The mobile phone proves itself to be the only way to guarantee efficiency in communication; therefore, it has its own process of rationalization of itself. The process is that it guarantees that the one with whom we are to connect is exactly the one with whom we want to connect, and that connection always transmits some significance, even when the connection is rejected or that it always guarantees the connectedness. What this means is that, despite the fact that we are in conversation with our partners and epitomize communicative rationality when we reach understanding with each other, we also are guaranteed by the mobile phone that our interlocutor is the individual exactly and always positioned in a communicative sphere in order for us to infer some meaning from our connection, and vice versa from the other person to us. In this system, we cannot escape from connectedness, because we already are positioned in the sphere. The technological system always sees us as an individual unit that is always available as a mechanic of the most efficient communication. As Habermas has said about communication in a social sphere, that we could infer our individuality as well as others' from the sphere, our being individual means very much to the technological systematizing that reveals itself very clearly in communication technology.

SELF-EXTENSION AND SELF-EXPRESSION: CONCRETE VERSION OF AUTONOMY THROUGH THE MOBILE PHONE

In the other aspect of the mobile phone, we know that it is a tool for communication. It is a technological tool that enhances our capacity to talk and hear; in other words, communication. This nature of the mobile phone is one of the technological characteristics—self-extension.

This characteristic of selfhood is a crucial aspect of autonomy; it is our self-governance or self-rules in accord with our social being with others (Berofsky, 2003). From this, I have a question: What is it about our autonomy when our selves are not restricted to our bodies, minds, or living conditions? Imagine a mobilization world in which there is a need for the best efficiency in communication; every single person must have a personal mobile phone. Inevitably, our mobile phone will be with us everywhere and every time when we are not in a face-to-face communication in order for us to remain in this social universe of communication. This is a universe in which the mobile phone is indispensable. Each of us is connected to our own mobile phone; it is like one of our organs. Surely, the concept of autonomy is involved with self-identity; therefore, questions arise. What will become of autonomy when self-identity is enhanced that way? Will the duty of our self-governance be left to the mobile phone because we begin to feel being monitored by anyone in this universe of communication who also knows where we are by tracking the position of the phone? If the mobile phone is highlighting this way of our self-governance, what will be the kind of autonomy from the influence of communication in this modernity? Is that still a so-called freedom? These are the other questions I try to answer.

I believe that a critique of technology by Heidegger (1977), his superb writing of *Being and Time* (1962), and a reinterpretation from Habermas' theory could provide an interesting answer. I do not agree with Myerson that Habermas' theory does not provide at all a plausible account for communication on mobile phones. If this relation of our being and the mobile phone is an inconvertible one, this means that the phone contains a technological essence in the Heideggerian sense—Enframing. I already have shown why I am inclined to think that the essence will rationalize the necessity of itself. If this is the case, some aspect of autonomy surely will be revealed, though not one that Habermas expects as a normative ideal or as freedom. From here, I offer the last argument developed from my own terminology: the abstract and the concrete version of autonomy and their difference in an ethical aspect of human nobility. I will show that the old concept of autonomy found in philosophers or in Buddhist culture is an abstract one and preserves human nobility. In contrast, the new concept of autonomy derived from a reflection of mobile phone technology is a concrete one that lessens human nobility and values the positioned individuality as indispensable in a control for technological systematization.

We may imagine people in the mobilization era having their own mobile phones. When they are in communication with each other, they become connected. Are they only connected to their interlocutor? No. It is also that sphere of communication with which they are connected. Besides communication with a friend, the speaker also communicates with the world. There is also an activity of reporting to the world that he or she is there in the sphere in which he or she and the friend intersubjectively guarantee each other as being in the world. This means that, at least, there is knowledge of their existence as unique and irreplaceable individuals in that sphere. Anyway, are our friends initiating themselves to have this knowledge as individual being by themselves? No. Human beings cannot conceive themselves as individuals solely standing without the world. But the answer is the communication technology—the mobile phone. It constitutes a way of being individuals as valuable for the most efficient characteristic in communication technology. It is valuable because anyone in the mobilization era must preserve it for himself or herself in order to continue one's own way of being in the sphere. From here, I have an interpretation developed from Heidegger's concept of the two interdependent ontological statuses of entities: presence-at-hand and readiness-to-hand (Heidegger, 1962). The characteristic I would

like to interpret from Heidegger is not about the theoretical and practical aspects of entities like those that other philosophers have done (Ihde, 1979) but is about the ontological characteristic of readiness-to-hand as unobtrusiveness (Dreyfus, 1991). Heidegger says that a hammer is always at work as the most efficient tool, if and only if it is an entity as or working as a hammer without our awareness that it is in the hand. But if it is broken, we have to look carefully at what was wrong with it. Nevertheless, when we hold it in our hand and conceive it as a hammer, we do not call it a hammer merely because we want to. Heidegger points out that because it is in a context of being an equipment for a carpenter that we conceive it as one of necessary tools in that context. A hammer has it own status as an individual entity, if and only if it is in a broader context than itself with the other tools. Moreover, the context itself never reveals itself to us; it is always unobtrusive, but it is where the hammer is conceived as an entity. In Harman (2002), the readiness-to-hand is called *tool-being*, because it is interpreted in his work that every aspect in Heidegger's work is about theory of objects, which are tool-beings. I use that interpretation for my purpose here to offer an argument that the individuality of an individual person is considered valuable in the context of the best communication when two persons do not have a face-to-face communication. Consequently, the two persons can claim that they are individuals connecting on the line, because the hi-tech communication creates a context for that communication to be possible. The exact positions of the two persons are only tools for the efficiency of communication, and the individuality is a valuable characteristic derived from the context. The value I mention here is a kind of instrumental value, the value that human beings would have for the era of hi-tech mobilization as tool-beings.

If this speculation turns out to be true, then in the new era of hi-tech mobilization, each of us is always seen as an individual unit in the context of efficient communication. The individuality is valued for every unit; therefore, we must be responsible to preserve our own individuality. (Always keep your mobile phone with you and never let anyone use yours). This positioned individual is considered concrete, because he or she always reports himself or herself to the sphere of communication. Furthermore, in terms of responsibility, individuals have to preserve the very individuality of themselves, I have to conclude that there also is a crucial aspect of autonomy in that activity. This is concrete autonomy. It means that when everyone has their own mobile phone and are positioned individuals in the sphere of communication, then at least one obligation they have is to promise to preserve their own positions in the sphere. This kind of promise is not merely some kind of thinking supported by pure reason in a person's mind without being seen by anyone else outside. The very act of position preserving when people are connected on mobile phones is the right act to show that this obligation is kept and not violated.

Therefore, this concrete version of autonomy can be defined as a self-responsibility in the mobilization era. Your position must be revealed in the sphere, and you must keep it for your own sake of living. The old version of autonomy is aimed at self-government and self-expression in a moral understanding of agency. We can find this concept in Kant; autonomy is a self-government according to the universal laws derived from pure reason. We accept those universal laws and comport ourselves to them, because they are from our own pure reason. This concept also can be found in Habermas. Nevertheless, universal laws are not derived from pure reason but from communicative rationality, according to Habermas. We have to comport ourselves to them, because they are from the consensus we had with the other people who are involved in a public sphere making a reasonable agreement. However, this version of autonomy cannot be seen by anyone outside. We just believe in each other that we have a promise

to keep according to what agreement we have. This is not a version of autonomy that we can see from outside a self-responsibility of a moral agent except when his or her behavior is expressed. I call this old version of autonomy *abstract* because of the reason I mention. Everybody who came out of the public sphere is credited as a citizen in a political system and knows well how to behave according to the agreement he or she made. To violate the agreement means that he or she is no longer credited as a citizen; therefore, in this version of autonomy, what it means is that the moral agent is preserving, in my word, his or her own status of nobility. He or she can choose his or her own style of living, as long as that choice of living is not against the agreement. His or her freedom also is derived from the public sphere.

Surprisingly, besides the notion of self-government, the concept of autonomy in many philosophies is involved with the notion of human nobility. Even in Thai culture, which derives the concept of autonomy mainly from Buddhism, to talk about autonomy consists of considering how someone comports himself or herself to the right course of living. In Phra Dhammapitaka (1995), there is a concept of *Attasammapanidhi*, which means the characteristic of a person who can set himself or herself on the right course, in the right direction of self-guidance or perfect self-adjustment. This moral characteristic is among the 38 highest blessings (Mangala, 38). Also, in a set of seven qualities of a good man (Sappurisa-dhamma, 7), there is a characteristic, *Atthaññuta,* which means to know the meaning and to know the purpose and the consequence of *dhamma* (Phra Dhammapitaka, 1995). From those characteristics, we have to understand more of a crucial aspect of the Buddhist concept of the self. The doctrine of *anatta* is the crucial doctrine, which encourages the Buddhist practitioner to detach himself or herself from clinging to his or her own individual self. In reality, the self has no existence of its own, because it is a construction from many causes and conditions. When those causes and conditions cannot engage with one another, the self can persist no more. So, to understand what the Buddhist teachings are telling us about self-adjustment or self-government is not to understand in such a way that those characteristics belong to a self or that there is a persistent person who acts as their bearer; it is to understand the matter in such a way that those characteristics are occurrences in a moment of conceiving by a person's mind (or *citta*). Autonomy, in Buddhist concept, is some noble characteristic that occurs with a mind that conceives *dhamma*. It is reflected in those noble persons who can adjust themselves to the right course of living. In order for other people to know that someone has autonomy, they also must be in that right course of living. Therefore, from my previous definition, I have to conclude that the version of autonomy in Buddhist concept is an abstract version involved with nobility of the mind. It can be conceived only in the minds of those who know the *dhamma*, or the Buddhist teachings, not a characteristic that we can see from outside of them.

Moreover, there also is a kind of freedom in the Buddhist concept of autonomy. In Phra Dammapitaka (1999), there is a notion of *Attanovatti*, which means the freedom of the mind that conceives the *dhamma* of three characteristics of existence: impermanence (*Anitya*), suffering (*Duhkha*) and not-self (*Anatta*). The mind of those who thoroughly understand the teaching will never cling to the impermanence of existence. This freedom (*Attanovatti*) of mind will guarantee that the person autonomously will never commit any immoral act. But if anyone who is still clinging to the existence of, say, wealth or social reputation, which actually are impermanent, he or she will never have that freedom, because his or her mind is still tied to those illusions, and he or she still has a chance of doing something immoral.

In sum, the Buddhist concept of autonomy has something involved with a nobility status of a person. We may have to conclude that the old concept of autonomy is an abstract version and

that we can find it in Kantian or Habermasian philosophies; namely, that nobility of person is still preserved in a political status as well as in Buddhist philosophy—that nobility of person can be found in the *dhamma*, or the teachings. Nevertheless, all of these philosophies conceive autonomy with a notion of freedom. Finally, what is left to say about these versions of abstract autonomy when one compares them with the concrete version that we have obtained from our reflective investigation of the mobile phone and connectedness? I will offer only a description of a near future that is predicted to occur as an answer to the question.

CONCLUSION

Based on the arguments I have offered so far, I am inclined to conclude that the mobile phone has its own way of developing autonomy in people, because it has its own rationalization in a technological system. The version of autonomy derived from it is a new way of self-expression in the sphere of communication. This is the case, because the self-identity of the mobile phone's owner is extended to the tool of communication; therefore, the position of any individual owner always is expressed by the guarantee of connectedness. Moreover, as the very being of a positioned individual is so valuable in the sphere in which people are responsible to preserve it, any individual person is considered only an instrument in the most efficient mobilization system. Coyne (1995) points out that this is a matter of our being controlled by hi-tech communication, which reveals the clearest essence of technology that Heidegger calls Enframing. However, here, I have more to say about it, for it is a new concept of autonomy that still involves freedom; this technology does not exclude all aspects of freedom. We still feel free to receive or not to receive connections from other people. But in the aspect of our status in the mobilization era, are we really free to escape from being a positioned individual? No. Are we free to leave the obligation of preserving our own individuality that guarantees the efficiency of hi-tech communication? No. It is just that in this time of modernity, we have a new way of thinking how it is to be an individual person. In the old concepts, we have political or ethical ways of thinking about it, and we have preserved human nobility in those ways of thinking. We find our freedom from ethical characteristics of life. In contrast, the new concept from the mobile phone replaces the old value with a new one: the instrumental value of individuality. In the mobilization era, people with their own mobile phones are the best mechanics that guarantee the best efficiency in the communication system.

People in that time of modernity may have very good behavior. It seems that all the time they respect the nobility of the other and of himself or herself. But in reality, they have it because they have to preserve their own positions and not to misbehave, for it could lessen the best technological efficiency in their societies. They do not do good deeds because they feel that it is morally good, but rather because they feel that it is technologically good instead.

REFERENCES

Berofsky, B. (2003). Identification, the self and autonomy. *Social Philosophy & Policy*, *20*(2), 199-220.

Coyne, R. (1995). *Designing information technology in the postmodern age: From method to metaphor.* Cambridge, MA: The MIT Press.

Dreyfus, H. L. (1991). *Being-in-the-world: A commentary on Heidegger's Being and Time, division I.* Cambridge, MA: The MIT Press.

Guyer, P. (2003). Kant on the theory and practice of autonomy. *Social Philosophy & Policy, 20*(2), 70-98.

Habermas, J. (1979). *Communication and the evolution of society* (T. McCarthy, trans.). Boston: Beacon Press.

Habermas, J. (1990). *Moral consciousness and communicative action* (C. Lenhardt & S. W. Nicholsen, trans.). Cambridge, MA: Polity Press.

Habermas, J. (1992). *Postmetaphysical thinking: Philosophical essays* (W. M. Hohengarten, trans.). Cambridge, MA: The MIT Press.

Harman, G. (2002). *Tool-being: Heidegger and the metaphysics of objects.* Chicago: Open Court.

Heidegger, M. (1962). *Being and time* (J. Macquarrie, & E. Robinson, trans.). New York: Harper & Row.

Heidegger, M. (1977). *The question concerning technology and other essays* (W. Lovitt, trans.). New York: Harper & Row.

Ihde, D. (1979). *Technics and praxis.* Dordrecht: D. Reidel.

Myerson, G. (2001). *Heidegger, Habermas and the mobile phone.* Cambridge, MA: Icon Books.

Phra Dhammapitaka (Payutto, P. A.). (1995). *Dictionary of Buddhism* (8th ed.). Bangkok: Mahachulalongkorn University Press.

Phra Dhammapitaka (Payutto, P. A.). (1999). *Buddhadhamma* (8th ed.). Bangkok: Mahachulalongkorn University Press.

Warren, M.E. (1995). The self in discursive democracy. In S. K. White (Ed.), *The Cambridge companion to Habermas* (pp. 167-200). Cambridge, MA: Cambridge University Press.

This work was previously published in Information Technology Ethics: Cultural Perspectives, edited by S. Hongladarom and C. Ess, pp. 68-80, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.28
The Sociotechnical Nature of Mobile Computing Work:
Evidence from a Study of Policing in the United States

Steve Sawyer
The Pennsylvania State University, USA

Andrea Tapia
The Pennsylvania State University, USA

ABSTRACT

In this article we discuss the sociotechnical nature of mobile computing as used by three policing agencies within the United States. Mobile devices, access, and service was provided via a third-generation wireless network to a focal application, Pennsylvania's ***J***ustice ***NET***work (JNET), a secure Web-based portal connecting authorized users to a set of 23 federated criminal justice and law enforcement databases via a query-based interface. In this study we conceptualize mobility and policing as a sociotechnical ensemble that builds on the social-shaping of technology perspective and the tradition of sociotechnical theorizing, focusing on the co-design of work practices and technologies to support work. Drawing from the social informatics tradition, we turn a critical, empirical, and contextual lens on the practices of mobility and work. Our analysis of the data leads us to observing that the social and the technical are still considered separately in the context of mobile work. This simple view of social and technical as related, but distinct, often leads to problems with collecting and interpreting evidence of ICT-based systems' design and use. We further note that this over-simplification of sociotechnical action is likely to continue unless more viable analytic approaches are developed and the assumptions of the current techno-determinist approaches are challenged more explicitly.

INTRODUCTION

One of the many alluring possibilities of mobile computing is that people will be able to access computing resources while on the move. In organizational contexts, mobile computing (or mobility as we refer to it here) engenders scenarios of increased productivity through instant access to computing resources at any time from anywhere. Here we explore the sociotechnical nature of this envisioned future for mobility. In the social informatics tradition, we turn a critical, empirical, and contextual lens on the practices of mobility (Kling, 1999, 2000; Sawyer & Eschenfelder, 2002).

We first explain why policing is an appropriate domain in which to explore mobility and work. We then conceptualize mobility as a sociotechnical ensemble. In subsequent sections we lay out the research, outline our data collection and analysis, and then present and discuss seven findings. We conclude by focusing on implications regarding sociotechnical analysis.

Why Focus on Policing?

There are at least three reasons why policing is an appropriate domain for studying mobility. First, police officers' work has always been highly mobile. It is also knowledge-intensive and pervasive (for more on this, see Manning, 2003). Second, there continues to be great interest in using ICT to better support police officers' information needs. For example Manning (1996), in his study of cellular phone take-up among police, reported on the long-standing disparity between police officers' information needs and the abilities of the ICT used to provide them that information.[1] Third, policing and criminal justice have long been a focus of academic study; that provides us with extensive literature on police work, the social norms, informal and formal organizational governance mechanisms, and an understanding of their technological basis (see Manning, 1977; Klockars & Mastrofski, 1991; Manning, 2003)[2].

Current research findings provide contrary views as to whether the take-up of ICTs is driving the organization and structure of police departments, or if it is the reverse (Manning, 2003; Lin, Hu, & Chen, 2004; NASCIO, 2003; Taylor, Epper, & Tolman, 1998). Evidence is clear that the uptake of new computer-based systems and uses of mobile technologies (beyond the nearly omnipresent radio communications suite in most cars and with most police officers in the U.S.) is accelerating in the U.S. (Nunn, 2001). Partly, this attention comes in response to the country's increased attention to Homeland Security (Rudman, Clarke, & Metzel, 2003), though efforts to improve policing through advanced computing pre-date current attention (Northrup, Kraemer, & King, 1995). The limited functionality and advanced age of many criminal justice and police systems further magnify this attention (Brown, 2001).

Contemporary research also suggests that police are open-minded about new technologies (wireless and otherwise) and generally view favorably the potential of new technologies to change policing (Nunn & Quinet, 2002; Lin et al., 2004). In fact, the evidence shows that most police departments across the United States have one- to three-year plans either to implement wireless technology or have already implemented some form of wireless technology (Dunworth, 2000). To support these efforts, both the United States departments of Homeland Security (DHS) and Justice (DoJ) provide a range of grants to support information technology innovations in police departments throughout the nation. In addition, there is funding by local jurisdictions and a variety of other sources, including internally generated revenue, such as fines, to support technological innovation.

MOBILE COMPUTING AS A SOCIOTECHNICAL ENSEMBLE

Sociotechnical perspectives focus both conceptual and analytical attention on three concepts: that which is social, that which is technical, and their inter-relations. In our study of mobile access to computing resources for police work, the sociotechnical perspective helps us to highlight that mobility is a complex and interdependent set of relations among people (workers and managers), their organizational rules and roles, and various computing resources (such as the technical aspects of the mobile infrastructure, devices used, information sources, and applications accessed). Following Orlikowski and Iacono (2001), we conceptualize mobile access to computing resources as an *ensemble* comprising the wireless network, access devices, applications being used, information and data (both structures and content), procedures followed, norms of behavior (relative to events, systems, and others), governance structures, and both institutional and environmental constraints.

Conceptualizing mobility as a sociotechnical ensemble helps highlight the nuanced and multifaceted interdependencies uniting people, what they do with computing resources, and how they are designed and used. We further argue that what is social and what is technical are engaged in certain times and places and in certain ways. Thus, we build on the work of policing by focusing on specific events and situate these events in specific times and places. This contextual frame provides us the means to ground the analysis of the sociotechnical interactions.

The particular interactions among these constructs will likely vary by situation. For example, in a common[3] event such as a traffic stop, these constructs are tied together in a prescribed way. There are policies regarding the use of the car and personal (attached to the officer's uniform) radio, a standard set of practices that guide the set of interactions the officer has with both the police dispatcher and with the driver of the car being stopped, particular rules about the information needed from police resources (such as registration, license plate numbers, car details, and even data on the driver based on the driver's license proffered to the officer), and what data the officer can and should collect. Escalation procedures are proscribed, and these vary based on time of day, assessments of the local situation, and other operational considerations.

For instance, imagine that a sergeant[4] sees a pick-up truck speeding down a breakdown lane to avoid stopped traffic in the travel lanes and gives a chase. The drivers of the truck see the police car chasing them and, as is customary in the U.S., pulls over to the side of the road. The sergeant sees that the driver is agitated to the point where he is cursing out the vehicle's window; the truck is shaking from "omnidirectional fury," and the sergeant calls for backup from his car radio. While waiting for backup, the officer puts on black leather gloves (in case they scuffle), unsnaps his weapon's securing strap (in case it goes beyond scuffling), calls in to police dispatch with vehicle information, and then switches to his body radio, talk activated. With the radio live (and all other officers on that frequency quiet, and the police dispatcher dispassionately updating time until backup arrives)[5], the sergeant approaches the upset driver and starts the (relatively prescribed) process of gathering particular information on the driver's identity as the first step in writing up a traffic citation. The backup officer arrives while the sergeant is confronting the driver, pulls up diagonally in front of the stopped pickup (to reduce the possibility of a "drive-off"), and stands in plain view and direct line of sight to the driver, weapon at the ready.

A more common traffic stop will have less drama for the driver (but perhaps some irritation),

may not require backup or bring out the visible presence of force, and likely does not escalate until the driver receives multiple citations. But both traffic stops engage the same set of devices, applications, network, common information, and data flows; draw on the same governance structures; follow the same set of procedures (albeit down differing paths, but paths stemming from the same procedural guides); and reflect common and well-developed norms of policing behavior (norms both explicitly taught through extensive training, and also learned and reinforced by doing policing).

Conceptualizing mobility and policing as a sociotechnical ensemble builds on the social-shaping of technology perspectives developed by Bijker (1995), Law and Bijker (1992), and Bijker et al. (1987). In making this point, we acknowledge that there are several active streams of sociotechnical research/theorizing (see Horton, Davenport, & Wood-Harper, forthcoming). For example, the European tradition of sociotechnical theorizing, which we build on here, takes a social shaping of technology (SST) perspective. The SST perspective highlights that the material characteristics and actions of any technology are shaped by the social actions of the designers, the specific uses of that technology, and the evolving patterns of use over time. A second, work-studies tradition of sociotechnical theorizing focuses on the co-design of work practices and technologies to support work. This co-design perspective has been taken up in North America and evolved in two ways. The first is a benign neglect of the interaction between what is social and technical, leading to an evocation of the concepts without a concomitant analytical activity (see Scacchi, 2004, for a critical discussion). The second, an SST approach, is more recent and reflects social informatics in that the efforts are focused on developing specific analytic approaches that make explicit aspects of the social, the technical, and their interaction (Kling, McKim, & King, 2001).

Rather than focusing on a specific theoretical approach to examining the sociotechnical action of policing and mobility, we use Bijker's (1995) principles of sociotechnical change theory to illustrate the generic goals of this approach, and to discuss the theoretical tensions that exist in sociotechnical IT research. These tensions provide a range of possibilities for specific sociotechnical research efforts. Here we use them as orienting principles for our conceptualization of mobility and the consequent design of our research, data collection, and analysis.

Bijker's (1995) four principles of sociotechnical change theory are derived from work in the sociology of technology. These four principles provide a set of goals for any theory that strives to take a sociotechnical perspective: the *seamless Web* principle, the principle of *change and continuity*, the *symmetry* principle, and the principle of *action and structure*. The seamless Web principle states that any sociotechnical analysis should not *a priori* privilege technological or material explanations ahead of social explanations, and vice versa. The principle of change of continuity argues that sociotechnical analyses must account for both change and continuity, not just one or the other. The symmetry principle states that the successful working of a technology must be explained as a process, rather than assumed to be the outcome of "superior technology." The actor and structure principle states that sociotechnical analyses should address both the actor-oriented side of social behavior, with its actor strategies and micro interactions, and structure-oriented side of social behavior, with its larger collective and institutionalized social processes.

While Bijker's principles provide a set of ideals for sociotechnical research to strive for, in practice they illustrate tensions to be managed in the research process. Given the space limitations, in the analysis to follow we focus on highlighting findings relative to our concepts and not specifically examining how the four principles guide this work.

EVIDENCE FROM A FIELD TRIAL OF POLICING, COMPUTING, AND MOBILITY

To explore the sociotechnical perspective on productivity and the effects on work due in part to pervasive access to computing resources, we report on a field study[6] of police officers' uses of an integrated criminal justice system accessed via the public wireless data network from laptops and personal digital assistants (PDAs) provided to the participants. Each element of our field trial is discussed below.

Mobile access for this trial was done via a third-generation (3G) data network. In the U.S., 3G networks are rolling out (typically based on population density) and mirror the cellular phone network in terms of coverage. However, 3G networks use Internet protocols, packet switching (and, thus, digital packets), spread-spectrum transmission (which is inherently more secure than cellular and 2G standards), and can sustain throughput speeds of up to 150 kilobits per second. The 3G data networks in the U.S. are private, and multiple providers compete directly in each market. While wireless coverage is extensive, no one carrier provides complete coverage of the geography of the U.S., and there may be gaps in service within covered areas. Moreover, collectively, all providers' coverage does not cover the geography of the U.S., and a service gap in one provider's coverage is not alleviated by the coverage of a second. The major carriers in the U.S. have deployed their 3G networks in different ways and at different rates[7]. Generally, though, they have focused on deploying in areas where that are most populated (cities and suburbs) and most traveled (along major highways). Costs, reliability, and coverage vary greatly in all other areas (Federal Communications Commission, 2002).

The focal application was Pennsylvania's *J*ustice *NET*work (JNET)[8], a secure Web-based portal connecting authorized users to a set of 23 federated criminal justice and law enforcement databases via a query-based interface. The JNET architecture is characterized by four elements. First, and as noted, for the user it acts as a portal to the criminal-justice-related databases of the Commonwealth of PA (and the U.S. Federal government). The data are owned by the relevant state or Federal agency (e.g., Pennsylvania's Department of Transportation, or PennDOT, maintains driver's license records and a picture database), and JNET provides query-based access to the driver's license photos. Second, JNET is a secure system. Users are carefully vetted before they get access, their use is tied to specific roles, and these roles grant them varying levels of access to the range of data available. Further, use is tied to secure connectivity (enabled through encryption and virtual private networks); this requires several forms of identification to be used[9]. Users must also re-authenticate periodically during their sessions in order to assure security during use. Furthermore, re-authentication is required when accessing certain specific databases through JNET. Until the field trial we report on here, there was no mobile access: thus, security was done via fixed lines and desktop computers. Third, JNET also provides electronic messaging, e-mail, and reporting functions for users. These functions serve as both a common message board across all criminal justice personnel in PA. The e-mail alerts provide a means for people to keep track of activities where they have some interests. For example, it is possible for a parole officer to set up a query on a particular name, social security number, or case number(s). If that name or those numbers come across the message board, she will be alerted and can more easily follow up on the parolee. Fourth, JNET has been operational since early 2000, and it supports thousands of queries each month (and use has grown by nearly 10% per month since inception) (JNET, 2004)[10].

The third part of the mobile access to JNET is the device being used to provide mobile access to JNET (and to the Internet more broadly). This device must have a special 3G modem card and

needs to be mobile. Most police cruisers have an integrated laptop, making this seemingly a trivial effort (put in the wireless modem card, load on the security software, and use a browser). However, there were a number of operational and legal issues that made this a nontrivial effort. For example, many of the laptops are not equipped with space to load the modem card. Battery draw on police cruisers is substantial, and this further limits laptop use (and the 3G modem cards draw substantial power to run the antenna and maintain connectivity). Moreover, some cruisers' laptops have other software whose security and operational/licensing requirements precluded additional applications from being added.

For officers not in a cruiser, the mobile device must be carried on their person. Again, this is not a trivial effort, considering that almost every square inch of the average police person's body is covered by some piece of gear. Moreover, the combination of current equipment (including communications, weapons, body armor, etc.) is nearly 25 pounds. This means that the mobile device must often displace something the officer already carries. We return to this discussion later in the article.

FIELD TRIAL DESIGN, DATA COLLECTION, AND ANALYSIS

The field trial's design focused attention to collecting data on the *wireless network's* use, *device* uses, *JNET* and other *applications*' uses, *information and data sharing*, changes or alterations to police officers' *work practices* (particularly changes to infield operations), *social norms* on computing/uses (particularly regarding the value and importance of both mobile access and JNET), and the officer's operational *governance* (particularly the role of dispatch). As we noted at the article's outset, in focusing on criminal justice, we leverage the extensive knowledge of policing and also partially control for industrial (extra-organizational) factors by staying within one work domain.

The field trial also served as an intervention: mobile workers[11] were provided with either a laptop or a personal digital assistant (PDA)[12] and secure access to the public 3G network. This was done in two phases for pragmatic reasons. The first phase lasted three months, included five participants, and focused on laptop usage. The small number allowed us to refine data collection protocols, and ensure that we could meet the technological demands of supporting the access, security, and application use demands of a demanding operational environment. The second phase began directly after the first phase's completion, involved 13 participants, lasted three months, and focused on PDA usage. All five of the participants in the first trial were part of the second trial. This provided us with a subset of users who were engaged in mobile access to JNET for six months. The two-phase trial's six-month duration was guided by practical constraints of users' ability to participate in a trial while doing their normal policing and official duties. The number included in the trial was constrained by the costs of providing devices, connectivity, and support to the officers.

Participants in both trials were police and other criminal justice officers from three organizations (one county level and two local level) located within one Pennsylvania county. Two incentives were used to motivate participants. First, we promised that all participants could keep the mobile device(s) they were given to use (late-model laptops and high-end PDAs, both equipped with 3G modem cards; and in the case of the PDA, an external sleeve and battery pack to support the modem card). Second, we made it clear that the participants' input would be used to drive the design of JNET for criminal justice uses, particularly for mobile access. Participants mentioned that both were important to their deciding to engage. In addition, we worked with the department heads and unit police chiefs to

ensure that officers were given official recognition for engaging in the field trial. Participating department heads and unit police chiefs were both enthusiastic and supportive.

We used seven forms of data collection. First, we did pre- and post-interviews (at the beginning and end of each trial periods) of all users. In phase one these were face-to-face, open-ended, and semi-structured interviews that lasted from 60 to 90 minutes. In phase two, we used a more structured, self-administered survey in place of some of the open-ended user interviews and followed up with a phone discussion. Second, we led focus groups of users following the trials. These were voluntary, and only two participants did not participate (for schedule reasons). Third, all users completed a one-week time diary of work behavior during the field trial. Fourth, members of the research team did ride-alongs with users. We chose to ride-along with both police and court officers, and with both supervisors and patrol officers. Fifth, we gathered documents during all interviews, observations, and visits (and did extensive Web and library research to support the field work). Sixth, we engaged in informal weekly interactions (via phone, e-mail, and in person) with users. Finally, we gathered data about laptop uses, wireless data transmission, and JNET usage via unobtrusive means (such browser logs, server logs, and telecom activity logs). Data from the first six sources were either transcribed into digital format or collected at source in digital format. Data from the usage logs came in digital format.

Our analysis focused on identifying issues with the 3G network's connectivity/reliability, speed and access, uses of JNET (and other sources/applications), information and data access, and the roles of the mobile devices. This was done through analysis of data drawn from the trouble-ticketing log, analysis of time use (drawn from the logs) regarding connection via 3G networks, volume of data transfer and time/usage of JNET, and through a series of topical analyses of the texts created from the six forms of intensive data collection.

Analysis of data regarding information and data sharing, work practices, social norms, and operational governance followed traditional qualitative data analysis approaches (see Miles & Huberman, 1994). In particular, we used three techniques: interim analysis of the data to guide both future data collection and its interpretation, explanatory even matrices, and content analysis of the transcripts, logs, and field notes.

FINDINGS

We present and discuss seven findings. We find that police officer's uses of 3G *wireless networks* is dependent more on coverage and reliability of access than on speed (bandwidth). Certainly, higher throughput speeds are better than lower speeds (particularly when transferring driver's license photos, as we discuss below). However, if coverage is not certain, then officers either forget to access the network, or become frustrated and actively choose to NOT access the network. Moreover, if an officer takes the time, cognitive energy, and effort to connect, and the access attempt fails (for any number of reasons), it appears they quickly cease trying.

We find that the police officers in our study do not value laptops as access *devices*. They do, however, appreciate these devices for other activities (such as writing up their incident reports and other tasks that did not require them to have wireless access). Police officers valued PDAs to an even greater degree. Again, these *devices* are valued for personal information management and not as connective devices to the 3G network. We did not attempt to trial pen-based or tablet computers: we suspect that these may combine the portability of a PDA with the power and screen size (an important issue for officers) of a laptop.

The mobile access to and uses of *JNET* and other *applications* was difficult to assess for two reasons. First, the low reliability of the network coverage made it difficult for officers to access

these applications. The officer had to become very familiar with coverage patterns (that is, where they could and could not gain access) and then be able to adjust their work patterns to accommodate this coverage. Second, authentication and security overhead in access complicated the logon procedures and caused connection drops. The two factor logon procedures made it difficult for officers in the field to manage both connection and conduct their work. The design of JNET (which asks for updates on passwords and re-authentication as different databases are searched) meant that it was easy for JNET to shut down the session unless the officer devoted considerable attention to managing the interaction. This considerable attention to JNET had to come at the expense of attention to other aspects of the officer's work. In any operational event (such as a traffic stop) the officer would not make this commitment.

Despite this difficulty, officers value JNET for its ability to provide them *information* about drivers, particularly the driver's license photos and drivers' records. On this (and limited evidence of this) alone, officers prized mobile access to JNET and found value in mobility. We did not see any changes in *information and data sharing* for at least two reasons. First, the design of JNET for mobile access is to provide it to officers, and not through police dispatch. Most all other information and data sharing, however, goes through police dispatch (both in a controlled voice-based interaction and via current text-based systems that come to the police vehicle's onboard laptop).

We saw little changes to police officers' *work practices.* Perhaps this is not surprising — the operational environment of policing is harsh, and sometimes fatal. Police train extensively, continually, and with great care to develop procedures to take an ambiguous situation and make it less so. Changes in operational procedures are, thus, slow to come, painstakingly thought out, and must be demonstrable improvements. If not, police are unlikely to risk their lives.

The great enthusiasm and interest on the uses of computing to improve policing seems to be one of the strong *social norms* that police carry forward (Manning, 2003). However, when confronted with changes to operational procedures and concerns with the computing system's reliability, the social norms of policing operations such as safety, professionalism, and force projection overwhelm the potential value of mobile access to computing resources that cannot be consistently demonstrated.

The trial of mobile access to JNET and other computing resources amplified the institutional embeddedness of the command and control structures in policing. In particular, the critical social, organizational, and technical roles that the police dispatcher plays came clear during this trial. The design of JNET for individual access does not work well within police officers' operational *governance*. Were JNET to be a dispatch-based access model, however, governance and information sharing would likely change more quickly.

A SOCIOTECHNICAL ACTION PERSPECTIVE

In this section we draw on Bijker's (1995) four principles of sociotechnical change theory to help us reflect on and interpret these seven findings (see Table 1). Building on this reflection and interpretation, we raise three points: that the sociotechnical principles were supported by these findings; that simplistic approaches to engaging the sociotechnical nature of mobility may make it hard to interpret the results; and that there is a dire need for more substantive sociotechnical analysis techniques.

The premise of this field trial was that technological factors, such as mobile connectivity and higher bandwidth, would be central to taking up mobility. This violates the seamless Web principle. The findings suggest that the institutional structures which help to govern the work of

Table 1. Sociotechnical analysis

Findings	Principles	Comments
Coverage and reliability of access more important than speed/bandwidth	Seamless web	Technological features (bandwidth) were seen as more central than operational needs of officers (operational reliability)
PDA valued for personal use, not for mobile access	Symmetry	Take up of the device is a social decision, shaped by technical characteristics, and often made for personal needs.
JNET and other *applications* are used when mobile	Change and continuity	The expectation that JNET would be valuable for mobile officers (as it has been for officers via fixed access) was borne out in the study.
Officers value *information* drawn from *JNET*	Change and continuity	The expectation that information received while mobile would be valued was borne out in the study.
No changes in *information and data sharing*	Actor and structure	Social and operational structures seemed to be resilient to new technologies of access and use.
No changes to police officer's *work practices* and *social norms*	Actor and structure	Work practices seemed to be resilient to new technologies of access and use.
No changes to work *governance*	Actor and structure	Governance structures seemed to be resilient to new technologies of access and use.

policing serve as powerful moderators to both taking up and taking advantage of mobility. It appears, however, that the technological belief of connection and bandwidth were privileged, relative to these institutional structures. It appears from these findings that moving out access to high-value resources (from fixed to mobile connection) is valued, supporting the principle of change and continuity. However, the structural and institutional forces severely constrain action, and the promised performance of the devices, mobile access, and information sharing require more agentic effort than police officers have.

Second, we observe that the current professional practice of evaluating new ICT does not seem to engage sociotechnical principles. For example, the failure to fully engage sociotechnical principles when designing and trialing mobile access to JNET reflects a naïve view of sociotechnical action: that social and technical are distinct of one another (and that change in one causes change in the other). The findings we note above are not surprising; current institutional structures in policing were not considered (or, worse, ignored — as was the case with dispatch) when designing new work technologies. And, the technological elements must be considered on par with social elements — had this been more carefully considered, bandwidth would not have been the focus; it would have been reliability.

The field trial design reflects the collaboration between wireless service providers, device manufacturers, local and state police and information technology leaders, and faculty. That the resulting trial underplayed the sociotechnical issues leads us to theorize that organizational decision makers, users, and technology evaluators' orientation towards problem solving will make it attractive to focus on matching technical features with work and organizational needs. In doing this they are not likely to address the systemic interactions or to consider extended interdependencies. In essence, this simplification in analysis comes at the cost of accuracy in implementation.

Sociotechnical approaches, such as Bijker's (1995) four principles, appear more likely to be applied in *post-hoc* analysis. They become a comfortable frame for scholars to use. However, they are at best a weak analytic structure to base proactive action. That is, the principles are useful to frame and interpret evidence, but are difficult to use in guiding specific designs. What is missing are the intermediate-level guidance linked to specific technologies or specific social actions. In the absence of this intermediate-level guidance, the principles are difficult to apply proactively.

Building on this, it seems important, if not imperative, that sociotechnical models provide more intermediate guidance. By this we mean support for constraints and enablers tied to particular social actions or that highlight elements of particular technologies. This intermediate level of sociotechnical knowledge is likely to be represented as contingent or localized models. In doing this, such localized models will help academics and practicing professionals more directly to dominant patterns of interactions and consequences, and make these findings available in ways that more directly influence ICT/systems design and organizational decision making.

Our final observation from this analysis is that the over-simplification of sociotechnical action is likely to continue unless more viable analytic approaches are developed and the assumptions of the current techno-determinist approaches are challenged more explicitly. Given this view, it seems likely that organizational decision makers, users, and ICT designers will have trouble making sense of evidence drawn from failed attempts to implement and use ICT based on their simple views of ICT use, cause and effect. We believe the inability to understand this data is driven by the unsound approach of invoking direct effects of ICT use, not by the measurements taken or instruments used to gather evidence (e.g., Sawyer, Allen, & Lee, 2003).

While the research literature focused on the effects of ICT highlights on the indirect and often nuanced relationships among use of ICT and performance, professional practice continues to press for the direct effects model of ICT value. This suggests that more robust system or contingency models of ICT effects are needed (e.g., Avgerou, 2002). This is one of the most active areas of scholarship in IT, and this activity needs to enter the texts, teaching cases, and classrooms of the next generation's IT leaders, organizational managers, and technology developers. For example, those who have focused specifically on the roles of mobile and fixed location uses of ICT in policing all note that the operational value derived from using new ICT-centric information systems is minimal, if discernible (Ackroyd, Harper, Hughes, Shapiro, & Soothill, 1996; Dunworth, 2000; Meehan, 2000).

What seems important to us is a more focused effort to engage the principles of sociotechnical action in direct comparison to the bases of direct effects models (e.g., Kling & Lamb, 2000). They develop a comparative analysis of tool and Web models of computing relative to organizational activity. In doing this, they highlight both the seamless Web principle (privileging neither the social nor the technical) and the principle of action and structure by highlighting the concept of a social actor — one that has agency, but constrained by institutional structures (Lamb & Kling, 2003). Building on these two principles,

in the work reported here we provide a means of representing the principle of change and continuity by explicitly linking elements of the technical structure of JNET, the institutional structures of police work, and the actions of police.

REFERENCES

Ackroyd, S., Harper, R., Hughes, J., Shapiro, D., & Soothill, K. (1996). *New technology and police work*. Buckingham: Open University Press.

Avgerou, C. (2002). *Information systems and global diversity*. Oxford: Oxford University Press.

Bijker, W. (1995). *Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change*. Cambridge, MA: The MIT Press.

Bijker, W., Hughes, T., & Pinch, T. (1987). *The social construction of technological systems.* Cambridge, MA: The MIT Press.

Brown, M.M. (2001). The benefits and costs of information technology innovations: An empirical assessment of a local government agency. *Public Performance & Management Review, 24*(4), 351-366.

Dunworth, T. (2000) Criminal justice and the information technology revolution. In Horney (Ed.), *Policies, processes and decisions of the justice system* (volume 3, pp. 372-426). Washington, DC: National Institute of Justice/Office of Justice Programs.

Horton, K., Davenport, E., & Wood-Harper, T. (2005). Exploring sociotechnical interaction with Rob Kling: Five 'big' ideas. *Information, Technology and People*, (in press).

JNET. (2004). Usage statistics. Retrieved from: *http://www.pajnet.state.pa.us/pajnet/site/default.asp*

Kling, R. (1999). What is social informatics, and why does it matter? *D-Lib Magazine, 5*(1). Retrieved from: *http://www.dlib.org:80/dlib/january99/kling/01kling.html*

Kling, R. (2000). Learning about information technologies and social change: The contribution of social informatics. *The Information Society, 16*(3), 217-232.

Kling, R. & Lamb, R. (2000). IT and organizational change in digital economies: A sociotechnical approach. In B. Kahin & E. Brynjolfsson (Ed.), *Understanding the digital economy: Data, tools and research*: Cambridge, MA: The MIT Press.

Kling, R., McKim, G., & King (2001). A bit more to IT: Scholarly communication forums as socio-technical interaction networks. Retrieved May 6, 2004 from *http://www.slis.indiana.edu/CSI/WP/wp01-02B.html*

Klockers, C. & Mastrofski, S. (Eds.). (1991). *Thinking about police: Contemporary readings.* New York: McGraw-Hill.

Lamb, R. & Kling, R. (2003). Reconceptualizing users as social actors in information systems research. *MIS Quarterly, 27*(2), 197-235.

Law, J. & Bijker, W. (1992). Technology, stability and social theory. In W. Bijker (Ed.), *Shaping technology/building society* (pp. 32-50). Cambridge, MA: The MIT Press.

Lin, C., Hu, P., & Chen, H. (2004). Technology implementation management in law enforcement. *Social Science Computer Review, 22*(1), 24.

Manning, P. (1977). *Police work: The social organization of policing.* Prospect Heights, IL: Waveland Publishing.

Manning, P. (1996). Information technology in the police context: The 'sailor' phone. *Information Systems Research, 7*(1), 275-289.

Manning, P. (2003). *Policing contingencies.* Chicago: University of Chicago Press.

Meehan, A. (2000). The transformation of the oral tradition of policing through the introduction of information technology. *Sociology of Crime, Law and Deviance, 2,* 107-132.

NASCIO (National Association of State Chief Information Officers). (2003). Concept for operations for integrated justice information sharing version 1.0. Retrieved from: *https://www.nascio.org/publications/index.cfm*

Northrup, A., Kraemer, K.L., & King, J.L. (1995). Police use of computers. *Journal of Criminal Justice, 23*(3), 259-275.

Nunn, S. (2001). Police information technology: Assessing the effects of computerization on urban police functions. *Public Administration Review, 61*(2), 221-234.

Nunn, S. & Quinet, K. (2002). Evaluating the effects of information technology on problem-oriented-policing: If it doesn't fit, must we quit? *Evaluation Review, 26*(1), 81-108.

Orlikowski, W. &. Iacono, S. (2001). Desperately seeking the "IT" in IT research — A call to theorizing the IT artifact. *Information Systems Research, 12*(2), 121-124.

Rosenbach, W. & Zawacki, R. (1989). Participative work redesign: A field study in the public sector. *Public Administration Quarterly, 43,* 111-127.

Rudman, W., Clarke, R., & Metzel, J. (2003, July 29). *Emergency responders: Drastically underfunded, dangerously unprepared.* Report of an independent task force sponsored by the Council on Foreign Relations. Retrieved from: *http://www.cfr.org/pdf/Responders_TF.pdf*

Sawyer, S. & Eschenfelder, K. (2002). Social informatics: Perspectives, examples, and trends. In B. Cronin (Ed.), *Annual review of information science and technology* (volume 36, pp. 427-265). Medford, NJ: Information Today Inc./ASIST.

Sawyer, S., Allen, J., & Lee, H. (2003). Broadband and mobile opportunities: A sociotechnical perspective. *Journal of Information Technology, 18*(2), 11-35.

Sawyer, S., Tapia, A., Pesheck, L., & Davenport, J. (2004). Observations on mobility and the first responder. *Communications of the ACM, 47*(2), 62-65.

Taylor, M., Epper, R., & Tolman, T. (1998). *Wireless communications and interoperability among state and local law enforcement agencies*. Report 168945 of the National Criminal Justice Clearinghouse, Washington, D.C.

ENDNOTES

1 Manning (1996) focused on the take up and uses of cellular phones by police. Personal cellular phone ownership and use is now common among criminal justice officers. While the take up and use of cellular phones is beyond the scope of this article, two attributes are worth noting. First, the officers use their own (personal) cellular phones and do not consider them as part of their professional equipment. Second, personal use has made officers aware of issues with wireless coverage, reliability, and use.

2 Given the extensive literature on policing, in this article we draw from but do not develop or discuss principle findings. Instead, we refer the interested reader to anthologies of such work (listed in our references and cited here). The interested reader can also find courses in crime, law, and justice offered in most sociology departments, and the extensive material on the Web in locations such as the U.S. Department of Justice, the UK Home Office, and the International Association of Chiefs of Police.

[3] Perhaps one of the more difficult parts of a police officer's job is to remember that even a seemingly common thing such as stopping a speeding car may lead to armed confrontation. Thus, training is focused on preventing common from becoming routine.

[4] Policing in the United States is organized along paramilitary lines. Thus, sergeants are senior/experienced officers, typically with both patrol and supervisory responsibilities.

[5] Most police in the United States work alone, which means: 1) they rely on the radio as a link to others and 2) the police dispatcher is a critical node in this linkage. The radio stays on, and no one else speaks so that all can listen for a gunshot or the words "officer down."

[6] Our research design here builds on previous public-sector field studies of work (Rosenbach & Zawacki, 1989).

[7] Details of the debate and key issues in wireless network deployment, coverage, access,and use are beyond the scope of this article.

[8] For more information about JNET, see www.pajnet.state.pa.us.

[9] Security in the trial was done via "two-factor" identification. This means having a physical key, called a dangle by the officers, that stores a digital record identifying the owner that is tied to a logical password that must be entered when the physical key is connected (via USB port) to the computer.

[10] JNET is one of the earliest and most visible examples of a small and growing number of these integrated criminal justice information systems that are a focus of homeland security efforts in the United States. Others include the Capital Area Wireless Integrated Network (CAPWIN, see www.capwin.org), the automated regional justice administration system (ARJIS, see www.arjis.org),and a fast-growing number of municipal efforts such as systems in Chicago, IL;Montgomery County, MD; and Los Angeles County, CA.

[11] Participants included one sergeant (or shift supervisor), nine patrolmen, and three deputies (of the court). Participants were male from ages 28 to 45. The average work experience was 11 years; the most experienced officer had 18 years of work and the least experienced officer had seven years of work.

[12] Laptops and PDAs were provided to officers as an incentive to participate and were paid for by the research team.

This work was previously published in the International Journal of Technology and Human Interaction, edited by B. Stahl, Volume 1, Issue 3, pp. 1-14, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 5.29
Social Context for Mobile Computing Device Adoption and Diffusion:
A Proposed Research Model and Key Research Issues

Andrew P. Ciganek
University of Wisconsin-Milwaukee, USA

K. Ramamurthy
University of Wisconsin-Milwaukee, USA

ABSTRACT

The purpose of this chapter is to explore and suggest how perceptions of the social context of an organization moderate the usage of an innovative technology. We propose a research model that is strongly grounded in theory and offer a number of associated propositions that can be used to investigate adoption and diffusion of mobile computing devices for business-to-business (B2B) interactions (including transactions and other informational exchanges). Mobile computing devices for B2B are treated as a technological innovation. An extension of existing adoption and diffusion models by considering the social contextual factors is necessary and appropriate in light of the fact that various aspects of the social context have been generally cited to be important in the introduction of new technologies. In particular, a micro-level analysis of this phenomenon for the introduction of new technologies is not common. Since the technological innovation that is considered here is very much in its nascent stages there may not as yet be a large body of users in a B2B context. Therefore, this provides a rich opportunity to conduct academic research. We expect this chapter to sow the seeds for extensive empirical research in the future.

INTRODUCTION

What causes individuals to adopt new information technologies (ITs)? How much influence do the perceptions of the social context of an organization have on the acceptance of new ITs? These questions are significant because systems that are not utilized will not result in expected efficiency and effectiveness gains (Agarwal & Prasad, 1999), and will end up as unproductive use of organizational resources. Academic research consequently has focused on the determinants of computer technology acceptance and utilization among users. Some of this research comes from the literature on adoption and diffusion of innovations (DOI), where an individual's perceptions about an innovation's attributes (e.g., compatibility, complexity, relative advantage, trialability, visibility) are posited to influence adoption behavior (Moore & Benbasat, 1991; Rogers, 2003). Another stream of research stems from the technology acceptance model (TAM), which has become widely accepted among IS researchers because of its parsimony and empirical support (Agarwal & Prasad, 1999; Davis, 1989; Davis, Bagozzi, & Warshaw, 1989; Hu, Chau, Sheng, & Tam, 1999; Jackson, Chow, & Leitch, 1997; Mathieson, 1991; Taylor & Todd, 1995; Venkatesh, 1999, 2000; Venkatesh & Davis, 1996, 2000; Venkatesh & Morris, 2000).

Individual differences indeed are believed to be very relevant to information system (IS) success (Zmud, 1979). Nelson (1990) also acknowledged the importance of individual differences in affecting the acceptance of new technologies. A variety of research has investigated differences in the perceptions of individuals when using TAM (Harrison & Rainer, 1992; Jackson et al., 1997; Venkatesh, 1999, 2000; Venkatesh & Morris, 2000); however, the perceptions and influences of the social context of an organization have not been widely examined in the literature. Hartwick and Barki (1994) suggest that it is imperative to examine the acceptance of new technologies with different user populations in different organizational contexts.

Although mobile computing devices have existed for several years, strategic applications of this technology are still in their infancy. Mobile computing devices (in the context of business-to-business—B2B) is treated as a technology innovation in this chapter due to their newness and short history. An investigation into the usage of mobile computing devices within a B2B context, which we define as two or more entities engaged within a business relationship, is of value because of its increasing popularity (March, Hevner, & Ram, 2000). As an emergent phenomenon, relatively modest academic literature has examined the nature of adoption and use of this technology. Mobile computing devices, which have been described as both ubiquitous (March et al., 2000) and nomadic (Lyytinen & Yoo, 2002a, 2002b), offer a stark difference from traditional, static computing environments. A good characterization of these differences is provided in Satyanarayanan (1996). New technology innovations typically require changes in users' existing operating procedures, knowledge bases, or organizational relationships (Van de Ven, 1986). Such innovations may even require users to develop new ways of classifying, examining, and understanding problems. The domain of mobile computing devices has the potential to become the dominant paradigm for future computing applications (March et al., 2000), and topics of such contemporary interest are recommended to be pursued in IS research (Benbasat & Zmud, 1999; Lyytinen, 1999).

The primary objective of this chapter is to examine whether and how perceptions of the social context of an organization moderate the adoption, use, and infusion[1] of mobile computing devices for B2B transactions. We extend TAM to include individuals' perceptions of the social context of their organization, which incorporates aspects of both culture and climate research as recommended in the literature (Denison, 1996; Moran & Volkwein, 1992). Aspects of the social context of an organization are suggested as having a significant role in the introduction of new

technologies (Boudreau, Loch, Robey, & Straub, 1998; Denison & Mishra, 1995; Legler & Reischl, 2003; Orlikowski, 1993; Zammuto & O'Connor, 1992), particularly with the introduction of mobile computing devices (Jessup & Robey, 2002; Sarker & Wells, 2003). Only a handful of studies in the past have specifically looked at the micro-level connections of these relationships (Straub, 1994); unfortunately, even this has not been within a mobile computing context. We argue that an organization's social context will have a significant moderating effect on the perceptions of employees considering adoption and use of mobile computing applications for B2B purposes.

The chapter proceeds as follows: the next section presents the background research in the domains (adoption and diffusion of technology innovations within the context of TAM, DOI, and social context) underlying this research. This will be followed by the presentation and discussion of our proposed model and accompanying propositions. A brief discussion of the types of B2B application domains that are relevant to mobile-computing and would be of (future) interest to our investigation is then presented, accompanied by one methodological approach to how such research can be conducted. This chapter concludes with some potential implications for research and practice, limitations of the book chapter, and potential future directions.

BACKGROUND RESEARCH

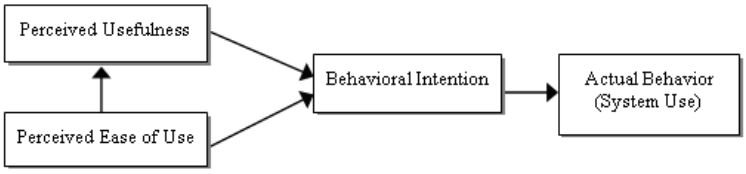
In this section, we first discuss the extant research connected with the technology acceptance model followed by research related to social context.

Technology Acceptance Model

The technology acceptance model proposed by Davis (1989) has its roots in the theory of reasoned action (TRA) of Fishbein and Ajzen (1975). As earlier alluded to, it is one of the most widely used models of IT acceptance. This model accounts for the psychological factors that influence user acceptance, adoption, and usage behavior of new IT (Davis, 1989; Davis et al., 1989; Hu et al., 1999; Mathieson, 1991; Taylor & Todd, 1995). The TAM model is displayed in Figure 1.

As is fairly well known in the IT literature, TAM specifies two beliefs—*perceived usefulness* (PU) and *perceived ease of use* (PEOU)—to be determinants of IT usage. It incorporates behavioral intention as a mediating variable in the model, which is important for both substantive and sensible reasons. In terms of substantive reasons, the formation of an intention to carry out a behavior is thought to be a necessary precursor to actual behavior (Fishbein & Ajzen, 1975). In terms of sensible reasons, the inclusion of intention is found to increase the predictive power

Figure 1. Technology acceptance model (Adapted from Davis, Bagozzi, & Warshaw, 1989)



of models such as TAM and TRA, relative to models that do not include intention (Fishbein & Ajzen, 1975). Perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance her/his job performance"; perceived ease of use is defined as "the degree to which a person believes that using a particular system would be free of effort" (Davis, 1989, p. 320).

The TAM model and other subsequent IT models of acceptance have largely ignored the influence that continued usage has on the acceptance of an IT. For example, Karahanna, Straub, and Chervany (1999) found differences in the determinants of attitudes between potential adopters and actual users of an IT. In particular, they found that perceived usefulness continued to play an important role in the attitudes of IT users, while ease of use ceased to be important over time. Consequently, the relationship between actual/demonstrated usefulness and continued use is added by us to the original TAM model. Once the actual/realized usefulness of an IT is confirmed by a potential adopter, it is likely to continue to play a significant role in the overall infusion of the technology.

Based upon conceptual and empirical similarities across eight prominent models in the user acceptance literature, Venkatesh, Morris, Davis, and Davis (2003) developed a unified theory of individual acceptance of technology (the unified theory of acceptance and use of technology, or UTAUT). The UTAUT theorizes four constructs as having a significant role as direct determinants of acceptance and usage behavior: performance expectancy (subsuming perceived usefulness), effort expectancy (subsuming perceived ease of use), social influence, and facilitating conditions. In addition, it considers four moderators—age, gender, voluntariness of use, and experience of the users to influence the relationship between the four direct antecedents and intentions to use (and in the case of facilitating conditions on actual use behavior). Although the UTAUT model explains a significant amount of variance in the intention to adopt an IT, the model lacks the parsimony and empirical replication of the TAM model. In this light, the modified TAM model that we propose may be considered a viable and prudent alternative to the UTAUT model. An empirical comparison between these two models is, of course, necessary.

Recent research employing the TAM model had identified individual differences as a major external variable (Agarwal & Prasad, 1999; Jackson et al., 1997; Venkatesh, 2000; Venkatesh & Morris, 2000). Individual differences are any forms of dissimilarity across people, including differences in perceptions and behavior (Agarwal & Prasad, 1999). For example, Agarwal and Prasad (1999) found that an individual's role (provider or user) with regard to a technology innovation, level of education, and previous experiences with similar technology were significantly related to their beliefs about the ease of use of a technology innovation. Agarwal and Prasad also found a significant relationship between an individual's participation in training and their beliefs about the usefulness of a technology innovation. Jackson et al. (1997) examined variables such as situational involvement, intrinsic involvement, and prior use of IT by users, and Venkatesh (2000) considered individual specific variables such as beliefs about computers and computer usage, and beliefs shaped by experiences with the technology in the traditional TAM. Both these studies found significant relationships among these individual differences and TAM constructs. Further, Venkatesh and Morris (2000) argue from their findings that "men are more driven by instrumental factors (i.e., perceived usefulness) while women are more motivated by process (perceived ease of use) and social (subjective norm) factors" (p. 129). Thus, while the various above-noted research studies have investigated the differences in the perceptions of individuals using TAM as the underlying theoretical basis, as noted earlier, perceptions of the social context of an organization is not

common in the literature. Most of these refinements to TAM and findings are accommodated in the earlier-noted overarching UTAUT model proposed by Venkatesh et al. (2003).

Social Context of an Organization and Innovativeness

As noted in the introduction, although the social context of an organization has been suggested as having a significant role in the introduction of new technologies (Boudreau et al., 1998; Denison & Mishra, 1995; Legler & Reischl, 2003; Orlikowski, 1993; Zammuto & O'Connor, 1992), particularly with the introduction of mobile computing devices (Jessup & Robey, 2002; Sarker & Wells, 2003), it has not been widely examined in the literature. In this chapter we extend the TAM to incorporate an individual's perceptions of the social context of their organization. The perceptions of the social context are of value to consider since they are likely to be fairly stable in the mind of the potential adopter and less subject to change than other perceived factors or the underlying technological innovation. As recommended in the literature, we examine the social context of an organization to incorporate aspects of both culture and climate (Denison, 1996; Moran & Volkwein, 1992). We take the stand that a study of organizational culture and organizational climate actually examine the same phenomenon—namely, the creation and influence of social contexts in organizations—but from different perspectives (Denison, 1996). Following the recommendation of prior research, we examine the broader social context in order to improve our understanding of the organizational phenomenon (Astley & Van de Ven, 1983; Denison, 1996; Moran & Volkwein, 1992; Pfeffer, 1982).

Organizational climate can be described as the shared perceptions of organizational members who are exposed to the same organizational structure (Schneider, 1990). Zmud (1982) suggests that it is not the structure of the organization that triggers innovation; rather, innovation emerges from the organizational climate within which members recognize the desirability of innovation, and within which opportunities for innovation arise and efforts toward innovation are supported. As summarized in Schneider (1990) and in Moran and Volkwein (1992), a number of different conceptualizations of organizational climate have been suggested over the years. Pareek (1987) advanced the idea that climate and culture can only be discussed in terms of how it is perceived and felt by individual members/employees of the organization, which is a perspective that is supported in the literature (Legler & Reischl, 2003). Thus, we are interested in capturing the perceptions of individuals within organizations. Since the unit of analysis (during empirical evaluation) in this chapter is the individual employees within organizations, appropriate measures of examining social context can be derived from psychological climate literature.

Rather than focusing on how the psychological climate of an organization gets formed and can be influenced (certainly important), of interest in this chapter is how the prevailing climate of an organization moderates the relationship between individuals' perceptions of an innovation's usefulness and ease of use, and their intentions to adopt and use the innovation. Psychological climate is a multi-dimensional construct that can be conceptualized and operationalized at the individual level (Glick, 1985; Legler & Reischl, 2003). In an attempt to integrate several different measures of psychological climate, Koys and DeCotiis (1991) derived eight summary dimensions—*autonomy*, *cohesiveness*, *fairness*, *innovation*, *pressure*, *recognition*, *support*, and *trust*. A brief definition/description of each of these dimensions is provided in Table 1.

In the next section, while presenting our research model and associated propositions, we will discuss how each of these dimensions would be expected to moderate the relationship between an individual's perceptions (of an innovation) and

Table 1. Dimensions of psychological climate (Adapted from Koys & DeCotiis, 1991)

Dimension Name	Definition
Autonomy	Employee's perception of their own sovereignty with respect to work procedures, goals and priorities.
Cohesion	Employee's perception of sharing and togetherness within their organization.
Trust	Employee's perception of freedom to communicate openly with members at higher organizational levels about sensitive or personal issues with the expectation that the integrity of such communications will not be violated.
Pressure	Employee's perception that time demands are incongruent with respect to task completion and performance standards.
Support	Employee's perception of the tolerance of their behavior by superiors, including the willingness to let employees learn from their mistakes without fear of reprisal.
Recognition	Employee's perception that their contributions to their organization are acknowledged.
Fairness	Employees' perception that their organization's practices are equitable and non-arbitrary.
Innovation	Employee's perception that change and originality are encouraged and valued within their organization, including risk-taking in domains where the individual may have little to no prior experience.

behavioral intention (to adopt and use it). Briefly, however, we will take a couple of these climate dimensions (*support* and *autonomy*) and discuss the relevance of these dimensions of organizational climate for the adoption of technological innovations.

Senior management's attitude toward change (consequential to the introduction of technology innovations) and thus the extent of their *support* impacts the adoption of these technology innovations (Damanpour, 1991). Senior management teams may be very conservative, preferring the status quo and using current or time-tested methods innovating only when they are seriously challenged by their competition or by shifting consumer preferences (Miller & Friesen, 1982). By contrast, they may be risk prone, actually encouraging and actively supporting the use of innovative techniques to move the organization forward, usually trying to obtain a competitive advantage by routinely making dramatic innovative changes and taking the inherent risks associated with those innovations (Litwin & Stringer, 1968). The potentially disruptive features typically associated with the adoption of (radical) innovations require an organizational context where managers encourage individual members of the organization to take (prudent levels of) risk, support adoption of technology innovations, and be supportive of changes in their organizations (Dewar & Dutton, 1986). Organizations should be wary, however, that a follower approach taken by employees may promote a "mindless" environment resulting in undesirable levels of risk-taking, which can cause significant problems (Swanson & Ramiller, 2004).

Organizational context/climate also reflects the extent of focus on autonomy/empowerment vs. control of its members. An organic organization as contrasted with mechanistic organization is

typically associated with open and free-flowing communication, sharing of necessary information and knowledge, flexibility, and absence of rigid rules and regulations; such an organization context is usually positively related to innovation (Aiken & Hage, 1971; Kimberly & Evanisko, 1981). Furthermore, an organizational climate that is geared toward and has built-in expectation of high levels of achievement and high standards of excellence nurtures a vibrant base of challenges posed to its members who have the freedom to apply innovative technologies, techniques, and procedures to effectively accomplish the tasks (Rosenthal & Crain, 1963). Such an organizational context will be more prone to encouraging its members to adopt technology innovations to accomplish high levels of performance.

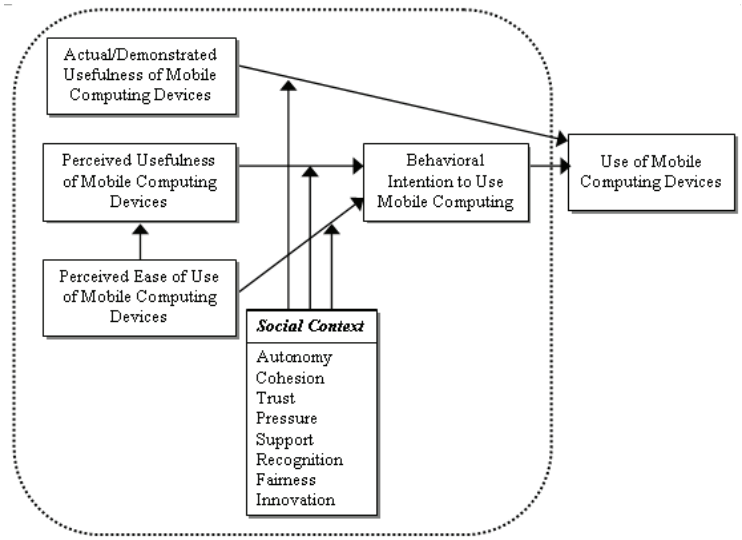
RESEARCH MODEL AND TENTATIVE PROPOSITIONS

Based on the foregoing brief discussion of the extant research, we extend the standard TAM model with social context dimensions as shown in Figure 2.

Traditional TAM Propositions

An individual's intention to adopt/use technology is driven by his or her perceptions of the usefulness of the technology (Davis et al., 1989). This contention, as noted in the background research section, has been supported extensively in previous research (Agarwal & Prasad, 1999; Davis et al., 1989; Hu et al., 1999; Jackson et al., 1997; Venkatesh, 1999, 2000; Venkatesh & Davis, 2000; Venkatesh & Davis, 1996; Venkatesh & Morris, 2000). A primary reason why individuals would intend to adopt/use mobile computing devices for B2B transactions is that they believe that this technology will provide them the flexibility to perform their job and enable their job performance enhancement (Davis, 2002; Intel, 2003). Furthermore, following the findings of Karahanna et al. (1999), the perceived usefulness of an IT influences the attitudes of both potential adopters and users of an IT. However, we contend

Figure 2. Research model



that when an IT has demonstrated its usefulness over time, it is likely to play a significant role in the overall infusion of the technology. Therefore, we propose:

- **Proposition 1:** Perceived usefulness will have a positive effect on organizational members' intention to adopt/use mobile computing devices for B2B transactions.
- **Proposition 2:** Actual/demonstrated usefulness will have a positive effect on organizational members' continued usage of mobile computing devices for B2B transactions.

As noted earlier, the second major determinant of behavioral intentions in the TAM model, *perceived ease of use*, has been observed to have both a (somewhat weak) direct influence on behavioral intention as well as a (strong) indirect influence through its effect on perceived usefulness (Davis, 1989; Davis et al., 1989; Hu et al., 1999; Jackson et al., 1997). This is understandable since a person who believes that a technology innovation is (relatively) easy to understand and use, and is less demanding of efforts, would likely believe that using such a technology is also more useful. While perceived ease of use may trigger users' intention to adopt/use the innovation (mobile computing devices for B2B), it is unlikely to play a key role in the spread/infusion since users would likely become more familiar with all the features of the innovation and gain significant expertise with time following the initial use. Hence, we propose the two following propositions:

- **Proposition 3a:** Perceived ease of use will have a positive effect on organizational members' intention to adopt/use mobile computing devices for B2B transactions.
- **Proposition 3b:** Perceived ease of use will positively influence organizational members' perceptions of the usefulness of mobile computing devices for B2B transactions.

Extended TAM Propositions

One of the key objectives of this chapter is to examine what role, if any, social context plays in the link between individuals' perceptions of usefulness/ease of use and behavioral intentions of the TAM model. We pointed out that social context, when conceptualized in terms of climate/culture, is a multi-dimensional construct composed of eight dimensions (Koys & DeCotiis, 1991). Since there has been no attempt to examine this additional set of dimensions within the context of TAM, many of the arguments and much of the rationale that we provide in the rest of this section while developing the propositions are likely to be tentative.

Autonomy

At one end of the spectrum, an organization can be extremely control and compliance oriented (*mechanistic* organizational context) in formulating, administering, and closely monitoring and enforcing a set of policies and procedures that guide employee work activities. At the opposite end of the spectrum, an organization can be performance and achievement oriented (*organic* organizational context) by empowering their employees to determine their task priorities and schedule, providing them the autonomy to make use of any and all techniques, tools, and technologies that they deem best for getting the work done, and being flexible with respect to adherence on the standard policies and procedures. Thus, organizations where the members perceive greater *autonomy* and flexibility being provided to them in making decisions and choices on their task-related activities are likely to more quickly exploit (any) opportunity that technology innovations offer. While this is fairly obvious when the technology is perceived to be useful and easy to use, even in instances where such perceptions (of ease of use and usefulness) may not be completely true, the

organizational members may still be more willing to make informed decisions that they are responsible and accountable for (Aiken & Hage, 1971; Kimberly & Evanisko, 1981). To become better informed, they may actively seek out knowledge from various pockets of the (internal) organization as well as from external sources (e.g., consultants, vendors, trade literature, etc.). Therefore:

- **Proposition 4a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions will be stronger in organizational contexts that provide greater autonomy to their employees.
- **Proposition 4b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions will be stronger in organizational contexts that provide greater autonomy to their employees.

Cohesion

As would be noted from the brief description provided in Table 1, *cohesion* refers to an organizational context/climate that fosters a sense of sharing, caring, accommodation, and togetherness among the members/employees (Koys & DeCotiis, 1991). Communication, sharing, and exchange of information and knowledge amongst the members is bound to be much more open in such a context. Employees would more willingly share their experiences and support one another when attempting to make decisions on complex and unknown topic areas (e.g., relevance and mastery of new technologies). It is, therefore, reasonable to expect that potential adopters of new technology innovations (mobile computing devices) would be more willing and prepared to assume any challenges posed by the new technology environment in view of the potential support that they can expect from their colleagues in their work environment. Therefore, we propose the following:

Proposition 5a: The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions will be stronger in organizational contexts that foster a greater sense of cohesion/cohesiveness among their employees.

Proposition 5b: The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions will be stronger in organizational contexts that foster a greater sense of cohesion/cohesiveness among their employees.

Trust

The third dimension of organizational climate, *trust*, refers to the extent to which employees within the organization can openly communicate with their superiors, seek their guidance and expertise, and be confident that the integrity of sensitive information will not be compromised (Koys & DeCotiis, 1991). It is easy to visualize that such expectations of trust work in both directions—from subordinate to superiors and vice versa. Trust also involves an expectation of confidence in the goodwill of others in the organizational context/environment, as well as the prospects for continuity of the relationship entered into (Hart & Saunders, 1997). It is normal to expect that in these trusting organizational contexts, employees will be more prepared to share their difficulties and concerns (work related and even personal), propose potential technology-based solutions, and seek approval/guidance/advice from their superiors and peers. This can be quite important as in the case of introduction of mobile computing devices where the work arrangements and workflows are bound to be disrupted and changed quite radically (e.g., employees may not have to

be always present on site and could increasingly work from off-site locations, at home, or on the move). Trust is a significant determinant of a stable relationship (Mayer, Davis, & Schoorman, 1995; McKnight, Choudhury, & Kacmar, 2002). Therefore, we propose:

- **Proposition 6a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions will be stronger in organizational contexts that promote and reinforce trust between employees and the organization.
- **Proposition 6b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions will be stronger in organizational contexts that promote and reinforce trust between employees and the organization.

Pressure

The fourth dimension of organizational climate, *pressure*, refers to the fact that the work context may not provide adequate time for the employees to accomplish their task-related activities and achieve the required standards of performance and goals (Koys & DeCotiis, 1991). Typically, it would be reflective of a situation of significant stress, perhaps hasty decisions and actions resulting in suboptimal results, and generally chaos. However, such a stressful environment may also be one that could spur the organizational members to creatively look for (technologically) innovative solutions to alleviate the difficulties and infuse some order. To the extent that the performance of tasks is not geographically constrained (e.g., assembly-line work in automotive manufacturing, patrons being serviced in a restaurant or a bank), it is possible that mobile computing devices may indeed alleviate the time pressure that is so rampant in the work context. For example, employees may become skillful in time management through the convenience of mobile computing devices in coordinating work and personal tasks (Davis, 2002; Intel, 2003). Therefore, surprising and counter-intuitive as it might sound, we propose:

- **Proposition 7a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that reflect one of (time) pressure for employees to accomplish their task and realize the set performance standards.
- **Proposition 7b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that reflect one of (time) pressure for employees to accomplish their task and realize the set performance standards.

Support

The fifth dimension of organizational climate, *support*, reflects an organizational context that is tolerant of errors and mistakes that employees may commit, and is supportive of them as long as they learn from these (Koys & DeCotiis, 1991). An environment that is permissive and lets its members learn from mistakes without fear of punishment and reprisal could engender deep-rooted learning, a "can-do" attitude to problem solving, and (reasonable) risk-taking orientation (Litwin & Stringer, 1968). As noted earlier, management's attitude toward change (often triggered by the introduction of technology innovations) and thus the extent of their support impacts the adoption and successful implementation of these technology innovations (Damanpour, 1991; Sanders & Courtney, 1985). The potentially disruptive

features typically associated with the adoption of (radical technology) innovations require an organization context where managers encourage individual members of the organization to take (prudent levels of) risk, support adoption of technology innovations, and be supportive of changes in their organizations (Dewar & Dutton, 1986). Supportive organizational context is also conducive to successful IT implementation (Ramamurthy, Premkumar, & Crum, 1999). Caron, Jarvenpaa, and Stoddard (1994) chronicle how CIGNA Corporation, due to its supportive and tolerance-for-failure environment, facilitated significant learning to accrue in the context of major disruptive and radical changes triggered by business process reengineering projects. Therefore, we propose:

- **Proposition 8a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are tolerant and supportive of employees in accomplishing their work.
- **Proposition 8b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are tolerant and supportive of employees in accomplishing their work.

Recognition

The sixth dimension of organizational climate, *recognition*, reflects an organizational context where employee achievements and accomplishments are acknowledged and recognized (Koys & DeCotiis, 1991). *Human relations management* and *job enrichment* literature (Hackman & Oldham, 1980) points out that intrinsic rewards (e.g., employee-of-the-month recognition) at times are more important than extrinsic rewards (e.g., salary raises, promotion). Extrinsic and intrinsic motivation literature has also been used significantly to explain adoption and use of innovations (Davis, Bagozzi, & Warshaw, 1992). Resource-based theory also acknowledges the vital role human assets/resources play in contemporary hyper-competitive external environments where progressive organizations strive to keep their employees satisfied and thus retain top talent. It is, therefore, natural to expect that organizations should strive to create a climate that spurs their employees to constantly look out for creative solutions (including new technology innovations) that foster excellence in achievement. Obviously, this is unlikely when such efforts and accomplishments go unrecognized. Thus, we would propose:

- **Proposition 9a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are open to acknowledge and recognize the accomplishments of their employees.
- **Proposition 9b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are open to acknowledge and recognize the accomplishments of their employees.

Fairness

The seventh dimension of organizational climate, *fairness*, reflects an organizational context where employees believe in equitable and non-arbitrary treatment (Koys & DeCotiis, 1991). This reinforces the notion that hard, sincere, and smart work pays off. Individuals that believe an inequity exists, for example, are likely to resent and resist organizational changes (Joshi, 1989, 1991).

Clearly an organization that does not design its workplace context with work/job assignments that are perceived to be fair and rewards that are perceived to be equitable for similar accomplishments would trigger significant discontent and distrust. Such an environment is hardly likely to evoke any voluntary or enthusiastic response to work-related organizational challenges, including searching for new technology innovations. Therefore, we would propose:

- **Proposition 10a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are deemed to be fair in the treatment of their employees.
- **Proposition 10b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are deemed to be fair in the treatment of their employees.

Innovation

The last (eighth) dimension of organizational climate, *innovation*, reflects an organizational context where employees believe change from status-quo can be good, that originality is valued, and risk taking will be encouraged (Koys & DeCotiis, 1991). As noted earlier, management's attitude toward change (often triggered by the introduction of technology innovations) impacts the adoption of these technology innovations (Damanpour, 1991). Some senior management teams may have conservative attitudes toward innovation and associated risk, preferring the status quo and using current or time-tested methods; such organizations innovate only when they are seriously challenged by their competition or by shifting consumer preferences (Miller & Friesen, 1982). By contrast, other senior management teams may be risk prone, actually encouraging and actively supporting the use of innovative techniques to move the organization forward. Such organizations usually try to obtain a competitive advantage by routinely making dramatic innovative changes and taking the inherent risks associated with those innovations. The potentially disruptive features typically associated with the adoption of (radical technology) innovations require an organization context where managers encourage individual members of the organization to take prudent levels of risk, support adoption of technology innovations, and be supportive of changes in their organizations (Dewar & Dutton, 1986). Thus, we would propose:

- **Proposition 11a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in progressive/innovative organizational contexts.
- **Proposition 11b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in progressive/innovative organizational contexts.

B2B APPLICATION DOMAIN AND SUGGESTED RESEARCH METHODOLOGY

Some of the broad domains of B2B application areas that are relevant for mobile-computing and of interest to us for this research would be inventory management, customer relationship and service management, sales force automation, product locating and purchasing, dispatching and diagnosis support to, say, technicians in remote

locations, mobile shop-floor quality control systems, as well as those applications and transactions in supply chain management (SCM) that facilitate the integration of business processes along the supply chain (Rao & Minakakis, 2003; Turban, King, Lee, & Viehland, 2004; Varshney & Vetter, 2001). An example of B2B transactions in the SCM context includes data transmission from one business partner to another through the typical enterprise resource planning (ERP) interactions. Other scenarios may involve the ability to continue working on projects while in transit or the ubiquitous access to documents via "hot spots" or wireless network access (Intel, 2003). Consequently, in light of the fact that a number of application domains have preexisted the Internet, the choice of application areas could be either Internet or non-Internet based.

As noted before, mobile computing is still in a very early stage of its evolution and use within organizations in a B2B context. Although a large-scale field survey would be required to test the research model that we presented, such an approach may not be appropriate in this context due to the exploratory nature of the inquiry proposed here. Therefore, the research methodology that we suggest and propose that researchers use at this stage is a combination of both qualitative and quantitative research for data collection. Rather than a large national random sample, we propose a purposive convenience-based sample of a few (say, 8-12) large and medium-sized corporations with almost equal composition of manufacturing and service sectors. Furthermore, based on secondary information and personal contacts, we would prefer that researchers select an equal mix of corporations that do not (yet) use and those that currently use mobile computing so that we can capture their "intention" and subsequently their "continued use." Although the "social context" or "climate" prevailing within each of these organizations may be a "given reality" at least at a point in time, as observed in most past research, it is the interpretations of this social context/climate that would drive individual actions, especially when the intended/actual behavior (in this case, adoption and use of mobile computing) is not mandatory (Moran & Volkwein, 1992). Thus, in-depth interviews coupled with a questionnaire survey from a number of focal members (about 20 to 25), sampled from multiple functional areas (that are amenable for use of mobile computing devices such as sales and marketing, purchasing, and operations) within participant organizations, should be used to capture individual perceptions of the mobile computing devices and their organization's social context. As argued above, since the rate of diffusion for mobile computing devices for B2B transactions is still relatively small, a convenient sampling approach among organizations that have and have yet to adopt these technologies is appropriate. To ensure relevance and reasonable generalizability of the study findings of the convenience-based sampling suggested by us, participants from each organization should be chosen randomly. A number of statistical techniques such as logistic regression (for the "intention to adopt" stage) and structural equation modeling or hierarchical moderated regression analysis (for the "infusion" stage) would be candidates for data analyses.

CONCLUSION

In this chapter we incorporated the social context of an organization into TAM and proposed an extended model to investigate adoption/use of mobile computing devices for B2B transactions as a technological innovation. We believe that such an extension is appropriate because aspects of social context have in general been found significant with the introduction of new technologies. In particular, a micro-level analysis of this phenomenon for the introduction of new technologies is rare. Since the unit of analysis of this chapter is individual employees, we utilized dimensions of psychological climate to represent

the social context of an organization. The primary objective of this chapter was to posit how perceptions of the social context of an organization would moderate the intention to adopt/use and infusion of a technology innovation.

A key feature of this study is that we examined an information technology that has the potential of becoming a dominant paradigm and platform for future computing applications. As we noted, although mobile computing devices have existed for several years, their use for business-to-business transactions or operating context has not been adequately or systematically explored in academic research. We drew upon theories from the diffusion of innovation, information systems, and organizational behavior literature, among others, to develop our research model and the associated 10 propositions. The model we proposed could serve as a foundation for one stream of IS research that integrates social context of an organization into TAM to examine the vital role of mobile computing devices in electronic commerce.

IMPLICATIONS, LIMITATIONS, AND FUTURE RESEARCH DIRECTIONS

Since the empirical segment of this research has not yet been conducted, we can only conjecture several potential research contributions for researchers and practitioners. One implication that this work has for future research is the exploration of how the social context of an organization may influence the acceptance and spread of an information technology innovation. The social context of an organization has not been applied to TAM, and an extension focusing on the micro-level aspects of the social context have not been widely examined in the literature. By explicitly investigating the social context of an organization, this study extends the innovation adoption and TAM literature base. Our model may be considered a viable and prudent alternative to the UTAUT model. Utilizing a (valid and popular base) model and measures that have become widely accepted among IS researchers allows for researchers in future research to replicate our study and examine other factors of interest. This chapter also addresses the need to explore technology that is close to the "leading edge" (Lyytinen, 1999, p. 26), which is recommended for maintaining the relevance of IS research (Benbasat & Zmud, 1999; Lyytinen, 1999; Orlikowski & Iacono, 2001). Obviously, considerable care and precautions (in the design of the study, operationalization, and evaluation of the measurement properties) will be needed in translating the theoretical model proposed in this chapter into a large-scale empirical investigation that can establish validity and reliability of its results.

The potential implication that this work has for IS practice is that it identifies a number of contextual factors that may influence the acceptance of a technological innovation that an organization wishes to introduce. Mobile computing devices can enhance employee productivity by granting them flexibility in work location and time management (Intel, 2003). Organizations that covet such gains in productivity are likely cognizant of the investments typically at stake when implementing IT innovations. Given that aspects of the social context of an organization are suggested as having a significant role in the introduction of mobile computing devices (Jessup & Robey, 2002; Sarker & Wells, 2003), it is desirable to understand the influence that the social context of an organization plays. Moran and Volkwein (1992) state that focusing on the micro-level aspect of the social context is appealing because it is relatively accessible, more malleable, and the appropriate level to target short-term interventions aimed at producing positive organizational change. This study helps to uncover several future opportunities for organizations since mobile computing devices have the potential to become the dominant paradigm for future computing applications (March et al., 2000; Sarker & Wells, 2003).

Although this chapter offers several potential contributions, several limitations exist. The social context of an organization is operationalized through psychological climate dimensions. The definition of social context that we adopted takes a much broader view than focusing on the individual incorporating traditions from research in the organizational culture literature as well. We feel that it is appropriate to use the social context of the organization to begin the integration of culture and climate literature. It is our opinion that the psychological climate research is the most appropriate theory to support the research model, which presents opportunities in future work to examine other aspects of the social context of an organization that may be influential in the acceptance of a technological innovation. Another limitation of this study (when an empirical investigation is conducted) is that it may obtain retrospective accounts/information from (current) users of mobile computing devices. Retrospective accounts are an issue because individuals may not be able to accurately recall the past. It would be necessary to consider preventive measures on this front to ensure validity and reliability of the results.

REFERENCES

Agarwal, R., & Prasad, J. (1999). Are individual differences germane to the acceptance of new information technologies? *Decision Sciences, 30*(2), 361-391.

Aiken, M., & Hage, J. (1971). The organic organization and innovation. *Sociology, 5*, 63-82.

Astley, W., & Van de Ven, A. (1983). Central perspectives and debates in organizational theory. *Administrative Science Quarterly, 28*, 245-273.

Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: The practice of relevance. *MIS Quarterly, 23*(1), 3-16.

Boudreau, M., Loch, K., Robey, D., & Straub, D. (1998). Going global: Using information technology to advance the competitiveness of the virtual transnational organization. *Academy of Management Executive, 12*(4), 120-128.

Caron, J., Jarvenpaa, S., & Stoddard, D. (1994). Business reengineering at CIGNA Corporation: Experiences and lessons learned from the first five years. *MIS Quarterly, 18*(3), 233-250.

Damanpour, F. (1991). Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal, 34*, 555-590.

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340.

Davis, F., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Davis, F., Bagozzi, R., & Warshaw, P. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology, 22*, 1111-1132.

Davis, G. (2002). Anytime/anyplace computing and the future of knowledge work. *Communications of the ACM, 45*(12), 67-73.

Denison, D. (1996). What is the difference between organizational culture and organizational climate? *Academy of Management Review, 21*(3), 619-654.

Denison, D., & Mishra, A. (1995). Toward a theory of organizational culture and effectiveness. *Organization Science, 6*(2), 204-223.

Dewar, R., & Dutton, J. (1986). The adoption of radical and incremental innovation: An empirical analysis. *Management Science, 23*, 1422-1433.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Glick, W. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review, 10*(3), 601-616.

Hackman, J., & Oldham, G. (1980). *Work redesign*. Reading, MA: Addison-Wesley.

Harrison, A., & Rainer, R. (1992). The influence of individual differences on skill in end-user computing. *Journal of Management Information Systems, 9*(1), 93-111.

Hart, P., & Saunders, C. (1997). Power and trust: Critical factors in the adoption and use of electronic data interchange. *Organization Science, 8*(1), 23-42.

Hartwick, J., & Barki, H. (1994). Explaining the role of user participation in information system use. *Management Science, 40*(4), 440-465.

Hu, P., Chau, P., Sheng, O., & Tam, K. (1999). Examining the Technology Acceptance Model using physician acceptance of telemedicine technology. *Journal of Management Information Systems, 16*(2), 91-112.

Intel. (2003). *Effects of wireless mobile technology on employee productivity.* Intel Information Technology White Paper (pp. 1-20), USA.

Jackson, C., Chow, S., & Leitch, R. (1997). Toward an understanding of the behavioral intention to use an information system. *Decision Sciences, 28*(2), 357-389.

Jessup, L., & Robey, D. (2002). The relevance of social issues in ubiquitous computing environments. *Communications of the ACM, 45*(12), 88-91.

Joshi, K. (1989). The measurement of fairness or equity perceptions of management information systems users. *MIS Quarterly, 13*(3), 343-358.

Joshi, K. (1991). A model of users' perspective on change: The case of information systems technology implementation. *MIS Quarterly, 15*(2), 229-242.

Karahanna, E., Straub, D., & Chervany, N. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly, 23*(2), 183-213.

Kimberly, J., & Evanisko, M. (1981). Organizational innovation: The influence of individual, organizational and contextual factors on hospital adoption of technological and administrative innovations. *Academy of Management Journal, 24*, 689-713.

Koys, D., & DeCotiis, T. (1991). Inductive measures of psychological climate. *Human Relations, 44*(3), 265-283.

Legler, R., & Reischl, T. (2003). The relationship of key factors in the process of collaboration. *The Journal of Applied Behavioral Science, 39*(1), 53-72.

Litwin, G., & Stringer, R. (1968). *Motivation and organizational climate*. Boston: Harvard University Press.

Lyytinen, K. (1999). Empirical research in information systems: On the relevance of practice in thinking of IS research. *MIS Quarterly, 23*(1), 25-28.

Lyytinen, K., & Yoo, Y. (2002a). Issues and challenges in ubiquitous computing. *Communications of the ACM, 45*(12), 63-65.

Lyytinen, K., & Yoo, Y. (2002b). Research commentary: The next wave of nomadic computing. *Information Systems Research, 13*(4), 377-388.

March, S., Hevner, A., & Ram, S. (2000). Research commentary: An agenda for information technology research in heterogeneous and distributed

environments. *Information Systems Research, 11*(4), 327-341.

Mathieson, K. (1991). Predicting user intentions: Comparing the Technology Acceptance Model with the Theory of Planned Behavior. *Information Systems Research, 2*(3), 173-191.

Mayer, R., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734.

McKnight, D., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334-359.

Miller, D., & Friesen, P. (1982). Innovation in conservative and entrepreneurial firms: Two modes of strategic momentum. *Strategic Management Journal, 3*, 1-25.

Moore, G., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

Moran, E., & Volkwein, J. (1992). The cultural approach to the formation of organizational climate. *Human Relations, 45*, 19-47.

Nelson, D. (1990). Individual adjustment to information-driven technologies: A critical review. *MIS Quarterly, 14*(1), 79-98.

Orlikowski, W. (1993). Learning from notes: Organizational issues in groupware implementation. *The Information Society, 9*, 223-250.

Orlikowski, W., & Iacono, C. (2001). Research commentary: Desperately seeking the "IT" in IT research—A call to theorizing the IT artifact. *Information Systems Research, 12*(2), 121-134.

Pareek, U. (1987). *Motivating organizational roles*. New Delhi, India: Oxford and IBH.

Pfeffer, J. (1982). *Organizations and organizational theory*. Boston: Pitman.

Ramamurthy, K., Premkumar, G., & Crum, M. (1999). Organizational and inter-organizational determinants of the EDI diffusion: A causal model. *Journal of Organizational Computing and Electronic Commerce, 9*(4), 253-285.

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Rogers, E. (2003). *Diffusion of innovations* (5th ed.). New York: The Free Press.

Rosenthal, D., & Crain, R. (1963). Executive leadership and community innovation: The fluoridation experience. *Urban Affairs Quarterly, 1*, 39-57.

Sanders, L., & Courtney , J. (1985). A field study of organizational factors influencing DSS success. *MIS Quarterly, 9*(1), 77-93.

Sarker, S., & Wells, J. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM, 46*(12), 35-40.

Satyanarayanan, M. (1996). Fundamental challenges in mobile computing. *Proceedings of the ACM Symposium—Principles of Distributed Computing*, Philadelphia.

Schneider, B. (Ed.). (1990). *Organizational climate and culture*. San Francisco: Jossey-Bass.

Straub, D. (1994). The effect of culture on IT diffusion: E-mail & fax in Japan and the U.S. *Information Systems Research, 5*(1), 23-47.

Swanson, E. B., & Ramiller, N. C. (2004). Innovating mindfully with information technology. *MIS Quarterly, 28*(4), 553-583.

Taylor, S., & Todd, P. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research, 6*(2), 144-176.

Turban, E., King, D., Lee, J., & Viehland, D. (2004). *Electronic commerce: A managerial perspective*. Upper Saddle River, NJ: Prentice-Hall.

Van de Ven, A. (1986). Central problems in the management of innovation. *Management Science, 32*, 590-607.

Varshney, U., & Vetter, R. (2001). A framework for the emerging m-commerce applications. *Proceedings of the 34th Hawaii International Conference on Systems Sciences.*

Venkatesh, V. (1999). Creation of favorable user perceptions: Exploring the role of intrinsic motivation. *MIS Quarterly, 23*(2), 239-260.

Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the Technology Acceptance Model. *Information Systems Research, 11*(4), 342-365.

Venkatesh, V., & Davis, F. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science, 46*(2), 186-204.

Venkatesh, V., & Davis, F. D. (1996). A model of the antecedents of perceived ease of use: development and test. *Decision Sciences, 27*(3), 451-481.

Venkatesh, V., & Morris, M. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly, 24*(1), 115-139.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Zammuto, R., & O'Connor, E. (1992). Gaining advanced manufacturing technologies' benefits: The roles of organization design and culture. *Academy of Management Review, 17*(4), 701-728.

Zmud, R. (1979). Individual differences and MIS success: A review of the empirical literature. *Management Science, 25*(10), 966-979.

Zmud, R. (1982). Diffusion of modern software practices: Influence of centralization and formalization. *Management Science, 28*, 1421-1431.

ENDNOTE

[1] We use the term *infusion* to refer to diffusion and spread of the innovation within an organization's internal environment.

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 675-693, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 5.30
Mobile Phone Use Across Cultures:
A Comparison Between the United Kingdom and Sudan

Ishraga Khattab
Brunel University, UK

Steve Love
Brunel University, UK

ABSTRACT

Recently, the ubiquitous use of mobile phones by people from different cultures has grown enormously. For example, mobile phones are used to perform both private and business conversations. In many cases, mobile phone conversations take place in public places. In this article, we attempt to understand if cultural differences influence the way people use their mobile phones in public places. The material considered here draws on the existing literature of mobile phones, and quantitative and qualitative work carried out in the UK (as a mature mobile phone market) and the Sudan (that is part of Africa and the Middle East culture with its emerging mobile phone market). Results indicate that people in the Sudan are less likely to use their mobile phones on public transport or whilst walking down the street, in comparison to their UK counterparts. In addition, the Sudanese are more willing to switch off their mobile phones in places of worship, classes, and meetings. Implications are drawn from the study for the design of mobile phones for different cultures.

INTRODUCTION

Economic globalization and the widespread use of mobile phones have changed the way people live and manage their lives, and cut down the virtual distance between countries, regions, and time zones. New ways of using mobile phones are constantly emerging (e.g., downloading music to listen to on the train), and the pervasive use of mobile phones in public places for private talk

(both business- and socially-oriented) is a clear example of how mobile phones are changing our economic and social lives. As a result of this, there is an emergent body of research on the use of mobile phones in social spaces. For example, Ling (2004) highlights how their use in public places has raised questions of what the appropriate or inappropriate behaviour is in public places. In this study, he found that people perceived mobile phone use in places such as restaurants as unacceptable, partly because mobile phone users tend to talk louder than usual so that people nearby feel intruded upon, embarrassed, and have a sense of being coerced into the role of eavesdropper on a private conversation.

Research has also shown that mobile phones can occupy concurrent social spaces, spaces with behavioural norms that sometimes conflict, such as the space of the mobile phone user, and the virtual space where the conversation takes place (Palen, Salzman, & Youngs, 2000). This feeling of conflict has led researchers in this area to propose that the use of mobile technology in public places is creating a new mixture of public and private space that has yet to be accommodated by for users of mobile technology and bystanders in terms of what is acceptable or unacceptable behaviour.

This phenomenon has been analysed predominately using concepts drawn from Goffman's analysis of social interaction in public places (Goffman, 1963). In this work, Goffman suggested that people have specific "public faces" and personas for different public social locations. The idea behind this is that individuals have rules that determine their behaviour in public places, or what Burns (1992) refers to as the "observance of social propriety." For example, Murtagh (2001) presented findings from an observational study of the nonverbal aspects of mobile phone use in a train carriage. Murtagh found that changing the direction of one's gaze—turning one's head and upper body away from the other people sitting next to you in the carriage—was a common feature of mobile phone behaviour on trains. These behavioural responses were seen as being indicative of the subtle complexities involved when using mobile phones in public locations. This study suggests that mobile phone users are actively engaged in trying to distance themselves from their current physical location in order to enter a virtual environment with the person they are having a mobile phone conversation. In relation to this, Love and Perry (2004) used role-play experiments to investigate the behaviour and attitudes of bystanders to a mobile phone conversation. They found that participants had strong views on embarrassment, discomfort, and rudeness. They also report that the actions of those who were put in the position of overhearers followed a pattern: they acted as though they were demonstrably not attending, even though they were all able to report accurately on the content of the conversation.

However, to date, most of the research reported in this area has tended to focus on what is termed the developed world. Mobile phones are also transforming people's lives in the developing world. In Africa, the unreliable and inefficient landline telecommunication infrastructure has made the mobile phone the solitary available communication tool for many people (BBC, 2003). However, as mobile phone use in Africa continues to grow, there is a need for mobile phone companies who are entering this market to consider the possible impact of cross-cultural differences in people's attitude towards mobile phone and service applications.

This article first briefly reviews relevant literature about the use of mobile phones in public places. The concept of culture and cultural models are explored in th second section. In the third section, the methods of this study are presented. Techniques of collecting the data and the procedure of this study are presented in the fourth and fifth sections, respectively. Some key findings from the study are presented and discussed in the sixth and seventh sections with reference to how cultural differences might affect mobile phone

use in public places. Finally, the conclusion of this study is presented in the last section.

WHAT IS CULTURE?

Culture is a complicated paradigm that is difficult to accurately define. According to some researchers, culture must be interpreted (van Peursson, in Evers & Day, 1997). Hofstede (1980) conceptualized culture as "programming of the mind," suggesting that certain reactions were more likely in certain cultures than in others, based on differences between the basic values of the members of different cultures (Smith, Dunckley, French, Minocha, & Chang, 2004). Culture can also be seen as a collection of attributes people acquire from their childhood training. These attributes are associated with their environment, surroundings that influence the responses of people in that culture to new ideas, and practices and use of new technology (such as mobile phones). Given that culture may affect the way people behave and interact in general, Ciborowski (1979) identified a close link between knowledge and culture. In the context of mobile phone communication, it may be argued that culture influences knowledge—or the individual's general experience—therefore affecting, in this instance, their attitude towards mobile phone use in public places.

Another explanation of culture has been offered by Hofstede (1980). He produced a cultural model that focuses on determining the patterns of thinking, feeling, and acting that form a culture's "mental programming." This model has been adopted for the study reported in this article, as researchers in the area of cross-cultural differences and technology use consider it a valid and useful measure of systematic categorization (e.g., De Mooij, 2003; Honald, 1999). In addition, it is also considered to be directly related to the relationship between product design and user behaviour (De Mooij & Hofstede, 2002). An explanation of Hofstede's cultural dimensions is as follows:

- **Power distance:** the extent to which less powerful members expect and agree to unequal power distribution within a culture. The two aspects of this dimension are high and low power distance.
- **Uncertainty avoidance:** discusses the way people cope with uncertainty and risk. The two faces of this dimension are high uncertainty avoidance and low uncertainty avoidance.
- **Masculinity vs. femininity:** refers to gender roles, in contrast to physical characteristics, and is usually regarded by the levels of assertiveness or tenderness in the user. The two aspects of to this dimension are masculinity and femininity.
- **Individualism vs. collectivism:** deals with the role of the individual and the group, and is defined by the level of ties between an individual in a society. The two aspects of this dimension are individualism and collectivism.
- **Time orientation:** deals with the extent to which people relate to the past, present, and future. The two aspects of this dimension are short-term orientation and long-term orientation.

A number of cross-cultural studies have investigated differences in attitudes towards new technology. Smith, French, Chang, and McNeill (2001) carried out a study using Hofstede's model. They adapted the Taguchi method—a partial factorial experimental design method—in order to investigate differences between British and Chinese users' satisfaction and preferences for Web sites. They found significant differences between British and Chinese users in their preference of detailed e-finance product information. For example, Chinese users tended to adopt a

more holistic approach to viewing Web content, as compared to British users.

In another study, Honald (1999) found that German mobile phone users preferred clearly-written and inclusively rich user manuals, whereas Chinese mobile phone users focused on the quality of the pictorial information.

Evers and Day (1997) found that there are clear cultural differences between user acceptance of interfaces for different cultural groups. In their study, they found differences between Chinese and Indonesian users. Indonesians were found to like soft colours, black and white displays, and pop-up menus more than Chinese users. Also, Indonesians seemed to prefer alternative input and output modes (e.g., sounds, touch screens, data gloves, and multimedia) in comparison to the Chinese who preferred the use of many different colours for the interface design.

Despite the importance and the relevance of cultural factors and its impact on the use of global products and services (such as mobile phones), little research has compared the effect of cultural differences on issues such as social usability of mobile phone use in the developing and the developed world. Sun (2003) argues that variation in cultural states will cause different attitudes or ways of using mobile phones.

The practice of the "missed call" is a clear example of how users from different cultures develop their own usage style. The missed call is when the caller places a mobile phone call and purposely hangs up before the recipient can answer the call. Donner (2005) investigated the phenomenon in Sub-Saharan Africa where the missed call is known as "Beeping." He found that users have produced elaborated codes and social messages to be exchanged over the network without bearing any cost—or at least not from those who are in a less secure financial situation.

Another exclusive mobile phone cultural attitude is evident in Bangladesh, Uganda, and Rwanda, where a woman, for example, borrows money to buy a special mobile phone designed for multiple user accounts and rural access. After this, she then buys minutes in bulk and resells them to customers in her village. This programme is funded by the Grameen Bank mainly for use in rural areas (Bayes, von Braun, & Akhter, 1999).

The same idea has mutated slightly in the big cities of Egypt and the Sudan where vendors who own fixed contract mobile phones buy a bulk of talk minutes and resell them in smaller chunks to individuals in order to make a profit. This practice is accommodated by their national phone service providers and is known as "balance transfer." Obviously, this practice cannot be seen in London or any of the developed world cities.

If mobile phone users have different usage patterns, the question that the study in this article addresses is: can we assume that people from different countries use mobile phones in the same way? Thus the question arises: are there any roles for cultural differences in the way people use their mobile phones in public places? Therefore, the attitude of the British (a mature mobile phone user market) and the Sudanese (an emerging mobile phone user market) were examined in relation to their attitudes towards the use of mobile phones in public places.

METHODOLOGY

Participants

88 participants took part in the study: 43 British (22 male, 21 female) and 45 Sudanese (20 male, 25 female), ranging in age from 15 to 63 years old, with the average age of 30 years. All participants were mobile phone users. The range of mobile phone use for the Sudanese participants was from 2-5 years, whereas the British participants had used mobile phones for 4-12 years.

Data Collection

Data was collected in this study using a questionnaire and an interview. The development of the questionnaire went through several stages. First, the generation of the questionnaire was collated by employing an exhaustive review of the literature generally on mobile phones, human-computer interaction (HCI), and cultural issues in mobile phone use. Second, an in-depth session was conducted with participants from both countries (the UK and the Sudan) to develop the questionnaire. Initially, a total of nine Likert-type questions were developed. The scale was then tested for content validity, which can be defined as the extent to which a test actually measures what it is supposed to measure (Rust & Golombok, 1989). This was undertaken using what is known as the judgemental approach, with three mobile HCI experts.

As a result of this process, the questionnaire was subsequently revised to consist of six Likert-type questions. The six Likert statements focused on attitudes towards the use of mobile phones in public places. An example of the Likert statement used in this study is as follows:

Mobile phones should not be switched off during meetings:

- □ Strongly agree
- □ Agree
- □ Neutral
- □ Disagree
- □ Strongly disagree

The attitude scale had a combination of positive and negative statements in order to control for any possible acquiescence effect from participants when they were completing the attitude questionnaire. This is a phenomenon whereby participants in a study may unwittingly try to respond positively to every question in order to help the investigator with their study. This type of questionnaire format is one of the most common methods used to elicit attitudes from users in HCI research (Love, 2005).

In addition to the questionnaire, a semistructured interview was carried out. The interview questions included open-ended and closed questions, and were designed to gather information on the use of mobile phones in public places, the practice of the missed call, and other features such as the use of mobile phone caller ID. The main points that were covered in the interview were:

1. Attitude towards the use of mobile phones in public places.
2. The use of the missed calls types in the two cultures. For example, the type of missed calls used and the social messages sent through the missed call, and how recipients differentiate between these types of missed calls.

Examples of questions covered in the interview were:

How do you feel about using mobile phones on public transport?
How do you feel about using mobile phones in school during classes?
How do you feel about using mobile phones in restaurants?

PROCEDURE

Participants were chosen from an opportunistic sample in both the UK and Sudan and asked to complete the questionnaire and return them to the researcher once they had completed them.

The questionnaires took approximately 15 minutes to complete. At this point, an arrangement was made to interview a subset of the participants who had been selected randomly and volunteered to answer the interview questions. Participants were informed from the outset that the results of

the study would be anonymous, and they would be able to obtain the results of the study from the researcher on request.

RESULTS

An independent sample T test was carried out to compare attitudes towards using mobile phones in public places in the UK and the Sudan. There was a significant difference found in the attitudes for using mobile phones in public transport between the British and the Sudanese ($t=5.99$, $p<0.001$). The British were more willing to use it on public transport than the Sudanese.

Another significant difference was noted between the two countries towards using mobile phones whilst walking on the street. Again, the

Table 1. Attitudes towards the use of mobile phones in public places in the UK and the Sudan

	COUNTRY	N	Mean	Std. Deviation	Std. Error Mean	t	df	P Value Sig 2 tailed	
I would be comfortable using my mobile phone in restaurants	British	42	2.83	1.146	.177	1.325	70.241	.189	
	Sudan	45	2.56	.755	.113				
I would not be comfortable using my mobile phone on public transport	British	42	3.29	1.175	.181	5.925	69.046	.000	***
	Sudan	45	2.02	.753	.112				
I would be comfortable using my mobile phone whilst walking down the street	British	42	3.69	1.070	.165	3.884	82.171	.000	***
	Sudan	45	2.84	.952	.142				
Mobile phones should be switched off in places of worship	British	42	4.45	.861	.133	3.094	51.314	.003	**
	Sudan	45	4.89	.318	.047				
Mobile phones should not be switched off during meetings	British	42	3.88	.968	.149	2.316	69.411	.023	*
	Sudan	45	4.29	.626	.093				
Mobile phones should be switched off in schools during classes	British	42	4.00	1.307	.202	2.552	61.278	.013	*
	Sudan	45	4.58	.690	.103				

**P<0.05, **P<0.01, ***P<0.001*

British were more favourable towards this than the Sudanese (t=3.884, p<0.001). The Sudanese were found to be more willing to switch off their mobile phones in places of worships, meetings, and in schools during classes. Please see Table 1 for a summary of the main results.

In terms of differences between the attitude of the British and the Sudanese males, an unrelated

Table 2. Attitude difference between the Sudanese males in using mobile phones in public places and the British males

	Gender	N	Mean	Std. Deviation	Std. Error Mean	t	df	P value sig 2 tailed	
Mobile phones should be switched off in places of worship	Sudanese Male	20	4.90	.308	.069				
	British Male	23	4.43	.992	.207	2.134	26.761	.042	
Mobile phones should be switched off during meetings	Sudanese Male	20	4.20	.523	.117				
	British Male	23	3.83	1.154	.241	1.397	31.583	.172	***
Mobile phones not to be switched on in schools during classes	Sudanese Male	20	4.50	.827	.185				
	British Male	23	4.17	1.403	.293	.942	36.374	.352	
I would be happy using mobile phones in restaurants	Sudanese Male	20	2.50	.688	.154				
	British Male	23	2.70	1.105	.230	-.706	37.389	.485	*
I would not be comfortable using a mobile phone on public transport	Sudanese Male	20	2.25	.786	.176				
	British Male	23	3.13	1.180	.246	-2.912	38.570	.006	**
I would be comfortable using a mobile phone whilst walking on the street	Sudanese Male	20	3.15	.813	.182				
	British Male	23	4.04	.825	.172	.869	40.330	.001	**

**P<0.05, **P<0.01, ***P<0.001*

T test revealed that the British males are more willing to use mobile phones on public transport and when walking on the street than the Sudanese males (t=-2.912, t=.869, p<.001). Please see Table 2 for a full summary of the results.

Comparing the attitudes of the British and the Sudanese females towards the use of mobile phones in public places—an unrelated T test revealed the British females are more relaxed using mobile phones in public transport than the

Table 3. Attitude differences between females in the UK and the Sudan

	Gender	N	Mean	Std. Deviation	Std. Error Mean	t	df	P value sig 2 tailed	
Mobile phones should be switched off in places of worship	Sudanese Female	25	2.60	.816	.163				
	British Female	19	3.00	1.202	.276	-1.248	30.068	.222	
Mobile phones should be switched off during meetings	Sudanese Female	25	1.84	.688	.138				
	British Female	19	3.47	1.172	.269	-5.408	27.256	.000	***
Mobile phones should not be switched in schools during classes	Sudanese Female	25	2.60	1.000	.200				
	British Female	19	3.26	1.195	.274	-1.955	34.863	.059	
I would be happy to use my mobile phone in a restaurant	Sudanese Female	25	4.88	.332	.066				
	British Female	19	4.47	.697	.160	2.348	24.196	.027	*
I would not be comfortable using a mobile phone on public transport	Sudanese Female	25	4.36	.700	.140				
	British Female	19	3.95	.705	.162	1.929	38.758	.061	
I would be comfortable using a mobile phone whilst walking on the street	Sudanese Female	25	4.64	.569	.114				
	British Female	19	3.79	1.182	.271	2.892	24.322	.008	**

**P<0.05, **P<0.01, ***P<0.001*

Sudanese females (t=2.348,p<.001). Please see Table 3 for a full summary of the results.

INTERVIEW RESULTS

The interview results corresponded with the questionnaire data, indicating that there is a difference between the British and the Sudanese attitudes towards the use of mobile phones in public places. Sudanese were found to be less willing to use mobile phones in public places than their British counterparts. In the interview, Sudanese participants revealed various reasons for their uncomfortable attitude towards the use of mobile phones in public places. For example, some of the participants felt that the use of mobile phones in public transport is unacceptable because it can be disturbing to other people in close proximity to the mobile phone user. As one of the Sudanese interviewees commented:

Using a mobile phone in public places, especially on public transport where you are closely surrounded by people, is not something that you can do comfortably. It is viewed as improper and unacceptable, as it disturbs others.

Another Sudanese interviewee added:

The use of mobile phones on public transport may be considered as a sign of disrespect to others. In particular, to older passengers who you have to respect and act quietly around them.

An added justification that was revealed by Sudanese participants for not feeling comfortable using mobile phones in public places was related to their tight rules in keeping private issues private, as one of the interviewees commented:

The use of mobile phones in public places to discuss private matters can put you in an awkward situation; because most of the people surrounding you will hear your conversation and this attitude in itself is not acceptable in our community. People are not expected to discuss private issues so publicly.

On the other hand, British participants were found to be more comfortable using mobile phones in public places as one of the interviewees commented:

I have no problems using my mobile phone in public places and especially on public transport, as I can make use of time while sitting there doing nothing.

Another British interviewee added:

I use my mobile phone in public places all the time and it does not bother me at all that people are listening to my mobile phone conversations. I do not know them and it is unlikely they are going to know more details about the topic I am discussing.

The results of this study also indicated that Sudanese females were less willing to use mobile phones in public places than British females. Sudanese females felt that the use of mobile phones in public places, especially on public transport, could attract unwanted attention to them in a society that expects females to keep a low profile. This was echoed in one of the Sudanese female interviewee's comments:

I do not like using my mobile phone in public places at all as it only magnetizes others' attention towards me. If you are on the mobile phone in a public place, people start gazing at you unappreciatively.

Another Sudanese female interviewee added:

Usually, I do not use my mobile phone in public places, I prefer to keep a low profile. For me, this attitude is a sign of respect for my self and others.

British females appeared to have different view—most of the interviewees were found to feel more comfortable using their mobile phones in public places. As one of the British interviewees commented:

I prefer to use my mobile phone in public places; it keeps me busy and in a way safe, for example when I want to get my car from the car park when it is dark, I always make sure that I am talking to one of my friends on the mobile phone just in case something happens.

DISCUSSION

The results from the study were interpreted in the light of Hofstede's cultural dimensions to try and gain some insight into the way culture may influence the use of mobile phones in public places.

It appears from the results that the British generally are more comfortable using mobile phones in public places than their Sudanese participants, who are more reluctant to use mobile phones in contexts such as public transport and whilst walking along the street.

The collectivistic culture to which the Sudan belongs to (using Hofstede's criteria) indicates an inclination toward a tightly-knit social framework (Hofstede, 1980). The priority is for the groups' needs, rather than the individual wishes. Therefore, perhaps the use of mobile phones in public places for private talks can be seen as a self-centred act, and quite impertinent for the group needs. The group expects the individual to be considerate to the established social etiquette. The mobile phone user in public transport is expected to adhere to the social protocol and to respect other people's privacy.

Another reason for the British comfortable attitude to mobile phone use in public places may be due to bystanders' nonverbal communication attitude. This concept is highlighted by Goffman (1963) where he refers to it as "civil inattention." Civil inattention refers to the ways in which people acknowledge the existence of others without paying them extra attention; he regarded this as a gesture of respect required from strangers. Lasen (2002a) found that "civil inattention" is clearly present in UK culture: the British tend to avoid open and straightforward looking at other people, and keep away from paying direct attention to others, especially on public transport, such as the Underground. He suggested that this attitude may encourage British mobile phone users to talk more freely outdoors without being concerned about others watching them.

In contrast, in the Sudan, it was noted that "civil inattention" is not clearly evident. Sudanese people tend to look at each other directly. Lasen (2002a) suggested that a lack of proper gaze in certain cultures where "civil inattention" does not rule may be viewed as a lack of respect or ignorance. This lack of civil inattention perhaps justifies the reason behind the Sudanese unwillingness to use their mobile phones in public places, as they are influenced by bystanders' nonverbal communication attitude. One can say the more civil inattention paid to others, the more free and relaxed they might feel towards using their mobile phones, and vice versa.

Another justification for not using mobile phones in public places might be due to the high score that the Sudan attained on Hofstede's uncertainly avoidance dimension. According to Hofstede, cultures with high uncertainty avoidance scores tend to be expressive—people talk with their hands, raise their voices, and show emotions. These characteristics can play a role in decreasing the need to carry out private conversations in public places because people in these cultures know that they tend to talk loudly and expressively, which attract bystanders' attention,

plus there is a high risk of being known to people around you. Another important point is that as Sudanese people in general talk loudly and in an expressive way, this tends to increase the level of external noise for mobile phone users. Therefore, people talking on mobile phones need to raise their voices more to win over competitive speakers. This loud talking may attract bystanders' attention and invite eavesdroppers, which can cause a feeling of embarrassment on the part of the mobile phone user. In addition, mobile phone users may feel that bystanders might disrespect them if they discuss their private matters publicly.

Additionally, the Sudanese attitude might be related to the high score obtained on Hofstede's power distance dimension, where a tight set of social rules are established, and people are expected to follow and respect these rules. For example, the social protocol for behaviour in public places is well recognized in the Sudan, and people are expected to behave in certain ways and not to speak loudly in front of others (especially older people). Private issues should be kept private and dealt with in a private manner and in private settings. It is considered improper to breach these norms. Although in the UK, a social protocol for behaviour in public places also exists, the maturity of the UK mobile phone market may have relaxed or altered people's expectations and acceptance behaviour in public places. Palen et al. (2000) found that a person's attitude towards public mobile phone use changes (becomes more accepting) as their mobile use increases. In addition, Palen (2002) predicted that as adoption of mobile phones increases, people will be less disturbed about proper use, but will still prefer to have "mobile free" zones.

In terms of specific gender differences, Sudanese females were found to be more uncomfortable about using mobile phones in public places in comparison to British females. This attitude fits in with the "feminine" attribute of the Sudan culture suggested by Hofstede (1980), where the prevailing value is caring for others. The UK, in contrast, is judged by Hofstede to be more masculine-oriented, and the dominant values are achievement and success.

Although the Sudanese females practice all their rights in terms of education, work, leisure, and the like, they are looked after and cared for by the whole society. As a result of this caring perception towards females in the Sudanese culture, their attitudes and behaviours are more controlled and guarded as they are expected to follow social protocols more than men. For example, Sudanese females are expected to keep a low profile and deflect attention from themselves by talking quietly—and preferably avoid talking—in public spaces.

On the other hand, according to the results of this study, British females are more comfortable using mobile phones in public places. This may be due to the feminine attribute of the UK suggested by Hofstede (1980) where women are seen as equal to men, and they are expected to look after and guard themselves more autonomously. In contrast to the Sudanese females, British females can use mobile phones in public places as "symbolic bodyguards" (Lasen, 2002b). In this context, mobile phones are used as a technique to defend your private space within areas that are heavily populated with unknown strangers (Cooper, 2000; Haddon, 2000). As Goffman (1963) has remarked, women especially do not like to show themselves alone in public places, because this may indicate that they are not in a relationship: a condition which (1) provides a bad impression of their social status and (2) leaves them in a vulnerable situation which can be acted upon by unknown males. To deal with these situations, the mobile phone is quite useful, as it works as a safety net and indicates that this person has their social networks and is not isolated (Plant, 2002).

The other significant result reported in this study is that the Sudanese are more likely to switch off their mobile phones in places of worship. Measuring these results against the Hofstede

typology, the Sudanese score high on uncertainty avoidance scale—religion is valued and greatly respected. People's attitude towards switching off mobile phones in places of worship in the Sudan is therefore expected. It is also related to the high scores Sudan has on power distance, as roles are set, and religious men are very much valued and respected in the society, so both the Muslim and the Christians in the Sudan tend to be aware of the importance of switching off their mobile phones in places of worship. This result could also be related to the reduced number of people in the UK attending places of worship.

The Sudanese also appear more willing to switch off their mobile phones during meetings than the UK participants. This attitude may be related to their high score in the power distance dimension where people are expected to respect the structure, rules, and the norms of the setting where they are currently present.

As for the British disinclination to switch off their mobile phones during meetings, it might be related to the individualistic feature of the British society, where time is valued, and there is a push for making good use of it. It may also be related to the maturity of British mobile phone adoption where mobile phones have blurred the borders between business and social rules. In relation to this, Churchill (2001) found that the mobile phones in the UK are used to form and maintain both work and leisure relationships.

CONCLUSION

The increased use of mobile phones by people from different cultural backgrounds has become an integral part of our world phenomena, yet to date the impact of cultural differences on the way people use their mobile phones—and its implications on mobile phone design—has failed to be investigated comprehensively. As this article illustrates, mobile phone users with cultural differences were found to use their mobile phones in different ways, and their attitudes may have been influenced by their cultural norms. Although one can argue that cultural norms can be reshaped by technology, results obtained from this study indicate that cultural heritage would appear to influence users' mobile phone behaviour. The results obtained from this study also suggest that mobile phone designers need to develop a richer understanding of culture in order to develop mobile phones that satisfy culturally specific needs, and thus support mobile phone users' in their current and potential future communication activities. This is an issue we intend to explore in the next phase of our research.

REFERENCES

Bayes, A., Von Braun, J., & Akhter, R. (1999). *Village pay phones and poverty reduction: Insights from a Grameen bank initiative in Bangladesh*. Bonn: Center for Development Research, Universitat Bonn.

BBC. (2003). *A report by the Worldwatch Institute in Washington. Mobile phone use grows in Africa*. Retrieved October 9, 2007, from http://news.bbc.co.uk/1/hi/world/africa/3343467.stm

Burns, T. (1992). *Erving Goffman*. London: Routledge.

Churchill, E. (2001). *Getting about a bit: Mobile technologies & mobile conversations in the UK* (FXPL International Tech. Rep. No. FXPAL. TR.01-009).

Ciborowski, T.J. (1979). Cross-cultural aspects of cognitive functioning: Culture and knowledge. In A.J. Marsella, R.G. Tharp, & T.J. Ciborowski (Eds), *Perspectives on cross-cultural psychology*. New York: Academic Press Inc.

Cooper, G. (2000). *The mutable mobile: Social theory in the wireless world*. Paper presented

at the Wireless World Workshop, University of Surrey.

De Mooij, M. (2003). *Consumer behavior and culture. Consequences for global marketing and advertising.* Thousand Oaks, CA: Sage Publications Inc.

De Mooij, M., & Hofstede, G. (2002). Convergence and divergence in consumer behavior: Implications for international retailing. *Journal of Retailing, 78*, 61-69.

Donner, J. (2005). *The rules of beeping: Exchanging messages using missed calls on mobile phones in Sub-Saharan Africa.* Paper presented at the 55th Annual Conference of the International Communication Association: Questioning the Dialogue, New York.

Evers, V., & Day, D. (1997). The role of culture in interface acceptance. In M.S. Howard, J. Hammond, & G. Lindgaard, *Proceedings of the Human Computer Interaction INTERACT'97 Conference* (pp. 260-267). Sydney: Chapman and Hall.

Goffman, E. (1963). *Behaviour in public places. Notes on the social organization of gatherings.* Free Press.

Haddon, L. (2000). *The social consequences of mobile telephony: Framing questions.* Paper presented at the seminar Sosiale Konsekvenser av Mobiltelefoni, organised by Telenor, Oslo.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills: Sage Publications.

Honold, P. (1999). Learning how to use a cellular phone: Comparison between German and Chinese users. *Jour Soc. Tech. Comm, 46*(2), 196-205.

Lasen A. (2002a). *The social shaping of fixed and mobile networks: A historical comparison*. DWRC, University of Surrey.

Lasen**,** A. (2002b). *A comparative study of mobile phone use in London, Madrid and Paris.*

Ling, R. (2004). *The mobile connection the cell phone's impact on society.* San Francisco: Morgan Kaufmann.

Love, S. (2005). *Understanding mobile human-computer interaction.* Elsevier Blueworth Heinemann: London.

Love, S., & Perry, M. (2004). Dealing with mobile conversations in public places: Some implications for the design of socially intrusive technologies. *Proceedings of CHI 2004*, Vienna (pp. 24-29).

Murtagh, G.M. (2001). Seeing the rules: Preliminary observations of action, interaction and mobile phone use. In B. Brown, N. Green, & R. Harper (Eds.), *Wireless world. Social and interactional aspects of the mobile age* (pp. 81-91). London: Springer-Verlag.

Palen, L. (2002). Mobile telephony in a connected life. *Communications of the ACM, 45*(3), 78-82.

Palen, L., Salzman, M., & Youngs, E. (2000). Going wireless: Behaviour and practice of new mobile phone users. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'00)* (pp. 201-210).

Plant, S. (2002). *On the mobile: The effects of mobile telephones on social and individual life*. Motorola, London. Retrieved October 9, 2007, from http://motorola.com/mot/doc/0/267_MotDoc.pdf

Rust, J., & Golombok, S. (1989). *Modern psychometrics: The science of psychological assessment.* New York: Routledge.

Smith, A., Dunckley, L., French, T., Minocha, S., & Chang, Y. (2004). A process model for developing usable cross-cultural Websites. *Interacting with Computers, 16*, 63-91.

Smith, A., French, T., Chang, Y., & McNeill, M. (2001). E-culture: A comparative study of efinance Web site usability for Chinese and British users. In

D. Day & L. Duckley (Eds.), *Designing for global markets. Conference (6th. 2001). Proceedings of the Third International Workshop on Internationalisation of Products and Systems* (pp. 87-100). Buckinghamshire: The Open University.

Sun, H. (2003). *Exploring cultural usability: A localization study of mobile text messaging use.* Paper presented at CHI 2003, Ft. Lauderdale, FL.

This work was previously published in the International Journal of Technology and Human Interaction, edited by B. Stahl, Volume 4, Issue 2, pp. 35-51, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 5.31
Mobile Phone Communication Innovation in Multiple Time and Space Zones:
The Case of Hong Kong Culture

Shirley Chan
City University of Hong Kong, Hong Kong

Douglas Vogel
City University of Hong Kong, Hong Kong

Louis C. K. Ma
City University of Hong Kong, Hong Kong

ABSTRACT

In most parts of the world, it is generally considered impolite or even rude to pick up an incoming mobile phone call and to have a longer (and loud) conversation in public places. Yet this type of interruption is generally acceptable in Hong Kong. This inspired the authors to ask: How does Hong Kong culture impact the perception of mobile phone interruption? This research note is about an ethnographical study on the culture in Hong Kong indicating a more positive perception towards mobile phone interruption. Their research results show that the cultural characteristics of fast pace, deal-making and sense of urgency explain why Hong Kong people are receptive towards such interruption and have the habit of participating in both the physical and mobile spaces at the same time. Their findings also challenge the engaging-disengaging paradox theory - that is, mobile phone users find it difficult to simultaneously engage in parallel activities.

INTRODUCTION

In most parts of the world it is considered impolite or even rude to pick up an incoming mobile phone

call and to have a longer (sometimes rather loud) conversation in public places such as restaurants or subways, but it is especially inappropriate during business meetings. However, this type of interruption is generally acceptable in Hong Kong, a phenomenon which has prompted the authors to ask the question: How does the Hong Kong culture impact the perception of mobile phone interruption?

This research note is an ethnographic study of the culture in Hong Kong, which indicates a more positive perception towards mobile phone interruption. There has been much research carried out on interruption in general as well as on how mobile phones cause interruption. Research has also been done on the role of culture in the usage of mobile phones and the Internet. Nevertheless, there appears to be little research on the relationship between all three elements: interruption, mobile phones, and culture. It is the intention of this study to fill this gap.

LITERATURE REVIEW

Given the attention of past research on the relationship between interruption and mobiles or other technology devices, as well as the role of culture in the usage of mobile phones or the Internet, the social aspect of mobile phone or device usage has generated attention from sociologists and other researchers.

Interruption

"Interruption" has long been a subject of study by communication or sociology specialists. Interruption has been said to be a power device (Ferguson, 1977). In some situations, interruptions create high involvement among the partners involved and help promote feelings of mutual interest, enthusiasm, and solidarity (Tannen, 1981 as cited in Li, 2001). Murata (1994) has identified two broad types of interruptions: intrusive and cooperative.

Interruption and Mobile/ICT Devices

Two relatively recent studies look at why and how mobile phone usage could be considered as unwelcome interruptions in public places. Monk, et al. (2004a) examine the reasons why mobile phones are annoying. Monk, et al. (2004b) find that hearing just one side of the conversation results in a publicly-conducted mobile phone conversation becoming more noticeable and intrusive.

Social Identity and Mobile Phone

Truch and Hulme (2004) opined that mobile phones have the effect of challenging the traditional definition of an individual's social identity in terms of location (the old location-based paradigm) and replacing this paradigm with the new social network-based paradigm. Under this new paradigm, individuals talking on mobile phones have a "second space" or "second identity" while simultaneously still having their "first identity" in the location where they are physically present. McGuigan (2005) considers the strengths and weaknesses of the methods of studying the sociality of the mobile phone as employed in a number of research papers.

Culture, Social Behaviour, and Mobile Phone/Internet

Several studies concern the impact of culture on the social behavioural aspects of mobile phone usage and the Internet. Lee, et al. (2002) conducted online surveys in Korea and Japan and reports cross-cultural differences in the usage patterns of mobile Internet. Hofvenschiold (2003) reports on a study looking at the possible differences in the use of and attitude to mobile phones of British and German university students and young professionals.

The literature review above reflects a definite knowledge gap and begs critique. In spite of the

research done on interruption and mobile phones, culture and mobile phone and Internet usage, few studies have examined the relationship between the three elements of interruption, mobile phone, and culture.

RESEARCH METHODOLOGY: ETHNOGRAPHY

Ethnography is a qualitative research approach setting out to understand the circumstances in which some sets of activities occur—the circumstances which give those activities meaning (Harper, 2000) or studying people in naturally occurring settings or "fields" by means that capture their social meanings and ordinary activities (Brewer, 2000).

Given that the ethnographical approach seeks to understand the circumstances surrounding certain activities that give meanings to those activities, our research question poses, "How does Hong Kong culture impact the Hong Kong population's positive perceptions or receptive attitudes towards mobile phone interruption?" This study explores how Hong Kong's cultural circumstances attribute certain meanings to mobile phone interruption as perceived by its people.

Participant Observation: Hong Kong Cultural Context

Ethnographers have different techniques in gathering data, the most common of which include participant observation and interviewing (Brewer, 2000), which is used in this study. From April to August 2005, the first author observed how Hong Kong people were using mobile phones in public areas such as the subway, buses, restaurants, and theatres, as well at social and business meetings. She paid particular attention to the situations, if any, where people picked up an incoming mobile phone call in the middle of an activity.

Interviews

While participant observation might give some inside views into how Hong Kong people behave in terms of using their mobile phones to interrupt their on-going activities in public places or during a meeting, it does not provide detailed information as to how Hong Kong people perceive such interruption. Interviews with some of these people overcame the limitations of participant observation in this regard. The data, as collected from both participant observation and interviews, thus complemented each other in addressing the research question.

The first author carried out unstructured interviews from July to August 2005 with ten individuals that resided or were residing in Hong Kong for a continuous period of at least 15 years. Five of them were aged between 25 to 39, and the rest between 40 and 65. They were five males and five females. She started off the interview with a "casual chat" tone: "In many parts of the world, people may look at you if you pick up an incoming mobile phone call and talk while you are in a public place or are in the middle of a social or business meeting. But we do this quite often here in Hong Kong, don't you think so? Do you think that might have something to do with our culture?"

DATA ANALYSIS AND FINDINGS

Ethnographic data analysis could be said to consist of various stages: data management, coding, content analysis, qualitative description, establishing patterns in the data, and looking for classification schema to explain the data (Brewer, 2000).

The data collected from participant observation and interviews were read and organised into suitable units on the basis of the research question. Finding the relevant concepts shed light on the data documented (O'Reilly, 2005). Data were

organised according to the three major elements or concepts making up this research question: Hong Kong culture, impact, and positive perception towards mobile phone interruption.

Content analysis of the data was followed by qualitative description–identifying the key events, people, and behaviour and providing vignettes (Brewer, 2000). Content analysis took the form of drawing together segments of data for sub-coding. For instance, data relating to the major code of "Hong Kong culture" was further divided into the sub-codes of "fast pace," "deal-making," and "sense of urgency." Data concerning the major code of "how the impact works" were split into the sub-codes of "pick up the call fast," "catching up with various things simultaneously," and "not miss a call." Data related to "positive perception towards mobile phone interruption" were broken up into the sub-categories of "instant attendance," "multiple role juggling," and "seize every opportunity." The qualitative description was noted in the fieldwork journal during participant observation and during or soon after the interviews. The following is a list of some of the major scenarios or vignettes as noted:

The cashier of a supermarket was working while talking on a mobile phone with her friend.

A gentleman was reading a newspaper in a bus and picked up an incoming call while continuing with his reading.

A teenager was in a cinema shortly before the movie was due to begin. He picked up an incoming mobile phone call while making body gestures to his friend sitting next to him with his hands.

It appeared from participant observation of these scenarios that there was the tendency in Hong Kong to engage in multiple tasks simultaneously with the aid of a mobile phone. Individuals as featured in the above vignettes were engaged in concurrent activities in both their physical space (where they were physically present) and cyber or mobile space (where they talked on their mobile).

Findings: Fast Pace, Deal Making, and Sense of Urgency

Establishing patterns in the data involves looking for recurring themes as well as relationships between the data and then developing the classification schema for explaining the data (Brewer, 2000). It was found that there were close relationships between these (sub-) codes identified during the earlier content analysis stages from which we derived findings that help answer the research question.

All ten informants said that the Hong Kong cultural hallmarks of "fast pace" and "sense of urgency" were the main reasons for picking up the incoming call fast. They did not want to miss a call in order to attain their goal of instantly attending to their business. All ten interviewees used the phrases "fast pace" or "urgent" as the very first words to describe or identify Hong Kong culture. Being "fast" or "urgent" seemed to be a major element of their social identity composition.

In addition, seven out of the ten informants opined that there was also the relationship between the Hong Kong cultural characteristics of deal-making and sense of urgency with the wish of catching up with various things simultaneously—and the mobile phone was a useful tool in reaching the related goals of seizing every opportunity and multiple role juggling.

These informants' responses appeared to coincide with what was observed during the participant observation as mentioned above—that people were engaged in one activity in their physical space (like a supermarket cashier checking out her client) and in another activity in their cyber space (talking to a friend on the mobile phone). The three other respondents' replies are elaborated upon in the next section.

DISCUSSION

The findings from our ethnographic study on the Hong Kong population's mobile phone behaviour show that they cherish multiple role-playing or juggling simultaneously with the intention of maximising the time and space resources they can get from both the physical and mobile spaces to survive in a highly competitive economy.

Culture is not static but dynamic (Tung, 1998) and has certain contextual ramifications. How would these Hong Kong people behave with their mobiles if they were in an environment or context where most people were not happy with such interruption or when they were outside of Hong Kong? The ease of mobile devices to take the user across contexts and cultures makes mobility research particularly challenging and interesting (Blom, et al., 2005).

CONTRIBUTION AND LIMITATIONS

There have been claims that mobile technology users are facing an "engaging-disengaging paradox," where they find it difficult to simultaneously engage in parallel activities, to engage in something new without disengaging from something else–when calls interrupt a conversation in the physical space, the person receiving the call will abruptly disengage from the current conversation and engage in a new one (Jarvenpaa et al., 2005). The findings of our study that Hong Kong people enjoy being engaged in both the physical and mobile spaces simultaneously, juggling between various roles, appear to challenge such claims regarding an "engaging-disengaging paradox."

There are numerous limitations to this study that bear recognition and set the stage for future research. Although two of this paper's authors are Hong Kong natives, differences on other cultural dimensions, e.g., professional culture (Vogel et al., 2001) easily lead to missed elements of importance and open the door to alternative explanations. This raises a number of issues with respect to cognitive processes and complexity that deserve special attention, e.g., the role of technology in knowledge management and collaboration (Stahl, 2006). These are just some of the many issues and questions that we can potentially consider for future research.

CONCLUSION

Hong Kong's cultural features of fast pace, deal-making, and sense of urgency impact on the Hong Kong population's positive perception towards mobile phone interruption. They have an innovative way of communicating on their mobiles and race with time and space–functioning simultaneously in both the physical and mobile spaces. Such "multi-engagement" findings challenge the Jarvenpaa et al. (2005) theory of "engaging-disengaging paradox" relating to mobile technology users' behaviour.

REFERENCES

Blom, J., Chipchase, J., & Lehikoinen J. (2005). Cultural and contextual challenges for user mobility research. *Communications of the ACM, 48*(7), 37-41.

Brewer, J . (2000). *Ethnography.* Buckingham and Philadelphia: Open University Press.

Ferguson, N. (1977). Simultaneous speech, interruptions and dominance. *British Journal of Clinical Psychology, 160*, 295-302.

Harper, R. (2000). *The organisation in ethnography: A discussion of ethnographic fieldwork program in CSCW.* Retrieved August 23, 2005, from http://lair.indiana.edu/courses/l701/papers/harper.pdf.

Hofvenschiold, E. (2003). Determining cultural issues in attitude to and use of mobile phones.

Proceedings of the First Annual GC-UPA Track. Retrieved January 31, 2005, from http://www.gc-upa.de/pdfs/UP2003_11_2_EHofvenschiold.pdf.

Jarvenpaa, S., Lang, K.R., & Tuunainen,V.K. (2005). Friend or foe? The ambivalent relationship between mobile technology and its users. In C.Sorensen, Y.Yoo, K. Lyytinen and J. DeGross (Eds.), *Designing ubiquitous information environments: Socio-technical issues and challenges* (pp. 29-42). New York, NY: Springer.

Lee, Y., Kim, J., Lee, I., & Kim, H. (2002). A Cross-cultural study on the value structure of mobile Internet usage: Comparison between Korea and Japan. *Journal of Electronic Commerce Research, 3*(4), 227-239.

Li, H. Z. (2001). Cooperative and intrusive interruptions in inter- and intracultural dyadic discourse. *Journal of Language and Social Psychology, 20*(3), 259-284.

McGuigan, J. (2005). Toward a sociology of the mobile phone. *Human Technology, 1*(1),45-57.

Monk, A. F., Carroll J., Parker, S., & Blythe, M. (2004a). Why are mobile phones annoying? *Behaviour and Information Technology*, 23(1), 33-41.

Monk, A., Fellas, E., & Ley, E. (2004b). Hearing only one side of normal and mobile phone conversations. *Behaviour and Information Technology, 23*(5),301-305.

Murata, K. (1994). Intrusive or cooperative? A cross-cultural study of interruption. *Journal of Pragmatics, 21*, 385-400.

O'Reilly, K. (2005). *Ethnographic methods.* Abingdon, Oxon: Routledge.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge.* Cambridge, Massachusetts: MIT Press.

Tannen, D. (1981). Indirectness in discourse: Ethnicity as conversational style. *Discourse Processes, 4*, 221-238.

Truch, A., & Hulme, T. (2004). *Exploring the implications for social identity of the new sociology of the mobile phone.* Retrieved August 23, 2005, from http://www.csmtc.co.uk/pieces/Mobile.pdf.

Tung, R. (1998). *Culture and international business.* Retrieved August 30, 2005, from http://www.cbe.wwu.edu/cib/papers/tung.PDF.

Vogel, D., Genuchten, M., Lou, D., Verveen, S., van Eekhout, M., & Adams, T. (2001). Exploratory research on the role of national and professional cultures in a distributed learning project. *IEEE Transactions on Professional Communication, 44*(2), 114-125.

This work was previously published in the Journal of Global Information Management, edited by F. Tan, Volume 15, Issue 4, pp. 79-85, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 5.32
Mobile Networked Text Communication:
The Case of SMS and Its Influence on Social Interaction

Louise Barkhuus
University of Glasgow, UK

ABSTRACT

This chapter introduces a qualitative study of the use of mobile text messaging (SMS) and reflects on how SMS influences social interaction. It describes how this new communication technology is used to maintain social relations and how it generally assists users in their everyday activities. Three issues are highlighted: how users use SMS to overcome shyness, how they use it for micro-grooming, and how they are able to control messages to their advantage. It is argued that SMS facilitates users in their everyday life through the ways it supports awareness and accountability. These characteristics make the communication channel a "social translucent" technology, contributing to its popularity. It is suggested that simple information and communication technologies such as SMS can provide powerful tools in new designs of information and communication technologies.

INTRODUCTION

Telephony is a communication technology that has altered our social practices in many ways, a change that has taken place over many decades (Fischer, 1992). The adoption of mobile telephony relied in many ways upon the century long diffusion of fixed line telephony. Still, researchers have been intrigued by the changing behaviour within many user groups that the mobile phone has brought about. Recent research in particular has looked at behavioural changes as people deal with being only "a phone call away" from each other (Brown, 2002; Katz & Aakhus, 2002). One of the most unlikely successes has been text messag-

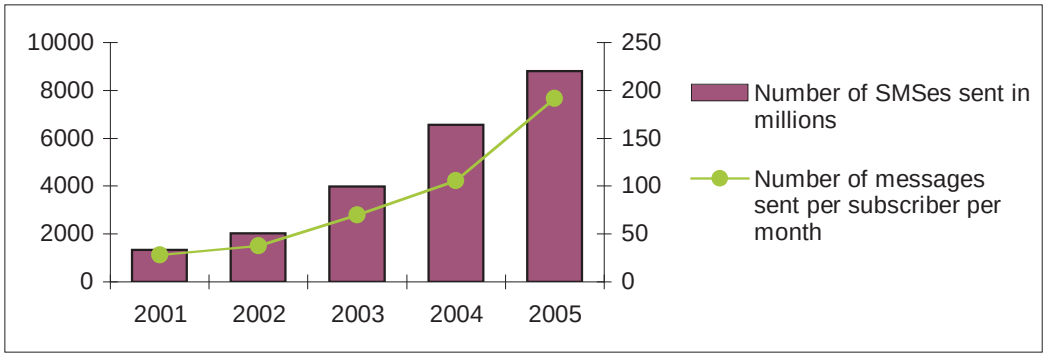
ing or SMS[1] (short message service), which, even with a limit of 160 characters, has become a very common medium of electronic communication in many parts of the world, particularly Europe and many parts of Asia. Text messaging has received considerable attention, with some researchers going so far as to argue that SMS—rather than voice calls—has been the major force in the adoption of mobile phones (Jenson, 2005). The mobile phone is not just acquired for keeping in touch with loved ones during the odd day away from home, but also for the practicalities it solves on an everyday basis, from reminders to buy milk, to arranging a birthday party for a friend. Early research on SMS use suggested that its popularity, especially among teenagers, was due to the controlled cost that SMS provides (Grinter & Eldridge, 2001). However, later research tends to differ from this, emphasizing the efficiency of the asynchronous communication model (Jenson, 2005).

Moving beyond questions of why SMS has become popular, this chapter focuses on how text messages fit into users' everyday lives and existing social practices. The chapter explores in detail how SMS is used among a group of young adults to manage the mundane activities of their lives by focusing on how text messages fit into the lives of users as well as how it both supports existing social practices and creates new ones. Instead of asking why users use "tedious" texting rather than "swift" phone calls (for example, Grinter & Eldrige, 2001; Ito & Okabe, 2005), we approach the medium with the view that mobile phones are now being bought and used as much for text messaging as for voice calls, especially in the Nordic countries where our study took place. This study provides support of how this seemingly simple communication medium is powerful enough to add new structures to users' lives without dominating their daily life. With the changing structures in users' lives, issues of design arise. Underestimating the simplicity of design within communication technologies is a threat to the potential creativity with which the user can shape the technology. Implications are therefore emphasized in relation to the design and adoption of information and communication technologies.

BACKGROUND AND RELATED RESEARCH

SMS was originally implemented into the GSM standard for mobile phone communication in the late 1980s as a replacement for pagers (Kopomaa, 2005). It was envisioned as an extra tool that business people would use on rare occasions to send messages, in a similar way as to how a pager sent a single phone number. The reasoning behind this was partly that messages were, and generally still are, limited to 160 characters and partly that the mobile phone manufactures and carriers could not imagine anyone wanting to type messages with a twelve-button keypad. However, after a slow start, SMS took off at incredible rates in the late 1990s in unison with teleoperators' subsidizing of handsets, making mobile telephones affordable for many people. In 1997, Finland, one of the earliest countries to adopt SMS, even offered it at no cost because of competition among teleoperators (Kopomaa, 2005). Figure 1 shows the increasing number of text messages sent in Denmark in the years before our study. Teenagers, in particular, represented a surprising group for the adoption of mobile telephony and, as will be elaborated upon later, much previous research has focused on this user group. A number of researchers have argued that SMS, rather than the possibility for mobile voice calls, was the main reason for teenagers' high adoption of mobile phones (Ling, 2004). Studies have looked into why teenagers have been so eager to use both mobile phones (Ling, 2004) and text messaging in Europe and Japan (Grinter & Eldridge, 2001; Ito & Okabe, 2005) as well as how they use this mobile communication technology.

Figure 1. The development of SMS traffic in relation to the number of subscriptions in Denmark between 2001 and 2005



Studies of Mobile Phones and SMS

Previous studies of SMS use have often been part of broader studies into the use of mobile telephony, with SMS considered as an alternative to voice communication rather than as a medium in its own right. However, several recent studies have looked specifically at the use of "text messaging". The book, *The Inside Text* (Harper, Palen, & Taylor, 2005) collects a number of studies of SMS use as well as design issues in relation to digital text communication in a broader sense. Several studies of mobile phone use (including SMS) and only SMS use concentrate their observations on teenagers. Grinter and Eldridge (2001), for example, were among the first to explore the use of text messaging among teenagers, investigating why they have been so eager in their adoption of mobile phones and in particular text messaging. They describe how text messaging helps teenagers retain their privacy in a parent-controlled life and how they maintain social relations outside school. Alternatively, Taylor and Harper (2002) focus on the significance teenagers give to text messages themselves, comparing their communication to "gift-giving" practices. Both studies emphasize the "leisure and fun" aspects of the medium among teenage user groups, although Ling (2004) later emphasizes how (virtually) all age groups in Norway use text messages for "micro-coordination" and organizational practicalities. Ling's work is important in how it connects text messaging to broader social practices (such as arranging to meet), yet there is little discussion of the broader social contexts where text messaging takes place, such as public places.

A common finding in the literature is that text messaging increases "ad hoc" coordination (Brown, 2002; Jenson, 2005; Ling, 2004). Ling calls this micro-coordination and describes how messages are often relied upon in situations of coordinating social life, not just for teenagers (Ling, 2004). Another well-cited finding is how text messaging is a tool for users to avoid surveillance or control over their relationships (particularly parental for teenagers) (Elwood-Clayton, 2005; Grinter & Eldridge, 2001; Ito & Okabe, 2005). Since the participants in our study were not under any parental control, this issue was not a factor. However, expectations from others

were found to manifest themselves within other areas of their social life, making this an issue worth exploring. Indeed, one neglected aspect of the earlier literature is how less direct social regulations such as social relationship principles also influence the use of SMS.

In relation to other communication technologies, SMS is a "lean" medium because messages are limited in length and as text only; a lean communication medium is here defined as single channel, compared to rich communication channels that contain, for example, both image and audio, or are synchronous. The limitations of texting make it difficult to compare and relate to multi-channel systems such as videoconferencing or synchronous voice communication. It appears that it would be difficult for such a lean medium to support a socially important and profound interaction of *social translucence.* Social translucence in a communication system is defined as support for coherent behaviour by making participants' activities visible and supporting communication effortlessly, for example. Erickson and Kellogg (2000) describe the advantages of a socially translucent system in that it provides the user with salient characteristics that support coherent behaviour and social activities. They mention three principles for obtaining social translucence: visibility, awareness, and accountability. A medium needs to afford visibility so that users can see the activities of each other; it needs to provide users with an awareness of the other people's presence; and finally, it needs to make users accountable for their interactions. Interestingly, two of these principles are in fact applicable to the SMS medium and thereby contribute to the medium's usefulness, particularly in social situations. Although the users' current activities are not visible to each other, as we will show in this study, messaging provides both awareness among users and accountability as the communication is instantly saved on the phone. As will be elaborated on through the present case study, SMS therefore provides, to a great extent, social translucence.

In our study, we aimed to look at the social management and general practices that govern the use of SMS by young adults in their everyday life. Young adults between 20 and 35 are an interesting group to study for two reasons. First, they have different life styles than teenagers, who live with their parents or at a dormitory in a frequently semi-controlled environment. Second, they often have fewer monetary concerns than teenagers, who most often depend on pocket money and jobs after school. Few studies of SMS include other age groups, noticeably because text messaging is used much less among people over thirty (although this is rapidly changing, as Ling (2004) points out). Moreover, studies that do include users in their early twenties often also include teenagers and thereby study a group with mixed concerns. Examples include socio-linguistic analysis of SMS messages among 12 to 25 year olds (Hård af Segerstad, 2005) and a cultural comparison between French and Japanese users where the French participants were between 15 and 28 years old (Riviere & Licoppe, 2005). The influence of text messaging on teenage culture is important to explore since new practices have been discovered although many of these practices are to be found among young adults as well. These are some of the issues that will be presented here, based on a qualitative study of text message use among young adults.

TEXT MESSAGING IN THE LIVES OF YOUNG ADULTS

Methods

The first part of this study was carried out over two weeks. Twenty-one participants kept a journal every evening, describing messages received and sent that day. In addition, the journal asked par-

ticipants to describe their location when messages were sent/received and the motivation for initiating messaging. Most of the participants had mobile phones that saved both outgoing and incoming messages. This enabled them to remember the messages for the diary in the evening, however, a few participants had to rely on memory in regard to outgoing messages. After two weeks, we conducted more in-depth interviews with seven of the participants, having them elaborate upon motivations and specifics of their SMS habits. These were selected from high-level users among the 21 participants and the aim was to interview a diverse subset of the group. All seven who were asked agreed to be interviewed.

The study took place in Denmark, where the rate of mobile phones was 85 phones per 100 inhabitants at the time of the study (Telecom Statistics, 2003). The participants were young adults, ranging from 21 to 36 years of age. Participants were recruited by way of e-mail lists and personal contacts (none of the participants were personally known to the author prior to the study), the main criterion being that the participants had a mobile phone.

We aimed for a diverse set of participants, not a representative set, and while this naturally limits generalization among SMS users, a purposeful selection enables insight into information rich cases, desirable in a qualitative study such as this. None of the participants had owned a mobile phone for less than two years. The participants were a mix of students (undergraduate and graduate) and young professionals. The students were mainly graduate students studying for their master's degree in subjects such as information science and political science while the professionals worked in jobs as varied as painter, waiter, and forester. A characteristic that we aimed for in recruiting participants was having an "adult life style"; this included having their own income and living either by themselves or with a partner or roommate. By studying independent adults, limitations that apply to teenage groups were minimized, and the study would provide insights into a group with a more consistent life style.

GENERAL TRENDS OF MOBILE TEXT COMMUNICATION

While the participant's diaries gave a good, if basic, impression of how SMS fits into their everyday life, the interviews provided a better understanding of motives for use. SMS was generally used for the coordination and up keep of social life, with some use of texting for work coordination. The more messages participants wrote per day, the more diverse uses text message were put to; in other words, when participants used many text messages, they communicated with a larger number of different people in different relationships

Table 1. Participant demographics

Participants	**Diary study**	**Interviews**
Number (male/female)	21 (9/12)	7 (3/4)
Age range	21-36	21-32
Students/non-students	14/8	4/3
Living with partner or room mate/living alone	9/12	3/4

such as work and family. All participants used SMS on an everyday basis. The average was four messages sent and four received per day; however the participant's level of use differed considerably. Five participants averaged only about three and a half sent and received messages per day; two participants sent and received on average 19 and 20 messages per day. Most participants sent as many messages as they received except for a few exceptions; one woman for example sent on average nine and received only four messages per day.

Relationships

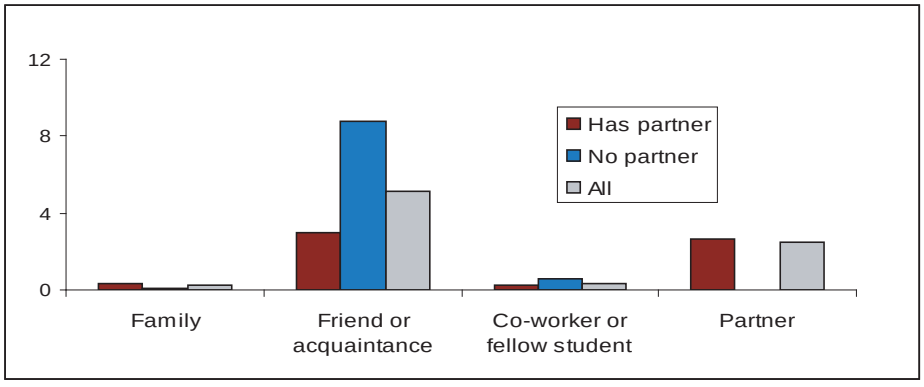
The participants were asked to record to whom they sent and received messages according to relationship. The graph in Figure 2 identifies the average number of messages per day that participants reported sending and receiving, according to whether they reported having a partner or not. As shown, participants mostly communicated with friends or acquaintances and significant others. Only six of the 21 participants texted with members of their family during the two weeks, most often siblings and in one case the participant's mother. However, all of the participants who had a significant other, communicated with him or her almost every day.

The interviews supplied a more nuanced description of the participants' core "SMS group". For one female participant, aged 25, the group was very broad; she was the head of a political organization and often coordinated meetings over SMS. Because of this coordination with other members, she received almost three times as many text messages as she sent. She said that SMS enabled her to control her communication and that she could not imagine having to coordinate this via voice communication or e-mail. A male participant, aged 21, on the other hand, used SMS to communicate mainly with friends of both genders to coordinate evening arrangements and to meet up with friends in a casual way. Another participant used SMS on several occasions to reach his manager when the manager was out of the office or away on an international business trip.

Messaging Purpose

Text messaging was used for both social organization and professional communication. A very general taxonomy of message topics was proposed in the diary where participants were asked to

Figure 2. The types of relationships SMS supported for our participants



characterize their messages as either coordination and practical information, social-based, or other. During analysis, this characterization was checked from the content of the messages. This showed that 55% of the messages were regarding coordination and practical information. Examples of this type of message include organization of future meetings or real-time meetings ("We are at the Mexi-Bar") and a participant coordinating gift buying with her sister. This finding is similar to what other studies report; Ling, for example, found that many users' messages mainly regard micro-coordination (Ling, 2004). Thirty-three percent of the messages were socially based and included goodnight messages and non-essential politeness ("Have a good vacation, see you soon"), and 12% were uncharacterized according to the two previously-mentioned categories.

Similarly to other communication technologies, the social context of the user influences how SMS is used. But where the context of stationary media such as PC-based e-mail or instant messaging is more predictable, mobile text messaging context is more varied. The mobile aspect in this regard facilitates a transformation in users' behaviour and attitude. One way this change was illustrated in the case study was through the messages that, according to the participants, would never have been passed on by other means if SMS had not been available. One interviewee, a 25-year old graduate student, said that she often sent little "I will be a bit late" messages, in situations where she would never have called. SMS was discrete and practical for such purposes, and the majority of participants confirmed this. SMS was used to manage participants' busy life style in relation to both work/study and social obligations.

SOCIAL CONTEXT OF SMS

In other studies, the use of SMS has been found to contribute to more loose social arrangements in the social lives of users (Ling, 2004; Ito & Okabe, 2005). We found that loose social arrangements have extensive consequences to the participants' behaviour and therefore focus the analysis on user's social life and everyday practice as well. Instead of looking at SMS messages in isolation, we find it essential to describe the participants' SMS messages in relation to their social practices.

SMS supports ordinary and existing social practices in new ways. For our participants, SMS was valuable not because it was hugely different from other existing communication technologies such as voice over phone or e-mail, but rather for how it supported more *subtlety*, *spontaneity*, and *mobility* in their existing communication. We found three characteristics of their life style were important to the SMS users, each one relating closely to how they perceive that they have changed since SMS became available to them. First, many participants used SMS to overcome shyness; second, most of them practiced social grooming with messages that would not have taken place without SMS; and third, participants used SMS to control their communication with others. Each of these parts is elaborated upon next.

Overcoming Shyness

Shyness is part of many people's personality, and our participants were not much different. Many of them were shy about talking in public on their mobile phone, and some expressed reservations about calling people they were not very close to or perhaps wanted to become more close to. SMS provided a simple and unobtrusive method of distance communication, which enabled them to maintain spontaneous as well as frequent contact with others.

Many of our participants commented that they would send an SMS rather than calling to avoid talking in public. As one female interviewee explains: "I don't like to talk on the phone when I am outside, I actually don't like to talk on the phone at all. I just think SMS is so much easier".

Although voice communication in public *was* used by the participants according to their diary and the interviews, several of our interview participants said that they tried to avoid this as much as possible by using SMS. In crowded public spaces, a spoken voice call made at least one side of the communication publicly available to others, and participants sometimes felt they were intruding on the privacy of others by talking on the bus or on the train for example. SMS messages, however, are essentially private. SMS thus offered advantages by supporting communication without attracting attention to the individual.

Another aspect of shyness concerned the actual communication content. In a number of cases, the participant found it easier to communicate invitations to acquaintances and friends by SMS rather than by phone. For example, one male participant aged 21 said that he would have found it difficult to casually call up a friend to ask if she wanted to come along to a club that night, but asking her via SMS was easy. The casualness of SMS corresponds well with the casualness of asking someone to an informal gathering among friends. The participant admitted in the latter part of the interview that he did not like phoning people at all and that he would even not see some of his friends as much if he was not able to SMS them on a regular basis. He explains:

"This girl, Linda, I would never call her, but we often SMS about where we are meeting, say, Saturday night, with the others. Sometime we also just chat, like, during the course of the day. But I don't think I would call her."

SMS enabled this participant to contact a friend where it would have been socially awkward for him to call her. In addition to other studies where SMS is found to assist the development of relationships (e.g., Elwood-Clayton, 2005; Grinter & Eldridge, 2003), SMS supports communication without the commitment and immediate reply required in a telephone call. One can send a one line SMS, or reply at leisure, without having to commit to a spoken conversation that can potentially be awkward.

In the earlier cases, SMS helped users who were shy (although most of our participants confessed to being shy in some way or another); in addition SMS allowed users to carefully manage their interactions turn by turn. This gave the users a sense of control when communicating, and they would have time to think before answering. The control that the asynchronous communication provided was essentially a "remedy" for the shyness expressed in relation to certain people.

The power of SMS is clearly illustrated in how it enables users to keep in contact with people with whom they would not keep as close contact with otherwise. However, participants may not be aware what they would do (or would have done) without SMS. Many studies have shown an increasing level of "ad hoc" arrangements with SMS due to its commonality and spread (Ling, 2004; Riviera & Licoppe, 2005), but no studies point to people being more social because of SMS. This was confirmed by our own study. One participant in the study even reflects on whether she actually *does* see her friends more often and has more "spur of the moment" meetings than she would have had without her mobile phone. First, the participant says that she feels that SMS gives her the possibility to be more spontaneous, but she then questions whether she *is* more spontaneous in regards to social activities:

"Hmmm, I don't know. I don't know if I would have called before [having the possibility to SMS]. I think we are [more spontaneous], but I am not sure if it is always possible, because people actually do a lot of things, so it is often difficult to get it coordinated. Even if I have an hour where I can drink a cup of coffee, it is not always that my girlfriends can do that [at the same time]. So I don't know if it is actually happening that much.

I think it stays with the agreement of 'lets SMS each other when we know we have ehm… some time'. And then time passes…"

Distinguishing between factors that influence users' motivation to use SMS and the consequences of it is important in this case. Where reasons for using SMS may be to better control communication (and thereby overcome shyness), the consequences for users' social life is not that the user socializes with more or different people, but merely the fact that they *experience* a more casual way of meeting up with others. By using SMS, they find that mundane communication in their everyday life is less complicated and intrusive.

Micro-Grooming

An important finding was that SMS was used to support "micro-grooming". Thirty-three percent of the messages in the diary were characterized as "social up-keeping" messages—messages that served no purpose in terms of planning or information aim, but were merely aimed at keeping up socially. Ling (2004) describes these messages as a form of social grooming. For example, the diary data contained messages such as "thank you for tonight" and a participant wishing a friend a good holiday. Because SMS is very "affordable" (both money-wise and effort-wise), our participants emphasized that these social maintenance messages would not have been expressed if they could not be sent via SMS. The "smallness" of SMS was a key aspect of the communication. In this way, these messages were more a form of micro-grooming—a wink or a small note—rather than the engaged level of interaction required by a call.

One example of this micro-grooming was a male interviewee who sent a message asking how his friend did at an exam. He explained that this was not something he would have called his friend about, but because he knew the friend had just had an important exam, he sent the SMS as part of "proper social behaviour". Another participant described in his diary how he and his friends competed to come up with the funniest movie quotes during the day:

"It is wonderful to have contact to a friend, also even if it is just gossip! It shows that they/I, during the course of the day, have thought about each other and done something about it. I got some laughs and so did my friends!"

Another example was a female participant, age 25, who sent a message to a friend living abroad and wished her a good holiday. She described that she would not have called her up or e-mailed her if SMS had not been available, but the possibility of SMS made her send an "extra message" in addition to the conversation they had had three days ago. She found messaging increased the closeness to her friend since they could communicate in an inexpensive and simple way.

These messages added to users' everyday lives and illustrate how people find it important and pleasant to stay in contact with both close and peripheral friends with "micro-grooming" messages. They are part of common politeness and have seemingly grown out of SMS technology. Comparable behaviour is one of "giving regards" to someone else; with SMS, this is being done directly rather than through someone else. It can lead to an awareness of other people's presence but also disappointment when users begin to expect these messages. Like other new communication media, SMS is still in its facilitation phase, meaning different things to different people.

Controlling the Communication

Participants' motivation for using SMS may seem straightforward; participants themselves emphasized the simplicity, discreteness, and

asynchronous aspects of using SMS. However, a closer examination of our interviews revealed an additional factor: the concern that SMS senders gave to how their messages would be received, and the situation the receiver was in when they received the message. This concern, in return, resulted in meticulous composing of messages.

SMS allowed users to request a different level of attention than that of a phone call. This different level could be used to change the *meaning* of a call—for example, from a call asking why someone is late to a message notifying the recipient where they are. One female participant, for example, describes a message in her diary:

[The message was regarding] where exactly we had arranged to meet. We were actually standing at two different entrances [to the theatre]. I SMSed because I didn't want to call in case she was just a bit late. … It was just to say where she could find me, without seeming too impatient.

This participant sent a message that from her point of view was a question asking where the other person was, but in the form of a message about where she herself was. This allowed her to avoid appearing impatient. In composing messages, users gave considerable thought to how they would be received, and this often made SMS the preferred medium for situations in which to get across important messages.

Another participant described texting her flatmate, telling her that she was not coming home that night. She explained in her diary that she used SMS because it seemed casual and it would have been "silly to call". She thereby controlled how her friend received the information by choosing SMS rather than a voice call. While messages are used for fun and non-essential information, such as indicated by for example Grinter and Eldridge (2001), they are also incredibly valuable in how they support this subtlety of communication and respect for social relations.

One of the more cited complications with mobile telephony is the constant availability that users feel they have to live up to, especially in the initial phases of getting accustomed to mobile phones (Brown, 2002; Gant & Kiesler, 2002). In contrast, for our participants, SMS helped adjust the need for availability. By not having to answer a ringing immediately, as is the case with voice communication, participants were able to manage their communication in a controlled way. Although efficiency was the most cited reason for this need to control availability, the desire *not* to talk with a specific person was also important; one participant explained that she had sent an SMS to her mother because she just "couldn't handle the talking". In other situations, knowledge of the receiver's situation influences the choice of medium; another participant, for example, knew that her friend was in a meeting and therefore felt an SMS was more appropriate. She did not address the possibility to postpone the communication, which shows that constant availability is often taken for granted.

Returning to the three criteria for social translucence—accountability, awareness, and visibility—, these characteristics are relevant in relation to the controlling that SMS encourages. First of all, users put such great trust in messages that they hold each other accountable for receiving messages. Although some participants expressed that there were several people they could not use SMS to communicate with (most often because the recipient's lack of texting skills, less often because their lack of a mobile phone), the ones they did SMS with were trusted to always receive and answer messages. It is a stored medium, where both sender and recipient can access messages later, thereby giving it more weight and accountability. In terms of awareness, SMS provides users with, if not a direct awareness, then a perceived awareness of always being close to one another. This is naturally tied closely to the mobile phone as such, where the fact it is mobile

is more important than the fact it is limited to text communication. However, the way SMS is used to keep in touch with micro-grooming messages supports the awareness factor to a great extent. The messages are often sent solely to make the recipient aware that the sender is thinking of him or her. Finally, the notion of visibility is blatantly lacking. The sender of an SMS has no idea of the visual context of the recipient, and the mobile factor means that they could be in any unusual situation. However, this also applies to voice phone calls, and we therefore argue that SMS in a sense allows for this lack of visibility by being discrete. When users are not sure what context the recipient is in, they send a message rather than risk a phone call that would be left unanswered. The SMS medium is therefore a good example of a social translucent system.

MOBILE TEXT COMMUNICATION AS A NEW TYPE OF COMMUNICATION

As described earlier, SMS communication not only adds new behaviour to users' social life style and assists in many mundane everyday social practices, it also functions as a social medium for general up-keeping among friends and sometimes colleagues. It is argued here that SMS constitutes a new type of communication that is already an important and integrated part of the lives of young adults. Although the design and implementation of SMS were not intended for this massive use, it is important to consider the next design directions that these types of communication technologies might take. The possibility to expand text messaging with pictures and even audio or video clips has not gone unnoticed by mobile phone providers in Europe and most mobile phones released today offer multimedia services (MMS) alongside SMS capabilities. These, however, have not proven as popular as plain text messages for a multitude of reasons (such as price, complicated functionality, and lack of interest, among other reasons outside the scope of this chapter), and the consequence is that SMS use is still rapidly increasing around the world. As described in the introduction and illustrated throughout this chapter, SMS is a powerful yet simple medium that affords many types of socially based communication.

Implications for Future Information and Communication Technologies

Designing information and computing technologies (ICT) requires insight into users' everyday practices with communication and where new technologies might spur new practices. It is not necessarily from rigid design considerations that this emerges, as the example of SMS shows. Had mobile phone manufacturers and service providers been able to predict the popularity of mobile text messaging, they would have focused both design and advertising on SMS much earlier than they did. Handsets with full keyboards would have been available and promoted earlier, and marketing would have focused on pricing schemes for texting rather than voice telephony as happened in the middle and late '90s when mobile telephony took off. Because SMS has been shown to alter users' way of communicating and plan social activities, it is imperative to recognize that designers and researchers should not underestimate the power of simplicity in communication technologies. Videoconferencing may seem rich and empowering in many situations because more information is available, through both visual means as well as audio, but the complexity in interpreting more than one channel (audio and video) often results in users rejecting it (Erickson & Kellogg, 2000). This is important to remember when mobile videophones become more common. With a single channel communication medium, limitations are used to the users' advantage, as has been shown here. Design and research should therefore embrace both rich and lean types of communication but in particular the dynamics that make them work for

users. For example, text messages can be defined easily by the features they do not have, such as fast text entry, long message length, tone of voice, and quick interaction. The feature of mobility, however, far outweighs the shortcomings of the medium. Therefore, it is important not to simply count the features of ICTs but rather weigh them in relation to each other.

SMS is just one example of how very simple networked communication can support users' everyday practices and social life. Other text communication such as instant messaging is another example of a communication channel that adds to the range of communication possibilities in many people's life. The synchronicity of this medium makes fluent conversations possible, which is a major difference compared to the message exchange with SMS. The advantage is a more smooth conversation, but the disadvantage in comparison to SMS is that immediate replies are required. Consequently, asynchronicity was one of the issues participants highlighted as a major benefit of SMS. All in all, the two communication media are not directly comparable in their use, but they both function as good examples of essential communication springing from simple media.

The last issue to emphasize in relation to communication technology design is therefore not to misjudge the creativity of the user. If designers assume that limited possibilities for rich communication will yield a limited amount of communication, the creativity of the user is underestimated. Users accept a technology if it supports a social or practical need and corresponds to their present life style. They have shown they are explorative and inventive in their way of using something as simple as texting.

CONCLUSION AND FUTURE DIRECTIONS

Together with the mobile phone, SMS has been one of the more distinct innovations of ICTs in the past century. Not only has SMS altered users' behaviour, it also works as an integrated part of users' social life with few disadvantages in regards to their increased availability (as compared with mobile phones). Although many proclaim that text messages facilitate an increasing spontaneity in the lives of users, it has not been shown that SMS actually improves social settings such as increasing spontaneous meetings. Without directly comparing their behaviour to pre-SMS use (which to the author's knowledge no studies do at present), participants were likely correct in asserting that spontaneity is merely a feeling and not actual behaviour. Still, the perceived value that stems from the use of text messaging should not be disregarded. The study presented here has shown that SMS is used to both build and maintain important social relationships and by doing so, adding value to the lives of the participants.

As described in the introduction, the SMS medium provides both awareness and accountability for users. Users feel that they have a sense of awareness of their friends and acquaintances when SMS is available to them. The expectations that they are only a text message away, as well as the many "micro-grooming" messages create a sense of awareness. Users are held accountable for their communication, since they know that the message was sent to a device that the receiver carries with him/her almost everywhere; the SMS medium is mutually agreed upon to be a legitimate communication channel. The channel has in fact many aspects of being socially translucent. Even the concept of visibility was found to be relevant to the users; by sending text messages to each other and being in constant touch, the users often know where their friends are and what they are doing. In this sense, the medium also supports partial visibility.

This chapter has described three different social contexts and uses of SMS and argued why they are important in relation to future design of ICTs. The controlling of shyness that characterizes SMS use reinforces the advantages of a limited

media of such short text. With limitations, the user does not have to excuse their brevity or find reasons *not* to use the potentially rich, audio or multi-channel medium to its limit; they are limited by the means of the technology. As illustrated by the findings, users do not necessarily want a rich communication medium to interact with on a daily basis. The exposure of private communication in the public sphere was to be avoided as much as possible by not talking on the phone but instead using silent text messaging.

Second, the study pointed to the concept of micro-grooming, a politeness focused type of communication that was part of the daily value that participants contributed to their use of SMS. These messages are not seen as essential for daily activities, but as essential for the maintenance of social relationships. Where users used the simplicity of text for simple but meaningful messages, it is important to realize that the power of the communication medium lies not only in its simplicity but also in its ubiquity. The mobile phone is carried everywhere, and the sender can be fairly certain that the receiver will get the message within a short time. The chance to wish an acquaintance good luck with an exam is only missed when the exam is over, so the greater time span that users have to wish good luck increases the chance that they will in fact do so by SMS.

Finally, we pointed to the controlling of the communication that the simplicity of SMS affords. Users can compose messages concisely without worrying about "accidentally" saying too much or saying the wrong thing, which they are concerned they might do in a voice conversation. They can phrase their messages to suit the situation and thereby control it more than in a synchronous conversation where speed is a factor.

In sum, SMS has shown itself to be a powerful communication technology, not only because of its mobility and simplicity but also because of the value users put into messages and the importance they attribute to this type of communication. The design considerations that arise from these findings are closely connected with the request for simple communication. ICTs might have much potential in relation to synchronous or video-based types of communication, but smaller more mobile devices become more powerful in social everyday settings, despite their communicative limitations.

ACKNOWLEDGMENTS

The author would like to thank the participants of the study for their time. She is also grateful to Anna Vallgårda for assistance with the research. This study was funded by the Danish National Center for IT Research (CIT#313).

REFERENCES

Brown, B. (2002). Studying the use of mobile technology. In B. Brown, N. Green, & R. Harper (Eds.), *Wireless world: Social and interactional aspects of the mobile age* (pp. 3-15). London; New York: Springer.

Elwood-Clayton, B. (2005). Desire and loathing in the cyber Philippines. In R. Harper, L. A. Palen, & A. Taylor (Eds.), *The inside text: Social, cultural and design perspectives on SMS* (pp. 195-218). Dordrecht: Springer.

Erickson, T., & Kellogg, W. A. (2000). Social translucence: An approach to designing systems that support social processes. *Transactions of Computer-Human Interaction, 7*(1), 59-83.

Fischer, C. S. (1992). *America calling: A social history of the telephone to 1940*. Berkeley: University of California Press.

Gant, D., & Kiesler, S. (2002). Blurring the boundaries: Cell phones, mobility, and the line between work and personal life. In B. Brown, N.

Green, & R. Harper (Eds.), *Wireless world: Social and interactional aspects of the mobile age* (pp. 121-131). London; New York: Springer.

Grinter, R., & Eldridge, M. (2001). y do tngrs luv 2 txt msg. In *Proceedings of ECSCW '01,* Bonn, Germany, September 16-20 (pp. 219-238). Bonn: Kluwer Academic Publishers.

Grinter, R. E., & Eldridge, M. A. (2003). Wan2tlk? Everyday text messaging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Ft. Lauderdale, FL, USA, April 5-10 (Vol. 441-448). New York: ACM Press.

Harper, R., Palen, L. A., & Taylor, A. (2005). *The inside text: Social, cultural and design perspectives on SMS*. Dordrecht: Springer.

Hård af Segerstad, H. (2005). Language in SMS – A socio-linguistic view. In R. Harper, L. A. Palen, & A. Taylor (Eds.), *The inside text: Social, cultural and design perspectives on SMS* (pp. 33-51). Dordrecht: Springer.

Ito, M., & Okabe, D. (2005). Intimate connections: Contextualizing Japanese youth and mobile messaging. In R. Harper, L. A. Palen, & A. Taylor (Eds.), *The inside text: Social, cultural and design perspectives on SMS* (pp. 127-143). Dordrecht: Springer.

Jenson, S. (2005). Default thinking: Why consumer products fail. In R. Harper, L. A. Palen, & A. Taylor (Eds.), *The inside text: Social, cultural and design perspectives on SMS* (pp. 305-324). Dordrecht: Springer.

Katz, J. E., & Aakhus, M. A. (2002). *Perpetual contact: Mobile communication, private talk, public performance*. Cambridge, UK; New York: Cambridge University Press.

Kopomaa, T. (2005). The breakthrough of text messaging in Finland. In R. Harper, L. A. Palen, & A. Taylor (Eds.), *The inside text: Social, cultural and design perspectives on SMS* (pp. 147-159). Dordrecht: Springer.

Ling, R. S. (2004). *The mobile connection: The cell phone's impact on society.* San Francisco, CA: Morgan Kaufmann.

Riviere, C. A., & Licoppe, C. (2005). From voice to text: Continuity and change in the user of mobile phones in France and Japan. In R. Harper, L. A. Palen, & A. Taylor (Eds.), *The inside text: Social, cultural and design perspectives on SMS* (pp. 103-126). Dordrecht: Springer.

Taylor, A. S., & Harper, R. (2002). Age-old practices in the "New World": A study of gift-giving between teenage mobile phone users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves,* Minneapolis, MN, USA, April 20-25 (pp. 439-446). New York: ACM Press.

Telecom Statistics (2003). *National IT and Telecom Agency Denmark*. Copenhagen: Ministry for Science, Technology and Development.

ENDNOTE

[1] SMS will throughout this chapter be used to describe text messaging on mobile phones; this is to distinguish between that text messages over computers (instant messaging) and instant messaging services (such as AIM) available through some phones in for example the U.S.

This work was previously published in Designing for Networked Communications: Strategies and Development, edited by S. Heilessen and S. Jensen, pp. 269-287, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Section VI
Managerial Impact

This section presents contemporary coverage of the managerial implications of mobile computing. Particular contributions address business strategies for mobile marketing, mobile customer services, and mobile service business opportunities. The managerial research provided in this section allows executives, practitioners, and researchers to gain a better sense of how mobile computing can inform their practices and behavior.

Chapter 6.1
Comprehensive Impact of Mobile Technology on Business

Khimji Vaghjiani
K & J Business Solutions, Australia

Jenny Teoh
K & J Business Solutions, Australia

ABSTRACT

In this chapter we explore the concept of enterprise, or organisational mobility. We examine how mobility in a business can provide a competitive advantage and enhanced sustainability. Potential industry applications for mobile technology are discussed. We delve further by exploring the growth areas of mobile technologies and outline key success factors for the stakeholders in the mobile technology arena. We assess the many opportunities mobile technology brings to various businesses. Furthermore the impacts of mobile technology on organisations and society are evaluated. We then conclude by outlining various competing mobile technologies available to the market both today and in the future.

INTRODUCTION

The business need for mobility and real-time connectivity are terms that are being used frequently in the technology industry but often without compelling business applications or concise and agreed-upon definitions. While it is important to note that technology on its own is only a means to an end, the purpose, or business objective, to which the most suitable technology is required has to be developed.

Mobility can enhance productivity, as workers are not constrained to their desk in order to perform everyday business tasks — for example, employees can still work whilst waiting in meeting rooms for a meeting to start. Furthermore, it can also help organisations enhance competitive

advantage by allowing the organisation to move toward the concept of real time enterprise (RTE) through real-time data input and quicker decision making regardless of location.

However, despite these benefits, mobility does have its disadvantages, namely blurring the divide between work and non-work life. This is especially evident in the Information Age.

Certain components of the value chain have leaped ahead of other aspects, prohibiting greater uptake of mobile technology. While mobile device manufacturers continue to produce devices at an alarming pace, uptake and adoption has slowed due to factors outside their control. Apart from commercial reasons such as cost, security fears (both real and unfounded) are inhibitors. There are also external factors that can inhibit the movement toward a truly mobile society. The limitations of carrier infrastructure and standardisation issues are just a few. Enablers to greater mobile uptake would be greater applications provided by a single device, with faster connectivity than the traditional GPRS technology.

The Internet has been a blessing in disguise to the apparent and recent surge in the mobile age. Mobile technologies leverage on the strengths of the Internet for services such as data communications and information services. Where will it lead to? What opportunities will it provide to businesses? How will mobile technology impact on daily life? These and other questions will be answered in this chapter.

MOBILE TECHNOLOGY

Mobile technology has evolved from the early '80s. It now includes wired LANs (local area networks), laptops providing a sense of mobility, and computing power in a handbag. In 2003 we saw more and more proliferation of wireless connectivity, and growing wireless hubs have brought with them multiple device manufacturers. Devices include laptops, phones, and PDAs (personal digital assistant), as well as those in the converged marketplace, that is, a PDA combined with a mobile phone. The predominance of higher transmission speeds will allow devices to be more useful in accessing the ever-growing applications.

The growth in devices, infrastructure, applications services, and consumer demand to be "always connected" will exponentially drive mobile technology needs.

Consumers will find greater availability to information, and opportunity to complete transactions such as purchasing goods and services within the mobile environment. This will become increasingly predominant and common over the next few years, before a slowing down or a catching up of one or more of the components of the value chain.

Some Industry Facts

Increased mobile technology and the desire for corporations and executives to be "always on" and "always connected" has led to some exciting industry developments; below are some extracts of these developments.

- Datamonitor (2003) claims that as "shipments of mobile hand-held devices will reach 300 million by 2006, the need for dedicated, specialised functions for business applications will increase for most corporations."
- Forrester (2001) claims between 2001 and 2003 corporations have become more mobile, with usage of certain corporate applications increasing up to 100%, with growth of 50% year-to-year.
- Kwikhand (2003), a Palm solutions provider, claims in its recent report entitled "Logistics & Materials Management," that, "To stay competitive, it is imperative that you drive down costs, accelerate productivity, and synchronise operations. The supply chain generates increasing 'data capture' require-

ments, across the corporation, instantly, and accurately. The corporation needs to be more mobile and aware."

Symbol, the largest worldwide scanning player with global sales of $1.5 billion in 2001, has generated sales of $600 million on scanning devices alone, of which mobile devices are only a small amount. Symbol has recently installed 600 mobile scanners in grocery stores in Europe, allowing customers to immediately scan and pay much faster while providing merchants immediate supply chain information. This application demonstrates benefits of immediacy, real time, and data quality integrity as benefits for corporations. This installation is just one of many reported globally that are being trialed and subsequently implemented. Symbol, a leader in scanning technology, has created market openings in many other industries such as law enforcement, health, and logistics.

- Forrester, in its report entitled "Doctors connect with Handhelds" (2001), claims the hand-held MD solutions will grow to $1.2 billion by 2006, with core applications of uplifting patient data and ordering prescriptions online in real time, reducing multiple handling and errors. Doctors claim, " we can reduce errors, and redundancies and communicate to staff better."
- eMarketer (2000) claims 23% of workers are now considered mobile and spending more then 20% of their working time outside of their offices. While the selling price of PDAs has thus far prohibited the diffusion of hand-helds, this will now change, with prices expected to fall down to $167 by 2004, making the devices more affordable to corporations and individuals. This has been proven partly by the number of PDA manufacturers entering the market, from Palm back in 1996 to Handspring, Sony, and Microsoft, who developed their own operating system in competition with Palm. Since the introduction of the PDA there are now 17 manufacturers operating on Palm, Microsoft, and the Symbian operating systems. The same eMarketer report claims approximately 1.3 million mobile bar code scanning PDA devices will be shipped in 2004.
- A company called Research in Motion (www.rim.com) has recently developed the Blackberry, a Personal Information Management (PIM) tool capturing the "always on" executive market. Blackberry has also been eyeing the mobile data capture market with great interest as an extension to its so far highly successful PIM market.

MOBILITY IN CONTEXT OF ORGANISATIONS

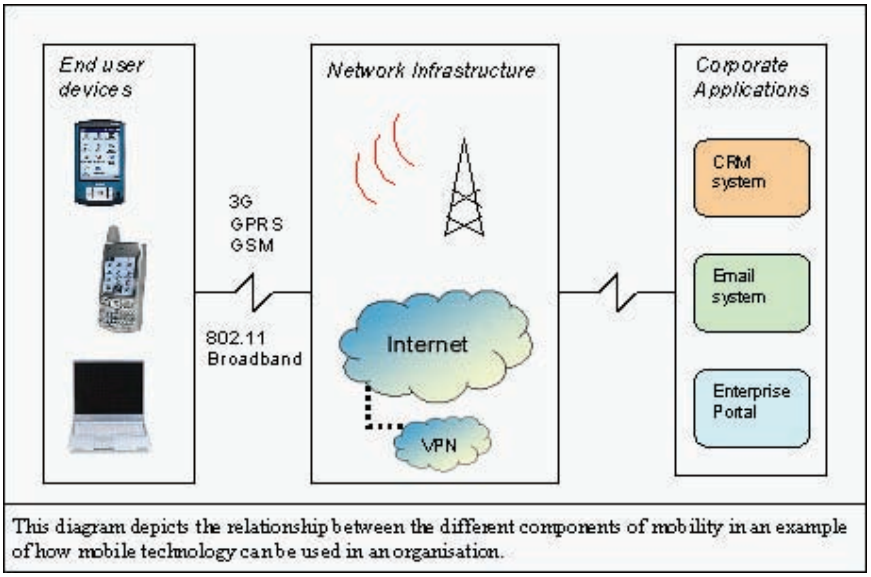
Enterprise mobility is the ability to work anywhere, anytime — at home, on the road, even in the office building away from your desk. To stay competitive, enterprises are mobilizing their businesses and workforce.

There are three components of mobility:

1. Mobile devices, for example, handheld devices, or points of access
2. Technology standards for transmission of data and/or voice, for example, Bluetooth, GPRS
3. Service providers and applications developers

A leading research firm, Roper/NOP, published findings in April 2003 based on a unique global research study (RoperASW & NOP World, 2003). The findings from the project "Business-Critical PCs" indicate that when small and medium businesses (SMBs) deploy wireless applications in innovative ways, whether using tablets, Pocket PCs, or laptop PCs, they are able to reduce costs,

Figure 1. The end-to-end mobile story



This diagram depicts the relationship between the different components of mobility in an example of how mobile technology can be used in an organisation.

provide their staff with more flexibility, and in many cases gain competitive advantage.

Organisations need to ensure that mobility solutions integrate into existing enterprise systems and are capable of extracting data from a wide variety of back-end systems, such as databases, content servers, e-mail systems, customer relationship management (CRM) applications, supply chain management (SCM) applications, and other enterprise software.

INDUSTRY APPLICATIONS FOR MOBILE TECHNOLOGY

Potential industry applications are varied and diverse; below is a list of potential mobile application for various industries:

- **Medical records** management – That is, records could be travelling to various departments, and the scanner can be used to log the exact location of such records.
- **Highway patrol** for on-the-spot-fines – Licences and registration can be scanned immediately, with fines being logged at the central computer. This market needs to be developed further, and much background effort is required by the authorities.
- **Retail stores** – It is believed that in the U.S. Home Depot is using the mobile devices to track and order stock immediately. This is potentially a huge area of the market that is currently being, albeit partially, addressed by Symbol. Other U.S.-based retailers have also started to explore wireless technology for stocktaking and automated ordering of depleted stock.
- **Asset management** – Large corporations that own PCs and printers and so forth need to locate and audit their equipment. These devices are all bar coded; it would be easy and less time-consuming if the auditor simply scanned the bar code rather then having to manually write the serial number for data entry later. The new process would eliminate duplicate serial numbers or equipment that is misplaced, improving data quality.
- **Tracking** – This category has many growing applications for mobile technologies. Track-

ing of stock and inventory, for example, for chain of custody or for the tracking and entry of people at sporting events. Tickets can be issued with bar codes, allowing wardens and staff to monitor and police entry to various events, say, at tennis matches.

- **Ordering –** At restaurants and sporting events. Mobile solutions are being developed for the food and beverages industry, for example, by allowing people at various remote locations or at large sporting arenas to order their meals whilst seated. Each seat could be bar-coded, with an attendant simply taking the order, which is wirelessly transmitted to the central kitchen, prepared, and taken to the customer without the customer having to walk to a take-away station.
- **Home delivery** – A mobile device with the capability to capture data such as credit card or depot card information would potentially have a large market with, say, pizza delivery companies; in fact, Domino's Pizza is known to be trialing the concept.
- **Military applications** – The battlefield could benefit from the ordering of supplies and products in the field directly from the main store, as well as providing inventory management capabilities.
- **Manufacturing** – Chrysler in the U.S. has shown interest in using these devices for the tracking and ordering of stock. This provides various employees the ability to order on-the-spot components when required as well as update stock levels.

These are but some of the industries that are looking at mobile technology as a means to deliver better customer service and enhanced profits. The maturity of end user, and infrastructure in mobile technology, will see greater usage, providing the industry with greater revenue opportunities.

What Predictions for Mobile Solutions?

Mobile technology will enable anytime, anywhere access and connectivity to information, and transactions, from ordering and automatically paying for pizzas to paying for a holiday apartment on the other side of the world to booking a flight with your favourite seat, on the fly, just before you get to the airport.

Businesses will be able to serve customers on site or from remote locations. Corporate staff will be able to access and perform business transaction such as interrogate customer relationship software while on the road, or simply to find a work colleague's contact details by connecting onto the corporate directory from the coffee shop. Placing customer requests from the retail location without having to return to the office will become more of a reality, hence increasing the "immediacy" of customer service. User interaction will be increasingly driven by speech recognition and could bring with it greater security.

WHAT IS THE GROWTH BUSINESS FOR MOBILE TECHNOLOGY?

Mobile technology will become more personalised. The device will no longer be just a mobile phone. It will now include a diary, an address book, a wallet (as seen by recent trial in Japan), and a mobile credit card. The technology will allow greater transmission speeds, lending it self to richer applications such as movies and live broadcast. Mobile technology will be ubiquitous to any specific device. Instead of asking for a Nokia or a Panasonic, consumers may be asking for a "communicator," similar to asking for a copier, rather then a Canon or a Hewlett Packard.

The competitive landscape is becoming customer-focused. Consequently, the customer base is becoming more demanding of the service it expects and receives. This leads organisations to

deliver more service more efficiently, with greater levels of quality, possibly at a higher cost.

More and more organisations are mobilising their resources to better deliver customer service. The core functions of mobility revolve around product ordering, data capture, data entry, and asset management as detailed below:

- Automated product ordering – based on data capture of stock numbers
- Field service ordering and data capture of product
- Sales force automation – product inquiry function
- Reporting – of asset management functions

The main value proposition that these core functions and values bring to mobile solutions is:

- Immediacy of action – data capture and information processing
- Data quality – due to "once-only" data entry
- Speed of service – via end-to-end automation rather then manual actions

The return on investment, or ROI (to be discussed in more detail later), starts to take effect once the above parameters are considered, particularly more for time-, quantity-, and quality-sensitive businesses. The benefits outweigh the cost of technology, cultural change, and training, as businesses become more responsive to customer need and customers in return reward through return business.

Hence the growth areas for mobile devices will be in the area of delivering better customer service and in innovative, and often unique, customer solutions that can be derived from these devices. Suppliers able to better meet the needs of customers around these core values will reap financial rewards in the mobile business.

Product differentiation is intrinsically linked to delivering innovative customer solutions. Consequently businesses will want to adapt new technologies in order to remain competitive and outgrow their competition.

What are the Inhibitors to Mobile Technologies, and How Will They Be Overcome?

Data transmission speeds will be the core inhibitor going forward for greater mobile usage, not applications. Mobile services will be limited by simply the type of application a consumer can possibly perform on a mobile device. For example, downloading a movie on a laptop provides a completely different experience than performing the same exercise on a mobile phone.

Cost of mobile transmission will also be a large inhibitor to the mainstream. While the early adopter and enthusiast will not be deterred by cost, the majority will exercise restraint based upon cost, until price reduction and levelling set in.

The usability aspects of mobile devices, or the lack thereof, will discourage mainstream uptake. However the onus to ease usability is not only on the mobile device designers but also on service providers who also have an equally important role to play.

Online services providers and their respective Web sites need to be re-designed for the smaller screens that are characteristic of mobile devices. It has been found that most are basically scaled-down versions of Web sites designed for PC users – that is, without the graphics and multiple columns. According to Nielsen (2003), to cater to mobile devices, Web sites and services should offer much shorter articles, dramatically simplified navigation, and highly selective features.

The message is clear for service providers — tailor online services and their presentation to the device or risk being left behind by the discriminating consumer.

Organisations currently face the issue of the trade-off between cost of supporting workers using mobile devices and the increased benefits gained through improved productivity. The total cost of ownership increases as mobile workers support their devices on an ad-hoc basis.

A way forward is to implement a centralized management solution. This means that support staff is able to deploy, configure, monitor, and troubleshoot the mobile device systems and applications from a central console manager.

Another key issue to consider is compatibility between different mobile devices. There is a variety of software available on the market, produced by different vendors, for different mobile devices. For example, Hewlett Packard's PDA uses Microsoft Windows CE operating system (OS), while Palm's PDA uses its own proprietary operating system Palm OS. Consumers are faced with many choices. The challenge is to ensure that different peripherals, file formats, and applications are compatible between different vendor's products as well as within a single vendor's product range.

When Will "Value-added" Mobile Technology Become Mainstream, As Distinguished from Voice Mobile Technology?

The notion of "immediacy" of information and transactions will drive the growth of mobile technologies. Mainstream adoption will surge once device prices reduce, similar to what happened with mobile phones, or be offered in different cost models for the consumer. The complexity of user functions needs to be made simpler, and there must be value in the information being accessed, and it must be worth paying for. In addition greater services proliferation will enable mainstream adoption, increase supply and influence cost. This argument is a contradiction in terms; greater functionality (which is a function of the technology) needs to be simple and intuitive.

Organisations' attitude toward adopting value-added mobile technology would be dependent upon several factors:

- **Speed of delivery of services for mobile devices** — Today's consumers understand that time equates to money. It is therefore imperative that services are delivered speedily with minimal delays.
- **Ease of use** — The workforce in organisations generally consists of users with different skill sets based on technology adoption curve. To ensure adoption of the technology, mobility solutions should be easy to use, requiring minimal effort by the consumer.
- **Reliability of service** — As workers have the ability to be always connected, they assume that the network will be available when demanded. Service providers must minimise network dropouts and implement policies to safeguard against any interference
- **Accuracy of transactions** — Service and application providers need to ensure that the integrity of transactions is maintained. For instance, if you asked for specific information, you are delivered that specific information. Or when you send information to a remote server, that information will not be altered during transmission.

Key Factors for Service Providers with Mobile Applications

While mobile applications and services offer enormous opportunities both to end customers and service providers alike, mobility brings with it many challenges. These challenges rest in the devices, carrier capability, infrastructure, and the application service provision.

The following text details the complexity of the wireless end-to-end solution. We can simplify the architecture around three main areas, the device, carrier capability, and application.

Below is an analysis of these challenges and factors that will affect take-up of mobile applications.

Mobile device end:

- **Length of battery life** – Mobile applications rely heavily on the capacity of the battery. Frequent charging deters users from using the technology. On the other hand, a longer life often implies a larger battery, implying a heavier and larger-sized device, which may inhibit take-up. As battery technology improves, devices will become smaller and more versatile.
- **Size of display area** – The size of the display on a phone, a PDA, or a smart phone is limiting for many applications such as watching a movie. However, a small screen size would be ideal for stock quotes and purchases or ordering commodity items.
- **Mechanism for data input** – Performing data entry onto a small handset would severely limit the mobile device usefulness for organizations that require large amounts of data entry. Automated bar code reading or Radio Frequency Identification Tags (RFID – an emerging technology) would provide fast and efficient data capture, reducing the need to manually enter the data and hence improving useability.
- **Overall form factor** – The form factor is the actual size, look, and usability of the device. Form factor used to describe the usability ("feel-ability") of the device measures how well the device performs the functions that it has been designed for, how well it does data entry, and how well it does data extraction work. All these issues are vital to ensuring the best device, with the best form factor, is chosen for the type of work being performed; that is, "fit for purpose."
- **Loss of device leading to loss of vital information** – Recently a device was stolen in the U.S., resulting in vital information of a financial organisation being lost to the underground market. This begs the question of whether the device should be an intelligent one or not – enabled to protect the data it holds through encryption. Should auto-synchronisation be a mandatory feature, although it would place demands for additional telecommunications and increasing costs? Should devices utilise their hard drives to allow for better availability, in case wireless connectivity is unavailable? Loss of mobile service should not prohibit a business from functioning, and while business operations would be best served in some cases with wireless connectivity, working offline (without connectivity to the network) should not stop an organisation from performing limited business functions. These questions must be answered, particularly where data quality, integrity, and immediacy of information is concerned.
- **Processing power** – Devices are becoming faster and smaller, though no sooner have the devices become faster and smaller than the application has become obsolete. This has lent itself to greater applications being available to consumers and will continue to increase with more players entering the mobile market, as profit potential materializes.
- **Cost of procurement** – Mobile devices, be it a PDA or a converged device (voice and data), are still in the early adopter phase, with manufacturers developing many varied devices, searching for an application that will capture the market. A converged device is one where the voice and data functionality are physically combined. It is early for these devices but the proposition to replace with a single voice and data end-user devices is high. Unlike mobile phones, these devices may take a while, say another one to two years before the majority takes up this

technology. The cost will not reduce until there is a larger adoption, which will create a gradual reduction in price.

Carrier end:

- **Types of wireless technology** – Over the next two to three years, the maturity of the technology will determine which is predominant. As noted earlier, a variety of mobile technologies are available from the carriers. Which is most suited to the particular application needs to be assessed to ensure adequacy of speed. Costs will continue to play their role in consumer take up.
- **Security factors** – While mobile phones have "crossed the chasm" as far as security fears are concerned, transferring data, making stock purchases, merchandising purchases, and performing corporate data transactions stills remains the domain of the "technology enthusiast." Security fears seem to be less of a concern, industry pundits talk about the general adoption of wireless devices, and limited trials currently underway are developing greater confidence in the business community and amongst users.

Application end:

- Third-party application availability – The richness of the applications and functionality will also impact the user uptake.

Key Success Factors for Mobile Industry Players

The following are excerpts from leading industry product and service providers addressing the need to assess and understand factors prior to penetrating the emerging mobile market. A short explanation is provided for each aspect covered below.

- **Targeting demographics for particular devices** – That is, wireless tablet PCs would be best suited in, say, an office environment, while a PDA would be best suited to personal use where carrying a large device is prohibitive. A wireless tablet is essentially a laptop-style computer with the capability to perform hand and speech recognition. Hand recognition is provided by simply using the touch screen to write on, as if it were a paper notepad, allowing greater flexibility. Tablet PCs are smaller and lightweight; most manufacturers now have a form of tablet PCs in the market as part of their emerging range of computer products.
- **Development of modular architecture and devices** – "Any device, to any application", a plug-and-play architecture, will be key to mobile technology uptake. Common industry standards will be vital to mobile technology adoption.
- **Developing relationships with service providers and systems integrators** – Industry players will need to work together, to enhance each other's capabilities. Working with others in the value chain to develop functionality would provide greater uptake of mobile applications.
- **Investigate advertising & corporate sponsorships opportunities** — Need to take advantage of various players operating in the mobile space working cooperatively to exploit market opportunities, not only for economic reasons but also for cross sell opportunities.
- **Producing small footprint applications devices** – That is, niche functionalities to cater to specific needs of various market segments. Manufacturers will need to investigate and fully understand market segmentation before investing large amounts of capital. A single standard mobile device may not suit all applications or business services, and hence compatibility and understanding business needs will be key.

- **Continue to develop next generation application** – Innovation is the key to sustainability. Manufacturers who continue to test and trial will eventually dominate the market. Continuous improvements to form factors and applications will reap technology providers with increased consumer uptake, market share, and profitability.
- **Cost of end-to-end solution** – Solutions need to be affordable and effective. Cost of telecommunications needs to be reduced for data transmission, as already evident in voice transmission.

What Will Kill the Potential "Killer Application?"

Mobile technology proliferation should learn from its voice predecessor. More compelling cost models and plans, suited to consumer lifestyles, will foster growth in demand. Poor data speeds and breaks between connections while moving (from mobile access point to another) will hinder the market. A non-ubiquitous network infrastructure, that is, one in which carriers do not cover a live session from each other's customers, will slow down adoption. An ubiquitous network could potentially allow any device from any carrier to be connected seamlessly, providing roaming from one carrier jurisdiction to another without loss of connectivity.

Other factors that will inhibit adoption:

- **Poor data transmission speeds** – Particularly for high imagery and interactive applications. Basic applications with limited graphics and written around simple text-based informational business functionality will develop the initial market. This will create market appetite and develop experience amongst the user community for these devices.
- **Cost of devices** – High end, for at least the foreseeable future, will prohibit mass uptake of usage amongst non-business markets. The business market will have more compelling reasons to adopt mobile technology. There will be some challenges faced by businesses to justify these costs, as seen by some organisations that wish to deploy PIM solutions via mobile PDAs.
- **Security of personal information-based applications** – For example, banking and other financial transactions. While the finance community have been looking for that elusive application that propels mobility, consumers are still, and will continue to be, reluctant to use mobile technology particularly for personal financial transactions. Informational interaction could be the driver for more meaningful uptake in the finance industry.
- **Loss of device** – Losing the mobile device can be considered a serious factor to faster mobile uptake. A mobile device can be a PDA, a smart phone, or a laptop/tablet PC. This can be considered similar to losing a mobile phone; however the above devices are still not mainstream and cost considerably more than a mobile phone, and, more importantly, contain valuable information that, if lost, could cause considerable financial harm or embarrassment to any organisation.
- And finally, a **lack of a simple set of valuable reasons for businesses to use mobile devices** – However this is unlikely as all the evidence suggests otherwise; applications are plentiful. It will nevertheless be imperative for businesses to ensure the application has specific business value, addressing propositions and real business needs. Wireless services will not be adopted simply because of "cool technology" reasons, or at least not by the masses. Compelling business value must be the core objective. Speed of customer service, data quality, time and cost savings, and innovative customer solutions will foster greater wireless uptake.

Social Implications of Mobile Technology

The ability to always "be connected" has many implications on societal values.

Organisations will be able to implement "work from home" programs helping employees optimise flexibility and, as a result, achieve an improved quality of life. The opportunity exists to place a greater emphasis on family society and see a shift in family values as employees find a suitable work-life balance, all the while maintaining or even improving productivity. Furthermore the widespread adoption of flexi-work programs could help reduce traffic congestion, air pollution, accidents, injuries, or deaths associated with commuting to work.

However mobile technology also has negative impacts on society by blurring the divide between work life and social life. The demise of standard working hours for full-time work is already evident in today's society — especially at the highly skilled spectrum of the job market. Furthermore employees are finding an increased need to be multi-skilled in the usage of technology, and there is a requirement to work longer hours in order for the organisation to stay competitive. It is no surprise that children of the Information Age see less of their working parents.

In an article by Shipley (2003), she acknowledges that it is not only the mobile device or the network that is "always on." Sometimes it is the employees who can never switch off and tune out of work. Whether habitual or not, this has the potential to increase stress for employees. This extra stress can have measurably adverse effects on our health, from insufficient sleep to chronic stress fatigue and even increased blood pressure.

Not only must society find a common ground for the definition of work-life balance, individuals, too, have to define the right mix of work and non-work life, bearing in mind that every individual has a different outcome. Individuals need to define their own priorities – the priorities of their careers, of their families, the time for themselves, and the time for others.

Organisational and Workforce Implications of Mobile Technology

There are three main areas that organisations will need to address when evaluating mobile technology's impact on the workforce:

- Cultural changes in the workforce,
- Segmentation of workers into different worker types,
- Occupational health and safety (OHS) issues surrounding the use of mobile technologies.

Enterprise mobility involves a cultural change in the workforce of an organisation. Adapting employees' attitudes toward working in new environments and working in ad-hoc manners requires change management strategies. Organisations need to assess how to train employees to use different types of technologies and how to design a best-fit training program. Organisations need to ensure that employees are able to knowledgeably retrieve, manage, and act upon information in a more flexible and efficient way, such that newer and more efficient business processes can be created, bringing increased levels of productivity.

As organisations move forward and embrace the concept of a truly mobile workforce, they need to understand the needs of their employees. These needs will vary, depending on factors such as role, location, network access, and types of information or applications required. Gartner (2003a) has classified mobile workers into five categories (see Table 1).

These categories all exhibit common patterns of mobility but distinct needs for information, devices, networking costings, support issues, and work patterns. Furthermore Gartner believes that by segmenting users into worker types, organi-

Table 1. Categories of mobile workers

Worker categories	Requirements and characteristics
Alerts workers	Require small amounts of data in short bursts, and one or two button responses. These workers generally use thin clients as their work tools. For example, service notifications via SMS used by field staff.
Message workers	Require high mobility, as these workers are e-mail-centric whether on-site or off-site at customer locations. For example, sales managers need to touch base with their team constantly, employing devices such as a Blackberry to access their email.
Forms workers	Require high degree of connectivity and clipboard or form replacement applications. Work tools can be either thin or thick clients. For example, geomatic engineers, medical staff.
Knowledge workers	Require heavier forms and generally have broader needs than form workers. For example, detailed blueprints and images sent by construction managers to construction workers.
Power workers	Require mobility and almost desktop-like performance in order to access e-mail, but using thick clients. For example, executives.

sations are able to create strategies around these groups, thereby creating effective solutions for the use of mobile technology.

Occupational health and safety (OHS) is one of the crucial objectives of training, in particular educating employees on the correct posture to adopt when using mobile devices and to conduct their own workplace assessment, whether at home or in the office. Research has shown that prolonged use of laptops while travelling on the road hinders blood circulation in the abdominal area and increases the likelihood of "economy class syndrome." It may also lead to chronic back problems and bad posture habits. Furthermore, poor lighting increases the risk of eyestrain.

All these health issues arise due to the extra mobility, allowing employees to work anywhere, especially in places that were not designed for working long periods.

ROI for Businesses to Provide Mobile Transactions and ROI for Customers' Uptake of Mobile Solutions

Consumers are looking for access to information and transaction capability. Businesses are only willing to provide mobile services if the value proposition is compelling enough — that is, there are significant and quantifiable financial improvements associated with the technology adoption.

Return on investments, or ROI models, will be driven by reduced infrastructure costs, both from carriers and applications providers. To date mobile services have been restricted due to these two core reasons. Overcoming consumer resistance to timed mobile "information calls" would encourage greater uptake and hence impact positively the ROI model of the service provider.

A report by Sage Research Inc. for Cisco Systems (*http://www.cisco.com*), prepared in 2001 and entitled "Wireless LANs: Improving Productivity and Quality of Life," outlined productivity benefits around three main areas:

- Time savings
- Flexibility
- Quality of work

Time-based savings: The Sage Research report claims "a wireless LAN use can save up to eight hours per week versus a wired LAN use." Time savings often result in dollar savings,

although how you measure these dollar savings can open a great debate. If these savings do not equate to revenue generation, their value can often be diluted. The time savings also have different values depending on the type of organisation and function an employee performs. Time-based savings with higher-paid employees potentially bring more value to the organisation being armed with mobile capability than does a lower-paid employee base. Time-based cost savings can be substantiated by idle time being better utilised by company executives; for example, a wireless device can provide connectivity while waiting at the airport or while in a cab.

On the other hand, service-based businesses such as tradespeople would benefit from "immediacy of action" by entering job data while on site, receiving payments while at the customers' premises rather than waiting for more traditional paper-based transactions.

Flexibility: Mobility brings the opportunity to, for example, remove your cabled tablet PC and seamlessly connect to the wireless connectivity, whether an employee is in the office, conference room, inventory area, training room, or even the café on ground floor. Apart from within the building, CDMA and other technologies will provide greater mobile range. A sales force can be connected via wireless connectivity while serving a customer many miles from base, hence eradicating the need to return to base to serve customer requests.

Quality of work: Sage Research claims "data can be fed directly from various locations instead of being manually entered at a later date." This improves data quality, reduced entry error and leads to potential cost savings.

Competing Technology Standards that Enable Mobility

Below we look at the various mobile technologies either in use now or slowly gaining momentum in the marketplace.

Wireless Local Area Network (WLAN) Technologies

Note that the transmission range figures (Deutsche Bank, 2003) provided for the IEEE (Institute of Electrical and Electronics Engineers) standards are a general guide only. In reality, it will depend on the mobile devices' antenna gain, the transmit power applied to the antenna, the reception sensitivity of the radio card, and the obstacles between end points.

- IEEE 802.11a

Frequency band: 5 Ghz
Transmission range: 50 ft
Transfer rate: 54 – 100 Mbps

Advantages: Higher transmission rates than other 802.11 standards. This would be an optimal choice for dense networks with bandwidth-intensive applications. The frequency band is not crowded, so there is less interference than in the 2.4 GHz band.

Disadvantages: 802.11a is not as popular as 802.11b. Currently this standard appeals only to niche markets. There is no backward compatibility with other IEEE standards. High capital cost to set up.

- IEEE 802.11b

Frequency band: 2.4 Ghz
Transmission range: 300 ft
Transfer rate: 11 Mbps

Advantages: This is the lowest-cost solution for small wireless networks. Decreasing chipset prices and increased volume of production will lead to notebooks embedded with chipsets. This is the most mature of the IEEE standards.

Disadvantages: Complex technology causes implementation issues such as security. Lack of support for quality of service, and the new 802.11g

has been approved by the IEEE standards committee (12 June 2003) and may shift worldwide acceptance.

- IEEE 802.11g

Frequency band: 2.4 Ghz
Transmission range: 150 ft
Transfer rate: 36 – 54 Mbps

Advantages: This standard allows for more demanding applications like wireless multimedia video transmission. This has a higher transmission speed than 802.11b. Provides interoperability as 802.11b and 802.11g devices can coexist in the same network.

Disadvantages: Total available bandwidth remains the same as 802.11b as restricted to three channels in 2.4 Ghz (this frequency band is getting crowded).

- IEEE 802.11i

This standard is currently under development by the IEEE 802.11 Task Group I. The driving objective behind this standard is to improve the standard and close gaps in current 802.11 WLAN IEEE standards. 802.11i provides a new authentication framework that encompasses several components to address and enhance the current security controls, including the integration of 802.1x (security for wired and wireless Extensible Authentication Protocol authentication).

An interim draft of IEEE 802.11i is now being circulated within the IEEE community, known as Wi-Fi Protected Access (WPA).

Wireless Personal Area Network (WPAN) Technologies

- *Bluetooth*

Frequency band: 2.4 Ghz
Transmission range: 10 m
Transfer rate: 1 Mbps

Advantages: The Bluetooth standard allows communication between mobile devices such as mobile phone and notebooks and peripherals. Users can communicate with another Bluetooth device without the need to configure the hardware or drivers.

Disadvantages: Short-range transmissions range. Despite tremendous momentum Bluetooth has not been adopted widely due to ease-of-use and interoperability issues.

Wireless Wide Area Network (WWAN) Technologies

- GPRS over GSM

Frequency band: Uses GSM's 900 MHz, 1800 MHz or 1900 MHz
Transmission range: Global
Transfer rate: Up to 170 kbps

Advantages: This standard allows for remote communication involving data. For example, PDAs or phones can be used to browse the Internet or e-mail on the road. Users are always connected and can send and receive data without the cost and delay of making a call each time. Users are charged by volume of data. Take-up of GPRS services has been slow in the consumer market but is steadily growing in the business market.

Disadvantages: Transfer rate usually slower as you share with other users within the range of the mobile transmitter.

- 3G

3G wireless systems largely revolve around two ITU (International Telecommunication Union)-approved standards, CDMA2000 and W-CDMA (Wideband CDMA), both of which are developments of CDMA (Code Division Multiple Access). The current dominant markets for 3G are in North America, South America, and parts of Asia-Pacific (South Korea and Japan) only.

Japanese giant NTT DoCoMo's brand name

for 3G W-CDMA services is FOMA (Freedom of Mobile Multimedia Access). In Europe 3G W-CDMA networks are known as UMTS (Universal Mobile Telephony System). In America the favoured technology is CDMA2000.

TD-SCDMA (Time Division Synchronous CDMA) is an upcoming wireless WAN broadband service that has recently attracted significant interest in China as an alternative to W-CDMA and CDMA2000. Universities in China, as well as research organizations, provided major contributions toward the development of TD-SCDMA. This is a major step in helping to bring China into the league of countries defining the future wireless industry, giving a boost to the Chinese wireless industry.

Transmission range: Within 3G network coverage areas.

Transfer rate: Ranges from 144 kbps in rural wide areas to 2.4 Mbps in stationary urban areas

Advantages: 3G allows for high-speed transmission of data and voice both for personal and business applications. Furthermore it supports enhanced multimedia, e-mailing, fax, video-conferencing, and Web browsing. The standard works by allowing multiple users to share radio frequencies at the same time without interfering with each other.

Disadvantages: There is currently uncertainty surrounding the 3G mobile services market as it is considered high risk with dubious returns. For example, in Hong Kong, Hutchison and CSL have taken opposite strategies in their 3G roll out, with the latter taking a "wedding cake approach," according to CSL Chief Executive Hubert Ng (Australian IT, 2003). The competing 3G standard W-CDMA has not found commercial application and has encountered numerous issues, such as expensive auctions for the use of new frequency spectrum, as well as difficult development of handset products. 3G poses significant challenges for call "hand-over" from 3G to 2G networks with factors such as different network configurations, vendor equipment, and even operating conditions making the task difficult. The signals are more prone to interference from hills, buildings, and other tall structures.

Today's market is still in the 2.5G arena, a "light" form of 3G. SMS is a big success factor for 2G and 2.5G, and the MMS market (offered through 2.5G) is following suit as evident with the growing number of photos sent from mobiles. However despite the slow uptake of 3G mobile devices today due to factors such as lack of a mobile handset or high rollout costs, we believe that the dominant uptake of 3G will lead to the phasing out of 2.5G in the upcoming years. Service providers and carriers need to find applications at the right price point that will attract consumers and therefore generate a return on their investment in 3G. The evolution of 2.5G to 3G and beyond into 4G is inevitable.

- WiFi with Wireless Broadband Services

Broadband wireless technology players and network infrastructure providers are seeking to take advantage of the growth and usage of WiFi. According to Gartner (2003b) they are looking at ways to integrate the WAN solutions with WiFi either as a complementary solution or as a backhaul of WiFi.

With the number of hotspots increasing globally, users can leverage the strengths of both technologies. For instance, 3G together with WiFi can ensure a user who is connected in a building can seamlessly roam onto a 3G network when he or she leaves the building and walks out to the street. Wide-area broadband wireless service providers are able to offer flexibility and convenience of access that WiFi technologies are lacking in.

The latter solution involves the use of broadband services as a backhaul for WiFi or PAN. The backhaul could be at a fixed location such as at a hotel or in a moving location such as on a train. According to Gartner the advantage of using this mobile solution is that a WiFi device

is much more universal and less expensive than a wide-area device.

Wide Area Network Technologies

- Fixed Broadband

Broadband allows for high-speed data transfer, providing fast-speed Internet access with high levels of interactivity, facilitating services such as digital video on demand, simultaneous phone and data, and a range of applications and content that can reduce the cost of performing or delivering business services.

Broadband technologies provide organisations the link between their corporate network and the other networks such as the Internet and those of their trading partners. Broadband technologies enable the employee to access their corporate network and perform work tasks from their own home.

Consumers also use broadband technologies such as ADSL, cable or broadband satellite as a mode of accessing the Internet from their homes.

The increasing take up of Digital Subscriber Lines (DSL) technology in Australia has contributed to the number of broadband connections reaching 500,000 and more than doubling from June last year.

RECOMMENDATIONS

Mobile technology, like most, will take time to become mainstream; however, as people have become accustomed to mobile phones, the new wave of wireless mobility will see a faster adoption of data- and voice-converged devices as users become more familiar with the technology and its potential uses. To this end the following recommendations are worth considering:

1. Devices will need to be made simple to use but highly interactive.
2. Fast and efficient logging on to the telecommunications services provider will be key to fast uptake. Delays in dial-up or access based around high security will hinder uptake.
3. Eloquent form factors, appealing and lifestyle-based products will be more successful. For example, Nokia has been the leader in mobile telephony products.
4. Applications that have purpose, that is, provide a real-time service and enhance knowledge will fast-track mobility uptake. Consumers will be more compelled to endure early adoption issues if there is value.
5. Process and time savings will increase mobility uptake. Organisations will constantly look for opportunities to reduce times to serve customers. On the way look for reduction of paper, hence improving the bottom line.
6. Regular new and innovative products and services will continue to drive innovative applications. Organisations will constantly challenge the usage patterns of products. Device manufacturers and service providers will need to be ahead of the industry demands to ensure customer demand for mobility does not wane.

In general wireless technology is just about to take off. As more and more organisations are taking the plunge to "try it out," some will benefit, some will not. If the above points are considered, many will come out singing the praise of wireless technology, but more importantly, their bottom line.

SUMMARY

Mobile technology will offer enormous opportunity for players up and down the value chain, from the device suppliers to carriers to the end

user. The Internet has entered the second phase, the mobile phase bringing with it a mature Internet platform and one where business models are based on customer value propositions and sound return on investments.

Amazingly different types of end-user devices are being developed and providing users with an array of choices. Some time during the maturity of mobile data devices, a set of particular characteristics will develop, forcing out many wild-end user device designs. Much potential exists for the end-user device manufacturer and application service provider who can develop characteristics and uses for mobile data devices.

As part of the evolution process, the technologies that fail to succeed and gain mass-market adoption due to lack of demand or other reasons will vanish from the market.

In this chapter we have covered many aspects of mobile technology with businesses. The core issues have been highlighted, from business applications, issues facing organisations, and benefits. Technology also plays a large role in the overall uptake, and this has been covered in this chapter with a view to highlighting the various types of technologies currently being developed by companies.

Over the next few years there will be winners, and some losers, a consolidation of manufacturers and service providers, a wide array of end-user devices and applications will continue to drive the organisation's need to explore new and innovative ways of serving their customers better. Innovative customer solutions and innovative organisations will continue to drive the market and set the pace of mobile and wireless technology adoption.

The only question is, will growing health concerns curb the enthusiasm?

REFERENCES

Datamonitor. (2003). *The future decoded – global devices to 2006 – A saturated world?*

Deutsche Bank AG. (2003). Wireless LAN, fool's gold? *Global Equity Research, Industry Focus.*

eMarketer. (2000). *PDA market report, global sales, usage, & trends.*

Forrester. (2001). *The Forrester Brief: Enterprise handhelds OS: Advantage Microsoft.* Retrieved November 13, 2003, from *http://www.forrester.com*

Forrester. (2001). *Doctors connect with handhelds.* Retrieved November 13, 2003 from *http://www.forrester.com*

Gartner. (2003a). Enterprises must plan for five categories of mobile workers. *Gartner Research, DF-19-0590.* Retrieved November 25, 2003, from *http://www.gartner.com*

Gartner. (2003b). Wireless WAN broadband service and technology alternatives. *Gartner Dataquest, Market Analysis.* Retrieved November 25, 2003, from *http://www.gartner.com*

Korporaal, G. (2003). Telstra's CSL puts 3G on ice. *Australian IT.* Retrieved November 2, 2003, from *http://australianit.news.com.au/articles/204,8033926%5e15320%5e%5enbv%5e15306,00.html*

Kwikhand. (2003) *Logistics and materials management.* Retrieved December 10, 2003, from *http://www.kwikhand.com/logistics.html*

Nielsen, J. (2003). *Mobile devices: One generation from useful.* Retrieved November 6, 2003, from *http://www.useit.com/alertbox/20030818.html*

Research In Motion & Ipsos Reid. (2001) *Analyzing the return on investment of a Blackberry deployment.* Retrieved November 10, 2003, from *http://www.rim.com*

RoperASW & NOP World. (2003). CMP/HP Technology Innovations Study. Retrieved November 12, 2003, from *http://cmp.agora.com/hp/pdf/research_2.pdf*

Sage Research, Inc., prepared for Cisco. (2001). *Wireless LAN's improving productivity and quality of life.*

Shipley, C. (2003). *Mobility changes everything.* Retrieved November 10, 2003, from *http://www.nwfusion.com/columnists/2003/0825shipley.html*

Symbol Technologies. (2003). Results from 2003 report. Retrieved November 10, 2003 from *http://www.symbol.com*

This work was previously published in Global Information Society: Operating Information Systems in a Dynamic Global Business Environment, edited by Y. Lan , pp. 214-239, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.2
Mobile Business Applications

Cheon-Pyo Lee
Carson-Newman College, USA

INTRODUCTION

As an increasing number of organizations and individuals are dependent on mobile technologies to perform their tasks, various mobile applications have been rapidly introduced and used in a number of areas such as communications, financial management, information retrieval, and entertainment. Mobile applications were initially very basic and simple, but the introduction of higher bandwidth capability and the rapid diffusion of Internet-compatible phones, along with the innovations in the mobile technologies, allow for richer and more efficient applications.

Over the years, mobile applications have primarily been developed in consumer-oriented areas where products such as e-mail, games, and music have led the market (Gebauer & Shaw, 2004). According to the ARC group, mobile entertainment service will generate $27 billion globally by 2008 with 2.5 billion users (Smith, 2004). Even though mobile business (m-business) applications have been slow to catch on mobile applications for consumers and are still waiting for larger-scale usage, m-business application areas have received enormous attention and have rapidly grown. As entertainment has been a significant driver of consumer-oriented mobile applications, applications such as delivery, construction, maintenance, and sales of mobile business have been drivers of m-business applications (Funk, 2003).

By fall of 2003, Microsoft mobile solutions partners had registered more than 11,000 applications including e-mail, calendars and contacts, sales force automation, customer relationship management, and filed force automation (Smith, 2004). However, in spite of their huge potential and benefits, the adoption of m-business applications appears much slower than anticipated due to numerous technical and managerial problems.

BACKGROUND

M-business applications can be classified into two distinct categories in terms of target groups: vertical and horizontal target group (Paavilainen, 2002). Vertical targets are typically narrow user segments, such as filed service engineers or sales representatives. On the other hand, horizontal targets are a massive number of users. For example, mobile e-mail, mobile bulletin

board, and mobile calendar are applications for a horizontal target group, while mobile recruitment tools, mobile sales reporting, and mobile remote control represent vertical applications (see Table 1). Generally, the goal of horizontal applications is to improve communication and streamlined processes in horizontal procedures, such as travel management and time entry. In contrast, the goal of vertical applications is to improve and solve business processes in more detailed and specific areas such as the needs of sales departments. Various vertical and horizontal applications are currently used in a number of industries. Table 2 provides examples of m-business applications in various industries.

THE IMPACTS OF MOBILE BUSINESS APPLICATIONS ON BUSINESSES

The advantages of using m-business applications are mobility, flexibility, and dissemination of m-business applications (Nah, Siau, & Sheng, 2005). Mobility allows users to conduct business anytime and anywhere, and flexibility allows users to capture data at the source or point of origin. In addition, m-business applications offer an efficient means of disseminating real-time information to a larger user population, which consequently enhances and improves customer service. According to Gebauer and Shaw (2004), users valued two things most in m-business applications use: notification, especially in connection with high mobility, and support for simple activities like tracking. The study suggested that the combination of mobility and the frequency with which each task occurred is a primary indicator of the usage of m-business applications.

M-business applications have shown significant impacts and created enormous business values. For example, m-business applications have improved operational efficiency as well as flexibility and the ability to handle situations to current operations (Chen & Nath, 2004; Gebauer & Shaw, 2004). In addition, m-business applications allow users to have access to critical information from anywhere at anytime, resulting in greater abilities to seize business opportunities.

It is very difficult to measure the direct impact of mobile business applications in *productivity* statistics, but according to an OMNI (2005) consulting report, financial services agents executed approximately 11.4% more trade options on an annualized basis with mobile business applications and achieved an average nominal improvement of 3.1% in overall portfolio performance. Also, health care and pharmaceutical filed sales representatives conducted an additional 8.3 physical briefings per week due to mobile business applications. Finally, insurance-filed claims adjusters handled an additional 7.4 claims per worker per week and improved payout ratios by an annual yield of 6.4% per adjuster using mobile business applications. Table 3 provides a list of values created by mobile business applications.

Table 1. Examples of vertical and horizontal mobile business applications (Paavilainen, 2002)

Vertical Mobile Applications	Horizontal Mobile Applications
• Mobile e-mail	• Mobile recruitment tools
• Mobile bulletin board	• Mobile tools for filed engineers
• Mobile time entry	• Mobile sales reporting
• Mobile calendar	• Mobile supply chain tools
• Mobile travel management	• Mobile fleet control
• Mobile pay slips	• Mobile remote control
	• Mobile job dispatch

Table 2. Examples of various mobile business applications (Sources: Chen & Nath, 2004; Collett, 2003; Dekleva, 2004)

	Examples
Hotel	• Embassy Suite: Maintenance and housekeeping crews are equipped with mobile text messaging devices, so the front desk can inform the crew of the location and nature of the repair without physically locating them. • Las Vegas Four Seasons: Customer food orders are wirelessly transmitted from the poolside to the kitchen. • Carlson hotels: Managers use Pocket PCs to access all of the information they need to manage the properties in real-time.
Hospital & Healthcare	• Johns Hopkins Hospital: Pharmacists use a wireless system for accessing critical information on clinical interventions, medication errors, adverse drug reactions, and prescription cost comparisons. • St. Vincent's Hospital: Physicians can retrieve a patient's medical history from the hospital clinical database to their PDA. • ePocrates: Healthcare professionals receive drug, herbal, and infections disease information via handheld devices.
Insurance	• Producer Lloyds Insurance: Field agents can assess the company's Policy Administration & Services System (PASS) and Online Policy Updated System (OPUS).
Government	• Public safety agencies can access federal and state database and file reports.
Manufacture	• General Motors: Workers can receive work instructions wirelessly • Celanese Chemicals Ltd.: Maintenance workers are able to arrange for repair parts and equipment to be brought to the site using wireless Pocket PCs. • Roebuck: Technicians can communicate and order parts directly from their job location instead of first walking back to their truck.
Delivery Service	• UPS & FedEx : Drivers can access GPS and other important information in real-time

FACILITATORS AND INHIBITORS OF MOBILE BUSINESS APPLICATIONS GROWTH

Several factors are expected to contribute to the continued growth of m-business applications. Across the globe, mobile devices such as Internet-enabled mobile phones and personal digital assistants (PDAs) are gaining rapid popularity among businesses and consumers. This rapid penetration of mobile devices can provide strong support for mobile business applications. Employees' demand to access critical business processes and services from anywhere at any time is also a significant driving factor for m-business applications (Chen & Nath, 2004). The traditional methods of wired communication, which have a limited reach and range, are no longer suitable for the fast-paced business environment. Finally, corporate and individual customers, who are demanding more channels for interaction and services, also contribute to the growth of m-business applications.

However, in spite of their huge potential and benefits, the adoption of m-business applications appears much slower than anticipated. Various factors have been offered as explanations for this slow growth, including the immaturity of the wireless technology, the existence of a chaotic array of competing technologies and standards, and the lack of killer applications (Chen & Nath, 2004). According to Gebauer and Shaw (2004), poor technology characteristics have inhibited application usage to a great extent. In addition, according to Nah et al. (2005), security, cost,

Table 3. Values of mobile business applications (Sources: Chen & Nath, 2004)

Value	
Efficiency	Reduce business process cycle time
	Capture information electronically
	Enhance connectivity and communication
	Track and surveillance
Effectiveness	Reduce information float
	Access critical information anytime-anywhere
	Increased collaboration
	Alert and m-marketing campaigns
Innovation	Enhance service quality
	React to problems and opportunities anytime-anywhere
	Increase information transparency to improve supply chain
	Localize

and employee acceptance are also significant barriers of the growth of m-business applications. Companies have been concerned about the loss or theft of mobile devices, which are easily misplaced or stolen, and their likelihood to contain sensitive or confidential data that can be accessed by unauthorized persons. Huge cost is also a concern to companies. To implement mobile applications, the company must invest in mobile devices, pay service fees for wireless access, and train employees. According to Lucas (2002), some U.S. firms are spending between $5 million and $50 million for mobile business applications. Finally, employee acceptance is also a big barrier. Not every employee is willing to embrace new technology, and some employees accustomed to standard operation procedures resist adoption of m-business applications.

FUTURE TRENDS

In the future, more customized and personalized business applications will be introduced. These applications are called context-aware or situation-dependent m-business applications (Figge, 2004; Heer, Peddemors, & Lankhorst, 2003). Currently, the majority of context-aware computing has been restricted to location-aware computing for mobile applications. However, more contextual information including spatial (e.g., speed and acceleration), temporal (e.g., time of the day), environmental (e.g., temperature), and social situation (e.g., office nearby) information will be added to increase the value of mobile business applications. In context mobile business applications, the most necessary information for the user to perform tasks will be provided in advance without the user's involvement. Therefore, in most cases, the user simply presses a single button rather than making several text inputs.

However, for m-business applications to grow, current limitations in technical and managerial issues should be resolved. Current technical limitations are mainly related to mobile devices such as small multi-function keypads, less computation power, and limited memory and disk capacity (Siau, Lim, & Shen, 2001). Other technical issues such as the lack of network standards and security problems also must be resolved (Chen & Nath, 2004). In addition, a clear understanding of the value of m-business applications is also very

important to grow m-business applications. The m-business development and adoption decision should always be based on clearly identified needs and business requirements (Paavilainen, 2002).

CONCLUSION

M-business applications have shown significant impacts on business processes. M-business applications not only increase productivity, but also develop new business processes that yield increased customer and job satisfaction as well as competitive advantage. In the future, richer and more efficient m-business applications will be introduced to attract more businesses. However, current technical and managerial limitations should be resolved to support continued growth of m-business applications. Especially, it is very important to understand the fundamental value derived from m-business applications before developing and adopting them.

REFERENCES

Chen, L.-D., & Nath, R. (2004). A framework for mobile business applications. *International Journal of Mobile Communications, 2,* 368-381.

Collett, S. (2003). Wireless gets down to business. *Computerworld, 37*(18), 31.

Dekleva, S. (2004). M-business: Economy driver or a mess? *Communications of the Association for Information Systems, 13,* 111-135.

Figge, S. (2004). Situation-dependent services: A challenges for mobile network operators. *Journal of Business Research, 57*(12), 1416-1422.

Funk, J. (2003). *Mobile disruption: Key technologies and applications that are driving the mobile Internet.* New York: John Wiley & Sons.

Gebauer, J., & Shaw, M.J. (2004). Success factors and impacts of mobile business applications: Results from a mobile e-procurement study. *International Journal of Electronic Commerce, 8*(3), 19-41.

Heer, J.D., Peddemors, A.J.H., & Lankhorst, M.M. (2003). *Context-aware mobile business applications.* Retrieved October 29, 2005, from https://doc.telin.nl/dscgi/ds.py/Get/File-25810/coconet.pdf

Lucas, M. (2002). Wireless financial apps grow slowly. *Computerworld, 36,* 14.

Nah, F.F.-H., Siau, K., & Sheng, H. (2005). The value of mobile applications: A utility company study. *Communications of the ACM, 48,* 85-90.

Omni. (2005). *Study finds 13.4 percent increase in worker productivity.* Retrieved October 10, 2005, from http://newsroom.cisco.com/dlls/2005/prod_020905.html

Paavilainen, J. (2002). *Mobile business strategies.* London: Wireless Press.

Siau, K., Lim, E.P., & Shen, Z. (2001). Mobile commerce: Promises, challenges, and research agenda. *Journal of Database Management, 12*(3), 4-13.

Smith, B. (2004). Business apps: Going for the tried and true. *Wireless Week, 10,* 22.

KEY TERMS

Horizontal Mobile Business Application: Mobile business application developed for a massive number of users to improve communication and streamline processes.

Location-Aware Computing: The capability of computing to recognize and react to location context. Global Positioning System (GPS) is the most widely known location-aware computing system.

Mobile Business Application: Mobile application used to perform business tasks such as sales force automation, customer relationship management, and filed force automation.

Situation-Dependent Mobile Application: Mobile application using various contextual information such as spatial, temporal, environmental, and social.

Vertical Mobile Business Application: Mobile business application developed for a specific target group such as filed service engineers and sales representatives.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 442-445, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.3
Business Model Typology for Mobile Commerce

Volker Derballa
Universität Augsburg, Germany

Key Pousttchi
Universität Augsburg, Germany

Klaus Turowski
Universität Augsburg, Germany

ABSTRACT

Mobile technology enables enterprises to invent new business models by applying new forms of organization or offering new products and services. In order to assess these new business models, there is a need for a methodology that allows classifying mobile commerce business models according to their typical characteristics. For that purpose a business model typology is introduced. Doing so, building blocks in the form of generic business model types are identified, which can be combined to create concrete business models. The business model typology presented is conceptualized as generic as possible to be generally applicable, even to business models that are not known today.

INTRODUCTION

Having seen failures like WAP, the hype that was predominant for the area of mobile commerce (MC) up until the year 2001 has gone. About one year ago however, this negative trend has begun to change again. Based on more realistic expectations, the mobile access and use of data, applications and services is considered important by an increasing number of users. This trend becomes obvious in the light of the remarkable success of mobile communication devices. Substantial growth rates are expected in the next years, not only in the area of B2C but also for B2E and B2B. Along with that development go new challenges for the operators of mobile services resulting in

re-assessed validations and alterations of existing business models and the creation of new business models. In order to estimate the economic success of particular business models, a thorough analysis of those models is necessary. There is a need for an evaluation methodology in order to assess existing and future business models based on modern information and communication technologies. Technological capabilities have to be identified as well as benefits that users and producers of electronic offers can achieve when using them.

The work presented here is part of comprehensive research on mobile commerce (Turowski & Pousttchi, 2003). Closely related is a methodology for the qualitative assessment of electronic and mobile business models (Bazijanec, Pousttchi, & Turowski, 2004). In that work, the focus is on the added value for which the customers is ready to pay. The theory of informational added values is extended by the definition of technology-specific properties that are advantageous when using them to build up business models or other solutions based on information and communication techniques. As mobile communication techniques extend Internet technologies and add some more characteristics that can be considered as additional benefits, a own class of technology-specific added values is defined and named mobile added values (MAV), which are the cause of informational added values. These added values based on mobility of mobile devices are then used to assess mobile business models.

In order to be able to qualitatively assess mobile business models, those business models need to be unambiguously identified. For that purpose, we introduce in this chapter a business model typology. Further, the business model typology presented here is conceptualized as generic as possible, in order to be robust and be generally applicable — even to business models that are not known today. In the following we are building the foundation for the discussion of the business model typology by defining our view of MC. After that, alternative business model typologies are presented and distinguished from our approach, which is introduced in the subsequent section. The proposed approach is then used on an existing MC business model. The chapter ends with a conclusion and implications for further research.

BACKGROUND AND RELATED WORK

Mobile Commerce: A Definition

Before addressing the business model typology for MC, our understanding of MC needs to be defined. If one does agree with the Global Mobile Commerce Forum, mobile commerce can be defined as "the delivery of electronic commerce capabilities directly into the consumer's device, anywhere, anytime via wireless networks." Although this is no precise definition yet, the underlying idea becomes clear. Mobile commerce is considered a specific characteristic of electronic commerce and as such comprises specific attributes, as for example the utilization of wireless communication and mobile devices. Thus, mobile commerce can be defined as every form of business transaction in which the participants use mobile electronic communication techniques in connection with mobile devices for initiation, agreement or the provision of services. The concept mobile electronic communication techniques is used for different forms of wireless communication. That includes foremost cellular radio, but also technologies like wireless LAN, Bluetooth or infrared communication. We use the term mobile devices for information and communication devices that have been developed for mobile use. Thus, the category of mobile devices encompasses a wide spectrum of appliances. Although the laptop is often included in the definition of mobile devices, we have reservations to include it here without precincts due to its special characteristics: It can be moved easily,

but it is usually not used during that process. For that reason we argue that the laptop can only be seen to some extend as amobile device.

Related Work

Every business model has to prove that it is able to generate a benefit for the customers. This is especially true for businesses that offer their products or services in the area of EC and MC. Since the beginning of Internet business in the mid 1990s, models have been developed that tried to explain advantages that arose from electronic offers. An extensive overview of approaches can be found in (Pateli & Giaglis, 2002). At first, models were rather a collection of the few business models that had already proven to be able to generate a revenue stream (Fedewa, 1996; Schlachter, 1995; Timmers, 1998). Later approaches extended these collections to a comprehensive taxonomy of business models observable on the web (Rappa, 2004; Tapscott, Lowi, & Ticoll, 2000). Only Timmers (1998) provided a first classification of eleven business models along two dimensions: innovation and functional integration. Due to many different aspects that have to be considered when comparing business models, some authors introduced taxonomies with different views on Internet business. This provides an overall picture of a firm doing Internet business (Osterwalder, 2002), where the views are discussed separately (Afuah & Tucci, 2001; Bartelt & Lamersdorf, 2000; Hamel, 2000; Rayport & Jaworski, 2001; Wirtz & Kleineicken, 2000). Views are for example commerce strategy, organizational structure or business process. The two most important views that can be found in every approach are value proposition and revenue. A comparison of views proposed in different approaches can be found in (Schwickert, 2004). While the view revenue describes the rather short-term monetary aspect of a business model the value proposition characterizes the type of business that is the basis of any revenue stream. To describe this value proposition authors decomposed business models into their atomic elements (Mahadevan, 2000). These elements represent offered services or products. Models that follow this approach are for example (Afuah & Tucci, 2001) and (Wirtz & Kleineicken, 2000). Another approach that already focuses on generated value can be found in (Mahadevan, 2000). There, four so-called value streams are identified: virtual communities, reduction of transaction costs, gainful exploitation of information asymmetry, and a value added marketing process.

In this work however, we are pursuing another approach: The evaluation of real business models showed that some few business model types recur. These basic business model types have been used for building up more complex business models. They can be classified according to the type of product or service offered. A categorization based on this criterion is highly extensible and thus very generic (Turowski & Pousttchi, 2003). Unlike the classifications of electronic offers introduced previously, this approach can also be applied to mobile business models that use for example location-based services to provide a user context. In the following sections, we are describing this business model typology in detail.

BUSINESS MODEL TYPOLOGY

Business Idea

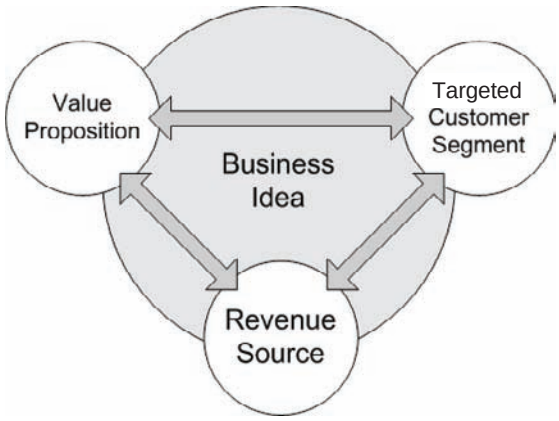
Starting point for every value creation process is a product or business idea. An instance of a business idea is the offer to participate in auctions or conduct auctions — using any mobile device without tempo-spatial restrictions. Precondition for the economic, organisational, and technical implementation and assessment of that idea is its transparent specification. That abstracting specification of a business idea's functionality is called business model. It foremost includes an answer to the question: Why has this idea the potential to be successful? The following aspects have to be considered for that purpose:

- Value proposition (which value can be created)
- Targeted customer segment (which customers can and should be addressed)
- Revenue source (who, how much and in which manner will pay for the offer)

Figure 1 shows the interrelationship between those concepts. It needs to be assessed how the business idea can be implemented regarding organisational, technical, legal, and investment-related issues. Further, it has to be verified whether the combination of value proposition, targeted customer segment and revenue source that is considered optimal for the business model fits the particular company's competitive strategy. Let's assume an enterprise is pursuing a cost leader strategy using offers based on SMS, it is unclear whether the enterprise can be successful with premium-SMS.

It needs to be pointed out that different business models can exist for every single business idea. Coming back to the example of offering auctions without tempo-spatial restrictions, revenues can be generated in different ways with one business model recurring to revenues generated by advertisements and the other recurring to revenues generated by fees.

Figure 1. Business idea and business model



Revenue Models

The instance introduced previously used the mode of revenue generation in order to distinguish business models. In this case, the revenue model is defined as the part of the business model describing the revenue sources, their volume and their distribution. In general, revenues can be generated by using the following revenue sources:

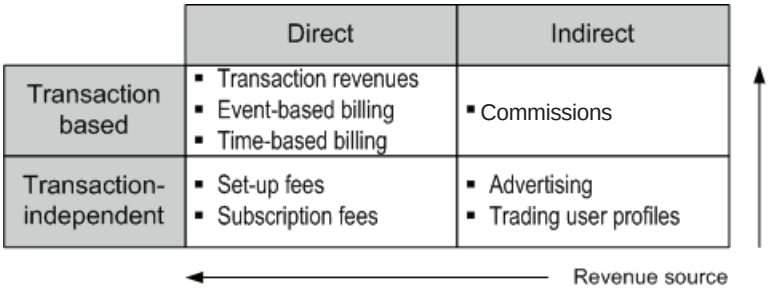
- Direct revenues from the user of a MC-offer
- Indirect revenues, in respect to the user of the MC-offer (i.e., revenues generated by 3rd parties); and
- Indirect revenues, in respect to the MC-offer (i.e., in the context of a non-EC offer).

Further, revenues can be distinguished according to their underlying mode in transaction-based and transaction-independent. The resulting revenue matrix is depicted in Figure 2.

Direct transaction-based revenues can include event-based billing (e.g., for file download) or time-based billing (e.g., for the participation in a blind-date game). Direct transaction-independent revenues are generated as set-up fees, (e.g., to cover administrative costs for the first-time registration to a friend finder service) or subscription fees (e.g., for streaming audio offers).

The different revenue modes as well as the individual revenue sources are not necessarily mutually excluding. Rather, the provider is able to decide which aspects of the revenue matrix he wants to refer to. In the context of MC-offers, revenues are generated that are considered (relating to the user) indirect revenues. That refers to payments of third parties, which in turn can be transaction-based or transaction-independent. Transaction-based revenues (e.g., as commissions) accrue if, for example, restaurants or hotels pay a certain amount to the operator of mobile tourist guide for guiding a customer to their locality. Transaction-independent revenues are generated

Figure 2. Revenue sources in MC (based on Wirtz & Kleineicken, 2000)



by advertisements or trading user profiles. Especially the latter revenue source should not be neglected, as the operator of a MC-offer possesses considerable possibilities for the generation of user profiles due to the inherent characteristics of context sensitivity and identifying functions (compared to the ordinary EC-vendor). Revenues that are not generated by the actual MC-offer are a further specificity of indirect revenues. This includes MC-offers pertaining to customer retention, effecting on other business activities (e.g., free SMS-information on a soccer team leading to an improvement in merchandising sales).

MC-Business Models

In the first step, the specificity of the value offered is evaluated. Is the service exclusively based on the exchange of digitally encoded data or is a significant not digital part existent (i.e., a good needs to be manufactured or a service is accomplished that demands some kind of manipulation conducted on a physically existing object)? *Not digital* services can be subdivided into *tangible* and *intangible* services. Whereas *tangible* services need to have a significant physical component, this classification assumes the following: The category of *intangible* services only includes services that demand manipulation conducted on a physically existing object.

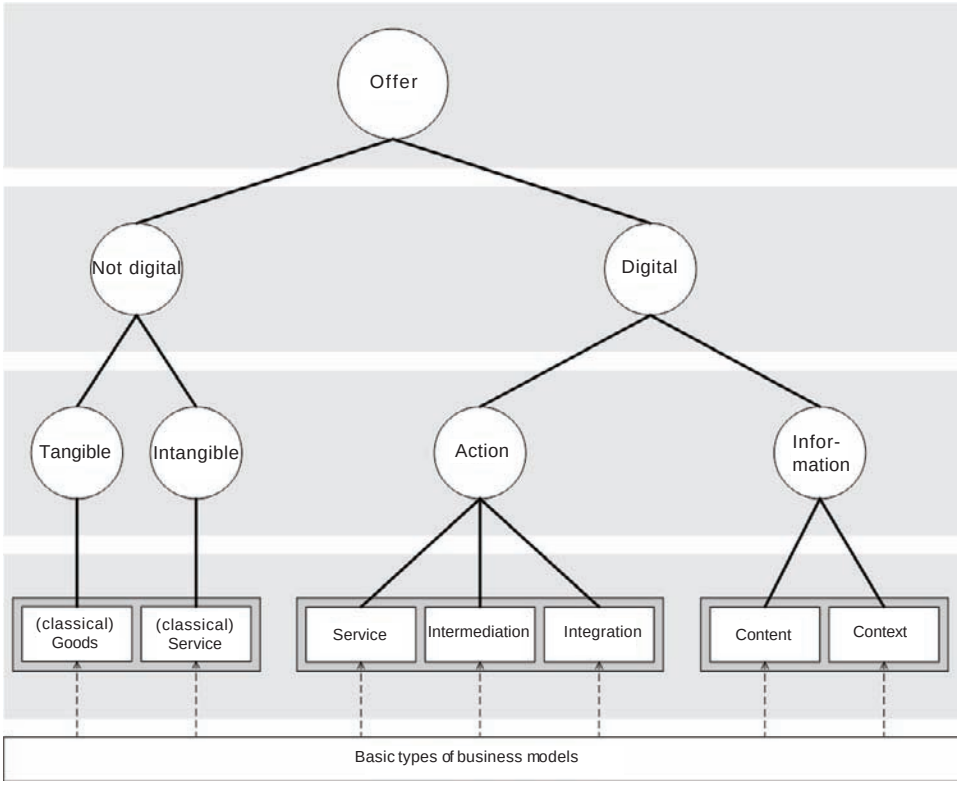
Services that can be created through the exchange of digitally encoded data are subdivided into *action* and *information*. The category *information* focuses on the provision of data (e.g., multi-media contents in the area of entertainment or the supply of information). Opposed to that, the category *action* includes processing, manipulating, transforming, extracting, or arranging of data.

On the lowermost level, building blocks for business models are created through the further subdivision according to the value offered. For that purpose, a distinction is made between the concrete business models that can include one or more business model types and those business model types as such. These act as building blocks that can constitute concrete business models.

The business model type *classical goods* is included in all concrete business models aiming at the vending of tangible goods (e.g., CDs or flowers, i.e., goods that are manufactured as industrial products or created as agricultural produce). Those goods can include some digital components (e.g., cars, washing machines). However, decision criterion in that case is the fact that a significant part of the good is of physical nature and requires the physical transfer from one owner to the other.

Concrete business models include the business model type *classical service* if some manipulation activities have to be conducted on a physical object. That comprises e.g. vacation trips and maintenance activities.

Figure 3. Categorization of basic business model types



The basic business model type *service* comprises concrete business models, if they comprise an original service that is considered by the customer as such and requires some action based on digitally encoded data as described previously, without having intermediation characteristics (c.f., basic business model type *intermediation*). Such services, e.g., route planning or mobile brokerage are discrete services and can be combined to new services through bundling. A typical offer that belongs to the business model type service is mobile banking. Further, it might be required (e.g., in order to enable mobile payment or ensure particular security goals (data confidentiality) to add further services, which require some kind of action, as described previously. As the emphasis is on the original service, these services can be considered as supporting factors. Depending on the circumstances, they might be seen as an original service. Due to that, those supporting services will not be attributed to a basic business model type. Rather, those services are assigned to the business model type service.

A concrete business model includes the business model type *intermediation* if it aims at the execution of classifying, systemising, searching, selecting, or interceding actions. The following offers are included:

- Typical search engines/offers (e.g., www.a1.net);
- Offers for detecting and interacting with other consumers demanding similar products;
- Offers for detecting and interacting with persons having similar interests;
- Offers for the intermediation of consumers and suppliers;
- Any kind of intermediation or brokering action, especially the execution of online auctions; and

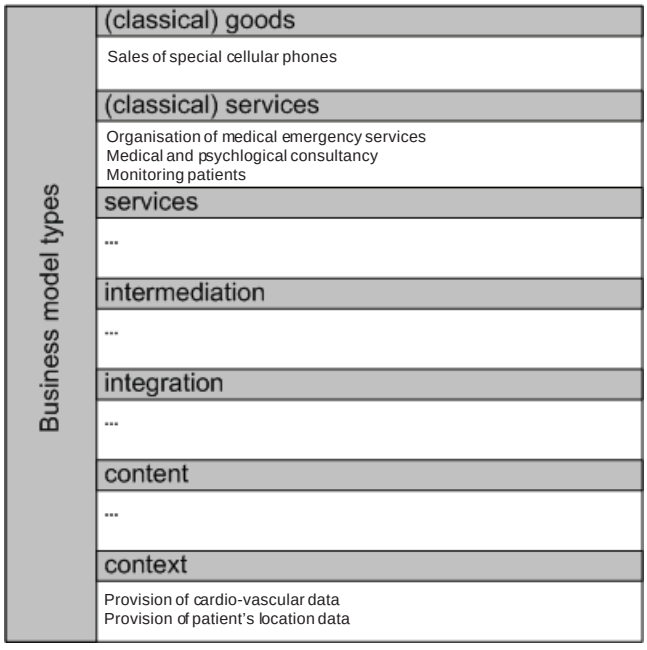
- In general the operations of platforms (portals), which advance, simplify or enable the interaction of the aforementioned economic entities.

Taking all together, the focus is on matching of appropriate parings (i.e., the initiation of a transaction). Nevertheless, some offers provide more functionality by for example supporting the agreement process as well (e.g., the hotel finder and reservation service (wap.hotelkatalog.de)): This service lets the user search for hotels, make room reservations, and cancel reservations. All the relevant data is shown and hotel rooms can be booked, cancelled, or reserved. The user is contacted using e-mail, telephone, fax, or mail. Revenues are generated indirectly and transaction-independent, as the user agrees to obtain advertisements from third parties.

The basic business model type *integration* comprises concrete business models aiming at the combination of (original) services in order to create a bundle of services. The individual services might be a product of concrete business models that in turn can be combined to create new offers. Further, the fact that services have been combined is not necessarily transparent for the consumer. This can even lead to user individual offers where the user does not even know about the combination of different offers. For example, an offer could be an insurance bundle specifically adjusted to a customer's needs. The individual products may come from different insurance companies. On the other hand, it is possible to present this combination to the consumer as the result of a customization process (custom-made service bundle).

The basic business model type *content* can be identified in every concrete business model that generates and offers digitally encoded multi-media content in the areas of entertainment, education, arts, culture sport etc. Additionally, this type comprises games. Wetter-Online (pda.wetteronline.de) can be considered a typical example for that business model type. The user can access free weather information using a PDA. The information offered includes forecasts, actual weather data, and holiday weather. The PDA-version of

Figure 4. Classification of Vitaphone's business model



this service generates no revenues, as it is used as promotion for a similar EC-offer, which in turn is ad sponsored.

A concrete business model comprises the basic business model type *context* if information describing the context (i.e., situation, position, environments, needs, requirements etc.) of a user is utilised or provided. For example, every business model building on location-based services comprises or utilises typical services of the basic business model type context. This is also termed context-sensitivity. A multiplicity of further applications is realised in connection with the utilisation of sensor technology integrated in or directly connected to the mobile device. An instance is the offer of Vitaphone (www.vitaphone.de). It makes it possible to permanently monitor the cardiovascular system of endangered patients. In case of an emergency, prompt assistance can be provided. Using a specially developed mobile phone, biological signals, biochemical parameters, and the users' position are transmitted to the Vitaphone service centre. Additionally to the aforementioned sensors, the mobile phone has GPS functionality and a special emergency button to establish quick contact with the service centre.

Figure 4 depicts the classification of that business model using the systematics introduced previously. It shows that vita phone's business model uses mainly the building blocks from the area of *classical service*. Those services are supplemented with additional building blocks from the area of *context*. This leads to the weakening of the essential requirement — physical proximity of patient and medical practioner — at least what the medical monitoring is concerned. This creates several added values for the patient, which will lead to the willingness to accept that offer.

Analysing the offer of Vitaphone in more detail leads to the conclusion that the current offer is only a first step. The offer results indeed in increased freedom of movement, but requires active participation of the patient. He has to operate the monitoring process and actively transmit the generated data to the service centre. To round of the analysis of Vitaphone's business model, the revenue model is presented in Figure 5.

Non MC-relevant revues are generated by selling special cellular phones. Further, direct MC revenues are generated by subscription fees (with or without the utilisation of the service centre) and transmission fees (for data generated and telephone calls using the emergency button).

CONCLUSION

This chapter presents an approach to classify mobile business models by introducing a generic mobile business model typology. The aim was to create a typology that is as generic as possible, in order to be robust and applicable for business models that do not exist today. The specific char-

Figure 5. Vitaphone's revenue model

MC-Business			Non-MC-Business
	Direct	Indirect	
Transaction based	• Communication with the service centre	...	Sales of special cellular phones
Transaction-independent	• Subscription fee	...	

acteristics of MC make it appropriate to classify the business models according to the mode of the service offered. Doing so, building blocks in the form of business model types can be identified. Those business model types then can be combined to create concrete business model. The resulting tree of building blocks for MC business models differentiates digital and not digital services. Not digital services can be subdivided into the business model types classical goods for tangible services and classical service for intangible services. Digital services are divided into the category action with the business model types service, intermediation, integration and the category information with the business model types content and context.

Although the typology is generic and is based on the analysis of a very large number of actual business models, further research is necessary to validate this claim for new business models from time to time.

REFERENCES

Afuah, A., & Tucci, C. (2001). *Internet business models and strategies.* Boston: McGraw Hill.

Bartelt, A., & Lamersdorf, W. (2000). *Geschäftsmodelle des Electronic Commerce: Modell-bildung und Klassifikation.* Paper presented at the Verbundtagung Wirtschaftsinformatik.

Bazijanec, B., Pousttchi, K., & Turowski, K. (2004). *An approach for assessment of electronic offers.* Paper presented at the FORTE 2004, Toledo.

Fedewa, C. S. (1996). *Business models for Internetpreneurs.* Retrieved from http://www.gen.com/iess/articles/art4.html

Hamel, G. (2000). *Leading the revolution.* Boston: Harvard Business School Press.

Mahadevan, B. (2000). Business models for Internet based e-commerce: An anatomy. *California Management Review, 42*(4), 55-69.

Osterwalder, A. (2002). *An e-business model ontology for the creation of new management software tools and IS requirement engineering.* CAiSE 2002 Doctoral Consortium, Toronto.

Pateli, A., & Giaglis, G. M. (2002). *A domain area report on business models.* Athens, Greece: Athens University of Economics and Business.

Rappa, M. (2004). *Managing the digital enterprise — Business models on the Web.* Retrieved June 14, 2004, from http://digitalenterprise.org/models/models.html

Rayport, J. F., & Jaworski, B. J. (2001). *E-Commerce*. New York: McGraw Hill/Irwin.

Schlachter, E. (1995). *Generating revenues from Web sites.* Retrieved from http://boardwatch.internet.com/mag/95/jul/bwm39

Schwickert, A. C. (2004). *Geschäftsmodelle im electronic business—Bestandsaufnahme und relativierung.* Gießen: Professur BWL-Wirtschaftsinformatik, Justus-Liebig-Universität.

Tapscott, D., Lowi, A., & Ticoll, D. (2000). *Digital capital—Harnessing the power of business Webs.* Boston.

Timmers, P. (1998). Business models for electronic markets. *Electronic Markets, 8*, 3-8.

Turowski, K., & Pousttchi, K. (2003). *Mobile Commerce—Grundlagen und Techniken*. Heidelberg: Springer Verlag.

Wirtz, B., & Kleineicken, A. (2000). Geschäftsmodelltypen im Internet. *WiSt, 29*(11), 628-636.

KEY TERMS

Business Model: Business model is defined as the abstracting description of the functionality of a business idea, focusing on the value proposition, customer segmentation, and revenue source.

Business Model Types: Building blocks for the creation of concrete business models.

Electronic Commerce: Every form of business transaction in which the participants use electronic communication techniques for initiation, agreement or the provision of services.

Mobile Commerce: Every form of business transaction in which the participants use mobile electronic communication techniques in connection with mobile devices for initiation, agreement or the provision of services.

Revenue Model: The part of the business model describing the revenue sources, their volume and their distribution.

This work was previously published in Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 114-21, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.4
Business Strategies for Mobile Marketing

Indranil Bose
University of Hong Kong, Hong Kong

Chen Xi
University of Hong Kong, Hong Kong

INTRODUCTION

With the appearance of advanced and mature wireless and mobile technologies, more and more people are embracing mobile "things" as part of their everyday lives. New business opportunities are emerging with the birth of a new type of commerce known as mobile commerce or m-commerce. M-commerce is an extension to electronic commerce (e-commerce) with new capabilities. As a result, marketing activities in m-commerce are different from traditional commerce and e-commerce. This chapter will discuss marketing strategies for m-commerce. First we will give some background knowledge about m-commerce. Then we will discuss the pull, push, and viral models in m-marketing. The third part will be the discussion about the future developments in mobile marketing. The last part will provide a summary of this article.

BACKGROUND

Popularity of Mobile Services

From the research done by Gartner Dataquest (BusinessWeek, 2005), there will be more than 1.4 billion mobile service subscribers in the Asia-Pacific region by 2009. Research analysts of Gartner Dataquest also estimated that China will have over 500,000 subscribers, and more than 39% of the people will use mobile phones at that time. In India, the penetration rate of mobile phones is expected to increase from 7% in 2005 to 28% in 2008. The Yankee Group has also reported a growing trend of mobile service revenues from 2003 to 2009. Although the revenue generated by traditional text-based messaging service will not change much, revenue from multimedia messaging services will rise to a great extent. Other applications of mobile services, such as m-commerce-based services and mobile enterprise

services, will continue to flourish. One thing that is very important in driving Asia-Pacific mobile service revenue is mobile entertainment services. Revenue from mobile entertainment services will make up almost half of the total revenues from all kinds of mobile data services from now on. Not only in the region of Asia-Pacific, but mobile services will increase in popularity in other parts of the world as well. In the United States, it is expected that the market for m-commerce will reach US$25 billion in 2006.

The Development of Mobile Technologies

Two terms are frequently used when people talk about mobile information transmission techniques: the second-generation (2G) and the third-generation (3G) wireless systems. These two terms actually refer to two generations of mobile telecommunication systems. Three basic 2G technologies are time division multiple access (TDMA), global system for mobile (GSM), and code division multiple access (CDMA). Among these three, GSM is the most widely accepted technology. There is also the two-and-a-half generation (2.5G) technology of mobile telecommunication, such as general packet radio service (GPRS). 2.5G is considered to be a transitional generation of technology between 2G and 3G. They have not replaced 2G systems. They are mostly used to provide additional value-added services to 2G systems. The future of mobile telecommunication network is believed to be 3G. Some standards in 3G include W-CDMA, TD-SCDMA, CDMA 2000 EV-DO, and CDMA EV-DV. The advancement in mobile telecommunication technology will bring in higher speed of data transmission.. The speed of GSM was only 9.6 kilobits per second (kbps), while the speed of GPRS can reach from 56 to114 kbps. It is believed that the speed of 3G will be as fast as 2 Megabits per second (mbps). The acceptance of 3G in this world began in Japan. NTT DoCoMo introduced its 3G services in 2001. Korea soon followed the example of Japan. In 2003, the Hutchison Group launched 3G commercially in Italy and the UK, and branded its services as '3'. '3' was later introduced in Hong Kong, China in 2004. Mainland China is also planning to implement 3G systems. Some prototypes or experimental networks have been set up in the Guangdong province. It is expected that 3G networks will be put into commercial use in 2007 using the TD-SCDMA standard that has been indigenously developed in China. Mobile information transmission can also be done using other technical solutions such as wireless local area network (WLAN) and Bluetooth. The interested reader may refer to Holma and Toskala (2002) for a fuller description of 3G systems, and to Halonen, Romero, and Melero (2003) for details of 2G and 2.5G systems.

The most popular mobile devices currently in use include mobile phones, wireless-enabled personal digital assistants (PDAs), and wireless-enabled laptops (Tarasewich, Nickerson, & Warkentin, 2002). Smartphones are also gaining favor from customers. Mobile phones are the most pervasive mobile devices. Basically, mobile phones can make phone calls, and can send and receive short text messages. More advanced mobile phones have color screens so that they can send or receive multimedia messages, or have integrated GPRS modules so that they can connect to the Internet for data transmission. PDAs are pocket-size or palm-size devices which do limited personal data processing such as recording of telephone numbers, appointments, and notes on the go. Wireless-enabled PDAs have integrated Wi-Fi (wireless fidelity)—which is the connection standard for W-LAN or Bluetooth—which helps them access the Internet. Some PDAs can be extended with GPRS or GSM modules so that they can work as a mobile phone. PDAs nowadays usually have larger screens than that of mobile phones and with higher resolution. They are often equipped with powerful CPUs and large storage components so that they can handle multimedia

tasks easily. Smartphones are the combination of mobile phones and PDAs. Smartphones have more complete phoning function than PDAs, while PDAs have more powerful data processing abilities. However, the boundary between smartphones and PDAs are actually becoming more and more fuzzy.

The Need for Mobile Marketing

The rapid penetration rate of mobile devices, the huge amounts of investment from industries, and the advancement of mobile technologies, all make it feasible to do marketing via mobile devices. Mobile commerce refers to a category of business applications that derive their profit from business opportunities created by mobile technologies. Mobile marketing, as a branch of m-commerce (Choon, Hyung, & Kim, 2004; Varshney & Vetter, 2002), refers to any marketing activities conducted via mobile technologies. Usually m-commerce is regarded as a subset of e-commerce (Coursaris & Hassanein, 2002; Kwon & Sadeh, 2004). That is true, but due to the characteristics of mobile technologies, mobile marketing is different from other e-commerce activities. The first difference is caused by mobile technologies' ability to reach people anywhere and anytime; therefore mobile marketing can take the advantage of contextual information (Zhang, 2003). Dey and Abowd (2001) defined context as "any information that characterizes a situation related to the interaction between users, applications, and the surrounding environment." Time, location, and network conditions are three of the key elements of context. The second difference is caused by the characteristics of mobile devices. Mobile devices have limited display abilities. The screens are usually small, and some of the devices cannot display color pictures or animations. On the other hand, mobile devices have various kinds of screen shapes, sizes, and resolutions. Thus, delivering appropriate content to specific devices is very important. Mobile devices also have limited input abilities, and this makes it difficult for customers to respond. Mobile marketing shares something in common with e-commerce activities. An important aspect of e-commerce is to deliver personalized products/services to customers. Mobile marketing inherits this feature. Mobile marketing also inherits some of the problems from e-commerce, especially the problem of spamming. Personalization in mobile marketing is to conduct marketing campaigns which can meet the customer's needs by providing authorized, timely, location-sensitive, and device-adaptive advertising and promotion information (Scharl, Dickinger, & Murphy, 2005).

MOBILE MARKETING

Benefits of Mobile Marketing

There are two main approaches to advertise and promote products in industry—mass marketing and direct marketing. The former uses mass media to broadcast product-related information to customers without discrimination, whereas the latter is quite different in this regard. Mobile marketing takes a direct marketing approach. Using mobile marketing, marketers can reach customers directly and immediately. Similarly, customers can also respond to marketers rapidly. This benefit makes the interaction between marketers and customers easy and frequent. Compared to direct marketing using mail or catalogs, mobile marketing is comparatively cost effective and quick. Compared to telephone direct marketing, mobile marketing can be less interruptive. Compared to e-mail direct marketing, mobile marketing can reach people anytime and anywhere, and does not require customers to sit in front of a computer. Therefore, to some extent, mobile marketing can be a replacement for other types of marketing channels such as mail, telephone, or e-mail. Advertisement or promotion information sent via the Internet can be sent via a mobile device. Mobile marketing can enhance marketing by adding new abilities

like time-sensitive and location-sensitive information. On the other hand, mobile commerce can generate new customers' data, like mobile telecommunication usage data and mobile Internet surfing data. Mobile marketing is the first choice for conducting marketing activities for m-commerce applications. However, due to limited size of screens of mobile devices, only brief information can be provided in mobile marketing solicitations, while e-mail or mail marketing can provide very detailed information. On the other hand telephone marketing requires the good communication skill of telesales. Once telemarketers have acquired this skill, the interaction between marketers and customers is quicker and more effective. It is not clear if mobile marketing is as effective or as popular as mass marketing agents like television and newspaper, but it can be said that it is indeed a powerful medium that is likely to gain in popularity in the future.

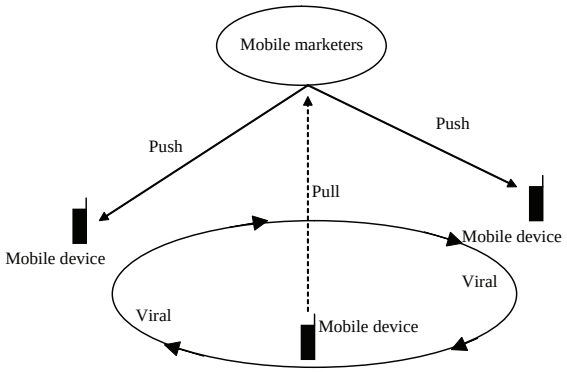
Models for Mobile Marketing

Mobile marketing usually follows one of the three kinds of models—push, pull (Haig, 2002; Zhang, 2003), and viral (Ahonen, 2002; Ahonen, Kasper, & Melkko, 2004; Haig, 2002), as illustrated in Figure 1.

Push Model

The push model sends marketing information to customers without the request of the customer. If the push model is used, besides knowing the targeted customers' interests, understanding the context of customers at the time the marketing activities are to be carried out is very critical. The timing in sending mobile information should also be appropriate. The content delivered to customers should also be displayable on their mobile devices. Permissions from customers are necessary before any solicitations can be sent. In the push model, marketers like to make it easier for customers to respond because of the poor input

Figure 1. Push, pull and viral models for mobile marketing



ability of mobile devices. G2000, a Hong Kong clothing chain, launched a mobile marketing campaign in November 2004. Mobile coupons in the format of SMS were sent to the mobile phones of selected customers. Customers could then use these coupons stored in their mobile phones when purchasing items in designated G2000 stores in order to get discounts. The campaign was considered to be a success because a number of customers responded to this program and used mobile coupons at G2000 stores.

Pull Model

In the pull model, the marketer waits for the customer to send a request for a solicitation. The marketer prepares the marketing information in a format that is displayable in all possible mobile devices and scalable for various connection speeds. The significance of the pull model is that the information from customers is very useful for understanding customers' preferences, such as the preferred marketing time and interests. Mobile marketing following a pull model can be conducted in many ways. One possible approach is to let the customers select and download coupons to their mobile devices. Mobile service providers can build a Web site using a mobile Internet protocol such as wireless application protocol (WAP), and

place various text-based coupons on this Web site. Customers can use their GPRS-enabled phone to browse the Web site and download coupons they like in the format of SMS. Each coupon will have a unique identity number. When the coupon is redeemed, related information, such as the phone number of the customer who downloaded the coupon and the time it was redeemed, is recorded. This kind of information is later used to analyze the behaviors of customers and build a profile for the customer. China Mobile, a mobile telecommunication operator in China, had established such a Web site for customers in Xiamen (a city located in southeast China) in 2005. The customers of China Mobile could download coupons displayed on this Web site onto their mobile phones using text messaging.

Viral Model

The phrase viral marketing was created by Steve Jurvetson in 1997 to describe the burgeoning use of Hotmail (Jurvetson & Draper, 1997). The principle of the viral model is based on the fact that customers forward information about products/services to other customers. The viral model enlarges the effect of other marketing activities while it costs the marketers very little in monetary terms. The viral model enables customer-to-customer communication. Like the pull model, the format of information that is delivered by viral model should be displayable in different devices and scalable for different connection speeds. Actually mobile marketing has the ability to be viral inherently because it is quite easy for people to forward mobile advertising or promotion information to their friends. However, viral marketing information has to be interesting and attractive enough to make the customers willing to forward it to other people. For example, "reply to this message in order to win $5000" may be a very attractive viral marketing message. Usually, viral marketing begins with push marketing activities to customers. According to Linner (2003), when the movie "2 Fast 2 Furious" was running in movie theatres, marketers tried to create a viral promotion using a mobile marketing strategy. Fans were asked to send SMS to enter a certain film-related competition. Besides inviting fans on every major phone network through advertisements on television, newspapers, and also through posters, a special code was designed and a low fee was offered to customers in order to encourage them to forward promotion information to their friends. Exciting gifts were offered as prizes in this competition (such as a replica of the vehicle, the EvO VII, that was used in the movie) to spur the enthusiasm of customers.

These three models of direct marketing can be complimentary to each other. Push-based mobile marketing can be used to stimulate pull-based marketing activities. For example, book marketers can send a short introduction to customers via SMS with a remark at the end saying "for more details, please reply to XXXXX." Once a customer responds to this by replying using SMS, more promotion or advertising information on this book can be sent to him. All three models—push, pull, and viral—can even be integrated together in a mobile marketing campaign. An example of an integrated mobile marketing approach was adopted by Fox Txt Club for the movie "phone booth" (Linner, 2003). At the beginning of the marketing campaign, members of the Fox Txt Club were sent invitations via SMS to a preview. The aim of this was to pull customers to the campaign. A competition that invited people to send SMS about questions pertaining to various details in the film was set up. The forwarding of SMS about the movie and the competition among club members and their friends, together with other media such as entertainment and event listings magazines and city-center posters, made the marketing campaign viral. The details of those who responded were recorded by Fox Txt Club, and this helped in building the database of customers for future release promotions. This could be used for push marketing for another movie in the future.

Strategies for Mobile Marketing

The most fundamental task for marketing activities following the push model is to send advertising or promotion information about products or services that the targeted customers once bought. This is the most direct and easy way to decide what is to be offered to customers in a solicitation. However, just marketing products already existing in customers' transaction records is not enough for marketers. It is necessary for marketers to explore the needs of customers. Two of the most commonly used marketing strategies are cross-selling and up-selling. Cross-selling is the practice of suggesting similar products or services to a customer who is considering buying something, such as showing a list of ring tones on a mobile Internet Web page that are similar to the one a customer has downloaded. Up-selling is the practice of suggesting higher priced, better versions of products or services to a customer who is considering a purchase, such as a mobile phone plan with higher fees and additional features. Two approaches can be used to find opportunities for up-selling or cross-selling. One is to find products or services that are similar to the ones a customer has bought. The other is to find people who have characteristics that are similar to a targeted customer. Products or services those people have bought and the targeted customer has not can be recommended to the target customers.

Pull-based marketing is relatively passive compared to push-based marketing. Usually in pull marketing, customers are responsible for searching for useful advertising or promotion information. The marketers' responsibility is to help customers find what they want more efficiently. Therefore, knowing what customers may request is very important in pull marketing. Instead of sending related information to customers like push marketing, marketers doing pull marketing can make information about products or services available on their mobile Internet Web site or ordinary Internet Web site. In viral marketing, marketers stand in a more passive position than even in pull marketing. However, for both pull and push marketing, some push activities should be carried out to start the marketing.

Whatever model one may use when carrying out mobile marketing activities, one issue must always be kept in mind and that is the necessity of obtaining explicit permission from customers (Bayne, 2002). Mobile technology makes connections so direct that it can interfere with customers' privacy very easily. Therefore, sending advertising or promotion information to people will cause trouble if permissions are not sought before solicitations or customers' wishes about not receiving a solicitation are not respected.

Understanding Customers in Mobile Marketing

All of the three models require good understanding of customers' needs. Marketing information that is not well designed will be regarded as spam by customers. Once a customer identifies some information from a company as spam, he or she will pay very little attention to or simply discard any information from that company. If a customer cannot find useful information on the Web site a company provides, it may be ok for the first time, a pity for the second time, but for the third time it will mean business lost forever. If information sent to customers is not interesting, customers may not want to forward them to their friends. All these situations may lead to failure of a marketing campaign. To avoid these situations, marketers need to understand customers well enough in order to send personalized marketing information. Customer profiling is a necessary approach to understand customers better. Customer profiling aims to find factors that can characterize customers. These factors are found by comparing customers to each other in order to discover similarities and differences among customers. Customer profiling encompasses two tasks—customer clustering and customer

behavior pattern recognition. Customer clustering aims to classify customers into different groups. Customers within the same group are said to be more similar to each other than to customers in different groups. Marketers cluster customers using various data. Traditionally, customers are clustered according to their geographic locations, demographic characteristics, and the industries they are working for. They can also be clustered based on information about their purchasing history, such as what they bought, when they bought, and how much they spent. With the appearance of mobile services and m-commerce, usage data of new customer data services can also be used for clustering. For example, messaging services that customers subscribed to, GPRS surfing and download records, the type of mobile devices the customers use, and monthly mobile phone usage including use of IDD and roaming can yield many interesting information about the customers.

Aside from these hard facts, marketers may also want to infer some soft knowledge about customers' behaviors as well. To recognize customer behavior the marketers must discover relationships between hard facts. For example, customers that download ring tones of game music may download games-related screensavers later on. Since mobile technologies can enable context-sensitive marketing activities, marketers should gather knowledge about customers' location preferences and time preferences. For example, when does a customer usually go shopping and which place does he/she visit on the shopping trips? Marketers can find this kind of soft knowledge from various mobile network usage data. Again, collecting information on location and time requires permission from customers. Based on customer profiling, more sophisticated personalized advertising or promotion information can be sent to customers.

FUTURE TRENDS

Mobile technologies will advance further in the future. New technologies will enable new kinds of marketing activities. For example, the implementation of fourth-generation (4G) wireless systems will make the bandwidth much larger than that in current networks. On the other hand, the mobile device will have larger screens with higher resolutions. These two factors together will make interactive audio and even interactive video marketing possible. Generally speaking, the limitation of current mobile technologies will be weakened or removed in the future. As a result, more emphasis may be put on time- and location-related marketing, as well as on better understanding customers' interests. The principle is not only to know what customers want, but also to know when and where they may have a certain kind of need. Data mining techniques can be used in the future to find customer behavior patterns with time and location factors. Data mining techniques have been used widely in direct marketing for targeting customers (Ling & Li, 1998). There are also data mining techniques for clustering customers such as self-organizing-map (SOM—Kohonen, 1995) and techniques for discovering customer behavior such as association rules mining (Agrawal & Srikant, 1994). In the future, the availability of huge amounts of data about customers will compel marketers to adopt strong data mining tools to delve deep into customers' nature.

CONCLUSION

Equipped with advanced mobile technologies, more sophisticated marketing activities can be conducted now and in the future. In this article, we have discussed the benefits of mobile marketing, the role of mobile marketing in m-commerce, and the models used in mobile marketing. Although mobile marketing is powerful, it cannot replace other methods of marketing and should only be used as a powerful complement to traditional marketing. Mobile marketing should be integrated into the whole marketing strategy of a firm so

that it can work seamlessly with other marketing approaches.

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very large Databases* (pp. 487-499), Santiago, Chile.

Ahonen, T. T. (2002). *M-profits: Making money from 3G services.* West Sussex, UK: John Wiley & Sons.

Ahonen, T. T., Kasper, T., & Melkko, S. (2004). *3G marketing: Communities and strategic partnerships.* West Sussex, UK: John Wiley & Sons.

Bayne, K. M. (2002). *Marketing without wires: Targeting promotions and advertising to mobile device users.* New York: John Wiley & Sons.

BusinessWeek. (2005). Special advertising section: 3G the mobile opportunity. *BusinessWeek* (Asian ed.), (November 21), 92-96.

Choon, S. L., Hyung, S. S., & Kim, D. S. (2004). A classification of mobile business models and its applications. *Industrial Management & Data Systems, 104*(1), 78-87.

Coursaris, C., & Hassanein, K. (2002). Understanding m-commerce. *Quarterly Journal of Electronic Commerce, 3*(3), 247-271.

Dey, A. K., & Abowd, G. D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction, 16*(2-4), 97-166.

Haig, M. (2002). *Mobile marketing: The message revolution.* London: Kogan Page.

Halonen, T., Romero, J., & Melero, J. (2003). *GSM, GPRS and EDGE performance: Evolution towards 3G/UMTS.* West Sussex, UK: John Wiley & Sons.

Holma, H., & Toskala, A. (2002). *WCDMA for UMTS* (2nd ed.). West Sussex, UK: John Wiley & Sons.

Jurvetson, S., & Draper, T. (1997). *Viral marketing.* Retrieved from http://www.dfj.com/cgi-bin/artman/publish/steve_tim_may97.shtml

Kohonen, T. (1995). *Self-organizing maps.* Berlin: Springer-Verlag.

Kwon, O.B., & Sadeh, N. (2004). Applying case-based reasoning and multi-agent intelligent system to context-aware comparative shopping. *Decision Support Systems, 37*(2), 199-213.

Ling, C. X., & Li, C.-H. (1998). Data mining for direct marketing: Problems and solutions. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (pp. 73-79), New York.

Linner, J. (2003). *Hitting the mark with text messaging.* Retrieved from http://wireless.sys-con.com/read/41316.htm

Scharl, A., Dickinger, A., & Murphy, J. (2005). Diffusion and success factors of mobile marketing. *Electronic Commerce Research and Applications, 4,* 159-173.

Tarasewich, P., Nickerson, R.C., & Warkentin, M. (2002). Issues in mobile e-commerce. *Communications of the Association for Information Systems, 8,* 41-84.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7,* 185-198.

Zhang, D. (2003). Delivery of personalized and adaptive content to mobile devices: A framework and enabling technology. *Communications of the Association for Information Systems, 12,* 183-202.

KEY TERMS

Bluetooth: Used mostly to connect personal devices wirelessly like PDAs, mobile phones, laptops, PCs, printers, and digital cameras.

Code Division Multiple Access (CDMA): A kind of 2G technology that allows users to share a channel by encoding data with channel-specified code and by making use of the constructive interference properties of the transmission medium.

Enhanced Data rates for GSM Evolution (EDGE): A kind of 2.5G technology. A new modulation scheme is implemented in EDGE to enable transmission speed of up to 384 kbps within the existing GSM network.

General Packet Radio Service (GPRS): Belongs to the family of 2.5G. GPRS is the first implementation of packet switching technology within GSM. The speed of GPRS can reach up to 115 Kbps.

Global System for Mobile (GSM) Communications: One of the 2G wireless mobile network technologies and the most widely used today. It can now operate in the 900 MHz, 1,800 MHz, and 1,900 MHz bands.

3G: The third generation of mobile telecommunication technologies. 3G refers to the next generation of mobile networks which operate at frequencies as high as 2.1 GHz, or even higher. The transmission speeds of 3G mobile wireless networks are believed to be able to reach up to 2 Mbps.

Time Division Multiple Access (TDMA): Divides each network channel into different time slots in order to allow several users to share the channel.

Time Division Synchronous Code Division Multiple Access (TD-SCDMA): A 3G mobile telecommunications standard developed in China.

2G: The second generation of mobile telecommunication technologies. It refers to mobile wireless networks and services that use digital technology. 2G wireless networks support data services.

2.5G: The second-and-a-half generation of mobile telecommunication technologies. 2.5G wireless system is built on top of a 2G network. 2.5G networks have the ability to conduct packet switching in addition to circuit switching. 2.5G supports higher transmission speeds compared to 2G systems.

W-CDMA: Developed by NTT DoCoMo as the air interface for its 3G network called FOMA. It is now accepted as a part of the IMT-2000 family of 3G standards.

Wireless Local Area Network (WLAN): Connects users wirelessly instead of using cables. WLAN is not a kind of mobile telecommunication technology. The coverage of WLAN may vary from a single meeting room to an entire building of a company.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 96-101, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.5
Applying Mobile Technologies to Banking Business Processes

Dinesh Arunatileka
University of Western Sydney, Australia

ABSTRACT

This chapter discusses the impact of mobile technologies on service delivery processes in a banking environment. Advances in mobile technologies have opened up numerous possibilities for businesses to expand their reach beyond the traditional Internet-based connectivity and, at the same time, have created unique challenges. Security concerns, as well as hurdles of delivering mobile services "anywhere and anytime" using current mobile devices with their limitations of bandwidth, screen size and battery life are examples of such challenges. Banks are typically affected by these advances as a major part of their business deals with providing services that can benefit immensely by adoption of mobile technologies. As an example case study, this chapter investigates some business processes of a leading Australian bank in the context of application of mobile technologies.

INTRODUCTION

Electronic commerce has become a dynamic force that has changed the way businesses operate on a global scale (Shi & Wright, 2003). Due to increased globalization, individuals, organizations, and governance frameworks have an increasing dependence on communication technologies. The Australian Communication Authority envisions that ubiquity is the "best possible outcome" in terms of the future of business and economy in the country. This ubiquity is based on the elements of technology, market dynamics, users, and rules and guidelines (ACA, 2005). All business organizations in this global context are forced to look at this "best possible outcome" in order to stay competitive. This gives rise to several research questions in the areas of business practices as well as workflow management, and affects the individual and collective social behaviour (Mylonopoulas & Doukidis, 2003). The research areas also focus on the mobile technologies and

their application to businesses, with particular emphasis on the method and manner in which services can be delivered using mobile processes. Mobile processes are business processes that are executed with the use of mobile devices such as PDAs (personal digital assistants), mobile phones, or mobile-enabled laptop computers. Thus, mobility, which is the ability to move freely while performing regular business activities, has become an extremely crucial aspect of today's business processes. Furthermore, as per Archer (2004), in order to incorporate mobility, business processes also have to undergo substantial changes themselves to make it essential that the changes are researched and experimented into.

Internet Usage in the Banking Sector

Banks, as primary institutions of service-oriented business, have increasingly leaned towards e-commerce-based operations. Emerging mobile technologies offer "anytime, anywhere" type of banking that results in better customer orientation and provides personalization of services to the customer. The concept of banking using handheld devices, such as PDAs or other mobile devices, is becoming popular as it enhances the Internet connectivity to the fingertips of the customer (Unnithan & Swatman, 2002). The Internet has also provided opportunities for service providers such as PayPal, an online payment processing company founded in 1999, to offer more cost-effective payment-related services similar to banking services to its customers. PayPal, after a mere four years of operation, has become the most used payment system for clearing auction transactions on eBay (Schneider, 2004), competing directly with the traditional banks. Banks thus face a major challenge and are forced to effect substantial cost reductions in order to be more competitive and offer cost-effective services to its customers. Banks aggressively push their customers to use electronic means for most of their banking, as these electronic transactions are far cheaper as compared to over-the-counter or ATM (automated teller machine) transactions. According to a recent study in the U.S., a teller transaction costs the bank U$1.07. as opposed to a telephone transaction costing 54 cents, an ATM transaction costing 27 cents, a software-based PC transaction costing 1.5 cents, and an Internet-based transaction costing a mere 1 cent (Money Central, 2005). Mobile devices enable secure and convenient use of e-banking, payments, brokerage, and other types of transactions which are part and parcel of the banking sector (Herzberg, 2003). Another study reveals that among the Internet-based banking users, there is a positive tendency to use mobile devices to do banking transactions (Coutts, 2002). Hence, factors determining the success or failure of the mobile business and how the corresponding mobile systems and applications are designed, in order to provide banks with cost-effective, flexible, and customer-oriented business processes, are of interest to the banking community. The fact that today's banking customers are more educated, along with increasing demand for state-of-the-art services, also add pressure and push the banks towards mobile technologies.

Global Banking Industry

The educated and technology-savvy customers demanding better service and state-of-the-art technology is a global phenomenon in the banking industry. For example, the banking industry in Europe is undergoing substantial changes as it looks to reduce costs and enhance the utility for customers through new technology. European banks are focused more on their core capabilities while exploring different sourcing options for non-core capabilities. They are disaggregating their value chain into independently operable functional units (Homann, Rill, & Wimmer, 2004). Furthermore, as communication capabilities reach higher levels of performance and reliability, these functional units are combined

across corporate borders, providing valuable e-collaborations and flexibility for the organization. The industry sectors are changing the way they do business by using many different collaborations with customers (B2C), service providers (B2B), funding organizations (e-payments), government (B2G), and even competitors (B2B) (Arunatileka & Arunatileka, 2003). The emerging mobile financial applications including both mobile payments and banking services are also being investigated, showing how new financial services could be deployed in mobile networks and also identifying key players in the mobile financing value chain (Mallat, Rossi, & Tuunainen, 2004). Mobile customer relationship management is another area that would personalise business processes, adding value to organisations (Unhelkar & Arunatileka, 2003).

This chapter specifically investigates the implications of service delivery using mobile technologies. Since the author of this chapter has been researching within a well-known Australian bank (name is withheld due to confidentiality issues), this chapter is based on the service-delivery challenges related to incorporation of mobile technologies within the banking environment.

STUDY OF TODAY'S BANKING NEEDS

This section starts with a brief introduction of banking requirements studied in a leading bank (referred to merely as 'the bank') in Sydney. The bank was seeking to create a policy to introduce mobile technology to its banking processes and staff. This policy for mobile services (services using mobile processes) is meant to evolve from a broad framework of existing policies defined for the operations of the entire bank that also encapsulated its values and objectives. The vision of the bank is to inculcate three great values—namely, *teamwork*, *integrity,* and *performance*. The vision drives the organisation purposefully towards its objectives based on the four foundations of *staff*, *customers*, *corporate responsibility,* and *shareholders* (internal documents of the bank). The bank comprises different business units divided based on functionality for management purposes. These various units have trained staff in different disciplines. For example, the financial markets division would have trained investment advisors, the institutional banking division comprising trained relationship managers and customer care personnel, the retail banking division having trained home loan advisors and customer care personnel, whereas the human resources division trained human resource and training personnel. Thus, the bank consists of a multi-disciplined, heterogeneous workforce. The processes and the work methods of separate units could be very different from each other as well. For instance, the corporate and institutional banking divisions work internationally, thus having a 24-hour operation, whereas the retail banking is more likely to be an office hour operation subject to few exceptions. As mobility options that could increase productivity in corporate banking, PDAs could be programmed to notify users of any new e-mails and SMS messages where the employees are notified immediately, and necessary action could be taken depending on the situation. This enables the employees to have more time with their families while still attending to urgent business. In retail banking, loan officers could use mobiles to be more competitive in the field. As parts of an organization, these various divisions have to work together as one entity in achieving the vision and goals of the entire organization.

The mobility policy, once accepted and incorporated, has to address the purpose and objectives underlined by the top management in the facilitation of better service delivery providing higher value to customers. This should fulfil most of the outlined areas by the management by facilitating the staff on better access, wider responsibility, and better tools, motivating them on better service delivery, achieving customer satisfaction, which in turn would fulfil the corporate objectives.

Business Units Under Study

There were four business functional units identified by the bank as the initial study areas for introduction of mobile technologies, namely: financial markets, institutional banking, retail banking, and small investor operations. Although the last two units have a high volume of transactions, the first two functional units were given priority by the bank due to their high-value transactions and the need for change by the employees and the customers of these units. Furthermore, financial markets and institutional banking also had a pressing need to change their existing processes due to some existing limitations in their operations. Before the effect of mobility is investigated, a brief summary of how these two units work is described here.

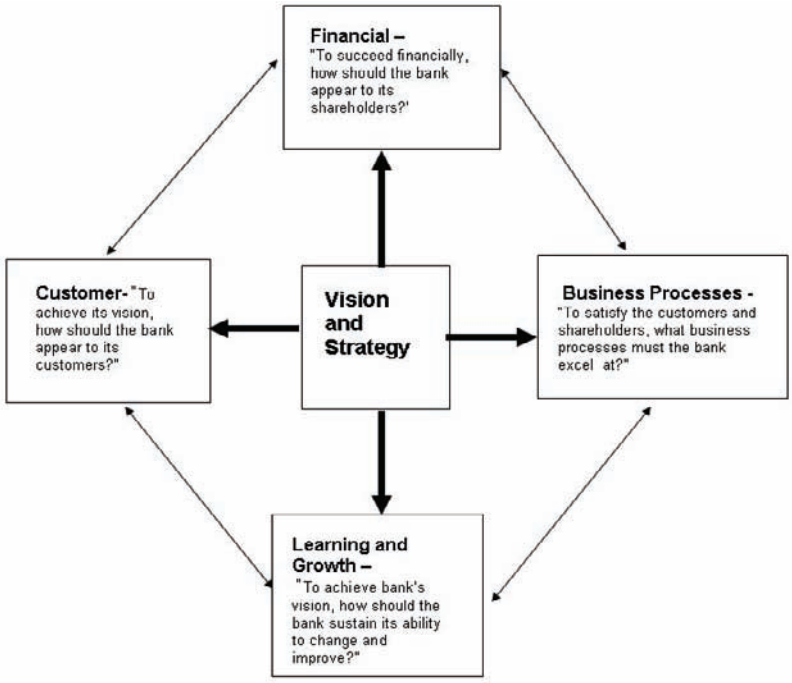
- **Financial Markets:** This is a highly specialised unit, which brings the bank high revenues and very high profitability. Although the number of customers may be lower, the revenues are very high and revenue per customer is very high, resulting in these customers demanding and warranting individual attention. The concerned managers would be international business managers and state managers who are handling time-sensitive corporate accounts. These managers travel a lot, meeting customers and looking after their specific interests.
- **Institutional Banking:** This is also a very highly specialised unit where institutional banking managers generally go to their client organisations in order to serve their needs. Although there are less volumes of transactions, the values could be very high, bringing very high profitability to the bank. Most accounts would belong to large business organisations having diversified needs. Mobile access to the systems would make it much easier for these managers to be in touch with the latest communications, rates, and so forth, which would be very essential in serving business customers in a highly competitive market.

BACKGROUND TO RESEARCH IN THE BANK

The background reasons for researching into the bank's processes in the context of mobility were based on the long-term expectations of the bank. Usually, the expectations of an organisation are summarized in the mission and vision statements. The bank had very specific values and purpose, spelled out in its mission and vision statements. The bank also used the concept of a balanced score card in order to maintain the policy balance and drive its strategies. The balance score card concept (Balanced Scorecard, 2005) is built on four perspectives—financial, internal business processes, customer, and learning and growth, with the vision and strategy in the centre of it.

The bank envisions and makes strategies in the business with respect to these four perspectives. The four perspectives were adapted from www.balancescorecard.org, which describes the perspectives with respect to the organization. All four perspectives are so tightly entwined that one leads to the other. Financials is the starting point, which is the very core of the objectives of a business, simply to pass a value to its owners/shareholders. However, since the banking industry is highly competitive, the customers should be well looked after in order to retain them. Business processes, learning and growth, and customer perspectives all contribute to make the bank more customer oriented. The main focus in this chapter is aimed at business process perspective, which refers to the business processes of the organisation. The measurement on this perspective allows the managers to know how well the business is running in terms of whether the products and services conform to customer requirements (the mission of the organisation). In addition to strategic manage-

Figure 1. A balanced score card for the bank



ment processes, two kinds of processes could be identified—mission-oriented processes and support processes (Balanced Scorecard, 2005). The applications of mobile technologies are looked at in all these processes.

The bank conceptually was looking at two major areas of mobility: mobility at the workplace and mobility in the delivery of service to customers. The mobility at the workplace is in line with a more futuristic plan for a new headquarters building with a smart office where the employees could work anywhere within the building. Mobility in delivery of service is more with the current business processes and how they could be improved to offer better service delivery to customers. In-depth knowledge of the existing work processes, coupled with a detailed plan to transform the existing organization into a mobile enabled (m-enabled) organization without disrupting the day-to-day functions of the organization, is one outcome expected of this exercise. Technology has made progress from earlier setbacks, but the methodologies to fully implement the existing technology into the business operations have to be done carefully in a systematic manner.

The initial expectations of the bank were to look at the business unit of financial markets. The decision was financial and also partly need driven since the financial markets were a big earner and the managers in the unit were very keen to move forward with new technology. Once the focus area was identified, the focus was concentrated on the fundamental questions.

TRANSFORMATION OF BUSINESS PROCESSES

The financial markets area was selected as the first unit for this study since the most pressing demand for change was persistent in this unit. There were several research questions arising with this selection. They were:

- How does the service-based industry change with application of mobile technologies? A generic question focused on the service-based industry as a whole investigating the possibilities to improve efficiency while cutting down on long-term costs.
- What are the changes happening in the banking industry? This is an industry-specific question to look at what is happening elsewhere in the banking industry which is relevant to the current timeframe.
- What should be the bank's response to the change in the service industry and specifically in the banking industry? This probing question is looking inward into the bank's own processes critically to decide how change could be facilitated to improve on the internal processes.
- What is the expected impact after mobile technology is introduced to a selected business unit in the bank? A specific question arising from mobile technology being introduced to the unit which is expected to create a positive impact.
- What would be the direct impact on the customers once the new technology is fully implemented? A probing question to understand the impact on the customers once the change is made.
- Would there be any anticipated problems during the changeover? A question to understand the management of change from the existing to the mobile-enabled organization. It is also important to measure this change in terms of cost factors, time factors, customer satisfaction factors, security issues, and other such factors which the bank thinks important to consider.

Let us look at transforming a simple business process, like the checking of an account balance by a customer as an example. The bank would concentrate on the cost of the process, the satisfaction of the customer in the process, assurance of the process in delivering the right information, security issues in providing the service, and timeliness of the information from the customer's point of view.

If the process is m-transformed from checking the balance through the Internet or at an ATM to use of a mobile phone, how will such a transition be measured in terms of this process? In order to understand the m-transition of the process, it is essential that the process is understood and examined critically. The first step in such a critical evaluation is the creation of a mobile transition roadmap. Such a roadmap is proposed in Figure 3, which shows m-transition applied to a bank focusing on the overall picture while the transition is in progress.

The mobile transition roadmap will capture all the areas that have to be considered in m-transforming an organization in order to become a mobile-enabled organization. The concept of the roadmap for mobile transformation, as shown in Figure 3, has been adopted and evolved from the electronic transformation roadmap (Ginige, Murugesan, & Kazanis, 2001). The mobile transi-

Figure 2. A business process transformation

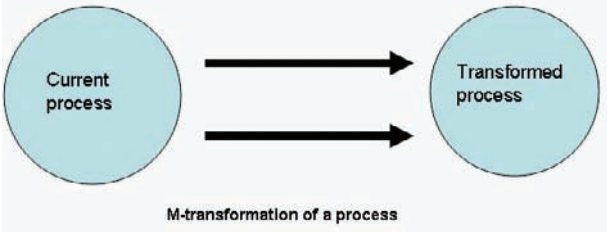
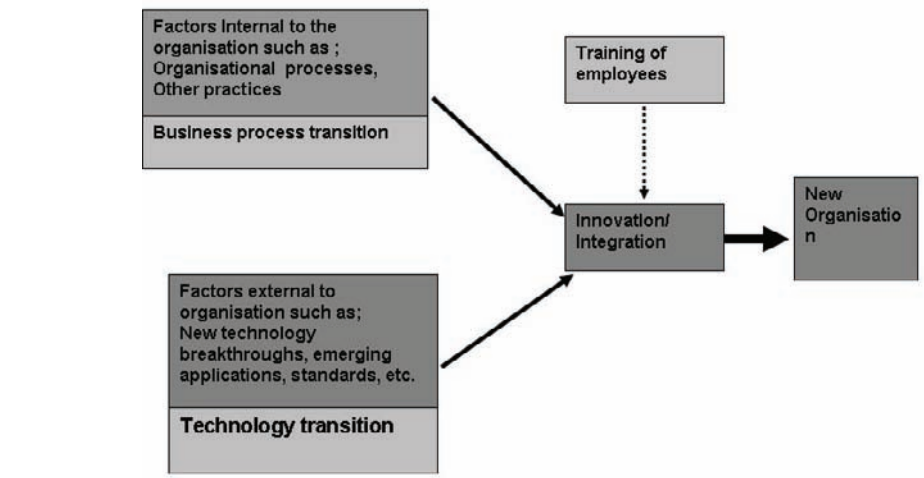


Figure 3. A mobile transition roadmap



tion roadmap is further analysed and investigated using different perspectives in Figure 8, later in the chapter.

The Business Processes in the Financial Markets Division

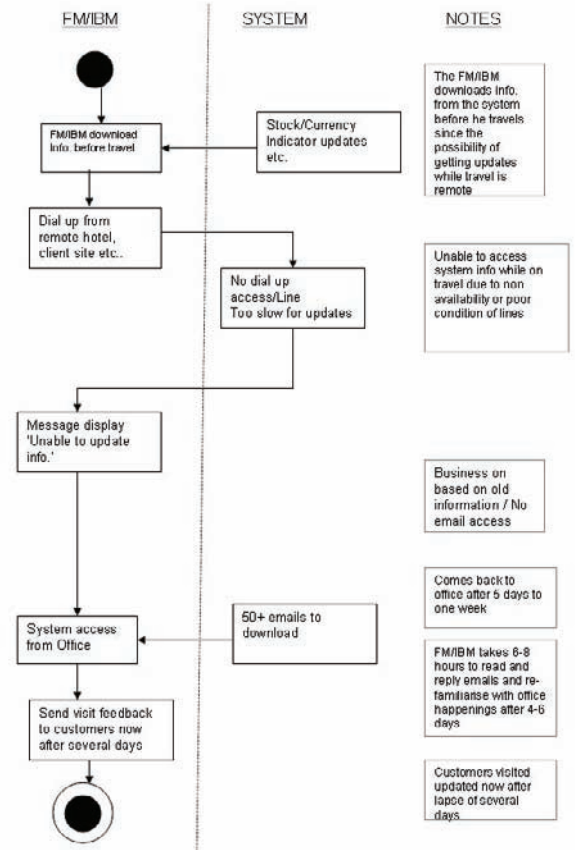
The Fund Managers and International Business Managers (FM/IBM) are called to service customers at anytime of the day since financial markets are a global business. Different market segments would be working in different time zones. It is important be on top, knowing what is happening all over, all the way from Sydney to New York through the other giants such as Japan, Hong Kong, and Europe. Therefore the managers working in this unit should be dynamically connected. This is one area where mobile technologies could be used very effectively to enhance productivity. The role of an FM/IBM in particular involves visiting clients and understanding their requirements in international business from telegraphic transfers to structuring major import/export deals, to financing solutions for the funding of these transactions to mitigate risks associated.

At present, laptops are being used which contain the data and programs to run the business that is provided online through the system. With the very mobile nature of the worker, being in the field for long periods visiting clients, the access to the system could pose problems. Dial-up access could be time consuming and frustrating. Response time would be one of the major factors for corporate customers in quantifying the service levels of the bank. The customers may be of a captive nature in the short term, and the exit may take some time due to contractual obligations. However, high service levels could keep the corporate customers on a long-term basis in line with the bank's objective of "achieving at least a 5% increase in agreed customer satisfaction measures." Thus, the retention and growth would be of great interest to the bank. Expensive acquisition is also valuable in this category of high value customers.

Example 1—Transition of Travel Process of FM/IBM

Travel process is discussed herein where the mobile-transformation was suggested. The process of travel in the financial markets division is undertaken by an FM/IBM. These managers travel often to meet customers all over Australia.

Figure 4. Activity diagram, FM/IBM travel—current process (before m-transition)



Two activity diagrams are drawn to show the travel process before and after mobile transition. The activity diagram which is a tool from the Unified Modeling Language (UML) is used herein as it is a typical analysis tool. Figure 4 depicts the picture before m-transition and the activities of the FM/IBM involved in the process.

In Figure 4, the current process for travel for the FM/IBM has been depicted. The FM/IBM visits the customers but does not have access to the bank's system most of the time due to bad lines and low line speeds. Therefore the FM/IBM would visit customers and be travelling for about five days every two months with no access to the bank's system most of the time. During this time, the FM/IBM is completely cut off from the news in the bank and e-mails from their own customers as well. Moreover, the customers visited during the trip will not get any feedback for a considerable time until the FM/IBM has access to the bank systems. When he or she finishes travelling and gets back to office, he or she would have to take six hours (approximated, based on current estimates) to read and respond to e-mails before starting on his or her other work such as making proposals for the customers visited.

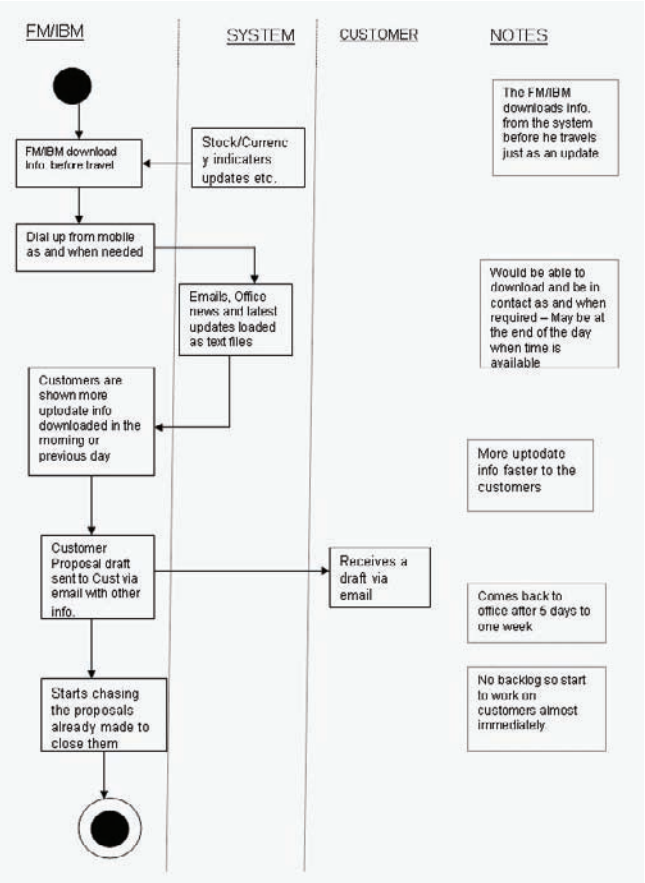
This process of travel is modelled using an activity diagram (which is like a flowchart, and is derived from www.omg.org) for easy comparison. Note that the customer, the most important person in the process, does not appear as an actor in the current process, since the FM/IBM has to wait until he or she gets back to the office to respond to the customers. Thus, the customer is not in the current process at all active.

Figure 5 depicts the scenario once the existing processes from Figure 4 are transformed, with mobile transition taking place. Each FM/IBM would each his or her laptop and could dial into the bank systems. The information would be initially text based so that speed problems could be overcome at initial stages until mobile technology is mature enough to deliver high bandwidth without problems.

The FM/IBM could download e-mails and relevant figures every night or every morning as and when he or she has free time before or after visiting customers while travelling. This would enable him or her to be in contact with the bank regularly. Customers could get draft proposals via e-mail since the FM/IBM could do the proposals quickly after visiting customers without having to wait until he or she gets back to the bank. It is also time saving since he or she is updated regularly and does not have to spend six long hours reading and responding to e-mails as per current estimates, since all that has already been done during his or her free time while travelling. Also note that the customer is an actor in this process. The customer gets the feedback while on travel.

Thus, mobile transition should save a lot of time for the managers in downloading and reading e-mails, and also should keep them in line with the current rates in the very volatile financial markets area. Also the entire sales process has been expedited, with the customers getting the draft version of their contracts very early. Thus the sales process would be shortened by several days, which could bring substantial income. Cus-

Figure 5. Activity diagram, FM/IBM travel—new process (after m-transition)



tomers should also perceive this state-of–the-art process positively, which appears to be very active in comparison to the existing process.

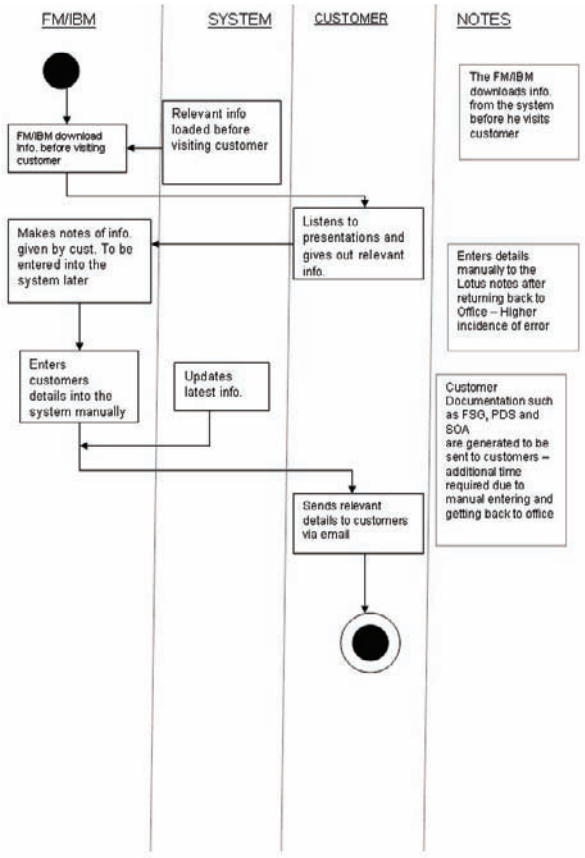
Example 2—Transition of Customer Meeting Process of FM/IBM

The second process discussed herein is that of customer meetings of the FM/IBM. This is also very important for the bank since these meetings could bring very high-value business. Most of these meetings would be happening during the travel process. Therefore, the access to bank computers and systems will only be available through dial up. Speeding up the process would be profitable to the bank on the one hand and also would give them an advantage on customer service on the other.

Customer Meeting Process Before M-Transition

Figure 6 shows the current process of customer meeting for the FM/IBM. The FM/IBM visits the customers, but does not have any forms in his or her laptop to enter any data at the customer's site. The FM/IBM must get information and then come back to the office and manually enter this data to generate relevant customer documentation. Thus, there is a considerable time gap until the customer receives final documentation. This could be subject to errors, and due to manual

Figure 6. Activity diagram, FM/IBM customer meeting—current process (before m-transition)



entry without verification, time is needed to get customer feedback after the entry.

Customer Meeting Process—After M-Transition

In Figure 7, the current customer meeting process has been changed with the introduction of mobile technology. The new process anticipated after the m-transition has taken place is shown. The FM/IBM would have his or her laptop preloaded with forms to enter customer data, which could later be loaded into Lotus Notes. Data is entered and verified at the customer's site, saving considerable time. The customer could get a draft report until final report (uploaded and updated with latest data) reaches him.

Considerable time savings is apparent with m-transition taking place. This would be crucial in higher productivity for the FM/IBM, while customers perceive higher service levels as well as remain ahead of the competition.

THREE PERSPECTIVES IN M-TRANSITION

To build on Figure 3, where transition of business processes and technology were merged in order to look at new business processes, Figure 8 describes the typical requirements on the bank's side, generalizing all the processes taken into consideration. The diagram considers the processes kept intact, changed, and scrapped, and also looks at the en-

Figure 7. Activity diagram, FM/IBM customer meeting—current process (after m-transition)

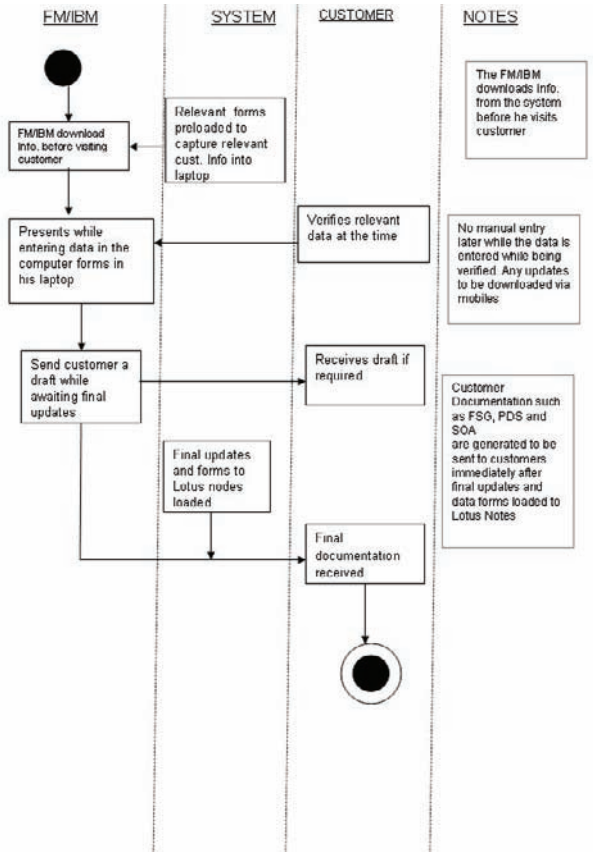
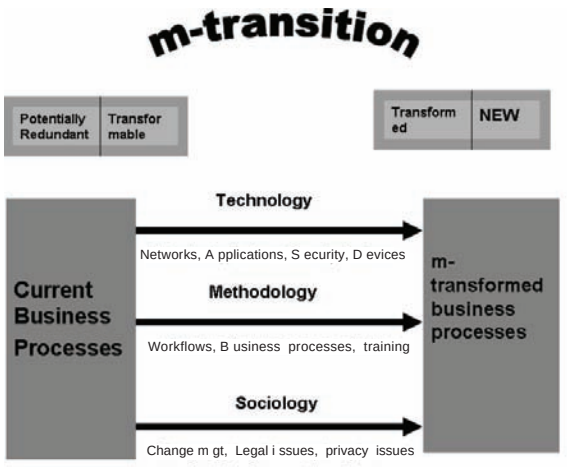


Figure 8. Mobile transition of business processes in technology, methodology, and sociology perspectives



tire change from technology, methodology, and sociology perspectives (Unhelkar, 2003).

These perspectives are useful in managing the entire organizational change due to m-transition.

Technology Perspective

The issues such as what applications should be used, how these applications should be integrated and the networking of these applications, the security issues in the new organization (the mobile transformed organization), and the devices that could be used to facilitate the employees in the service delivery aspects must be decided. In the FM/IBM travel example, the laptops and how to deliver connectivity while travelling has to be looked at from the technology perspective. The aspect of altering certain bank systems' fit enough for the mobiles to download them into a laptop computer has to be investigated. This technology is currently available.

Methodology Perspective

All the procedures that should be followed and adhered to in adopting the new business processes and the approach to be followed in order to transform into the new organization should be discussed. It is also very important that the employees are trained on the new security measures, devices, and new business processes. In the travel example, the new processes followed in order to download e-mails, delivery of business proposals, and so forth are the considerations in the methodology perspective.

Sociology Perspective

Wider issues include the management of the entire change process along with any legal implications and privacy issues in the new organization. Providing training to employees in these areas is also of utmost importance. In the travel example, the

Table 1. Comparison of the current processes with the proposed processes with regard to travel of the FM/IBM

Activity/Attribute	Current Process P	roposed Process
1. Downloading information before travel	Stock/currency indicator updates at office	Stock/currency indicator updates at office
2. Dial up for updates during travel	Too slow and difficult for updates	Mobile dial up for updates as frequently as necessary
3. Liaison with office I	nfrequent/almost nil due to bad lines, etc.	Always in touch with frequent periodic updates
4. E-mail contacts A	lmost nil	Periodic updates every evening with office
5. Feedback to the customers visited	Several days since the visit T	he proposal could be delivered within a day
6. Timeliness of information on hand	Could be several days old O	nly few hours since last update

Table 2. Comparison of the current processes with the proposed processes with regard to customer meeting of the FM/IBM

Activity/Attribute	Current Process P	roposed Process
1. Downloading information before visiting customer	Relevant forms are downloaded at office	Relevant form downloaded at office is updated with current information just before visit
2. Verification of customer information	Makes notes to do corrections later while entering	Updates and verifies data at current time before providing a draft to customer at his/her site
3. Re-verification of customer information	May be needed to be done via e-mail due to manual entering	Not required since it is already done
4. Liaison with office	Infrequent/almost nil due to bad lines, etc.	Always in touch with frequent periodic updates
5. E-mail contacts A	lmost nil	Periodic updates every evening with office
6.Feedback to the customers visited	Several days since the visit to enter the information and verifications	The proposal could be delivered within a day
7. Timeliness of information on hand	Could be several days old O	nly few hours since last update

training of the FM/IBM on the new processes, any legal issues falling therein in giving the managers the authority to provide preliminary proposals, and so on will fall within the sociology perspective.

Table 1 shows that considerable time savings could be achieved in m-transition. There are other measures as well that are significant to consider. The customers who are mostly corporate clients would like to get up-to-date information. This falls within the customer perspective of the balanced score card of the bank.

Table 2 shows the comparison of the current and the proposed processes for customer meetings. Considerable time savings appear on delivery of proposals, and so forth. This also leads to better accuracies, as the data is entered at the customer site and verified then and there. Thus considerable monetary savings result due to sped-up process, and also additional revenue results due to customers being signed up earlier than the current system.

There would be additional costs for laptop and software upgrades in the proposed processes. However, benefits would be much more compared to costs involved since large time savings and additional revenue would compensate for one-time costs of upgrades. There will also be intangible benefits such as better customer satisfaction due to timely delivery and better employee satisfaction due to saving of considerable time, which could be used to make more customer visits.

The main concentration was on the business process perspective of the balanced score card. However, the correct implementation of this perspective would lead to improvements on the customer perspective leading to learning and growth and financial perspectives in a rolling effect.

CONCLUSION AND FUTURE DIRECTIONS

The transformation of the travel process and customer visit process of fund managers and international business managers introducing mobile technology into the bank created a situation wherein the bank stood to gain significantly in terms of savings of time and money. The mobile transition also highlighted potential intangible benefits such as better customer focus, timely proposals boosting customer satisfaction, and

significant time savings, leaving the managers with more time to focus on their customers. The outlook created by the intangible benefits should also lead to gain more customers. However, a systematic approach is suggested, introducing the change gradually while educating the employees to be aware of the change. The customers are also being included in this process of gradual change towards an m-enabled organization.

Similar processes in the financial markets division and other divisions are currently under investigation with a view for further m-transformation. Once the other significant processes have also been transformed across all units, to use mobile technology effectively, the chances of the bank achieving a perfect balanced score card will be significantly enhanced.

REFERENCES

Archer, N. (2004). The business case for employee mobility support. In *Proceedings of the IADIS International Conference in E-Commerce*, Lisbon, Portugal.

Arunatileka, S., & Arunatileka, D. (2003, December). E-transformation as a strategic tool for SMEs in developing countries. In *Proceedings of the 1st International Conferences on E-Governance*, New Delhi, India.

Australian Communications Authority. (2005). Vision 20/20: Future scenarios for the communications industry—implications for regulation. *Final Report*, (April).

Balanced Scorecard. (2005). *What is the balanced scorecard?* Retrieved April 17, 2005, from http://www.balance scorecard.org

Coutts, P. (2002). *Banking on the move*. White Paper, Communications Research Forum.

Ginige, A., Murugesan, S., & Kazanis, P. (2001, May). A roadmap for successfully transforming SMEs in to e-businesses. *Cutter IT Journal, 14.*

Herzberg, A. (2003). Payments and banking with mobile personal devices. *Communications of the ACM*, *46*(5), 53-58.

Homann, U., Rill, M., & Wimmer, A. (2004). Flexible value structures in banking. *Communications of the ACM*, *47*(5), 34-36.

Mallat, N., Rossi, M., & Tuunainen, V. K. (2004). Mobile banking services. *Communications of the ACM*, *47*(5), 42-46.

Money Central. (2005). *MsMoney.com—online banking—online fees.*Retrieved April 16, 2005, from http://www.moneycentral.com

Mylonopoulas, N. A., & Doukidis, G. I. (2003). Mobile business: Technological pluralism, social assimilation and growth. *International Journal of Electronic Commerce*, *8*(1), 5-21.

Schneider, G. P. (2004). *Electronic commerce: The second wave* (5th ed.). Thomson Course Technology.

Schwiderski-Grosche, S., & Knospe, H. (2002). Secure mobile commerce. *Electronics & Communication Engineering Journal*, *14*, 228-238.

Shi, X., & Wright, P. C. (2003). E-commercializing business operations. *Communications of the ACM*, *46*(2), 83-87.

Unhelkar, B. (2003). *Process quality assurance of UML-based projects.* Reading, MA: Addison-Wesley.

Unhelkar, B., & Arunatileka, D. (2003, December). Mobile technologies, providing new possibilities in customer relationship management. In *Proceedings of 5th International Information Technology Conference*, Colombo, Sri Lanka (pp. 23-31).

Unnithan, C. R., & Swatman, P. M. C. (2002). Online banking vs. brick and mortar—or a hybrid model? A preliminary investigation of Australian

and Indian banks. In *Proceedings of the 7th CollECTeR Conference,* Melbourne, Australia.

This work was previously published in Handbook of Research in Mobile Business, edited byB. Unhelkar, pp 778-792, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.6
Consumers' Preferences and Attitudes Toward Mobile Office Use:
A Technology Trade–Off Research Agenda

Xin Luo
Virginia State University, USA

Merrill Warkentin
Mississippi State University, USA

ABSTRACT

Consumer preferences, attitudes, and behavior concerning product choice can be of vital importance in the development process and implementation of innovative products or services. The mobile office (MO) is becoming achievable in the business-to-employee (B2E) arena as more work is completed outside the office and the fixed office boundaries extend well beyond the spectrum of the desktop. Potential MO providers (e.g., employers) will encounter adoption resistance as users experience uncertainty. This paper investigates the critical factors in the decision models of consumers when evaluating the acceptance and intention to use MO. It will provide research guidelines for MO designers and developers, IT/IS managers, and IS researchers.

BACKGROUND

Mobile business (m-business, also known as mobile commerce or m-commerce), an emerging extension of electronic business, has received considerable interest among IS researchers, developers, service providers, and end users. Varshney and Vetter (2002) anticipate that the next phase of e-business will be in the area of m-business with the widespread deployment of wireless technologies. Mobile services have penetrated many leading-edge personal markets such as mobile

SMS, mobile games, mobile handset icons, and ring tones. Wireless computing is now becoming widely deployed in the business arena as managers have appreciated the significant added strategic value of having instant access to business information that can enhance work productivity, efficiency, and decision-making, ultimately leading to competitive advantage for the firm. Businesses that cater to consumers' preferences and needs and that capitalize on expanding opportunities, which arise with new technologies, can sustain competitive advantages in today's fiercely competitive marketplace. Deployment of mobile technology infrastructure, along with mobile devices, enables employee mobility and mobility of IT functions. This is transforming businesses processes by enhancing communication, information access, and business transactions from any device anywhere and anytime. Performance benefits from wireless technology adoption are being realized in the business-to-employee (B2E) domain as corporations seek to achieve their business goals by growing their capabilities.

The rapid development of innovative mobile technologies, along with better integration with the existing network infrastructure, presents new challenges for the enterprise. Thanks to existing wireless technologies, such as 2G and 2.5/2.75G, which introduced GPRS (general packet radio service) and EDGE (enhanced data rates for global evolution), new business opportunities are emerging through new value-added services. 3G services are beginning to receive acceptance in such Asian countries as China, South Korea, and Japan. The technological trend and challenge that mobile users are facing is how to better integrate between wireless services, as 3G technologies are being increasingly revamped and further evolved. For the 3G-based CDMA evolutions, handsets will support CDMA, CDMA 1xRTT, and CDMA 1xEV-DO with three kinds of spectrum including 850/1, 900/2, and 100MHz. For the GSM evolution, handsets will support GSM, GPRS, EDGE, and WCDMA, operating in five bands (850/900/1 800/1 900/2 100MHz). In the near future, 4G will surface as a collection of services combining existing technologies, such as 3G and WiFi, with other types of wireless technologies including WiMAX and future evolutions of 3G. 4G will be featured by high usability anytime, anywhere, and with any technology; support for multimedia services at low transmission cost; personalization; and integrated services. As such, 4G will be less disruptive and more widely accepted if the promise is delivered upon. It is expected that 4G networks will be all-IP-based heterogeneous networks that allow users to switch any system at any time and anywhere. 4G systems will not only support data telecommunication services, but also multimedia services. And users in widely diverse locations will use the services, as users can use multiple services from any service provider at the same time. Though 4G mobile technologies may offer even greater opportunities, the gradual maturation and deployment of 3G technologies makes MO become an achievable goal as more work is completed outside the office and the fixed office boundaries extend well beyond the desktop.

There is considerable prior IS research about m-business and wireless technologies (Featherman & Pavlou, 2003; Kleijnen & Ruyter, 2003; Liang & Wei, 2004; Muthaiyah & Ehsan, 2004; Suoranta & Mattila, 2004; Varshney & Vetter, 2002; Zellweger, 1997). However, these research studies have mainly shed light on areas, such as technology acceptance and penetration, as well as technology trends and issues, leaving the domain of consumer preferences and attitudes towards the adoption of innovative products, specifically MO, relatively unexplored. More research is needed to explore the factors that constitute ultimate MO adoption and use, as well as the relative importance of these factors for further diffusion of innovation. In consideration of this objective, we investigate the critical factors in the decision models of consumers when evaluating the acceptance and intention to use MO. Further, we provide research guidelines for MO designers and developers, IT/IS managers, and IS researchers.

INTRODUCTION: MOBILE OFFICE TECHNOLOGY

Most traditional business applications are developed and deployed for use within fixed office boundaries–using hardware that is not mobile. This confinement results in a wide range of limitations and difficulties if employees cannot access needed information whenever and wherever they want, causing postponement in responding to *customer requests, dissemination of inaccurate information, and delivering lower-quality work output* (Intel, 2004a). Advancements in wireless technologies have triggered a proliferation of mobile devices and broadened the spectrum of solutions for new business applications and services. In the post-2G era, where the business mobile information environment is comparatively dynamic, traditional mobile voice services cannot adequately meet customers' business requirements. 3G networks' throughputs are fairly equivalent to the early DSL networks that revolutionized the home office (Gruman, 2005; Varshney & Vetter, 2002). Notably, according to Gruman (2005), 3G will reduce the expectation gap and delivery gap between wireless and wired connections. For businesses, there is increasing demand for mobile access to multifunctional services that can enhance communication and collaboration as well as management of business information. Liang and Wei (2004), Muthaiyah and Ehsan (2004), and Gruman (2005) indicate that emerging 3G technologies, such as CDMA-based EvDO (evolution, data optimized) and GSM-based UMTS (universal mobile telecommunications system), and HSDPA (high-speed downlink packet access), have the potential to revolutionize MO users using notebook computers and handsets over the high-speed wide area network (WAN).

Due to the dynamic nature of today's business environment, employees are spending less time fixed to their desks and more time in collaborative work meetings, telecommuting, and working in remote locations to accomplish their job objectives. Unlike a fixed office where employees are restricted in a limited environment, MO, including *on the road, at home, and at work* (Cisco, 2002; Gruman, 2005; IBM, 2004; Intel, 2004a, 2004b; North-Smith, 2002), expands the reach of the office environment and provides employees with access to their information, applications/services, and teams, in an anytime and anywhere model, thereby eliminating the obstacles of fixed office boundaries. As more work is completed outside the office and as office boundaries extend well beyond the spectrum of desktop computing, many of the solid business benefits from wireless technology adoption are being realized in the B2E domain. According to Kleijnen and Ruyter (2003), MO has great potential to become one of the most widely utilized m-business solutions with the global user base potentially exceeding 100 million in 2004. It can beef up productivity for employees, since having real-time access to business information is key to increasing productivity and corporate profitability as a whole. The congruence of the findings of Kleijnen and Ruyter (2003), Cisco (2002), IBM (2004), North-Smith (2002), Liang and Wei (2004), Gruman (2005), and Muthaiyah and Ehsan (2004) is that unique MO services, thanks to the revolutionarily enhanced 3G technologies, consist of *accelerating mobile communication and collaboration services* (e-mail, e-fax, unified messaging, groupware messaging), *mobile business information management services* (real-time calendar events, address books, to-do task lists, calculator, word processor), and *mobile information access services* (access to CRM, access to corporate files and corporate databases via secure mobile portal, access to external business information services). These process facilitation services are increasingly becoming incorporated in a mobile corporate portal that is a combination of hardware and software with integrated network development, timely information management, and seamless security mechanisms to enable communications between wireless networks and devices.

Being able to create new expectations among business users who want to constantly maintain work sessions without disruption and disconnection *on the road, at home, or at work*, the emerging deployment of the 3G-powered MO initiatives will greatly transform and improve the way employees work and communicate with colleagues, customers, suppliers, and vendors. These improvements also contribute to rapid responsiveness, decreased costs, improved productivity and work efficiency, and better work/life balance in terms of more flexibility and choices. Managers, however, must understand whether and how MO would be accepted and ultimately adopted by employees/users in order to help companies achieve organizational objectives and obtain competitive advantage. Also, potential providers of MO will encounter a high uncertainty about consumers' acceptance and intention to use. A lack of studies directly investigating the adoption and diffusion patterns of MO is to be expected due to the newness of the MO initiatives *per se*. Employee/user behavior in the MO context also has remained rather uncharted territory, which leads to an important topic for further research within the MIS discipline.

THEORETICAL FOUNDATION

The theoretical framework of this chapter is grounded in the innovation diffusion theory (IDT) and perceived characteristics of innovating (PCI). Despite the fact that there is little empirical research conducted on MO, there is a plethora of adoption theories and models that investigate and capture user behavior characteristics. In IS research area, the landmark is the technology acceptance model (TAM), proposed by Davis (1989) and Davis (1993), that identified ease-of-use (EOU) and usefulness as the two key determinants influencing user adoption. However, Plouffe et al. (Plouffe, Hulland, & Vandenbosch, 2001) indicate that TAM's parsimony makes individual responses to new technologies differ depending on the context. In a bid to integrate the main user acceptance models, Venkatesh, Morris, Davis, and Davis (2003) formulated the unified theory of acceptance and use of technology (UTAUT), which exhibits significantly enhanced predictive value for adoption intention, with an adjusted R square of approximately 70%. Yet, one weakness of the UTAUT model is that the empirical base did not include e-commerce or m-commerce technologies, which Venkatesh et al. (2003) identified as needing further investigation and testing. Consistent with Plouffe et al. (2001), Kleijnen and Ruyter (2003) argue that the narrow focus of the adoption concepts hinders us from identifying other potential drivers of m-commerce adoption. User acceptance of m-commerce-oriented MO can be identified as a technology adoption. Following the recommendation of Kleijnen and Ruyter (2003), we thus focus on the adoption process in search of valuable insights for building a theoretical framework for critical success factors of MO.

In the domain of adoption process, innovation, and diffusion (ID) is extensively researched and is "*perhaps one of the most widely researched and best documented social phenomena*" (Mahajan & Peterson, 1985). In ID research, IDT, proposed by Rogers (1995), is the most acceptable and reliable framework that has been fairly widely validated in sociology, psychology, and communications as well as IS to explain user adoption of technical innovations. According to Rogers (1995), innovation is "*an idea perceived as new by the individual*" and diffusion is "*the process by which an innovation spreads*." As a consequence, diffusion processes result in the acceptance or penetration of a new idea, behavior, or physical innovation (Rogers, 1995). To make an innovation successful, Rogers' IDT has identified five critical characteristics: relative advantage, compatibility, complexity, communicability, and trialability. Further, Moore and Benbasat (1991) expanded IDT by proposing perceived characteristics of

innovating (PCI) in which three additional constructs, including voluntariness, image, and result demonstrability, were identified for ID research. As the key antecedents to technology adoption decision (Plouffeet al., 2001), these PCI factors, along with the additional constructs resided in MO context, must be explored and explained.

- **Relative advantage (perceived usefulness):** The degree to which an innovation is perceived as being better than its precursor (Moore & Benbasat, 1991; Rogers, 1995; Venkatesh et al., 2003). This construct, particularly in MO study, contains issues such as usability and availability. Here, usability relates to enhanced 3G network throughput for wireless business applications or services and application design to deliver the right information to the right users; availability relates to assured network that is reliable in the wireless network in terms of seamless service coverage and handy mobile access.
- **Compatibility:** The degree to which an innovation is perceived as being consistent with the existing values, needs, and past experiences of potential adopters (Moore & Benbasat, 1991; Rogers, 1995; Venkatesh et al., 2003). This relates to the issue of relevance of technology (Wang & Butler, 2003) and interoperability, as well as integration in terms of open standard, of the MO environment with mainline business and office support systems.
- **Complexity**, also referred to as perceived ease-of-use (PEOU), is the degree to which an innovation is perceived as being difficult to use (Davis, 1989, 1993; Moore & Benbasat, 1991; Rogers, 1995; Venkatesh et al., 2003). This can relate to ease of accessing business information, the amount of effort it takes to comprehend the functionality of mobile devices and programs, and how easy it is to retrieve and send information in 3G networks.
- **Communicability:** The extent to which the innovation lends itself for communication, particularly the extent to which the use of the innovation is observable by others (Moore & Benbasat, 1991; Rogers, 1995; Venkatesh et al., 2003). This relates to social influence (also known as "social norm"), since use of an innovation is often influenced by a social context, including supervisors, peers, and others that are highly regarded (Karahanna & Straub, 1999; Kleijnen & Ruyter, 2003). Users might perceive the need to use MO services to achieve work objectives with job-related participants.
- **Image:** The degree to which use of an innovation is perceived to enhance one's image or status in one's social system (Moore & Benbasat, 1991; Rogers, 1995; Venkatesh et al., 2003). According to Plouffe et al. (2001), it signifies the extent to which a user believes an innovation will add social prestige or status.
- **Result demonstrability (RD):** The tangibility of the results of using the innovation, including their observability and communicability (Moore & Benbasat, 1991; Rogers, 1995; Venkatesh et al., 2003). This relates to visibility, which Wang and Butler (2003) consider as the degree to which change is apparent to users. The more visible and more accessible technology changes are, the more likely individual users are to be aware of them and, therefore, more likely to carefully evaluate them. This is in line with Zaltman, Duncan, and Holbek (1973), and Moore and Benbasat (1991) that "*the more amenable to demonstration the innovation is and the more its advantages are, the more likely it is to be adopted*."

Additionally, we propose that *perceived risk* (PR) and *perceived security* (PS) be included into the ID taxonomy for wireless computing. Despite significantly improved technical advancements

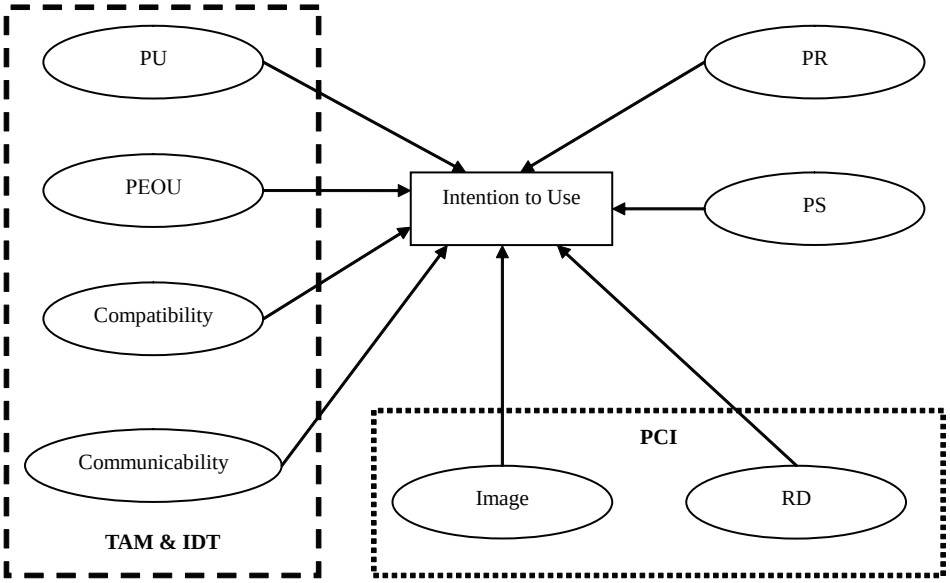
in 3G security mechanisms, MO users might still be concerned about sensitive information transmitting in the open airwaves. In fact, one of the most pressing concerns for businesses considering wireless computing relates to the security in operations. *PR* refers to the extent to which a functional or psychosocial risk a consumer feels he/she is taking when purchasing and use a product. Kleijnen and Ruyter (2003) further define perceived risks as the extent to which risks are attributed to the mobile services. It greatly affects a user's intention to use a particular product/service (Featherman & Pavlou, 2003; Pavlou & Gefen, 2004). User's perception of unsatisfactory security on the Internet is one of the primary reasons hindering online operation (Zellweger, 1997). From a user's perspective, adapted from Chellappa (2005), we define *PS* as the subjective probability with which users believe their sensitive information (business or private) will not be viewed, stored, and manipulated during work sessions by unauthorized parties in a manner consistent with their confident expectations. And we deem that the relationship between perceived risk and perceived security for user adoption is reversed: the more risks perceived by users, the less likely users are to adopt the MO services; the more security perceived by users, the more likely users are to adopt the MO services.

Based on TAM, IDT, and PCI, as well as such new constructs as perceived risk and perceived security, a conceptual framework for mobile office adoption is proposed (see Figure 1). We argue that PR and PS are negatively related to the intention to use mobile office. Further, other constructs, such as perceived usefulness, complexity, compatibility, communicability, image, and result demonstrability are positively related to the intention to use mobile office.

Implications and Research Agenda

The rapid and innovative developments in wireless technologies can be of significant contribution to current and future professional communications. Mobile office technology has the strong potential to extend the office boundary toward an anywhere/anytime model. It is of crucial importance to gauge how MO users would make the decision for technology acceptance and what key

Figure 1. Proposed framework for mobile office adoption



factors influence their decision-making. Table 1, containing guidelines for research and practice, is thereby proposed for both researchers and practitioners for further analysis and investigation on MO acceptance and trade-off research. It is necessary for both practitioners and researchers to understand which factor is the most important and which factor is comparatively the least important toward technology adoption.

IS researchers should develop and pursue sound empirical research based on the IS acceptance theories, including TAM, IDT, and PCI, combined with practical field-based research utilizing the case method and surveys of MO users. Such research would introduce a new avenue for innovative technology adoption research and would be of help for practitioners in the field of professional communications to understand the trade-offs that consumers are willing to make for the technology adoption decision.

Mobile office must be a strategic part of a company's IT portfolio, rather than simply a technological tool for tactical productivity gains. Furthermore, the workplace of the future will be an open, collaborative realm, with less reliance on geographic limitations between the physical location of the organization and its employees. We must also pursue practical questions such as (1) which factors are consumers most concerned with when electing to adopt MO for professional communications?, and (2) what is the trade-off that consumers might make for accepting and adopting MO technology?

The application of emerging technologies must be accompanied by careful organizational research that allows managers to understand the user's requirements, the impact of the technology on teams and organizations, the factors that lead to greater success with the technology, and the expected outcomes of various implementation scenarios. Without careful *a priori* and *ex poste* analysis, technology can have unintended consequences. The research factors and preliminary agenda detailed in this paper will facilitate greater understanding of this important socio-technical phenomenon, and will contribute to its success. The findings of this stream of research will enable future managers to confidently implement MO with a presumption of positive strategic outcomes for the individual users and also for the entire organization.

REFERENCES

Chellappa, R. K. (2005). *Consumers' trust in electronic commerce transactions: The role of*

Table 1. Guidelines for research and practice

What are the objectives?
• Understand how users make the decision for MO technology acceptance.
• Understand which key factors influence MO users' decision making when electing to adopt MO for professional communications. What are the trade-offs?
• Understand the user's requirements, so that appropriate MO technologies can be adopted.
• Understand the impact of the technology on teams and organizations.
• Train individuals and teams so the technology can be applied most effectively.
• Understand the factors that lead to greater success with the technology.
• Train to leverage the technology for the organization's benefit.
• Understand the expected outcomes of various implementation scenarios.
• Provide awareness to employees so that individuals use technologies in best ways.
• Develop "best practices" for MO implementation to guide further usage patterns.

perceived privacy and perceived security. Under review.

Cisco. (2002). *Cisco mobile office.* Retrieved August 15, 2005 from http://www.cisco.com/asiapac/mobileoffice/index.shtml

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 318.

Davis, F. D. (1993). User acceptance of information technology: System characteristics, user perceptions and behavior impacts. *International Journal of Man-Machine Studies, 38,* 475-487.

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-services adoption: A perceived risk facets perspective. *International Journal of Human-Computer Studies, 59,* 451-474.

Gruman, G. (2005). Taking it to the streets: 3g arrives. *InfoWorld, 27*(10), 30.

Heijden, H. v. d. (2004). User acceptance of hedonic information systems. *MIS Quarterly, 28*(4), 695-704.

IBM. (2004). *Mobile office: The next breakthrough in professional productivity gains.* Retrieved August 15, 2005 from http://www.incentric.com/solutions/solutions/Mobileoffice_The_next-breakthrough.pdf

Intel. (2004a). *Anytime, anywhere mobile office*. Retrieved August 15, 2005 from http://cache-www.intel.com/cd/00/00/10/28/102829_pp022001_sum.pdf

Intel. (2004b). *Solutions for mobile network operators*. Retrieved August 15, 2005 from http://www.cisco.com/en/US/netsol/ns341/ns396/ns177/networking_solutions_white_paper09186a00801fc7fa.shtml

Karahanna, E., & Straub, D. W. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly, 23*(2), 31.

Kleijnen, M., & Ruyter, K. d. (2003). Factors influencing the adoption of mobile gaming services. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce technology, theory and applications* (Vol. 11, pp. 202-217). Hershey, PA: Idea Group Publishing.

Liang, T.-P., & Wei, C.-P. (2004). Introduction to the special issue: Mobile commerce applications. *International Journal of Electronic Commerce, 8*(3), 7.

Mahajan, V., & Peterson, R. A. (1985). *Models for innovation diffusion (quantitative applications in the social sciences).* Beverly Hills, CA: SAGE Publications.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 173-191.

Muthaiyah, S., & Ehsan, S. D. (2004). *Readiness towards 3g: Antecedents of 3g adoption and deployment in Malaysia*. Paper presented at the Wireless Information Systems.

North-Smith, L. (2002). *Mobile office Solutions: Real-time access to PIM/e-mail.* Retrieved August 15, 2005 from http://www03.ibm.com/industries/wireless/doc/content/bin/real-time_access.pdf

Pavlou, P. A., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information Systems Research, 15*(1), 37.

Plouffe, C. R., Hulland, J. S., & Vandenbosch, M. (2001). Richness versus parsimony in modeling technology adoption decisions—understanding merchant adoption of a smart card-based payment system. *Information Systems Research, 12*(2), 208.

Rogers, E. M. (1995). *Diffusion of innovation* (4th ed.). New York: Free Press.

Suoranta, M., & Mattila, M. (2004). Mobile banking and consumer behavior: New insights into the

diffusion pattern. *Journal of Financial Service Marketing*, *8*(4), 354-366.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7*, 185-198.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425.

Wang, X., & Butler, B. S. (2003). *Individual technology acceptance under conditions of change*. Paper presented at the International Conference on Information Systems, Seattle, WA.

Zaltman, G., Duncan, R., & Holbek, J. (1973). *Innovations and organizations.* New York: Wiley and Sons.

Zellweger, P. (1997). Web-based sales: Defining the cognitive buyer. *Electronic Markets, 7*(3), 10-16.

This work was previously published in E-Business Process Management: Technologies and Solutions, edited by J. Sounderpandan; T. Sinha, pp. 175-284 , copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 6.7
Customer Relationship Management on Internet and Mobile Channels:
An Analytical Framework and Research Directions

Susy S. Chan
DePaul University, USA

Jean Lam
IBM, USA

ABSTRACT

The Internet has served as an effective channel for companies to build and manage relationships with customers. The mobile channel, emerging from the convergence of wireless communications and the mobile Web, promises to deliver additional support to meet consumer needs. This chapter examines features of customer relationship management (CRM) as implemented on the Internet (eCRM) and the mobile channel (mCRM) from the customer's perspective. It further explores how companies can better coordinate their CRM strategies between these two channels to support e-commerce customers. We propose an analytical framework to examine the current eCRM and mCRM practice in terms of customer loyalty, branding, customer satisfaction, customization, and trust. These five factors affect customer acquisition, sales and services, and customer retention. A checklist was developed to guide the evaluation of CRM practice for e-commerce sites. Several examples and research directions are discussed in the chapter.

INTRODUCTION

Customer relationship management (CRM) involves the deployment of strategies, processes, and technologies to strengthen a firm's relation-

ship with customers throughout their lifecycle – from marketing and sales, to post-sales service. The motivation for CRM stems from companies' desire to increase their revenues and profitability through improved customer satisfaction and retention (Reichheld, 1996; Reichheld & Sassar, 1990; Winer, 2001). Internet technology has transformed CRM into electronic CRM (eCRM), because companies can use Internet technologies to capture new customers, track their preferences and online behaviors, and customize support and services. Furthermore, the convergence of wireless communication and mobile Internet provides companies with opportunities to interact with their customers through a new mobile channel.

Despite the potential growth of mobile commerce for location-aware and customer-aware services (Varshney, 2003), recent research points out that most mobile sites were designed primarily for supporting existing e-commerce customers (Chan et al., 2002). Customers who are already familiar with the interface and services provided on a company's Web site are likely to benefit more from its mobile site. Therefore, out of a wide range of mobile services (Varshney, 2003; Varshney & Vetter, 2001), it is logical to consider the mobile channel as appropriate for building and retaining relationships with existing customers. Because of current technology and usability barriers (Chan & Fang, 2003; Ernst & Young, 2001; Shim et al., 2002), businesses and consumers are hesitant to adopt the mobile channel. Research is needed to examine how the mobile channel can be effectively leveraged to attract and retain e-commerce customers.

The main purpose of this chapter is to provide an analytical framework for examining how companies can build and manage relationships with their e-commerce customers by leveraging the Internet and the mobile channels. We take a customer's perspective in examining the firm-customer interactions through these two channels. The chapter focuses on the features of content and services presented on companies' Web and mobile sites. Our intent is to identify (a) how CRM can be effectively coordinated between these two channels, and (b) key research questions pertinent to the eCRM and mCRM coordination. Our proposed framework examines CRM implementation across three phases of an e-commerce site's interactions with its customers – acquisition, sales/service, and retention. Interactions in each phase are also examined along five factors that are essential to Internet-based CRM solutions: (1) customer loyalty, (2) branding, (3) customer satisfaction, (4) customization, and (5) trust. We apply this framework to several e-commerce sites and their corresponding mobile sites to explore how CRM features are currently incorporated into these sites. A checklist, derived from the framework, was used for the site analysis. From this exploratory work, we identify commonalities between eCRM and mCRM, and the respective roles played by each channel. Furthermore, we propose a set of research questions for future investigation. This chapter contributes to a better understanding of mobile commerce technology and strategies. In particular, it addresses how organizations can optimize CRM by leveraging the unique characteristics of Internet and wireless technologies.

CRM AND E-COMMERCE

CRM Research

CRM is a strategy for companies to build and manage long-term relationships with their customers. Researchers have shown that CRM implementation can provide better customer service, as well as improvement and management of customer expectations and loyalty (Cho et al., 2001; Reichheld, 1996; Reichheld & Sassar, 1990; Romano, 2001; Winer, 2001). CRM also complements a firm's capability to present products, quality, and services to its customers (Chen & Sukpani, 1998). By implementing CRM solutions, many

firms expect to improve profitability by gaining customer loyalty, customizing offerings, and lowering costs.

The increasing pressure on profitability has motivated companies across different industry sectors to invest in CRM solutions. An Internet impact study shows that CRM applications are the most widely adopted e-business solutions (Varian et al., 2002). On the average, 71% of companies in this study have adopted Internet-based solutions for customer service and support, 68% adopted e-marketing for customer development, and 52% adopted e-commerce for sales and transactions. Generally, an investment in retaining repeat customers contributes more to a company's profitability than do marketing expenditures for attracting new customers. Reichheld and Sassar (1990) have demonstrated that the overall profit generated by existing customers over seven years exceeded those generated by new customers. For e-commerce companies, the need to expand customer base and attract repeat customers may be equally important for their sustainability. Forrester Research (2003) has projected online retail sales to grow to $96 billion in 2003, a 26% increase from 2002. However, this growth only represents 4.5% of total retail sales in 2003. E-commerce still has potential for further growth. Therefore, a dual emphasis on customer acquisition and retention is important to achieve profitability for e-commerce companies.

CRM approaches are built on the concept of relationship marketing, which emphasizes building a long-term relationship with individual customers. In contrast, traditional transaction marketing maintains a short-term focus on the transaction of products. Relationship marketing embraces strategies of personal and ongoing exchanges with customers for brand management, feedback, knowledge acquisition, and customer differentiation (Moon, 2002). Knowledge acquisition enables companies to gather better information about their customers through some type of self-disclosure. Customer differentiation allows companies to offer services that match different customer needs and customer values. Essential to relationship marketing is the strategy of customizing the marketing mix – products, services, communications, channels, and price. Thus, "the relationship marketing process involves an iterative cycle of knowledge acquisition, customer differentiation, and customization of the entire marketing mix" (Moon, 2002).

Researchers and industry practice tend to adopt a suppliers' (or firms') perspective of relationship marketing by emphasizing the goal of customer retention and profitability (Hennig-Thurau & Hansen, 2000; Hennig-Thurau & Klee, 1997). Most of the relationship and loyalty programs tend to focus on the company's drive for transforming relationships into profit (Winer, 2001). In contrast, less attention has been devoted to understanding customers' motives and wishes regarding their relationships with the companies.

The IT approach to CRM stems from early research on customer resource life cycle (CRLC). Different life cycle modes have been used for analyzing how a company can strengthen its relationship with customers through the application of information technology (Burnstine, 1980; Ives, 1984). Ives (1984) expands IBM's four-stage model into 13 steps to: (1) establish customer requirements, (2) specify requirements, (3) select sources, (4) order products or services, (5) authorize and pay for product/services, (6) acquire products/services, (7) test and accept products/services, (8) integrate products/services into existing processes, (9) monitor product/service performance, (10) upgrade products/services, (11) maintain the condition of products/services, (12) transfer or dispose of products/services, and (13) account for the products/services. In practice, this CRLC model may be simplified into three broad phases of interactions between a firm and its customers – acquisition, sales/service, and retention.

For e-commerce, the acquisition phase emphasizes marketing activities that are based on personalization technology to facilitate the customer

decision process in the pre-sales phase. During the sales phase, creating customized transactions makes a customer's shopping and purchasing experience more efficient and satisfactory (Lee & Shu, 2001). An e-commerce site can enhance customer retention by building customer trust and loyalty through a variety of online features (Hoffman et al., 1999; Lee & Shu, 2001; Papadopoulou et al., 2001). These features enable customers to check the status of transactions, shipments and orders, and to work collaboratively with the sales force. Incentives for repeat visits through push e-mails and other loyalty programs can also enhance customer trust and loyalty.

Electronic CRM

Internet technology enables companies to capture new customers, track their preferences and online behaviors, and customize communications, products, services, and price. The mass customization concept, or the one-to-one approach, promoted by writers such as Peppers and Rogers (1993), has become the "mantra" of eCRM (Winer, 2001). A company's e-commerce Web site integrates marketing, sales/service, and post-sales support as a seamless front-end to meet customer needs. Therefore, e-commerce Web sites have become viable channels for customer acquisition, sales/service, and retention.

The Internet plays an active role in customer acquisition via e-marketing, which emphasizes proactive and interactive communications between companies and their customers. Companies can provide information on products and services on their Web sites for prospective customers. Advanced searching capability and functions for product and service inquiry can attract new and repeat customers to visit, compare products and prices, and reach decisions for purchase. Companies also create online communities to facilitate social groups among existing and prospective customers. Online product discussions and reviews encourage customer-initiated communications between firms and customers and among fellow customers (Strauss, 2000). These online communities improve customer loyalty, branding, and trust, which can lead to increased sales and improved customer relationships (Lee & Shu, 2001).

The Mobile Channel

The convergence of mobile Internet and wireless communication technology has promised users "anytime anywhere" access to information for their work and personal communication. Mobile services support m-commerce transactions and improved management of personal activities, mobile office, and mobile operations (Alanen & Autio, 2003). Among many mobile applications proposed by wireless researchers (e.g., Kannan et al., 2001; Mannecke & Strader, 2001; Varshney & Vetter, 2002), mobile financial applications, location-aware and context-aware advertising, and location-based services seem to hold special promise (Varshney, 2003). These mobile services may provide customized support for individual users.

Many researchers point to four reasons that the mobile channel could be used to build relationships with customers. The mobile channel and wireless technology enable companies to: (1) personalize content and services; (2) track consumers or users across media and over time; (3) provide content and service at the point of need; and (4) provide content with highly engaging characteristics (Kannan et al., 2001). Anckar and D'Incau (2002) point out that consumers are most interested in services with high mobile values that meet spontaneous and time critical needs, such as checking stock quotes, driving directions, and short messages.

A recent study indicates that, at present, most of the available mobile sites tend to share similar interfaces with their corresponding Web sites and primarily support existing customers (Chan et al., 2002). For example, Amazon only offers

the 1-click order option for purchasing from its wireless site. This feature does not allow customers to review order details before submitting the order. Once an order is submitted, it is difficult for customers to navigate to the right screen on the handheld device to cancel the order. Therefore, only experienced mobile customers who have already built trust in Amazon and the interface of the 1-click order option would find it efficient to order products from the mobile Amazon site. In comparison, new customers would be hesitant to use the mobile channel. In the case of accessing eBay by a wireless PDA device, users often encounter a large number of results from a product search. The high volume of transferred data can result in connection errors and frustrate new customers. Only seasoned eBay customers are more likely to benefit from using a handheld device to monitor a bid in progress.

These findings imply that current mobile sites have been designed primarily to support existing e-commerce users. The inherent difficulties using the wireless technology may discourage prospective customers from exploring a new mobile site. These barriers include limited bandwidth and poor connectivity, small screen display, and difficulty in input formats of wireless handheld devices (Chan & Fang, 2003). The study by Anckar and D'Incaur (2002) indicates that e-commerce users are more likely to adopt m-commerce services. Their finding further confirms the proposition that the mobile channel is more relevant to customer support and retention than acquisition.

AN ANALYTICAL FRAMEWORK

Based on the above review, we propose an analytical framework for examining how e-commerce sites implement CRM strategies online and on the mobile channel. This framework views eCRM and mCRM across three phases of customer interactions with an e-commerce site — acquisition, sales/service, and post-sales retention. In each phase, the framework also examines CRM implementation according to five inter-related factors — customer loyalty, branding, customer satisfaction, customization, and trust. These five factors represent the salient characteristics of relationship marketing, as emphasized by Winer's (2001) customer relationship model, Lee and Shu's (2001) framework of American Customer Satisfaction Index (ACSI), and Andaleeb's (1992) research on trust in relationship marketing. Winer's model (2001) identifies customer satisfaction as the key to establishing customer relationships. Customer loyalty, customization, community building, and unique services with branding contribute to high customer satisfaction and retention. Winer further emphasizes that delivering a high level of customer satisfaction that exceeds customer expectation increases profitability – a key objective of relationship management strategy. Lee and Shu's (2001) ACSI framework explains the importance of customization and brand building to raise customer perception of quality and value of products and services. A higher level of perceived quality and value of products and services contributes to customer satisfaction and customer loyalty in a multi-layer fashion. Andaleeb and Anwar (1996) point out that trust is one of the most widely confirmed factors in relationship marketing. Table 1 provides an overview of the five CRM factors and their roles in the three phases of firm-customer interactions. The ensuing sections discuss the proposed framework in greater detail.

Customer Loyalty

Dick and Basu (1994) conceptualize customer loyalty as the strength of the relationship between an individual's relative attitude towards an entity (brand, service, store, or vendor) and repeat patronage. The work of Lowenstein (1997) further introduces the concept of commitment into the relational paradigm through the identification of what he termed "commitment-based" companies.

Table 1. An analytical framework for CRM

Factors\ Phases	**Acquisition**	**Sales & Service**	**Retention**
Customer Loyalty	Loyalty program details Loyalty program enrollment Loyalty program status display Loyalty program reward Custom status customer display	Custom service for member and status customer Capability to redeem reward Membership convenience service	Delivery options Order tracking Help desk service Product review and discussion group Customer feedback/survey Return policy
Branding	Large customer community Unique branding product/service Exclusive product	Exclusive interface for transaction support	Exclusive product
Customer Satisfaction	Information consistency Product variety Product and price comparison Attractive graphic interface Self-management capability Company details Efficient and accurate search engine Product review	Easy to use transaction interface Alternate product and pricing recommendations Payment options	Delivery options Order tracking Help desk service Product review and discussion group Customer feedback/survey Return policy
Customization	Profile and preference self-manage capability Self-help, FAQ Personal custom display Preference product suggestion	Question posting/ inquiry capability Use customer profile information to complete product transaction Fast check-out service	Profile and preference self-manage capability Self-help, FAQ Customer purchase history, detail billing, delivery history, and status Delivery tracking Custom incentive Custom services. E.g., personal reminder E-mail promotion notification
Trust	Information consistency Privacy statement for customer profile Authentication mechanism Authorization mechanism Third party signature	Payment options Order confirmation Security measurements, digital certification, SSL transmission, encryption, non-repudiation Authentication mechanism	E-mail order notification Help desk support

These are firms that adopt a proactive approach to creating customer value and loyalty management by constantly anticipating and responding to latent customer needs (Lowenstein, 1997). According to Aakar (1991, 1996), customers who exhibit the highest level of commitment to a brand will also demonstrate a high level of loyalty. Dekimpe et al. (1997) emphasize that companies should treat their loyal customers as a competitive asset. Indeed, customer loyalty represents a basis for charging price premiums and a barrier to competitive entry (Aaker, 1996). Accordingly, companies can provide unique customer benefits that are difficult for competitors to match in order to achieve a higher level of customer loyalty (Evans & Laskin, 1994).

Relationship marketing strategy includes introducing customer loyalty programs, like

frequent flyer and reward programs, membership, and online community. For example, American Airlines offers the AAdvantage program for its frequent travelers. This program encourages customers to accumulate mileage from traveling with American Airlines to redeem free plane tickets for future trips. Similarly, Starwood Hotel Group has implemented the Starwood Preferred Guest program for repeat customers to accumulate hotel points with Starwood-chain hotels and redeem these points for automatic upgrades and free vacations.

E-commerce players can achieve customer loyalty by providing the following CRM features:

- Detailed information about the loyalty program;
- Incentives for joining the loyalty program;
- Instructions for creating a personal account;
- Detailed information about a personal account with purchase history and loyalty status information;
- Personalized services for repeat customers;
- Frequent buyer incentives such as discounts or personal upgrade services;
- A status page on customer loyalty status, upgrade options, redeem procedures, and special discounts/promotions;
- Special services for frequent buyers—no cost delivery, priority seating, and/or 1-click checkout; and
- Online mechanisms to actively collect feedback from frequent customers.

Companies have used loyalty programs for marketing and attracting new customers. These programs are also important for repeat customers who value the effectiveness and convenience for registered members to redeem rewards and updates. An e-commerce site can also enhance customer loyalty through retention efforts such as customer feedback, status information about loyalty programs, and help desk services. Therefore, loyalty programs are important for all three phases of firm-customer interactions.

Branding

The efficient use of branding can increase product differentiation (Aaker, 1991, 1996) and build customer relationships by influencing a customer's attitude towards the brand. A customer's perception of the functional, experiential, and symbolic aspects of the product can strengthen customer loyalty to the company. Good branding tactics include selling exclusive products and services and having a large e-community of customer participants.

In an e-commerce environment, branding involves a number of strategies:

- Building a large customer community through online chat rooms, discussion sessions, and product reviews (e.g., online chat rooms on MSN.com and Amazon's community of online reviewers) to accentuate the customer's experience with the brand;
- Providing unique branding products or services (e.g., eBay's auction trading) to differentiate a site from its competitors;
- Providing exclusive brand name products, such as Gap.com and JCrew.com;
- Providing supplementary services to enhance the main business and raise the barrier to entry, such as Citibank's online personal banking services through citi.com; and
- Providing unique interfaces to support customer shopping experiences (e.g., Amazon's one-click ordering interface and Peapod's grocery shopping interface).

Large customer communities, unique branding of products and services, and exclusive brands help to attract new customers. Exclusive products and services help to build long-term customer loyalty and retention. Internet technology has

also enabled companies to create brand recognition through their unique user interface design for transaction support.

Customer Satisfaction

Customer satisfaction is a major factor in retaining long-term customers and can indirectly attract new customers through referral. Researchers have used the confirmation/disconfirmation (C/D) paradigm to explain customer perception of performance and quality (Anderson & Sullivan, 1993; Fournier & Mick, 1999). The C/D paradigm states that customer satisfaction stems from a customer's comparison of post-purchase and post-usage evaluation of a product with the expectation prior to purchase (Achim et al., 2001). Oliver and Swan (1989) suggest that customer satisfaction occurs when the purchasing experience and after-sales service meet the customer's expectation. Customer satisfaction is often viewed as a cumulative experience, measured as the general level of satisfaction based on the overall experience with the firm (e.g., Garbarino & Johnson, 1999). So CRM tactics, implemented across multiple channels, can form a cumulative customer experience.

Silk and Kalwani (1982) suggest that fairness and ease in the ordering process affect consumer satisfaction after purchase. If customers feel they are being treated fairly and feel easy with the ordering process, they are more likely to be satisfied with the products. Extending this finding to the e-commerce context, one can suggest that user interface and usability are factors that contribute to good customer satisfaction. There exists a high correlation between perceived convenience and customer satisfaction with the products and services sold on the Internet (Lee & Ahn, 1999). For e-commerce, low price, low asset specificity, and clear description are important product and service characteristics that attract online shoppers.

Therefore, an e-commerce site should incorporate the following features to build customer satisfaction:

- Wide variety and lower price products;
- Useful descriptions and price comparison for products and services;
- Self-service capability;
- Self-help, FAQ, and help contact services;
- Easy-to-use transaction interface;
- Easy-to-understand text, images, and animation to communicate with the customers;
- Accurate information about products and services to support pre-purchase services;
- Company details;
- Search engines for information searching;
- Product reviews and discussion;
- Different payment and delivery options;
- Recommendations for alternate product and services;
- Comparable products and services with lower prices;
- Purchase and delivery confirmation;
- Follow-up e-mail notification for product and service status;
- An order tracking method;
- Follow-up surveys for customer feedback; and
- Easy options for product return.

The quality of information and interface design for information search on an e-commerce site helps to draw new customers. For transaction and service support, good interface design and usability of the shopping cart are critical to a customer's shopping or service experience. Availability of timely post-sales support, such as order tracking and response to customer inquiry, contributes to customer satisfaction and retention.

Customization

Customization in CRM refers to the entire marketing mix – communications, products, services, processes, prices, and channels. Lee and Shu (2001) emphasize that the level of customization helps to shape customers' perception of quality in products and services. By tailoring products

and services to meet individual customers' needs and preferences, a company can fulfill and exceed customer expectations and increase their perception of product quality. By using the ACSI model, Lee and Shu (2001) demonstrate how the perceived quality and perceived value of a product contribute to customer satisfaction in a multi-layer fashion.

Mass customization tactics, such as personalized direct e-mails and product recommendations, are essential to eCRM. As acquiring information about customers is essential to relationship marketing, the Internet technology has made it easier for companies to collect data about customer profiles and online activities. Winer (2001) emphasizes that building a customer database is the first step towards an eCRM solution. His model involves the following steps: (1) build a database of customer activities, (2) analyze customer activities, (3) determine the target customers, (4) develop tool to target these customers, (5) implement privacy issues, and (6) define metrics for measuring the success of CRM program. After understanding the customer activities and selecting target customers, the company can proceed to creating products and services. Companies should characterize their customers as product makers rather than product takers.

Personalization techniques can be used to customize online interactions with e-commerce customers. Common techniques involve collaborative filtering, rule-based, and intelligent agent-based methods. Amazon.com has applied these techniques not only for pre-sales product recommendations, but also for one of their loyalty programs in the form of Gold Box special promotions. A registered customer has opportunities to receive discount promotions in a timed presentation, but only once. The Gold Box service remembers what items have already been shown to the same customer. Customers can also configure products and services that they are interested in purchasing. For example, Peapod.com allows customers to create personalized shopping lists, which, in turn, enables customers to tailor their shopping experiences and product choices. These forms of customization allow companies to capitalize on customer-initiated communications and interactions. In the long run, both firm-based and customer-initiated customization approaches can lead to cumulative positive customer experience with the products, services, and the Web site.

An e-commerce site can customize its content, products, and services by providing the following features:

- A personal page display, such as "my bookstore" and "my news box";
- Self profile and preference update with self-management capability;
- Self-help, FAQ, and question posting capability;
- Recommendations for products and services based on the customer's personal profile;
- Customer purchase history, delivery history, and account status;
- Incentives according to customer preference; and
- Personal services – remembering the customer's delivery address, personal reminders, previous search results, contact lists of friends and families, and so forth.

Customization features can be implemented in all three phases of CRM. To support pre-sales activities, customization can be applied by providing customers with product recommendations and the capabilities to create their own profiles and preferences. Customized order transaction processes facilitate the sales phase. Features that help to customize post-purchase support, such as e-mail promotion notifications, delivery tracking, and self-management capability for updating profiles and preferences are important for customer retention.

Trust

Trust is one of the most widely examined and confirmed constructs in relationship marketing research. There is the notion that trust constitutes the belief, attitude, or expectation of a party that the relationship partner's behavior or its outcomes will benefit the trusting party itself (Andaleeb & Anwar, 1996). Trust is built on the level of risk, which can be determined by network infrastructure, Web and mobile applications, customer privacy issues, security of data transfer, and system authentication (Lee & Ahn, 1999). On one hand, easy-to-use system interfaces, consistent and complete information, reliable connectivity, and sufficient customer support ensure customer trust. On the other hand, a high level of perceived risk associated with these system features may result in customer hesitation for performing transactions via the mobile channel (Chan & Fang, 2003).

Online trust is based on the user's Internet experience. Reputation contributes to "trust belief" and "trust intention". Thus, third-party endorsement and icons placed on e-commerce sites can affect consumer trust (McKnight, Choudhury & Kacmar, 2000). Trust is "the willingness of a party to be vulnerable to the actions of another party based on the expectations that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control the other party" (Mayer, Davis & Schooman, 1995, p. 712). This definition accentuates vulnerability, which is not just risk-taking but the willingness to take risks. Ambrose and Johnson (1999) have applied this definition to the online retailing environment. In this environment, the absence of face-to-face interaction between the buyer and the seller increases the buyer's vulnerability. Customer perceptions of a site's assurance of privacy and security influence trust. Thus, a high level of perceived risk affects a customer's intention to carry out the transactions online.

Trust also affects customer satisfaction and customer loyalty, and it directly influences the effectiveness of the eCRM strategy. Therefore, an e-commerce site needs to build customer trust by incorporating the following features:

- Consistent product and service information;
- Product information that embodies brand equity, transience, quality, variety, availability, as well as competitiveness and options for customization;
- Price information and payment options;
- A privacy statement to guarantee that customer information is kept confidential;
- Security measurement such as a digital certificate, public-key cryptography, authenticity, integrity, confidentiality, non-repudiation, and third-party verify signature, and SSL;
- Proper authentication mechanism;
- Secure payment and transmission; and
- Help desk support.

Trust features are important to all three phases of CRM. Privacy statements should be thorough, and authentication mechanisms as well as third-party signatures should be prominently demonstrated for new customers. Secure transactions are essential for bringing back repeat customers.

APPLYING THE FRAMEWORK

To explore the applicability of the proposed framework, we conducted a cognitive walkthrough of e-commerce sites and their corresponding mobile sites for the Palm OS platform. We also developed a checklist (in Appendix) based on the framework to guide the cognitive walkthrough. The choice of Palm OS version of mobile sites allowed us to evaluate a wider range of CRM features, because Palm handheld devices have relatively larger screens and support more interface features than WAP phones do. On the Palm.Net site, we downloaded the wireless applications

for selected sites onto the Palm VII device prior to the evaluation.

For illustration purpose, this chapter includes three examples — Amazon.com, United Airlines (united.com), and USA Today (usatoday.com). These three sites represent the retail, travel, and news portal industries. Tables 2, 3, and 4 summarize observations generated from the three cognitive walkthrough studies. Common features appearing on both the Web and the mobile channels are noted as "C". Features only available on the Web channel or the mobile channel are noted as "W" or "M" respectively.

From these three examples, we observe that eCRM supports all three phases of firm-customer interactions — acquisition of new customers, sales/services, and retention of existing customers through cross-sell, sales promotion, and loyalty programs. In contrast, mCRM focuses primarily on supporting and retaining existing e-commerce customers; little attention is focused on acquisition of new customers.

Mobile sites require customers or subscribers to register online first, particularly for sites involving transactions (Figure 1). It is not easy for new customers to initiate relationships with a

Table 2. Summary of eCRM and mCRM – Amazon.com (books)

Factors\ Phases	**Acquisition**	**Sales/Service**	**Retention**
Customer Loyalty	(**W**) Online book community	(**C**) Coupon available for redeem	(**W**) Gold box (**C**) E-mail for purchase discount and promotional free shipping
Branding	(**W**) Online community for review and discussion (**W**) Purchase certificate	(**W**) Used books and price info to facilitate other buying options (**C**) One-click order	(**M**) Simple product browsing access anywhere anytime (**W**) Amazon credit card
Customer Satisfaction	(**C**) A Variety of products are available. (**W**) Promotion product - books, music, special deals, electronic, games (**W**) Price comparison	(**C**) Book search (**W**) New hard copy, paper back and used books are available with price comparison. (**C**) Book review (**C**) Cross sales - customer also buy items)	(**C**) E-mail confirmation (**W**) Purchase tracking (**W**) Full online support, FAQ and contact number (**M**)Simple FAQ
Customization	(**W**) News, preferences, and personal recommendation display on the first page (**W**) Personal wish list (**W**) Provide baby and wedding registry services	(**C**) Require sign on for purchase (**C**) Access profile from the web (**C**) Able to modify delivery information	(**W**) Provide friend and family occasion reminder (**W**) Personal order and personal recommendation available at sign in (**M**) Simple book purchase link on top of the first page
Trust	(**C**) Security guarantee on personal profile	(**C**) Sign on required for purchase (**C**) Security indication	(**C**) Profile is saved on the Web (**C**) Address and purchase information can be modified (**W**) Preference can only be modified on the Web

(C) Common Feature *(W) Web Feature Only* *(M) Mobile Feature Only*

Table 3. Summary of eCRM and mCRM – United Airline (Flight)

Factors\ Phases	**Acquisition**	**Sales/Service**	**Retention**
Customer Loyalty		(**C**) Mileage plus program is associated with the customer purchase (**W**) Redeem award	(**C**) Mileage Plus summary (**C**) Award availability (**C**) Upgrade status
Branding	(**C**) Flight schedule and arrival/ departure detail (**W**) About United, united product and service, contact United		(**M**) Upgrade, travel awards, and red carpet club are on the first page
Customer Satisfaction	(**W**) Promotion travel packages, special deals, and cruise (**W**) Price comparison (**W**) Spanish version support (**W**) Service information (W) Company details	(**C**) Book/purchase a flight (**C**) Flight availability (**W**) Seat selection (**W**) Detailed price comparison (**W**) Electronic and non-electronic tickets (**M**) Only electronic tickets	(**C**) My itinerary (**C**) Flight status (**C**) Flight paging request registration
Customization	(**W**) After sign on, preference page display with personal preference of price alert	(**C**) Sign on or fill in mileage plus member number is necessary for both platforms.	(**W**) E-mail promotion registration and preference change (**M**)Book a flight, flight status, my itinerary, travel awards are on the first page
Trust	(**W**) Sign up can only perform on Web site	(**W**) Customer address, profile and form of payment can only be changed on Web site (**C**) Both platforms indicate secure transaction	(**C**) Profile is saved on the Web (**M**) Required to be mileage plus customer with current profile for access

(C) Common Feature *(W) Web Feature Only* *(M) Mobile Feature Only*

company on the mobile site. However, news sites, like USA Today, due to the time-sensitive nature of their services, seem better positioned to attract new customers. For all three sites, the mobile channel provides limited customer support. Other than limited product and service information, customer self-help and self-configuration delivery are not available on the mobile site. Mobile customers need to refer problems or questions back to the Web site.

In general, the mobile site emphasizes information delivery. Transaction and registration functions are carried out on the Web sites. Among the three sites illustrated in this chapter, the mobile site of USA Today, because of its focus on content, offers the least amount of services for retention purpose. In comparison, the mobile sites for Amazon and United Airlines include more mobile services for sales transaction and post-sales support. United Airlines provides a more complete range of mobile services to meet the needs of its mobile customers.

Therefore, the mobile channel supplements, rather than substitutes, the Internet channel for supporting and retaining existing e-commerce customers. mCRM targets existing customers

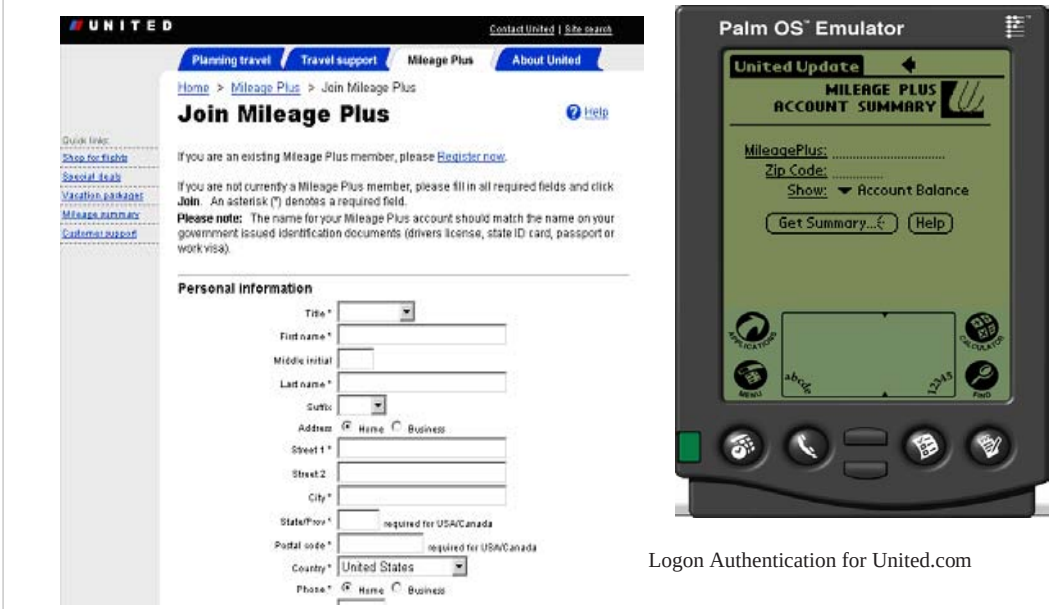
Table 4. Summary of eCRM and mCRM – USAToday

Factors\ Phases	**Acquisition**	**Sales/Service**	**Retention**
Customer Loyalty	(**W**) Provide incentive for online subscriber (**W**) Online subscribe with American Express, earn member reward	(**W**) With sign on id, the Web site recognizes subscriber	
Branding	(**C**) Well known and reputation newspaper (**W**) Online archive search	(**W**) Can online purchase full archive USAToday article or get free version of the brief highlight	
Customer Satisfaction	(**W**) Attractive front screen design (**C**) One click to the latest news (highlights) (**W**) Search engine available (**C**) About USAToday (**W**) Provide quick tour and sample complete online paper for new customer	(**W**) Search engine available (**C**) News display by category (**W**) Online stock quote inquiry (**W**) More real time news update (**W**) Provide online subscription process	(**W**) FAQ and feedback are available (**W**) Provide electronic version of USAToday complete copy online (with subscription) (**W**) Online report with delivery problem (**W**) Confirmation to subscription
Customization	(**W**) Customize to favorite columnist (**W**) Customize local weather display	(**W**) Can subscribe as a gift to someone else	(**W**) Online profile and preference management (**W**) Online address and password maintenance (**W**) Current subscriber can retrieve past issues of paper online (**W**) Online and e-mail reminder for when subscription is up
Trust	(**W**) Provide quick tour and sample complete online paper for new customer (**W**) Subscriber is provided with decure login authentication (**W**) Partner with American Express	(**W**) Provide multiple payment options (**W**) Online bill pay services (**W**) Confirmation to subscription	(**W**) E-mail notification for subscription

(C) Common Feature *(W) Web Feature Only* *(M) Mobile Feature Only*

who are: (a) frequent users with a high purchase rate and strong brand loyalty, and (b) in need of "on the move" services and spontaneous shopping. As illustrated in Figure 2, United Airlines' mobile site contains the essential information and features for a frequent traveler who is already a registered Mileage Plus member. Figure 3 shows that customers who chose to access Amazon's mobile site must overcome many interface barriers to access the mobile services. These customers may already have a strong commitment to the brand of Amazon.

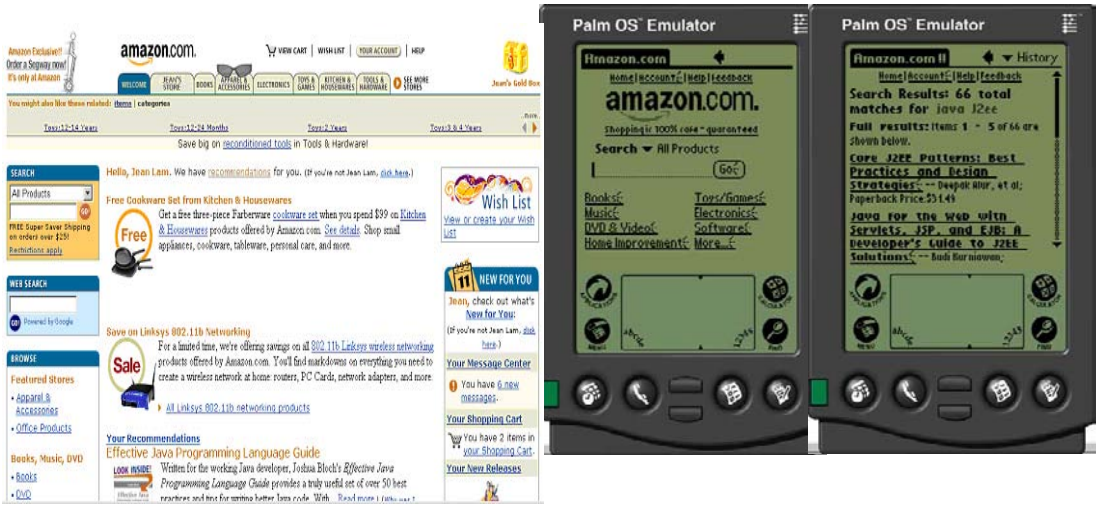
Figure 1. Web customer registration interface and mobile logon authentication for United.com



Logon Authentication for United.com

Figure 2. Web and mobile front screen for United.com

Figure 3. Web and mobile book search and book matching display for Amazon.com



RESEARCH DIRECTIONS

Drawing from the literature review and the analysis of eCRM and mCRM presented in this chapter, we identify several implications and issues for future research.

The Relationship between eCRM and mCRM

A company's e-commerce Web site serves as the primary channel for building and managing relationships with e-customers. The current state of wireless communication technology limits the role of the mobile channel in supporting customer needs. It seems more effective to use the mobile channel for retention of a small number of frequent and loyal customers who have mobile needs. For e-commerce sites that emphasize complex transactions and interactions, a simpler version of these interactions and information delivery should be provided for mobile users. Companies should use the profile and preference data from registered frequent users in designing appropriate mobile services. Further research is needed to validate the relationship between eCRM and mCRM in several areas:

- How are eCRM and mCRM features implemented in other sectors of online retailing, service, and content portals? How are these features implemented across three CRM phases?
- What services are provided on both channels? How consistently are these services implemented on both channels?
- What are characteristics of best practice for eCRM and mCRM?
- Are mCRM features most often implemented for retention purpose?
- How do companies map their mCRM to different wireless technology platforms?

Coordination of CRM across Multiple Channels

Beyond the Internet and the mobile channels, companies should develop a synergistic approach to coordinating CRM strategies across multiple channels in order to optimize customer satisfaction. For example, a hotel could provide a Web site for guests to conduct product and service search, compare prices and services, and make reservations. Registered guests can use their mobile devices to check and update reservations, and for advance room check-in when they are on the road. The hotel could send e-mail alerts and location-based information according to the guest's profile and preference. These interactions could be coordinated with the traditional in-hotel services to offer the frequent guests an integrated experience. Further research should examine multi-channel CRM in several areas:

- How should the firm-customer interactions be mapped across the entire process of customer life cycle to identify touch points for interactions?
- What criteria should be considered to guide the process mapping and requirement analysis?
- How could the information gathered from different channels be integrated to form a comprehensive customer profile?

The Tradeoff among CRM Factors

The five CRM factors emphasized in the proposed framework seem to play different roles in eCRM and mCRM; each factor also affects the firm-customer interactions differently. Our limited examples show that customer loyalty, customer satisfaction, and customization factors are more prominently presented on transaction-based mobile sites than branding and trust factors. These five CRM factors are inter-related. Future research should empirically examine their individual and collective impact of these factors on eCRM and mCRM. Researchers should construct and test the underlying model in the mobile environment to examine which factors are most important to mobile customers. Findings on specific mCRM features for transaction support and retention will improve the understanding of specific CRM tactics.

- How do the five CRM factors relate to one another in supporting customer acquisition, shopping experience, and customer retention?
- Which factors are most important for mCRM?
- Which factors are most important for eCRM?
- Are loyalty, customization, and customer satisfaction factors more important than trust and branding factors for mobile customers?
- What kind of trade-offs among CRM factors should be considered to strengthen long-term customer relationship?

mCRM and Customer Acquisition

Our analysis reveals that the mobile channel currently plays a limited role in customer acquisition. Advantages of location- and context-based marketing and mobile commerce remain conceptually sound but are not substantiated. Location-aware advertising is primarily text based. However, the introduction of third-generation mobile network and multimedia-enabled mobile devices may change the mobile commerce environment. A recent study (Oh & Xu, 2003), an exploratory simulation, shows that multimedia location-aware advertising messages led to favorable attitudes and increased intention to reuse the mobile advertising service. More creative mobile services for attracting new customers will emerge. In the meantime, researchers will need to address:

- How could location-aware and context-aware technology be effectively used to attract new customers?
- What are key concerns of new customers in selecting mobile commerce sites?
- How can multimedia technology and short text messages be best designed to attract new mobile customers?

Usability and Personalization Issues for mCRM

As technology advances, a wider range of wireless applications may be introduced. Future research on usability for wireless applications (Chan et al., 2002; Chan & Fang, 2003) and personalized content adaptation (Zhang, 2003; Zhou & Chan, 2003) may contribute to more effective use of the mobile Web for relationship building with customers. Unique mobile features appear to be implemented mostly by content adaptation so the mobile users can access essential services and information more efficiently on their handheld devices. Future research will need to address:

- How could content and services be personalized for CRM on the mobile platform? To what extent are current personalization techniques useful to mCRM?
- Would personalization be more important for mCRM in terms of information content, transaction support, or services?
- How does the flow of shopping experience using wireless devices differ from the online experience? What are the implications of such differences on interface design?

REFERENCES

Aaker, D.A. (1991). Measuring brand equity across products and markets. California Management Review, 38(3), 102-120.

Aaker, D. (1996). Managing brand equity: Capitalizing on the value of a brand name. New York, NY: Free Press.

Achim, W., Thio, M., & Helfert, G. (2001). The impact of satisfaction, trust, and relationship value on commitment: Theoretical considerations and empirical results. In D. Ford, S. Leek, P. Naude, T. Ritter & R. Turnbull (Eds.), IMP Conference, Vol. CD. Bath.

Alanen, J. & Aution, E. (2003). Mobile business services: A strategic perspective. In B. Mennecke and T. Strader (Eds.), Mobile commerce: Technology, theory, and applications (pp. 162-184).

Ambrose, P., & Johnson, G. (1999). A trust based model of buying behavior in electronic retailing. Proceeding of 1998 Americas Conference.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. Journal of Information Technology Theory & Application, 4, 43-64.

Andaleeb, S. & Anwar, S. (1996). Factors influencing customer trust in salespersons in a developing country. Journal of International Marketing, 4(4), 35-52.

Anderson, E. & Sullivan, N. (1993). The antecedents and consequences of customer satisfaction for firms. Marketing Science, 12(2), 125-143.

Burnstine, D.C. (1980). BIAIT: An emerging management engineering discipline. Working paper.

Chan, S., & Fang, X. (2003). Mobile commerce and usability. In K. Siau & E. Lim (Eds.), Advances in mobile commerce technologies (pp. 235-257). Hershey, PA: Idea Group Publishing.

Chan, S., Fang, X., Brzezinski, J., Zhou, Y., Xu, S., & Lam, J. (2002). Usability for mobile commerce across multiple form factors. Journal of Electronic Commerce Research, 3(2), 187-199.

Chen, L., & Sukpani, N. (1998). Assessing consumers' involvement in Internet purchasing. Proceedings of the Fourth Americas Conference on Information Systems, 281-283.

Cho, Y., Im, Hiltz, R. & Fjermestad, J. (2001). Causes and outcomes of online customer complaining behavior: Implications for customer relationship management (CRM). Proceedings of the Seventh Americas Conference on Information Systems, 900-907.

Dekimpe, M.G., Steenkamp, J., Mellens, M., & Vanden. A. (1997). Decline and variability in brand loyalty. International Journal of Research in Marketing, 14(5), 405-420.

Dick, A., & Basu, K. (1994). Customer loyalty: Toward an integrated conceptual framework. Journal of the Academy of Marketing Science, 22(2), 99-113.

Ernst & Young. (2001). Global online retailing: An Ernst & Young special report. Retrieved September 5, 2003, from http://www.ey.com/global/Content.nsf/International/Industries_-_RCP_-_Global_Supply_Chain_ Survey .

Evans, J., & Laskin, R. (1994). The relationship marketing process: A conceptualization and application. Industrial Marketing Management, 23(5), 439-452.

Forrester Research. (2003). The state of online retailing 6.0. Retrieved September 5, 1993, from http://www.shop.org/press/03/051503.html.

Fournier, S., & Mick, D. (1999). Rediscovering satisfaction. Journal of Marketing, 63(4), 5-28.

Garbarino, E., & Johnson, M.S. (1999). The different roles of satisfaction, trust, and commitment in customer relationships. Journal of Marketing, 63(2), 70-87.

Hennig-Thurau, T., & Hansen, U. (2000). Relationship marketing – Some reflections on the state of the art of the relational concept. In T. Hennig-Thurau & U. Hansen (Eds.), Relationship marketing: Gaining competitive advantage through customer satisfaction and customer retention (pp. 3-27). New York: Springer.

Hennig-Thurau, T., & Klee, A. (1997). The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. Psychology & Marketing, 14(8), 737-765.

Hoffman, D., Thomas, P., & Peralta, M. (1999). Building consumer trust online. Communication of ACM, 41(3), 44-51.

Kannan, P., Chang, A., & Whinston, A. (2001). Wireless commerce: Marketing issues and possibilities. Proceedings of the 34th Hawaii International Conference on System Sciences. Los Alamitos, CA: IEEE Comp Soc.

Lee, D., & Ahn, J. (1999). An exploratory study on the different factors in customer satisfaction with e-commerce between in the United States and in Korea. Proceedings of the Second International Conference on Telecommunication and Electronic Commerce.

Lee, S., & Shu, W. (2001). An integrative and complementarity-based model for the design and adoption of customer relationship management technologies. Proceedings of the Seventh American Conference on Information Systems, 854-856.

Lowenstein, M. (1997). The customer loyalty pyramid. Westport, CT: Quorum Books.

Mayer, R., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709-734.

McKnight, D., Choudhury, V., & Kacmar, C. (2002). The impact of the initial consumer trust on intention to transact with a Web site: A trust building model. Journal of Strategic Information Systems, 11(3-4), 297-323.

Mennecke, B., & Strader, T. (2001). Where in the world on the Web does location matter? A framework for location based services in m-commerce. Proceedings of the Seventh Americas Conference on Information Systems, 450-455.

Moon, Y. (2002). Interactive technologies and relationship marketing strategies. Marketing strategies business fundamental series. Boston: Harvard Business School Publishing.

Oh, L., & Xu, H. (2003). Effects of multimedia on mobile consumer behavior: An empirical study of location-aware advertising. Proceedings of Twenty-fourth International Conference on Information Systems, 679-691.

Oliver, R., & Swan, J. (1989). Equity and disconfirmation perceptions as influences on merchant product satisfaction. Journal of Consumer Research, 16, 372-383.

Papadopoulou, P., Andrew, A., Kanellis, P., & Martakos, D. (2001). Trust and relationship building in electronic commerce. Internet Research: Electronic Networking Applications and Policy, 11, 322-332.

Papers, D., & Rogers, M. (1993). The one to one future: Building relationships one customer at a time. NY: Currency Doubleday.

Reichheld, F. (1996). The loyalty effect: The hidden force behind growth, profits, and lasting value. Boston: Harvard Business School Press.

Reichheld, F., & Sasser, E., Jr. (1990). Zero defection: Quality comes to services. Harvard Business Review, 68(5), 105-111.

Romano, N., Jr. (2001). An agenda for electronic commerce customer relationship management research. Proceedings of the Seventh Americas Conference in Information Systems, 831-833.

Shim, J.P., Bekkering, E., & Hall, L. (2002). Empirical findings on perceived value of mobile commerce as a distributed channel. Proceedings of the Eighth Americas Conference on Information Systems, 1835-1837.

Silk, A., & Kalwani, M. (1982). Measuring influence in organizational purchase decisions. Journal of Marketing Research, 43(2), 165-181.

Strauss, B. (2000). Using new media for customer interaction: A challenge for relationship marketing. In T. Hennig-Thurau & U. Hansen (Eds.), Relationship marketing: Gaining competitive advantage through customer satisfaction and customer retention (pp. 3-27). New York: Springer.

Varian, H., Litan, R., Elder, A., & Shutter, J. (2002). The net impact study: The projected economic benefits of the Internet in the United States, United Kingdom, France and Germany. Retrieved August 25, 2003, from http://www.netimpactstudy.com

Varshney, U. (2003). Mobile wireless information systems: Applications, networks, and research problems. Communications of the Association for Information Systems, 12(11), 155-166.

Varshney, U., & Vetter, R. (2001). A framework for the emerging mobile commerce applications. Proceedings of the Thirty-fourth Hawaii International Conference on System Sciences. Los Alamitos, CA: IEEE Comp Soc.

Winer, R. (2001). A framework for customer relationship management. California Management Review, 43(4), 89-105.

Zhang, D. (2003). Delivery of personalized and adaptive content to mobile devices: A framework and enabling technology. Communications of the Association for Information Systems, 12(13), 183-202.

Zhou, Y., & Chan, S. (2003). Adaptive content delivery over the mobile Web. Proceedings of the Ninth Americas Conference in Information Systems, 2009-2019.

APPENDIX

CRM Feature Checklist

Customer Loyalty		
❑	L1	Does the site provide enough information about the loyalty program?
❑	L2	Does the site provide ways to join the loyalty program?
❑	L3	Does the site provide a personal account?
❑	L4	Does the site provide information about purchase history and shipping status?
❑	L5	Does the site provide personalized services for repeat customers?
❑	L6	Does the site provide frequent buyer incentives (discount or personal upgrade services)?
❑	L7	Does the site provide a special page for status customers – Status page, upgrade options, redeem procedures, and special discount / promotion?
❑	L8	Does the site provide special services for frequent buyers – no cost delivery, priority seating, or 2 clicks checkout?
❑	L9	Does the site actively collect feedback from frequent customers?
Branding		
❑	B1	Does the site have large customer community and provide chat room, discussion sessions, and product reviews?
❑	B2	Does the site provide special service that differentiates it from other sites?
❑	B3	Does the site provide exclusive brand name products?
❑	B4	Does this site include exclusive interfaces to support order processing?
❑	B5	Does the site provide supplementary services, which are associated with the traditional business services and high barrier of entry: Banking, Ameritrade, and American Express Credit Card?
Customer Satisfaction:		
❑	S1	Does the site provide good variety and lower price products?
❑	S2	Does the site provide good descriptions and price comparison for product and services (extensive feature description and product performance)?
❑	S3	Does the site provide self-management capability?
❑	S4	Is the transaction interface design easy to understand?
❑	S5	Does the site provide good text, images, and animation to communicate with their customers?
❑	S6	Does the site provide pre-sale services – accuracy product and services information?
❑	S7	Does the site provide company detail?
❑	S8	Does the site provide search engine for information searching?
❑	S9	Does the site provide product review or discussion groups?
❑	S10	Does the site provide different payment options?
❑	S11	Does the site provide different delivery options?
❑	S12	Does the site provide alternate product and services suggestions?
❑	S13	Does the site provide comparable product and services with lower price?
❑	S14	Does the site provide confirmation with purchase and delivery?
❑	S15	Does the site provide self-help, FAQ, and help contact services?

continued on following page

APPENDIX CONTINUED

❑	S16	Does the site provide follow-up email notification for product and services status?
❑	S17	Does the site provide order-tracking method?
❑	S18	Does the site provide follow-up survey for customer feedback?
❑	S19	Does the site provide easy way for product defective or unwanted return?
Customization		
❑	C1	Does the site provide personal page display?
❑	C2	Does the site provide self-profile and preference update (self-management capability)?
❑	C3	Does the site provide self-help, FAQ, and question posting capability?
❑	C4	Does the site provide recommendations for product and services based on personal profile?
❑	C5	Does the site provide customer purchase history, delivery history, and account status?
❑	C6	Does the site provide incentives according to customer preference?
❑	C7	Does the site provide personal services – remembering delivery address, personal remainder, previous search result, friends and family contact list, etc?
Trust		
❑	T1	Does the site present consistent information?
❑	T2	Does the site present product information with brand equity, transience, quality, variety, customization, competitiveness and availability?
❑	T3	Are the price and payment options available?
❑	T4	Does the site present a privacy statement to guarantee that customer information confidential?
❑	T5	Does the site present security measurements such as: digital certificate, public-key cryptography, authenticity, integrity, confidentiality, non-repudiation, attributes of the system (benevolence, competency, predictability), third party verifies signature, and SSL?
❑	T6	Does the site present the proper authentication mechanism?
❑	T7	Is secure payment (payment gateway, firewalls and encryption) and transmission available?
❑	T8	Does the site provide help desk support?

This work was previously published in E-Commerce and M-Commerce Technologies, edited by P. Deans, pp. 1-31, copyright 2005 by IRM Press (an imprint of IGI Global).

Chapter 6.8
Exploring Mobile Service Business Opportunities from a Customer-Centric Perspective

Minna Pura
HANKEN—Swedish School of Economics and Business Administration, Finland

Kristina Heinonen
HANKEN—Swedish School of Economics and Business Administration, Finland

ABSTRACT

Mobile services have evolved into an important business area and many companies in various industries are offering mobile services. However, formal classifications or user-centric categorizations of mobile services are still scarce. This chapter develops a conceptual classification for mobile services that illustrates the characteristics of mobile services and gives indications on how to describe mobile business opportunities and categorize services from a customer-centric perspective. The classification scheme, grounded in previous research, is based on the type of consumption, context, social setting, and customer relationship with the service provider. The explorative classification is illustrated with two case studies of existing mobile services in the European market. The theoretical contribution to service management research involves how to describe mobile services from a customer perspective. Managerially, the classification helps marketers, service developers, and stakeholders to evaluate, differentiate, group, and market mobile service offerings more effectively.

INTRODUCTION

Mobile services differ from traditional services in their ability to provide service offerings regardless of temporal and spatial constraints. The benefits of mobile services are often summarized by four factors: (1) ubiquity, (2) convenience, (3) localization, and (4) personalization that differentiate mobile services from online services (Clarke & Flaherty, 2003). Mobile services are also different from traditional interpersonal services that are

delivered face-to-face, or from other types of e-services, such as wireless online services, where the service delivery is linked to a specific fixed local area network or specific location. Mobile services can be accessed on the move, where and whenever the need arises. In this paper, mobile services are defined as "all services that can be used independently of temporal and spatial restraints and that are accessed through a mobile handset (mobile phone, PDA, smart phone, etc.)" Examples of most popular BtoC mobile services in Europe include logos, ring tones, games, address inquiry, account balance inquiry, paying for parking, vending machines, subway tickets, finding the nearest service location, maps, directions, and so forth.

Although an increasing number of academic studies are starting to focus on mobile services from a service management perspective rather than a technology perspective (e.g., Balasubramanian, Peterson, & Järvenpää, 2002; Heinonen & Andersson, 2003; Nysveen, Pedersen, & Thorbjørnsen, 2005a, 2005b; Pura, 2005), formal classifications or categorizations of mobile services are still scarce. Previous studies clearly indicate that specific categorizations are needed, and especially categories of mobile services have been called for (e.g., Rodgers & Sheldon, 2002). Additionally, so far theories used to analyze mobile business stem from information systems literature and often treat mobile services as a category as such compared to Internet and brick and mortar services. Aspects that would allow us to categorize different types of mobile services have remained largely unexplored, and future research has been encouraged in the field (Okazaki, 2005).

Many service classifications in earlier literature stem from traditional service environments that distinguished services from products. They attempt to offer managerial insights on how to organize and classify services in order to serve customers better. Lovelock's (1983) service classification of traditional interpersonal services is one of the notable classifications. It suggested a need to move away from the industry-specific classifications by exploring managerially relevant service characteristics. However, previous service classification models incorporating several fields of industry are quite generic, and more specific classifications are needed to depict the nature of the new electronic channels, especially in order to identify the specific characteristics of mobile services.

Some attempts have already been made to develop service categorizations that depict the special nature of electronic services in general (e.g., Angehrn, 1997; Dabholkar, 1996; Meuter, Ostrom, Roundtree, & Bitner, 2000). However, they have not acknowledged the mobile nature of delivering services. For example, Meuter et al.'s categorization of technology-based service encounters does not include services provided through a mobile interface. Hence, existing e-service categorizations do not identify the special nature of mobile services in comparison to other e-services.

Furthermore, most existing mobile service categorizations tend to focus on the providers' perspective rather than the customer or user perspective (e.g., Giaglis, Kourouthanassis, & Tsamakos, 2003; Hyvönen & Repo, 2005; Mitchell & Whitmore, 2003; Mort & Drennan, 2005). Although some previous research on mobile services does incorporate a customer perspective of mobile services, the focus of this group of studies has not been on classifying mobile services, but on some specific aspect of mobile services, such as intentions (e.g., Nysveen et al., 2005a, 2005b) or motivations (Pura & Brush, 2005) to use, segments of users, value (Anckar & D'Incau, 2002; Van der Heijden, 2004), user acceptance (Van der Heijden, Ogertschnig, & Van der Gaast, 2005), or sociability (Heinonen & Andersson, 2003; Järvenpää & Lang, 2005). To our knowledge, there are no studies that specifically attempt to provide a solid ground for categorizing mobile services, and most existing mobile service categorizations are mainly a by-product of the study. The study by Nysveen et al. (2005b) represents an exception,

as it was one of the first to compare adoption of different types of mobile services, but the grounds for service categorization remain to be explored further. Thus, further conceptualizations are needed in this area.

The aim of the chapter is to develop a conceptual classification for mobile services that illustrates the characteristics of mobile services and gives indications on how to describe mobile business opportunities and categorize services from a customer-centric perspective. The chapter contributes mainly to service marketing research with its classification of mobile services, and hence it differs from most studies on mobile services positioned in information systems research. By focusing on mobile service value from a customer perspective and building on previous research of mobile services, we propose a classification of mobile services based on the type of consumption, context of use, social setting, and relationship with the provider. These concepts are described more in detail by breaking them down into classification grids that differentiate mobile services from one another by describing the aspects in a two-dimensional way. The resulting typology and suggested managerial questions, as well as the preliminary set of questions for evaluating perceptions of service users, give implications for further empirical research in the mobile area. They also help managers and service developers to differentiate and group their mobile service offerings in a meaningful way that is especially useful for marketing purposes. Managerial challenges and questions related to each aspect are also discussed.

The chapter has the following structure. First, we introduce the theoretical background and discuss how the chapter combines views from previous service management literature. The main thrust of the chapter consists of the proposed framework and two case studies that illustrate the practical use of the framework. We also discuss the solutions based on the case studies and suggest recommendations for evaluating business solutions based on the classification scheme. We conclude the chapter with future trends, suggestions for future research, and concluding remarks that summarize the contribution of the chapter. While discussing previous theories, we provide concrete examples of existing European mobile services in each category of the classification grid. Furthermore, two case studies (a mobile fishing permit and a mobile adventure game) are used to illustrate how the framework can be applied to assess specific mobile services.

THEORETICAL BACKGROUND

What distinguishes different mobile services from each other and from services offered through other channels? What kind of value do users perceive in different mobile services? Based on a review of previous research in the service management and relationship marketing literature, as well as on literature on mobile services and mobile technology, we can identify four main aspects that represent and summarize the special nature of mobile services (see Table 1).

- What is the type of consumption?
- What is the temporal and spatial context of service use?
- What is the social setting in service use situations?
- What is the relationship between the customer and the service provider?

Based on a review of the literature, we argue that these aspects are of key importance with regard to categorizing mobile services from a customer-centric perspective, and hence we build the classification of mobile services on these four aspects. Next we describe in more detail the four aspects that form the foundation of our conceptual framework and provide examples of existing mobile services. The two case studies will further elaborate on how these four factors

Table 1. Relevant research on mobile services

Authors	Focus	Type of Consumption	Context	Social setting	Relationship
Isoniemi & Wolf 2001	Segments of mobile service users			●	●
Anckar & D'Incau 2002	Value creation in mobile commerce	●	●		
Balasubramanian et al 2002	Mobile commerce		●		
Pura (2003b)	Nature of loyalty in mobile services		●		●
Heinonen & Andersson 2003	Use of mobile services			●	
Nysveen et al 2005 a,b	Intentions to use mobile services	●		●	
Pura & Brush 2005	Motivations for mobile service use	●	●		
Pura 2005	Value and loyalty in mobile location-based services	●	●		
Järvenpää & Lang 2005	Mobile technology			●	
Van der Hejden et al. 2005, Koivumäki, Ristola and Kesti 2006	User acceptance of mobile information services	●	●		
Laukkanen & Lauronen 2005	Customer value creation in mobile banking	●			
Lin & Wang 2006	Loyalty in m-commerce	●			

are linked together into a hierarchical framework that illustrates the business opportunities from a customer point of view.

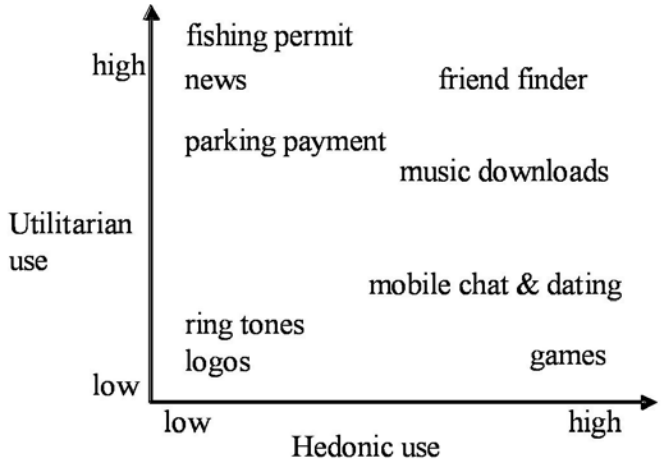
Although these four key aspects have been indicated in earlier literature by several individual authors, as summarized in Table 1, so far the four aspects have not been discussed in conjunction with each other. However, many of the previous studies do focus on one or two of the aspects in different combinations, and hence the studies indicate the importance of integrating them into one conceptual framework.

Next, the four key areas are discussed individually and combined into a classification framework presented later in Figure 5. Each area includes two dimensions and they are illustrated in four schemes.

Type of Consumption

Different types of services are used for many purposes based on customers' individual consumption values (Sheth, Newman, & Gross, 1991). Different channels may be used for various types of tasks (Neslin et al., 2006). Motivations

Figure 1. Consumption types

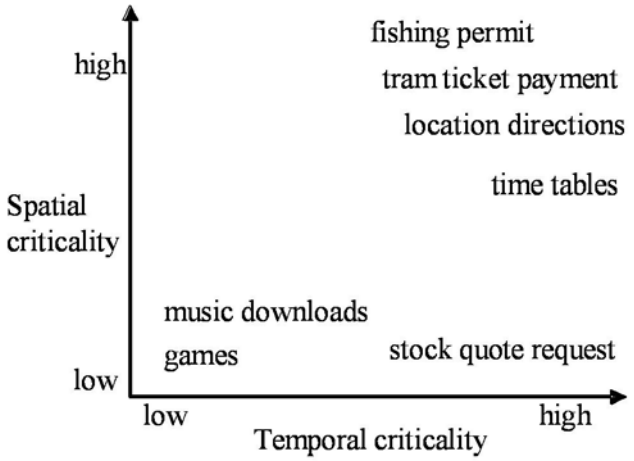


for consumption are often divided into hedonic and utilitarian value in the marketing literature (Babin, Darden, & Griffin, 1994; Chaudhuri & Holbrook, 2002; Koivumäki et al., 2006; Van der Heijden, 2004). Utilitarian value refers to extrinsic motivation that exists in goal-directed service use (Babin et al., 1994). Hedonic value means intrinsic motivation that exists in experiential, fun, and enjoyable service use as such (Novak, Hoffman, & Duhachek, 2003). Similar categorization into goal-directed and experiential services has been used also in the mobile field (Nysveen et al., 2005b; Okazaki, 2005). It has been suggested that hedonic value refers to entertainment needs and utilitarian value to efficiency needs (Anckar & D'Incau, 2002; Cotte, Tilottama, Ratneshwar, & Ricci, 2006; Pura & Brush, 2005). The use of new technology to access the services can be seen as fun and exciting, as such services can also create both utilitarian and hedonic value.

The first classification scheme depicting consumption types is based on utilitarian and hedonic use. In this study, utilitarian use denotes the level of goal-oriented value a user receives from using the service. Examples include information-based services, for example, news, weather reports, timetables, traffic information, and search services (address and number inquiry, nearest service location, search for stolen vehicles, routes, etc.) are examples of services that create high utilitarian value and help users to achieve a goal effectively and conveniently. Transaction-oriented payment services (mobile banking, parking payment, paying for fishing permit) are also used for utilitarian reasons, such as saving time and providing an efficient and convenient way to do transactions (Laukkanen & Kantanen, 2006; Laukkanen & Lauronen, 2005).

In this study, highly hedonic use involves services that create fun experiences and that are used for the sake of the experience. Examples include entertainment-oriented services such as mobile chat, games, and music downloads.

Figure 2. The temporal and spatial criticality of service



Finding examples for the category that presents low value both on utilitarian and hedonic aspects is not easy, but they may include many of the most popular services currently offered by the majority of mobile service providers: logos, ring tones, pictures that may be perceived "nice to have", but something that you can also manage without.

Temporal and Spatial Context

The temporal and spatial context of service use also differentiates mobile services from other types of services (e.g., Anckar & D'Incau, 2002; Balasubramanian et al., 2002; Mennecke & Strader, 2003; Yoo & Lyytinen, 2005). It influences the value of mobile services (e.g., Heinonen, 2006; Nysveen et al., 2005b; Pura & Brush, 2005; Van der Heijden, 2004) and, as the criticality of time and space of service use is a factor mostly considered in previous literature (Mennecke & Strader, 2003), it is essential also when categorizing mobile services. Balasubramanian et al. (2002) proposed a space-time matrix for tasks that could be used in the mobile environment. Other researchers have also used time and location to conceptualize products and services based on the relative immediacy of the task of the user and the relative location of the user when the service is used (Heinonen, 2004a, 2004b, 2006; Mennecke & Strader, 2003; Pura, 2003c).

We use a similar approach and introduce a classification scheme based on temporal and spatial criticality. In this scheme, temporal criticality depicts a time dimension of how urgently the customer needs the service. Spatial criticality indicates whether the use situation is location non-critical, that is, if the service can be used anywhere, for example, at work or at home, where there are other alternatives to the mobile device such as fixed Internet connections. Alternatively, the use location can be critical, that is, the customer is on the move using the service in the street or in other places where there are no other alternatives

to the mobile service, or location-based information is needed.

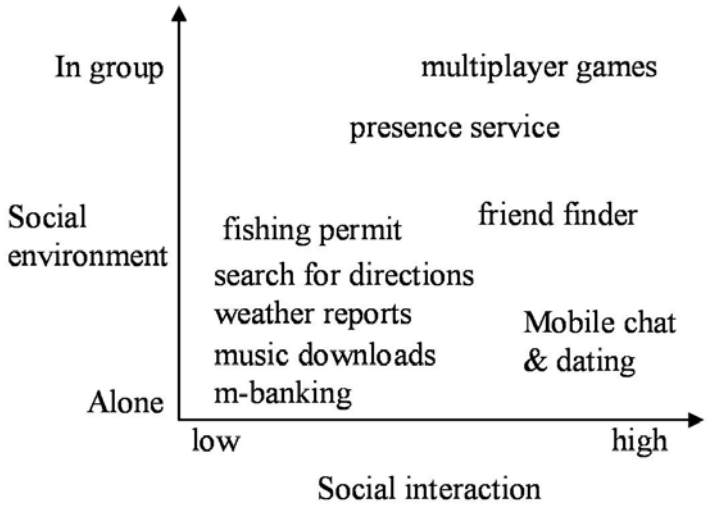
Although temporal and spatial elements of the use context represent a main benefit of technology-based services (Meuter et al., 2000) and mobile services in particular, it can be argued that a broader perspective on the use context should be considered. This larger context incorporates aspects other than time and space, namely the social setting where the service is used (Celuch, Goodwin, & Taylor, 2007). Researchers have acknowledged sociability as a purpose for using mobile technology (Yoo & Lyytinen, 2005), and it has been suggested that social aspects influence the use of mobile services (e.g., Heinonen & Andersson, 2003) or intentions to use them (Nysveen et al., 2005b). Traditionally, social pressure of others to use new technology has gained attention in the mobile context (Kleijnen, Wetzels, & De Ruyter, 2004; Venkatesh, Morris, Davis, & Davis, 2003). However, the social setting itself has not explicitly been taken into consideration in this field. In a retail context, social setting has been defined as "the social setting focuses on the presence or absence of others, together with their social roles, role attributes and opportunities for interaction" (Nicholson, Clarke, & Blakemore, 2002, p. 134) This definition also applies well to the social setting where mobile services are used, because it encapsulates opportunities for interaction with friends and family through to the presence of other people at work and even proximity to total strangers. The social setting can enhance or inhibit the use of mobile services in a certain situation (Isoniemi & Wolf, 2001). The social setting is expected to have a greater influence on the usage of services in the mobile environment than the Internet, because mobile services are often used in a social environment that involves interpersonal influence (Mort & Drennan, 2005).

The type of social setting in which mobile services are normally used is very important in defining what kind of interactivity or possibility of lack of interactivity is offered by specific mobile services (Okazaki, 2005). The third classification scheme depicting the social setting is based on the social environment and social interaction involved when using the service. The social environment denotes the environment in which the user is when the need to use the service arises. It is depicted by the continuum alone vs. in-group, as illustrated in Figure 3. The social environment in a group may motivate to socialize with people by playing multiplayer games, location-based games that involve social interaction. Alternatively, social environment with other people present may also motivate people to use mobile services, because they can be used discretely without disturbing others, for example, ordering tickets or paying for parking during a meeting at work. Similarly, one reason for using mobile services in the presence of strangers, for example, on public transport is the need to create a personal space for social interaction, communicating with friends discretely in mobile chat rooms without disturbing others, or playing a game in order to kill time. These types of services are normally used alone.

The other axis in Figure 3 depicts the desired state of social interaction through the mobile service, named social interaction. Mobile banking services are an example of services that are typically intended for use in situations involving low social interaction with others. Because of the private nature of the financial information, the user normally wishes to be left alone. In contrast, people who use mobile chat services or search the whereabouts of their friends or family members seek social interaction (high social interaction).

A similar conceptualization has been previously mentioned by Nysveen et al. (2005b), who differentiate between machine-interactive and person-interactive services. In our opinion, machine-interactive services, that is, interactivity between the medium and the user, are services that aim at low social interaction and thus people wanting to be alone. In contrast, person-interactive services that occur between people through

Figure 3. The social setting



a medium are, in our model, services that aim at using services in a highly social setting, or that aim at interaction with other people either through the mobile media or in a real environment, for example, playing multiplayer mobile games in a group.

The social setting is the core of the benefit offered by newly launched mobile phone "presence" services that enable the person to specify criteria of how he/she wishes to be contacted at a specific time and social environment. It is also possible to indicate wishes for social interaction. For example, the user can indicate his/her current environment to others wishing to contact him/her: "I am now at a meeting, but can read text messages." Others can check who else is available for free-time socializing or work-related negotiating at that specific time, or how some particular person is best reached in the near future. This way they can acknowledge the other person's social environment and wishes for the level of social interaction, and proceed in an appropriate manner (for more information on presence applications, see e.g., http://europe.nokia.com/A4170049).

Relationship Between Customer and Service Provider

The relationship between the customer and service provider represents another important aspect of mobile services. Relationships can be used to categorize mobile services, because mobile services are especially effective in reaching individual customers, and they involve different types of relationships (Pirc, 2006). It has been argued that mobile services are considered more personal than any other remote service (Kleijnen, De Ruyter, & Andreassen, 2005), and that they can easily be personalized for specific customers (e.g., Watson, Pitt, Berthon, & Zinkhan, 2002).

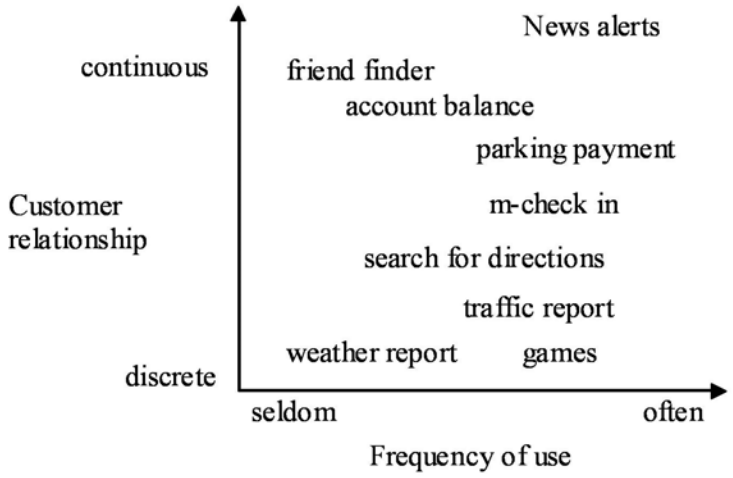
However, since customers cannot be expected to engage in a long-lasting relationship with every service provider, companies need to customize their service offerings according to the desired depth or length of relationship between the service provider and the customer (Pura, 2003a).

Thus, relationships represent the fourth classification scheme. The two ends of the customer relationship axis can be described as discrete transactions or continuous relationships (see Figure 4). In his study on relationship switching in a mobile service context, Pirc (2006) defined the mobile service relationship as involving either contract or pre-paid transactions. Discrete transactions can be seen as episodes that represent a set of different interconnected actions, and at the other extreme, there is a continuous relationship based on a set of interconnected sequences (Holmlund, 1997). Whereas the former constitutes a customer's discrete transactions, a continuous relationship often involves some kind of agreement between the customer and service provider. Subscription-based services entail opportunities for closer relationships with the service provider. However, most services offered in Europe today are based only on occasional transactions initiated by the customer and invoiced on the customers' monthly telephone bill (Pura, 2003b).

In this study on mobile services, continuous relationships are mainly related to services that are based on a contract. For example, mobile check-in requires a membership in an airline loyalty program. Services like checking for account balances, receiving an short message service (SMS) from the library when books are due, an SMS reminder the day before the dentist appointment, or ordering a security alert message to a mobile phone if someone happens to break in the summer cottage, all involve a subscription with a specific company and therefore represent continuous relationships.

In contrast, many services are offered primarily to unidentified occasional users. For example customers without prior relationships with a

Figure 4. Relationship



specific company can use m-payment for public transport, buying a fishing permit, logos, and weather reports, or buying products from vending machines. These types of services represent discrete transactions.

Frequency of use represents another perspective on the relationship between the customer and service provider. It is relevant when considering that a service that is paid continuously (continuous relationship) is not necessarily used often, or that a discrete relationship still can involve very regular use. The frequency of use may be linked to spontaneous and mobility needs (Anckar & D'Incau, 2002) and need for planning and improvising (Järvenpää & Lang, 2005). Some mobile services are used infrequently, when the need arises in a specific situation, for example, a real-time weather report may be found necessary while sailing and when the weather conditions are changing. Similarly, a fishing permit can be first purchased while at sea, without having to plan in advance where and when to go fishing. These types of services may be accessed via other channels in normal situations, but when other channels are not available, mobile services are used. This means that some services are used infrequently and sporadically. Furthermore, discrete services can be used without permission from the other party. In contrast, some other services are used often, for example, mobile games. Games are often included in a service portfolio offered by mobile portals and these types of services may be used based on a monthly subscription. However, this does not necessarily mean that the services are also used often.

The previous examples illustrate that creating value to the user by offering mobile services requires consideration of several aspects. These aspects may influence the choice to use mobile services instead of other electronic or traditional service delivery channels. In an attempt to offer a classification framework for mobile services, we summarize the previous grids into a framework that is discussed next.

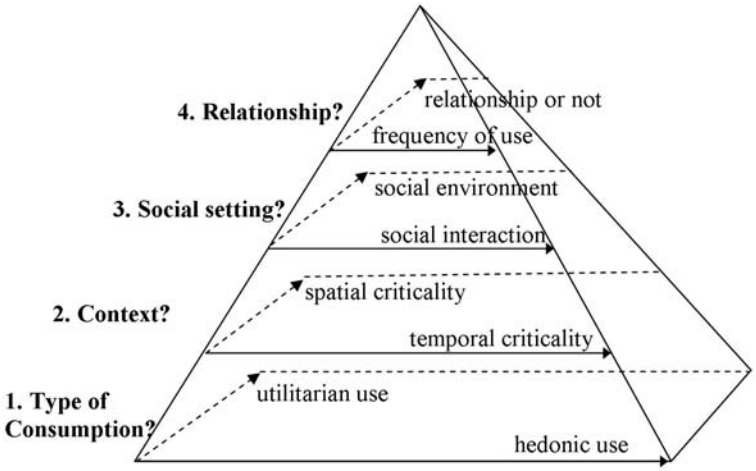
Classification of Mobile Services

Based on previous research, we now combine each of the classification schemes and introduce a framework for classifying mobile services by focusing on mobile service value (Figure 5). It is proposed that the assessment of the mobile service offering is hierarchical, based on the service value for the customer.

Hence, the four levels are related to each other and each level contributes to offering value-in-use to the customer. 1) The type of consumption related to the specific mobile service represents the underlying core of the value proposition; why the mobile service is used, whereas the following three levels provide additional value to the user relating to more specific aspects of mobile service use situations and service providers. The next consideration is 2) the context; when and where services may be used, that is, the temporal and spatial context of use, followed by the 3) social setting in which they may be used, and finally, 4) the relationship between the customer and service provider. These aspects represent factors that distinguish mobile services from other types of electronic or interpersonal services.

Thus, all four levels should be taken into account when categorizing and designing mobile services, pricing services, and segmenting services into bundles that offer similar value propositions. The pattern of each mobile service as depicted by the positioning in the classification framework can then be used for grouping similar services together and targeting service bundles to potential user groups. Next, we will illustrate how the classifying framework can be used to assess current or future mobile services. We will present two case examples that are different in nature.

Figure 5. Classification of mobile services



CASE STUDIES

The case studies illustrated in this chapter describe mobile services that are aimed at a specific interest group rather than the mass of people using a mobile device. We believe that the maturing global mobile market will slowly evolve into offering more targeted services for niche interest groups. This requires developing an understanding of hedonic and utilitarian needs, context of use, social setting, and the requirements of customers in their relationship with service providers. Better understanding of these issues should result in more effective marketing strategies by segmenting customers and different markets and targeting the services and marketing messages to the right potential user groups. The first case study used to depict the use of the conceptual framework is a mobile transaction service that allows purchase of a lure-fishing permit by text message, and the second case study describes a mobile adventure game. Next, the case studies are described in detail and summarized in the classifying framework to illustrate the differing nature of these two mobile services.

Case 1: Mobile Fishing Permit

To the authors' knowledge, a mobile fishing permit is currently a unique service offered in Northern Europe. The basic idea of the service is that the customer orders a regional fishing permit by text message when the need arises. The service has been offered as a pilot service in Northern Finland since 2003. Two documents are required when fishing by means other than basic angling or ice fishing. The first is a receipt for payment of the fishing management fee, the second the actual fishing permit. The actual fishing permit is region-specific and must be purchased before fishing. It can be bought traditionally in post offices or by electronic banking and in the region also with the help of a mobile service. The permit fee

Figure 6. Mobile fishing permit (Nikulainen, 2003)

is charged on the monthly phone bill. A receipt for the transaction is received as a text message. The provincial lure fishing fee is about 27 euros for a year or 6 euros for seven days.

The service has several implications for different stakeholders: unauthorized fishing is in decline, because of the convenient way of paying for the permit. Furthermore, easy remote controlling of permits through mobile devices saves controlling resources. The service has been successful and a large proportion of the permits have been purchased via a mobile device. This is an example of a niche service targeted at a hobby-related interest group. Since the pilot has been successful, similar services could be developed for various purposes, for example location-based services for hunters and pet owners.

Case 2: Mobile Game: Can You Confront Your Worst Nightmare?

Mobile games [1] have developed quickly in less than a decade. Nevertheless, the unique capabilities of wireless mobile devices have not yet been exploited much, with the exception of highly developed markets like Japan and Korea (Hämäläinen, 2006). The majority of the popular mobile games played in North America and Europe are basic games targeted at casual players, such as Tetris and the Who Wants To Be A Millionaire quiz (Segerstrale, 2006). As the mobile industry is working to provide faster access and improved ease of use, combined with more transparent and user-friendly pricing, game developers can start envisioning games that make best use of the features of mobile devices and combine those with the recent developments in the Web and Internet environment (Hämäläinen, 2006).

Rovio Mobile (www.rovio.com) is a Finnish mobile game developer and publisher founded in late 2004 by industry veterans. It is acknowledged as a leader in both product quality and game play innovation. The Rovio games are an example of developments into gaming solutions that have adopted views from PC and console-based game contexts. Rovio games are aimed at a demanding niche target group that looks for more action and experiences than the basic mobile games usually provide. The company focuses on developing strategy, adventure, role-playing and action games, and the games are story-driven, easy to pick up, and offer in-depth, extensive gaming experiences. A newly launched Rovio adventure game titled Darkest Fear 3: Nightmare is used to illustrate the classification scheme. A more detailed description of the course of the game may be found in Table 2 and a screenshot of the game is illustrated in Figure 7.

Rovio Mobile games are being sold through wireless application protocol (WAP) portals. The games can be ordered by clicking the "buy" tag on a WAP-page, or alternatively by sending an SMS that downloads a link to the mobile device screen in SMS format. The games are usually downloaded over the air (OTA), using the operator's network, directly to the user's mobile phone. In some cases games can be preinstalled or downloaded over Bluetooth to the phone. The game is usually charged on the users' monthly

Table 2. Description of the Darkest Fear™ 3: Nightmare game

Darkest Fear™ 3: Nightmare is the final part of Rovio Mobile's award winning Darkest Fear™ horror trilogy. The game offers completely new lightning effects like never seen on mobile. Thomas Warden has found his daughter Helen while searching for survivors of the horrible events at Grim Oak's Hospital. A monstrous bacterium has taken over her body, giving Helen unique powers but also making her extremely sensitive to light. Now it is up to you to find the ingredients for an antidote. As Thomas, you are forced to come up with ways to lighten your pathway with fire sparkles, water reflections and burning objects. When controlling Helen however, light is the deadliest of enemies. The game's ingenious puzzles, horrifying atmosphere, fifteen different endings and Helen's new monster abilities offer a unique experience. The game throws flesh-eating creatures, prowling zombies and three monstrous bosses against you. Can you confront your worst nightmare?
ROVIO MOBILE Copyright © 2006 Rovio Mobile Ltd.

phone bill. However, alternative payment methods, such as credit card or PayPal, may also be used in different markets.

Logic of the Case Analysis

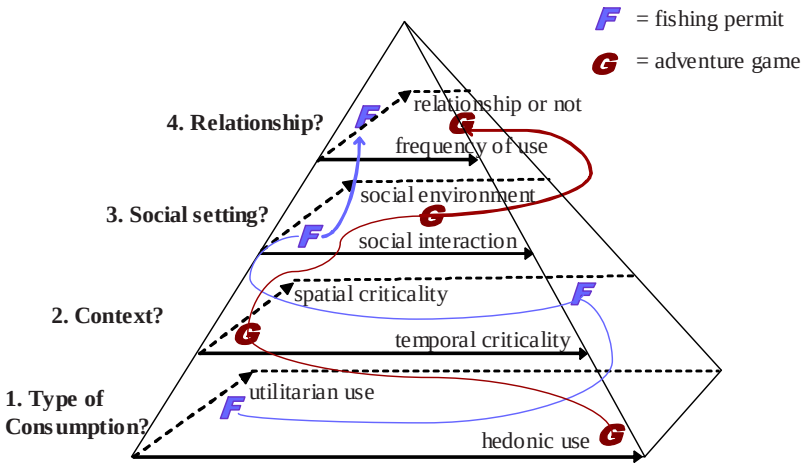
The proposed conceptual framework presented in **Figure 8 is expected to ease constructive evalu**ation of mobile service business opportunities and to describe new services by clarifying the intended target group and the type of situations in which the services will most likely be used. We illustrate the possible applications by evaluating the case study services with the help of the framework.

The Fs in the framework in Figure 8 represent the fishing permit service and the Gs represent the adventure game case. They are located on the four grids in places where they are most likely positioned from a customer perspective, in most typical use situations. The positioning of the fishing permit and adventure game was done by the authors, and it is based on discussions with service users and our perceptions of typical users of the specific services in typical usage situations. The service providers were able to comment on

Figure 7. Mobile game: Darkest Fear 3: Nightmare (horror/adventure/puzzle) (Source: Rovio & N70 image retrieved from the Nokia press site)

(Copyright: 2006 Rovio Mobile Ltd. Used with permission)

Figure 8. Classification of the case study examples



our preliminary suggestions and changes were made accordingly.

DISCUSSION OF THE SOLUTIONS BASED ON THE CASE STUDIES

The framework presented in Figure 8 helps to clarify the value of the mobile service from a customer's point of view. A presentation follows of a managerial list of questions about issues to consider while planning or evaluating a mobile service. The value of the fishing permit can be communicated as a convenient, easy, and quick way to buy the permit. The service should be mainly marketed proactively, so that the potential customers know where to get it when the need arises spontaneously, on the move. Reaching potential users in the natural environment, alone, out of the reach of other media, is demanding. Nevertheless, regarding the social appreciation of others, fishing communities can be effective peer marketers. Payment on the mobile phone bill seems an appropriate choice, since it is a discrete transaction. However, there is also potential for continuous subscriptions for current permit holders, because payment on the phone invoice is more convenient than paying for a permit the traditional way in a post office.

The adventure game case represents another type of service that is purely hedonic in nature. Therefore, the consequences of use important in the previous example, such as convenience, speed, and good value for money, are not that important in a gaming context. Instead, people evaluate the service based on how entertaining the game is and what experiences they go through while playing the game. Games are mostly accessed via mobile portals anywhere and any time the inspiration arises to play a game. Therefore the portal plays an important role in reaching the customers wherever they are. The customer relationship is also formed primarily with the mobile portal, and payment for the service is thus dependent on the customers' type of contract with the portal. Furthermore, in a portal context, service marketers face a challenge

in attracting potential customers while competing with other interesting services offered.

It should be noted that the mobile service examples mentioned in the classification grids may be placed in several positions, depending on the use situation and the individual customer's preferences. For example, the mobile fishing service as a discrete transaction meets especially well the needs of tourists or occasional fishermen, who may not otherwise pay for the permit. However, fishermen who go regularly to the same place to fish may consider using the mobile service only if it is a more convenient way to pay for the permit than the traditional ways. These regular service users may also consider spatial and temporal issues less critical, but may face more social pressure to pay for the permit in the local community. Similarly, games are normally used in non-temporally and spatially critical situations, for example at home. However, games can also be played at highly context-critical places, if they are mostly used while waiting for public transport or while traveling on it. Furthermore, basic games are used alone. However, the multiplayer games present opportunities for use of games in a group, even when the group of people is not geographically in the same place.

A summary of the main managerial questions for mobile business stakeholders and conclusions on how the two case examples of mobile services are positioned in the framework are presented in Table 3. The classification exercise can be done for different types of services, and it is expected to give indications on how to group similar services together and how to communicate the value of the service to the customers.

In order to be able to position different services relative to each other, we also propose a preliminary practical approach for evaluating, comparing, and positioning different mobile services. Therefore, a short set of eight possible questions is presented in Table 4. The questions were developed based on previous literature and they should give indications as to what aspects of mobile services are important to the user. A Likert scale or other type of scales could be used to illustrate the positioning of a service relative to other services. However, it should be noted that a low vs. high figure on the scale does not necessarily mean that one service is better than the other, but only indicating that they differ in how and when they are used.

The proposed set of questions illustrated in Table 4 should be developed further and tested with mobile service users. The eight questions summarize the four important factors that influence the use of mobile services: (1) type of consumption, (2) temporal and spatial criticality, (3) social setting, and (4) relationship between the customer and service provider. The results of this type of questionnaire should ease the task of differentiating mobile services from each other with regard to the factors that are important to customers.

RECOMMENDATIONS BASED ON THE CLASSIFICATION SCHEME

The proposed classification schemes of mobile services contribute to marketing practice in several ways. The proposed categorizations have implications for communication, design, and pricing of mobile services and for segmenting customers using the services. They provide insights for marketers on how to differentiate and group mobile services based on criteria that are relevant from a customer's point of view. The framework might be used to evaluate potential customer perceptions of specific mobile services and to understand the types of mobile services that customers are likely to perceive valuable and use in different situations.

Thus, in the end, the customers will evaluate the overall value of the services based on the type of consumption, context of use, social setting, and relationship with the provider. The classification framework helps service providers

Table 3. Summary of the positioning of the case studies in the classification framework

MANAGERIAL QUESTIONS	FISHING PERMIT	ADVENTURE GAME
1. Type of consumption • Is the service used for efficiency needs to achieve a task? (utilitarian) • Is the service used just for its own sake, for experience? (hedonic)	Utilitarian task of paying for permit effectively	Hedonic, fun experience
2. Temporal and Spatial Context • When and where is the service used? (spatial) • How spontaneously is the service used? (temporal)	In time and place critical situations, while at sea. A need for personal permit is mainly spontaneous.	The user can play a game anywhere, anytime in non-time and place critical situations. Games are usually played often, but can be used also sporadically e.g. while waiting and 'killing time'
3. Social Setting • Is the service used alone or in a group? (social environment) • Is the service normally used for creating personal space or mainly for the purpose of socializing? (social interaction)	The service is normally used alone, avoiding social interaction especially with the permit controllers.	This game is usually used alone with no network connectivity, but can create social interaction.
4. Relationship • How does the service support the different phases of customer relationship from new customer acquisition to customer retention? (relationship or not) • Does the nature of the service support better regular customers or mainly sporadic use of services on the move? (frequency of use) • How do customers prefer to pay for services? (discrete transactions / subscription-based use) • Does the delivery of the service require permission from the customer, peer customers, other stakeholders? (e.g. location-based information and advertising)	The permit is a discrete transaction for new customers. It is used sporadically while fishing. It is a discrete transaction that is paid once a year or for a week. It does not require permission, but continuous users could opt-in for e.g. reminders from the provider when the permit expires.	Because Rovio does not offer games directly to end-customers, but uses mobile portals as distribution channels, customers do not necessarily know the Rovio brand while playing the game. New customers are acquired via portals. Games are often used regularly. This game is a discrete transaction, because it is delivered to the end-users via mobile portals and invoiced on the monthly phone bill.

Table 4. Proposed approach for evaluating different services

PROPOSED QUESTIONNAIRE ITEMS	SCALE		
1. Type of consumption			
The added value gained by using the service effectively is...	*low*	□□□□□	*high*
The added value gained by enjoying the use experience as such is...	*low*	□□□□□	*high*
2. The temporal and spatial criticality of service			
How critical is the place where the service is used?	*low*	□□□□□	*high*
How urgent is time of service use or service delivery?	*low*	□□□□□	*high*
3. The social setting			
The social environment where the service is primarily used is...	*alone*	□□□□□	*In group*
The social interaction with other users related to service use is...	*low*	□□□□□	*high*
4. The relationship			
The type of the relationship between the user and the service provider is...	*discrete*	□□□□□	*continuous*
The service is used...	*seldom*	□□□□□	*often*

to assess the critical success factors of mobile services compared to other electronic channels from a customer-centric viewpoint and to identify situations where the service is mostly used. As such, the classification schemes can be used to develop strategies to create value in the mobile channel in situations in which it represents most added value for the customers compared to other channels.

Because customers may use services for various reasons, customers may be segmented based on the relative balance of what is important to the customer in different situations. These key benefits can be communicated to the right customer segments in order to attract the right customers. To date, mobile services are quite often marketed as a bundle of services that people can use via their mobile devices. However, this type of homogenous communication undermines the real potential of using mobile services for different needs and preferences, in different situations. Different types of services should be marketed based on their potential value to the customer.

The **hedonic vs. utilitarian** categorization of the service types gives implications on how to communicate the value of some specific services or service types to the customers. Hedonic value is often gained from entertainment-related services, for example, games, and thus can be communicated as fun and an enjoyable service experience. This can be facilitated with pictures and describing what kinds of feelings the entertaining content is expected to evoke in the user. Utilitarian services represent information-based services,

such as weather reports, timetables, and search services that aim at achieving a task effectively, maybe saving time and finding information easily. Therefore the communication of the value of these services should also reflect the consequences of use; effectiveness; saving time and effort; good value for money compared to alternatives; and convenience of accessing the service regardless of temporal and spatial constraints.

Moreover, the **context** of use has several implications. The fact that there may be other alternative services that the customer can use influences pricing and service design. Mobile services that are offered as the only alternative can be priced higher than services that are competing with other service delivery alternatives. Correspondingly, when there are other service alternatives, the mobile service must be priced competitively.

Furthermore, it is important that services aimed for urgent or highly critical spatial context are designed in a simple way, so that the customer can use the service easily without extra effort or time needed. Point-of-sale advertising, for example, at subway stations, informing how to pay for the subway ticket with the mobile phone, are good examples of how the use of mobile services can be promoted on the spot. Services used in an urgent situation need to rely on logical user interfaces and customers' recall of how to access the services. In this respect, word-of-mouth may have an important role in marketing strategies. Moreover, marketers need to acknowledge that the customer's location may change even during a mobile service use session and network availability may pose some problems, for example, if sitting on a train. Ability to save information for later use or for continuing to use the service at a later stage may therefore be a useful feature. On the other hand, services aimed at less time and place-critical contexts, where the customer has more time to use the service, can be designed with additional service elements to increase the value of the service. It is easier to reach customers in a state when they have time to read instructions and marketing communication. They may also have access to other channels, for example, digital television, the Internet, radio, and magazines that can serve as cross-marketing channels.

The social setting of mobile service usage situations also has implications for marketing strategies. Social pressure from friends and family may exist when using services, especially in group usage situations. Moreover, services used for socializing purposes have high interaction with participants, and therefore marketing efforts should seize the opportunity of using these networks of customers interacting with each other. The interaction may even result in sharing downloaded digital content and received messages (Maier, 2005; Van Camp, 2005) Motivating current customers to spread positive word-of-mouth is often an essential part of service providers' marketing strategy in mobile environments (Barnes, 2002).

Customers who have an ongoing **relationship** with a specific service provider are more attractive to companies than customers who use services in an ad hoc manner. Obviously, it is less expensive to cross-sell to existing customers than to attract new customers. However, it is important to understand that there may also be customers with an ongoing relationship, who still use the service infrequently. In addition, there may be unidentified users who may use the service regularly and wish to stay anonymous. The discrete use of services is often a benefit sought compared to other service channels for people who do not wish to reveal their identity or use services discretely without others noticing (e.g., adult entertainment, chat services, searching for others' taxed income information, car owner information based on the registration plate, etc.). Avoiding personal contact has also been suggested as a motive for using mobile services, which offers new perspectives on how to market these services. They may often be an alternative to other channels that require personal interaction. In the mobile context, people can act

anonymously and plan better what to say and how to respond to others' comments than in personal interaction situations (Aminuzzaman, 2005). Thus, a viable option of seeing services as connecting people is to market some mobile services as a planned choice of avoiding personal contact and offering a way to interact anonymously and discretely with the service provider.

In addition, a much-debated issue in the literature is how to invoice services so that customers feel they gain value for money. Many companies have ended up offering alternatives to access the services, either with transaction-based invoicing on the phone bill; a short trial subscription for free or for a small amount of money; or a monthly subscription. The choice depends on the nature of the service and no definite recommendation can be made (Munnukka, 2005). However, service providers should also acknowledge that the best payment method for the service provider, one that creates steady and secure cash flows (subscription-based or payment from a service account), may not always be the preferred way from the customer's point of view. Invoicing per usage on the monthly phone bill is regarded as convenient and may even motivate people to use mobile services (Pura, Viitanen, & Liljander, 2003). Alternative pricing strategies are common in different markets. For example in the gaming context, some offer subscription models such as i-Mode. Some offer try-and-buy solutions, where the game is distributed free and the first level or one minute of game play is free. Afterwards, the game needs to be registered and paid for. The Darkest Fear game case example presented in this study is invoiced on the monthly phone bill after downloading the game. These transaction-based invoicing methods are common in Europe. Eighty-three percent of the services used in the focal market are post-paid. However, other new methods are being tested, such as a mobile wallet and chip cards in the mobile phone that can be used as debit cards. On a global scale, we believe that pre-paid contracts are more common, and therefore package deals and monthly subscriptions are also more common than in the market in question. Nevertheless, we suggest that the classification grid can be used in all markets, regardless of how the service is paid for, because the relationship grid in the framework encourages constructive analysis of the payment and contract options.

FUTURE TRENDS AND SUGGESTIONS FOR RESEARCH

The typology and the proposed set of questions give indications for further empirical research in the mobile sector. On a global scale, mobile services are currently primarily offered by network operators and service portals. However, especially in markets where the majority of services are post-paid based on transactions, the role of individual service and content providers may become more important in the future. This kind of development impacts on all the layers of the classification framework presented in this study. The customer relationship with the service provider in particular could evolve into a more personal relationship, and information about real use of services and customer attitudes towards individual services could be tracked more accurately. In general, the transaction vs. relationship nature of service use requires further research. Some research results on the loyalty of mobile services indicate that customers who are restricted by a contract to stay with a service provider are more likely to switch service providers after the contract period is over than those who do not have a contract (Fullerton, 2005; Libai & Nitzan, 2005). Thus, subscription-based services may create a falsified feeling of loyal customers who are only committed to use the same service provider because of constraints that prevent them from changing providers. Fullerton (2005) even claims that customers' feelings of being stuck in the relationship tend to override the positive feelings of attachment to the provider.

Moreover, the discrete transactions are very important as they also potentially represent a door opener to more frequent use of services and can be used to increase the awareness of the service provider. Mobile services represent a new range of services for many customers, and offering customers the opportunity of trying new mobile services and thus indicating the potential benefit of the mobile device is expected to be successful.

Future research needs to empirically explore and develop reliable scales for measuring the proposed conceptualization, in order to validate the proposed classification schemes. Although future research may need to structure the classification schemes according to industries, we feel that it is important to move beyond the traditional industry-specific classification towards a more generalizable and ultimately more descriptive categorization of mobile services.

CONCLUSION AND CONTRIBUTION

In this chapter, we focus on what users value in mobile services and how mobile services are used in typical use situations, in order to better understand the underlying reasons why mobile services are used. This chapter extends past classification schemes of traditional interpersonal services and e-services by examining the characteristics of mobile services and focusing on service value from a customer perspective. Based on a literature review, we introduce four different classification schemes that can be used to understand the consumption type, use context, social setting, and customer relationship. The resulting classification framework gives a holistic view of mobile services from a value-in-use perspective, and it is considered essential in differentiating and grouping mobile service offerings in a meaningful way. It also describes the distinct characteristics of mobile services that should be considered in developing, assessing, and planning marketing communication of mobile services.

The classification scheme differs from existing research on mobile services in the respect that emphasis is moved away from the prevailing focus on what developments of mobile applications are technologically possible. The focus is on what may be offered and how customers perceive the value of these offerings. This service- and customer-oriented perspective on mobile services is hence achieved by acknowledging why current and potential customers might use services in concrete situations, and how services might create value for customers. Because the focus is on different characteristics that describe mobile services, the classification schemes can also be used to explore new avenues for mobile services and to create new types of services.

In conclusion, the classification scheme can be used to evaluate potential customer reactions to specific mobile services and to understand the types of mobile service that customers are likely to try and use.

ACKNOWLEDGMENT

The authors would like to thank the Foundation for Economic Education for its financial support. We also appreciate the comments of the reviewers who have reviewed previous versions of this paper. Our sincerest thanks also to Jan Bonnevier who provided us with the Rovio Mobile Ltd case information and to Lauri Haapanen for providing information about the mobile fishing permit. Both authors have contributed equally to the chapter.

REFERENCES

Aminuzzaman, S. (2005). Is mobile phone a socio-cultural change agent? A study of the pattern of usage of mobile phones among university students in Bangladesh. In *Proceedings of the International Conference on Mobile Communi-*

cation and Asian Modernities II, Information, Communications Tools & Social Changes in Asia, Beijing, China.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. *Journal of Information Technology Theory & Application, 4*(1), 43-64.

Angehrn, A. (1997). Designing mature Internet business strategies: The ICDT Model. *European Management Journal, 15*(4), 361-369.

Babin, B. J., Darden, W. R., & Griffin, M. (1994). Work and/or fun: Measuring hedonic and utilitarian shopping value. *Journal of Consumer Research, 20*(4), 644-656.

Balasubramanian, S., Peterson, R. A., & Järvenpää, S. L. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science, 30*(4), 348-361.

Barnes, S. J. (2002). Wireless digital advertising: Nature and implications. *International Journal of Advertising, 21*(3), 399-421.

Celuch, K., Goodwin, S., & Taylor, S. A. (2007). Understanding small scale industrial user Internet purchase and information management intentions: A test of two attitude models. *Industrial Marketing Management, 36,* 109-120.

Chaudhuri, A., & Holbrook, M. B. (2002). Product-class effects on brand commitment and brand outcomes: The role of brand trust and brand affect. *Brand Management, 10*(1), 33-58.

Clarke, I., & Flaherty, T. (2003). Mobile portals: The development of m-commerce gateways. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce, technology, theory and applications* (pp. 185-201). Hershey, PA: Idea Group.

Cotte, J., Tilottama, G. C., Ratneshwar, S., & Ricci, L. (2006). Pleasure or utility? Time planning style and Web usage behaviors. *Journal of Interactive Marketing, 20*(1), 45-57.

Dabholkar, P. A. (1996). Customer evaluations of new technology-based self-service options: An investigation of alternative models of service quality. *International Journal of Research in Marketing, 13,* 29-51.

Fullerton, G. (2005). How commitment both enables and undermines marketing relationships. *European Journal of Marketing, 39*(11/12), 1372-1388.

Giaglis, G. M., Kourouthanassis, P., & Tsamakos, A. (2003). Towards a classification framework for mobile location services. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce: Technology, theory, and applications* (pp. 67-85). Hershey, PA: Idea Group.

Hämäläinen, M. (2006). Enabling innovation in mobile games—Going beyond the conventional. In *Proceedings of Mobility Round Table*, Finland: Helsinki School of Economics.

Heinonen, K. (2004a). Reconceptualizing customer perceived value: The value of time and place. *Managing Service Quality, 14*(2/3), 205-215.

Heinonen, K. (2004b). Time and location as customer perceived value drivers. (doctoral thesis No. 124), HANKEN, Swedish School of Economics and Business Administration, Finland.

Heinonen, K. (2006). Temporal and spatial e-service value. *International Journal of Service Industry Management, 17*(4), 380-400.

Heinonen, K., & Andersson, P. (2003). Swedish mobile market: Consumer perceptions of mobile services. *Communications & Strategies, 49,* 151-171.

Holmlund, M. (1997). Perceived quality in business relationships (Doctoral dissertation No. 66). HANKEN, Swedish School of Economics and Business Administration, Finland.

Hyvönen, K., & Repo, P. (2005). Mobiilipalvelut suomalaisten arjessa (Mobile Serives in the everyday life of Finns). In J. Leskinen, H. Hallman, M. Isoniemi, L. Perälä, T. Pohjoisaho & E. Pylvänäinen (Eds.) *Vox consumptoris—Kuluttajan ääni,* . Kerava: Kuluttajatutkimuskeskus.

Isoniemi, K., & Wolf, G. (2001). Three segments of mobile users. In *Proceedings of the Seamless Mobility Workshop,* Stockholm, Sweden.

Järvenpää, S. L., & Lang, K. R. (2005). Managing the paradoxes of mobile technology. *Information Systems Management, 22*(4), 7-23.

Kleijnen, M., De Ruyter, K., & Andreassen, T. W. (2005). Image congruence and the adoption of service innovations. *Journal of Service Research, 7*(4), 343-359.

Kleijnen, M., Wetzels, M., & De Ruyter, K. (2004). Consumer acceptance of wireless finance. *Journal of Financial Services Marketing, 8*(3), 206-217.

Koivumäki, T., Ristola, A., & Kesti, M. (2006). Predicting consumer acceptance in mobile services: Empirical evidence from an experimental end user environment. *International Journal of Mobile Communications, 4*(4), 418-435.

Laukkanen, T., & Kantanen, T. (2006). Customer value segments in mobile bill paying. In *Proceedings of the 3rd International Conference on Information Technology: New Generations 2006 (ITNG 2006),* Las Vegas, NV: IEEE Computer Society Press.

Laukkanen, T., & Lauronen, J. (2005). Consumer value creation in mobile banking services. *International Journal of Mobile Communications, 3*(4), 325-38.

Libai, B., & Nitzan, I. (2005). Customer profitability over time in the presence of switching costs. In *Proceedings of 14th Annual Frontiers in Services Conference,* Arizona: The Center for Service Leadership, W.B. Carey School of Business, Arizona State University.

Lin, H-H., & Wang, Y-S. (2006). An examination of the determinants of customer loyalty in mobile commerce contexts. *Information & Management, 43,* 271-282.

Lovelock, C. H. (1983). Classifying services to gain strategic marketing insights. *Journal of Marketing, 47,* 9-20.

Maier, M. (2005, March 1). *Song sharing for your cell phone.* Retrieved November 10, 2006, from http://money.cnn.com/magazines/business2/business2_archive/2005/03/01/8253120/index.htm

Mennecke, B., & Strader, T. (2003). *Mobile commerce: Technology, theory, and applications.* London: Idea Group.

Meuter, M. L., Ostrom, A. L., Roundtree, R. I., & Bitner, M. J. (2000). Self-service technologies: Understanding customer satisfaction with technology-based service encounters. *Journal of Marketing, 64*(3), 50-64.

Mitchell, K., & Whitmore, M. (2003). Location based services: Locating the money. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce, technology, theory and applications* (pp. 51-66). Hershey, PA: Idea Group.

Mobile game. (n.d.). Retrieved September 28, 2006, from http://en.wikipedia.org/wiki/Mobile_game

Munnukka, J. (2005). Dynamics of price sensitivity among mobile service customers. *Journal of Product & Brand Management, 14*(1), 65-73.

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., Teerling, M. L., Thomas, J. S., et al. (2006). Challenges and opportunities in multichannel customer management. *Journal of Service Research, 9*(2), 95-112.

Nicholson, M., Clarke, I., & Blakemore, M. (2002). One brand, three ways to shop: Situational variables and multichannel consumer behaviour. *The*

International Review of Retail, Distribution and Consumer Research, 12(2), 131-148.

Nikulainen, K. (2003). *Fishermen hooked in mobile in Oulu (Oulussa kalastajat iskivät mobiilikoukkuun).* Retreived January 26, 2006, from http://www.digitoday.fi/page.php?page_id=11&news_id=20035391

Novak, T. P., Hoffman, D. L., & Duhachek, A. (2003). The influence of goal-directed and experiental activities on online flow experiences. *Journal of Consumer Psychology, 13*(1&2), 3-16.

Nysveen, H., Pedersen, P. E., & Thorbjørnsen, H. (2005a). Explaining intention to use mobile chat services: Moderating effects of gender. *Journal of Consumer Marketing, 22*(5), 247-56.

Nysveen, H., Pedersen, P. E., & Thorbjørnsen, H. (2005b). Intentions to use mobile services: Antecedents and cross-service comparisons. *Journal of the Academy of Marketing Science, 33*(3), 330-46.

Okazaki, S. (2005). New perspectives on m-commerce research. *Journal of Electronic Commerce Research, 6*(3), 160-64.

Pirc, M. (2006). Mobile service and phone as consumption system—The impact on customer switching. In *Proceedings of the Helsinki Mobility Round Table*. Helsinki: Helsinki School of Economics.

Pura, M. (2003a). Case study: The role of mobile advertising in building a brand. In B. Mennecke & T. Strader (Eds.), *Mobile commerce: Technology, theory, and applications* (pp. 291-308). London: Idea Group.

Pura, M. (2003b). Linking perceived value and loyalty to mobile services. In *Proceedings of the ANZMAC 2003,* Adelaide, Australia.

Pura, M. (2003c). Measuring loyalty to mobile services. In *Proceedings of the Third International Conference on Electronic Business (ICEB),* National University of Singapore.

Pura, M. (2005). Linking perceived value and loyalty in location-based mobile services. *Managing Service Quality, 15*(6), 509-538.

Pura, M., & Brush, G. (2005). Hedonic and utilitarian motivations for mobile service use. In *Proceedings of SERVSIG,* National University of Singapore.

Pura, M., Viitanen, J., & Liljander, V. (2003). Customer perceived value of mobile services. In *Proceedings of the 32th EMAC Conference*, Glasgow: University of Strathclyde.

Rodgers, S., & Sheldon, K. (2002). An improved way to characterize Internet users. *Journal of Advertising Research, 42(*5), 82-94.

Segerstrale, K. (2006). *Enabling innovation in mobile games—Going beyond the conventional.* Helsinki: Helsinki.

Sheth, J., Newman, B., & Gross, B. (1991). *Consumption values and market choices, theory and applications.* South-Western.

Mort, G. S., & Drennan, J. (2005). Marketing m-services: Establishing a usage benefit typology related to mobile user characteristics. *Journal of Database Marketing & Customer Strategy Management, 12*(4), 327-41.

Van Camp, S. (2005, May 1). 15 minutes with Derek Pollock on markets for tablet PC. *Adweek Magazines' Technology Marketing.*

Van der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS Quarterly, 28*(4), 695-704.

Van der Heijden, H., Ogertschnig, M., & Van der Gaast, L. (2005). Effects of context relevance and perceived risk on user acceptance of mobile information services. In *Proceedings of 13th European Conference on Information Systems.* Regensburg, Germany: Institute for Manage-

ment of Information Systems at the University of Regensburg.

Venkatesh, V., Morris, M., Davis, G., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Watson, R. T., Pitt, L. F., Berthon, P. R., & Zinkhan, G. M. (2002). U-commerce: Expanding the universe of marketing. *Journal of the Academy of Marketing Science, 30*(4), 333-47.

Yoo, Y., & Lyytinen, K. (2005). Social impacts of ubiquitous computing: Exploring critical interactions between mobility, context and technology. *Information and Organization, 15*, 91-94.

ENDNOTE

[1] The definition of a mobile game on Wikipedia ("Mobile game," n.d.) reads: "A mobile game is a computer software game played on a mobile phone, smartphone, PDA or handheld computer. Mobile games may be played using the communications technologies present in the device itself, such as by text message (SMS), multimedia message (MMS) or GPRS location identification. More common, however, are games that are downloaded to the mobile phone and played using a set of game technologies on the device."

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, and J. Day, pp. 111-133, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.9
Strategy Aligned Process Selection for Mobile Customer Services

Ragnar Schierholz
University of St. Gallen, Switzerland

Lutz M. Kolbe
University of St. Gallen, Switzerland

Walter Brenner
University of St. Gallen, Switzerland

ABSTRACT

In this chapter we analyze how companies define their customer value proposition and how the selection of successful mobile customer services is done in alignment with this strategic positioning. We derive a set of five different strategic goals (price leadership, product quality leadership, customer intimacy leadership, accessibility leadership, innovation leadership) and apply this classification to case studies we analyzed. We show interdependencies between the strategic premises and the processes selected for being supported by mobile technology, resulting in typical properties which qualify processes for mobilization. These are used to derive guidelines for strategy aligned process selection when implementing mobile customer services.

INTRODUCTION

Mobilizing Customer-Oriented Business Processes

Technological advancements in mobile communications enable new ways of doing business (Feldman, 2000, pp. 26; Stafford & Gilleson, 2003), often referred to as mobile business (MB) or mobile commerce (MC). While Turowski and Pousttchi (2003, p. 3) do not distinguish between the two but rather use the term *mobile commerce*,

Lehner (2003, pp. 6-8) and Zobel (2001, pp. 2-3) define *mobile business* as the application of mobile technologies to improve or extend business processes and open new market segments and distinguish it from mobile commerce. Here, the latter is rather a subordinate field of MB, focusing of the handling of transactions. In this chapter we will follow the understanding of Lehner and Zobel and concentrate on the application of mobile technologies to support customer-oriented business processes.

The research field dealing with the interaction of businesses with their customers and the related back-end processes within the businesses, such as marketing, sales, and service processes has often been referred to as customer relationship management (CRM) or, when supported by Internet technologies, e-commerce CRM (eCRM) (Romano & Fjermestad, 2002, 2003). Gebert, Geib, Kolbe, & Brenner (2003) classify CRM processes as knowledge-intensive processes, managing knowledge for customers (e.g., knowledge about products and services), knowledge from customers (e.g., customer experience with products and services), or knowledge about customers (e.g., knowledge about customers' preferences and histories). Geib, Reichold, Kolbe, & Brenner (2005) provide a framework identifying major CRM processes in the fields of marketing, sales, and service and point out their interdependencies.

An empirical analysis addressing 1,000 subjects with CRM responsibility in large companies (82% with a revenue > € 100 million) and 89 respondents (9%) was conducted in the authors' research team. Eight percent of the respondents indicated that they already have a mobile CRM solution, a further 22% are currently working on a mobile CRM solution, and 30% are planning to do so (Dous, Salomann, Kolbe, & Brenner, 2004).

Combining the concepts of CRM and MB allows new types of interaction between companies and customers. To leverage investments in IT, the investment has to be aligned with the business strategy (Bakos & Treacy, 1986; Brynjolfsson & Hitt, 2000; Hitt & Brynjolfsson, 1994; Weill & Broadbent, 1998; Weill, Subramani, & Broadbent, 2002). A recent survey conducted by the German Society for Management Research investigated major success factors and success barriers for MB initiatives. The top success barrier was a lack of strategic vision and the initiatives' alignment with corporate strategy (Wamser & Buschmann, 2006).

Obviously, companies face the question of how to select the right MB investment to support their business strategy and how to identify potentials that can be exploited using mobile communication and transaction channels. Depending on the strategic premises different alternatives of mobilizing customer-oriented business processes must be chosen (Weill & Vitale, 2002). The goal of this chapter is to provide assistance in making this decision.

Research Goals and Structure

In this chapter we show interdependencies between the strategic premises and the processes selected for being supported by mobile technology. We explicitly do not analyze the process of defining the strategy but rather rely on existing work of strategy research. Therefore we answer the questions:

- *What are the typical characteristics of business processes chosen for mobile technology support?*
- *What are the interdependencies between these characteristics and the companies' market strategy?*

First, the second section gives an overview of existing research in the field of MB and CRM and identifies the gap of customer-focused research the authors see. In the third section we briefly describe 10 cases, where companies have successfully introduced mobile solutions to support business processes in alignment with their

strategic positioning towards customers. In the fourth section we introduce the classifications of different strategic focuses from Crawford and Mathews (2001) and Treacy and Wiersema (1994). We derive a set of five different strategic goals which are specifically focused on the company's interaction with the customer and apply this classification to the analyzed cases. We analyze the selection of processes for the support by mobile technology in the cases and identify the relationship between the strategic premises according to the framework derived from Crawford and Mathews (2001) and Treacy and Wiersema (1994). This results in typical properties which qualify processes for mobilization. The final section summarizes the findings and gives an outlook on further research to be done in this field.

Research Methodology

Our research approach follows the concept of case study research as described by Eisenhardt (1989), Stake (1995), and Yin (2002). The cases (see the third section) have been selected from available published material, in the case of the Helsana health insurance, cologne public transport authority, eBay, and Lufthansa airline the authors have been involved in-depth through a long-term research partnership. Selection was based on the following criteria: a) availability of information about the company's strategic orientation towards customers, b) the case deals primarily with the introduction of mobile technology (be it cellular, synchronization, or other), and c) the process(es) affected by the introduced technology is a customer-oriented business process as defined in the process model developed by the authors' research team and described in Geib et al. (2005). Data were collected by analysis of the published available material about the projects and the companies in general as well as by semi-structured interviews with employees involved in the projects. Only the core aspects from the previously published cases are summarized in this chapter.

The data from each case was analyzed following the strategy suggested by Yin (2002). The analysis had the primary objective of understanding the process selection and the influence which the corporate strategy had in this process. The findings finally have been integrated into a generalization of strategy's implications for the process selection and design.

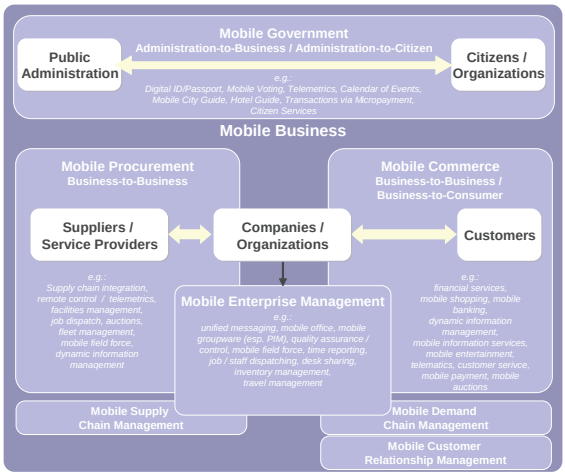
BACKGROUND

Mobile Business

Technological advancements in mobile communications enable new ways of doing business (Raisinghani, 2002), often referred to as mobile business or mobile commerce. While Turowski and Pousttchi (2003) do not distinguish between the two but rather use the term *mobile commerce*, Lehner (2003) and Zobel (2001) define *mobile business* as the application of mobile technologies to improve or extend business processes and open new market segments. They differentiate between MB and MC, the latter being a rather subordinate MB field focusing on the handling of transactions. With a similar understanding of the term, Möhlenbruch and Schmieder (2001) conceptualize MB in analogy to electronic business and distinguish fields such as mobile supply chain management, mobile procurement, mobile customer relationship management, and so forth (see Figure 1). We follow this more general understanding and concentrate on mobile CRM (mCRM), which we define as mobile technologies' application in order to support CRM processes such as marketing, sales and service delivery.

The research development in MB and MC can be compared to the development in electronic business and e-commerce. It can be structured into multiple stages. The first stage begins with the technological foundation in IT and infrastructure. It is followed by simple consumer-focused application and service concepts along with business models

Figure 1. Conceptualization of mobile business (cp. Möhlenbruch & Schmieder, 2001)



for technology and base service providers. These applications and services are being advanced further, until they reach a maturity level appropriate for business use. Finally the technology is applied to support business processes and entire business models for businesses other than technology and base services are developed. The shift of focus in MB from simple applications for end consumers to advanced applications for business is a condition that hints for mobile technologies to be disruptive technologies in the sense of Christensen (1997): "technologies [that] underperform established products in mainstream markets […] but […] have other features that a few fringe (and generally new) customers value. Products based on disruptive technologies are typically cheaper, simpler, and, frequently, more convenient to use" (p. 15). Funk (2003) analyzes this issue in great detail.

The subject of MB and related subjects has gained a substantial interest in the research community, which can be seen in the emergence of new journals focusing on this particular subject (e.g., the *International Journal of Mobile Communications* or the *International Journal of Mobile Computing and Commerce*), conferences explicitly dedicated to the subject (e.g., the International Conference on Mobile Business sponsored by the IEEE being held for the fifth time in 2006) or special issues of well-established journals in the field of IS (see Liang & Wei, 2004; Mylonopoulos & Doukidis, 2003; Schierholz, Kolbe, & Brenner, in press; Urbaczewski, Valacich, & Jessup, 2003).

Current research has certainly passed the mere technology-focused stage, even though advancements in technology are still a subject (e.g., Fritsch & Rossnagel, 2004; Turowski & Pousttchi, 2003). Consumer applications and services are well established and drive impressive markets as well (Funk, 2003, p. 20). Past and current research is further advancing this field as well (Ali, Torabi, & Ali, 2006; Amberg, Figge, & Wehrmann, 2003; Figge, 2001, 2002; Figge, Schrott, Muntermann, & Rannenberg, 2002; Kunze, Zaplata, & Lamersdorf, 2006; Mallat, Rossi, & Tuunainen, 2004; Paavilainen, 2002; Rannenberg, 2004; Reichwald, 2002; Sheng, Nah, & Siau, 2005; Silberer, Wohlfahrt, & Wilhelm, 2001; Stender & Ritz, 2006; Tarasewich, 2003; Titkov, Poslad, & Tan, 2006).

There is a plethora of publications regarding typical benefits of mobile technologies and MB or MC. The subject has been approached both from a more technical point of view as well as from a business perspective. Therefore, we derive two classifications of typical benefits from available literature. The first classification focuses on the technical benefits of mobile technologies (see Table 1).

While these benefits are proven, they must be transformed into improvements of business performance in order to justify investment in MB technologies. Table 2 gives an overview of typical business process benefits which can be realized by leveraging technological benefits in alignment with business process goals and requirements.

Table 1. Classification of technical benefits of mobile technologies

Benefit	Definition	References
Ubiquity	Mobile technologies allow for IS to become accessible from virtually any place and at virtually any time.	Anckar & D'Incau, 2002a, 2002b; Balasubramanian, Peterson, & Jarvenpaa, 2002; Clarke III, 2001; Laukkanen, 2005; Laukkanen & Lauronen, 2005; Lehner, 2003, 11ff.; Pousttchi, Turowski, & Weizmann, 2003; Wohlfahrt, 2001
Context sensitivity	Mobile technologies allow for the contextualization of IS. The context may include the identification of the individual user as well as geographic position and physical environment.	Clarke III, 2001; Laukkanen, 2005; Laukkanen & Lauronen, 2005; Lehner, 2003, 11ff.; Pousttchi et al., 2003, 11ff.; Siau, Sheng, & Nah, 2004; Skelton & Chen, 2005; Wamser, 2003; Wohlfahrt, 2001
Interactivity	Mobile technologies allow for greater interactivity in IS, since they typically provide an "always online" connectivity and have shorter set-up times (e.g., for booting, "instant on").	Anckar & D'Incau, 2002a, 2002b; Clarke III, 2001; Hartmann & Dirksen, 2001; Laukkanen, 2005, 11ff.; Laukkanen & Lauronen, 2005; Lehner, 2003
Convenience and familiarity	For certain tasks, mobile technologies can offer a higher degree of convenience as compared to standard desktop or laptop PCs. This is partially due to limited functionality, thus reduced complexity and higher ease of use. For example, most users are capable of using most features of their cell phones (voice and text communication, address book, etc.), while most users only use a fraction of their PCs functionality.	Anckar & D'Incau, 2002a, 2002b; Gebauer, 2002; Gebauer & Shaw, 2004; Kenny & Marshall, 2000; Lehner, 2003, 11ff.; Perry, O'Hara, Sellen, Brown, & Harper, 2001; Siau, Sheng, & Nah, 2004; Van der Heijden & Valiente, 2002; Wohlfahrt, 2001
Multimediality	Mobile technologies have gained multimedia functionality over the years, for example, most cell phones shipped today include a digital camera, current models even with sufficient resolution for quality snapshots.	Han, Cho, & Choi, 2005; Kung, Hsu, Lin, & Liu, 2006; Pousttchi et al., 2003; Wamser, 2003; Wolf & Wang, 2005

In order to actually improve the business performance through IT investments, these investments and the expected benefits should be aligned with the business performance metrics and the overall strategy (Bakos & Treacy; 1986; Brynjolfsson, 1993; Brynjolfsson & Hitt, 2000; Hitt & Brynjolfsson, 1994; Kohli & Devaraj, 2004; Weill, 1992; Weill & Vitale, 2002; Weill et al., 2002). Despite the accepted importance of strategy alignment of IT investments there is only little research addressing the strategic aspects of applying MB (Amberg & Remus, 2003; Clarke III, 2001; Sadeh, 2002; Sheng et al., 2005; Wamser & Buschmann, 2006). Even fewer research addresses strategic potentials of *mobile business to businesses* whose core competencies are outside of the technology or base service field, such as financial service providers (Looney, Jessup, & Valacich, 2004).

Customer Relationship Management

The origins of CRM can be traced back to the management concept of Relationship Marketing (RM) (Levitt, 1983). RM is an integrated effort to identify, build up, and maintain a network with individual customers for the mutual benefit of both sides (Shani & Chalasani, 1992, p. 34). RM is of largely strategic character and lacks a holistic view on business processes, although they are regarded as important (Parvatiyar & Sheth, 2000).

Table 2. Classification of business benefits of mobile business

Benefit	Definition	References
Flexibility	The ubiquity and interactivity of MB applications allows for the break-up of process structures. Activities in processes, which were previously bound to location or time constraints, can now be dispatched more flexibly. Unforeseeable events can be responded to more flexibly and timely, since decision makers and action takers can be informed and immediately wherever they are and can be involved in the emergency response interactively.	Anckar & D'Incau, 2002a, 2002b; Fleisch, 2001; Fleisch & Bechmann, 2002; Fleisch, Mattern, & Österle, 2002; Gebauer, 2002; Gebauer & Shaw, 2004; Hartmann & Dirksen, 2001; Humpert & Habbel, 2002; Laukkanen, 2005; Laukkanen & Lauronen, 2005; Nah, Siau, & Sheng, 2004, 2005; Perry et al., 2001; Reichwald & Meier, 2002; Siau et al., 2004; Van der Heijden & Valiente, 2002; Wamser, 2003; Wohlfahrt, 2001
Organizational efficiency	The ubiquity and interactivity of MB applications allows for higher operational efficiency since the gaps between information's point of creation and its point of action can be bridged. For example, field agents can enter information electronically and directly to corporate IS, thus duplicate entry can be eliminated and back-end processing of the information can begin immediately. Information is available ubiquitously and immediately and can be used in geographically dispersed processes and activities.	Anckar & D'Incau, 2002a, 2002b; Fleisch & Bechmann, 2002; Fleisch et al., 2002; Gebauer, 2002; Gebauer & Shaw, 2004; Hartmann & Dirksen, 2001; Humpert & Habbel, 2002; Kadyte, 2005; Laukkanen, 2005; Laukkanen & Lauronen, 2005; Nah et al., 2004, 2005; Perry et al., 2001; Siau et al., 2004; Skelton & Chen, 2005; van der Heijden & Valiente, 2002; Wamser, 2003; Wohlfahrt, 2001
Individual productivity and effectiveness	Context sensitivity and interactivity as well as convenience and familiarity of MB applications allow for a greater level of effectiveness of business processes and a higher individual productivity. Interactive and context-sensitive offerings can increase the effectiveness of marketing campaigns. Interactive and ubiquitous control mechanisms can increase effectiveness of machines since they can send alerts in case of errors. Similarly individual productivity of employees can be increased since they can use waiting time more effectively (e.g., in airport terminals).	Anckar & D'Incau, 2002a, 2002b; Gebauer, 2002; Gebauer & Shaw, 2004; Kadyte, 2005; Nah et al., 2004, 2005; Perry et al., 2001; Siau et al., 2004; Skelton & Chen, 2005; Van der Heijden & Valiente, 2002; Wamser, 2003; Wohlfahrt, 2001
Transparency	Ubiquity and interactivity of MB processes allow for the increase of process transparency. This decreases costs for process control and customer satisfaction. Transparency of information can lead to higher market transparency and thus more efficient market mechanisms, for example, customers can compare prices online while in a retail store.	Chen, 2005; Kadyte, 2005; Laukkanen, 2005; Laukkanen & Lauronen, 2005; Reichwald & Meier, 2002; Wamser, 2003; Wohlfahrt, 2001
Entertainment	Especially multimediality but also ubiquity and interactivity increase the entertainment gained from MB applications. Entertainment content typically is multimedia-based in nature, thus entertainment devices need to be multimedia enabled. Additionally, mobilization of everyday life leads to more mobile and spontaneous entertainment needs.	Anckar & D'Incau, 2002a, 2002b; Dickinger, Arami, & Meyer, 2006; Han et al., 2005; Humpert & Habbel, 2002; Park, 2006; Reichwald & Meier, 2002; Wolf & Wang, 2005; Wong & Hiew, 2005

Advances in IT had a significant influence on CRM, focusing mainly on the IS layer in the past. The goal was to support the existing isolated approach of dealing with customer relationships. With the CRM philosophy aiming at creating an integrated view of the customer across the enterprise, these systems were connected and today form the building blocks of comprehensive integrated CRM systems.

We consider CRM to view the customer relationship as an investment, which is to contribute to the bottom line of the enterprise. The design and management of customer relationships is to strengthen the competitive position of an enterprise by increasing the loyalty of customers. While this extends beyond the use of IT, IT is still an important enabler of modern CRM.

Apart from the strategy-oriented concept of RM and systems oriented concepts, there are several CRM approaches with special focus on business processes (Schulze, Thiesse, Bach, & Österle, 2000). However, these approaches are based on the separation of the functional areas of marketing, sales, and service, which by itself does not provide a cross-functional process view.

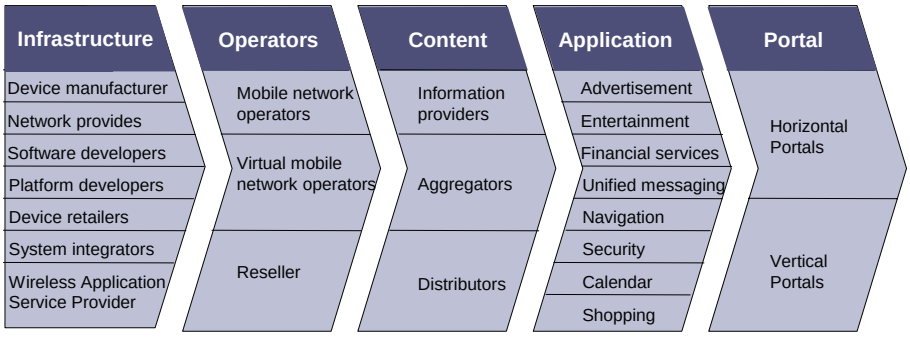
CRM processes typically require not only transactional data, which can be automatically collected and stored in relational databases, but also a significant amount of knowledge. Also, CRM processes are typically complex and only structured to a certain extent. Hence, they can be considered knowledge-intensive processes (Eppler, Seifried, & Röpnack, 1999). Besides developing an integrated view of CRM processes, it is therefore critical to address the management of knowledge flows from and to the customer across all communication channels as well as to enable the use of the knowledge about the customers.

Customer Focus in Mobile Business

When trying to achieve strategy alignment of mobile solutions in CRM, the analysis of the customer value proposition (i.e., the market strategy) is a crucial step. Some researchers have given guidance for a company's positioning in a mobile value chain, but guidance on how to analyze different customer value propositions and how to support these by mobile support for business processes still is lacking.

Zobel (2001) introduces a value chain of MC (see Figure 2). It begins with network infrastructure providers, providing, for example, IP infrastructure and devices. On the second stage, operators provide mobile infrastructure for such cellular networks. On the third stage, content is provided, for example, by news agencies, media owners, and so forth. On the fourth stage mobile applications (such as payment solutions, security solutions, etc.) are built on top of the network infrastructure and content. Players on this stage are service operators (e.g., mobile ticketing services), transaction clearing centers, and so forth. Finally

Figure 2. Mobile business value chain (Zobel, 2001, p. 122)



an interface to the MC is provided by mobile portal providers. A similar value chain can be found in Paavilainen (2002). For each stage, Paavilainen explains strategies and business models showing opportunities for market players.

Published MB value chains do not cover the user of the output of the value chain though. Works analyzing consumer value have mostly focused on mobile services provided via the mobile network operator. What is lacking in published research is a concept how businesses should go about using mobile solutions in customer processes, for example, companies offering mobile services such as banking or businesses using mobile sales support to improve and extend the way they are doing business. In these cases, the strategic options of the MB value chain are less important, because these businesses are primarily positioned in a different industry's value chain. For example, Lufthansa is primarily an airline and as such positions itself strategically in the airline industry's value chain rather than as a mobile content provider, even though part of its service offerings is mobile content for customers. Also, while still very important, consumer value of the services are not the only aspect in the decision process, because the strategic goals of the company offering the service provide the business background and thus have to be taken into account as well. There is little research published on how companies can support their strategic customer value proposition using mobile services. This is the focus of this chapter.

Strategic Frameworks

A substantial amount of research has been done in the field of business strategy or strategic management. While many of the foundational works were developed in the 1960s through the 1980s (Andrews, 1969; Ansoff, 1965, 1969; Anthony, 1965; Chandler, 1962, 1977; Mintzberg, 1980, 1987; Porter, 1979, 1996, 1998; Wernerfelt, 1984), it has been argued how far these frameworks are still helpful in facing today's business challenges. Cummings and Angwin (2004) analyze this in detail and come to the conclusion that the traditional frameworks are somewhat helpful still, but need to evolve to handle multi-dimensional today's strategies. Hartline, Maxham, and McKee, 2000; Rust, Inman, Jia, and Zahorik (1999); Rust, Zeithaml, and Lemon (2000); Rust, Lemon, and Zeithaml (2004); and Colgate and Danaher (2000) suggest a stronger focus on the customer. Luo and Seyedian (2003) go along with this argumentation suggesting a contextual marketing strategy.

Crawford and Mathews (2001) break the customer strategy concepts down to an operational applicability and describe the strategic aspects of customer focus for business strategy. In an empirical research addressing 5,000 American consumers, they find that consumers do not simply look for lowest prices, best products, and best services, but rather have a desire for more complex values (rather than simple value). They discovered the following five attributes which customers demand and successful companies focus on (Crawford & Mathews, 2001, p. 23):

- **Price:** Customers want a transparent, fair price (not necessarily the lowest price).
- **Customer service:** Customers want a hassle-free fulfillment of their basic needs and want to be recognized as individuals.
- **Accessibility:** Customers want simple access to the products, including clearly distinguished products and clear channels to interact with businesses.
- **Customer intimacy:** Customers want a unique experience in the interaction with businesses, that is, they want personalized offers and they want to be treated as a human individual.
- **Product quality:** Customers want an overall good product, not necessarily the single-one best product, but the best value for their money.

Crawford and Mathews (2001, p. 26) point out four levels to which a company can perform concerning these attributes. These range from level 0, where a customer avoids the company; level 1, where a customer trusts the company for everyday business; level 2, where a customer prefers the company over others to; level 3, where the customer only accepts this company, even if it means, for example, waiting for a product not yet shipped.

Crawford and Mathews (2001, p. 33) suggest that reaching level 3 in all attributes is unrealistic. They propose the successful strategy to be selecting one attribute to focus on to achieve level 3 (dominant position in the market), to focus a secondary attribute to reach level 2 (differentiating from competitors), and to maintain level 1 (market average) on the remaining three attributes.

In a different study analyzing multiple cases from market leaders in several branches (such as Casio, Kellogg's, FedEx) Treacy and Wiersema (1994) identify the following three main strategic goals (Treacy & Wiersema, 1994, p. 29):

- **Operational excellence:** Focus on the combination of quality, price, and ease of purchase without being exceptionally innovative in products or customer service.
- **Product leadership:** Focus on exceptional innovation in product features.
- **Customer intimacy:** Focus on the intimate one-to-one relationship to the customer.

Combining the results of Crawford and Mathews (2001) and Treacy and Wiersema (1994) it becomes obvious that the aspects of "Product Quality" (Crawford & Mathews) and "Product Leadership" (Treacy & Wiersema) as well as "Customer Service" and "Customer Intimacy" are almost identical in concept. Further more, the aspect of "Price" (Crawford & Mathews) is included by "Operational Excellence" (Treacy & Wiersema). Thus the list of attributes can be consolidated to:

- Price
- Customer intimacy
- Accessibility
- Product quality

STRATEGY ALIGNED MOBILE BUSINESS: INSIGHTS FROM THE REAL WORLD

In order to identify interdependencies between corporate market strategy (following the framework explained previously) and the selection of business processes to be supported by mobile solutions, we analyzed 10 cases. As mentioned in the second section, the cases have been published before and the analysis was mainly based on the published material. In the following each of the cases will be briefly summarized, pointing out the aspects of most relevance to our analysis. For further details about the cases, please refer to the original publications. For each of the cases, we briefly describe the company background, followed by the specific challenge that lead to the introduction of a mobile solution to support a business process. We also briefly describe the implemented solution and the characteristics of relevance for our analysis.

Helsana/Progrès: Mobile Marketing

Company Background

Helsana (http://www.helsana.ch) is the largest health insurance provider in Switzerland with about €2.5 million annual premium yield (2002). Its brand Progrès (http://www.progres.ch) represents affordable offerings for young customers. The brand strategy focuses on *maximum availability* and *competitive prices.*

Challenge

- There is only a short time frame for contract switching, thus high marketing efforts by all competitors overload the customers' perception.
- Customers show a high price sensitivity and low interest in the product itself thus they need a spontaneous and instant trigger.

Solution

Customers can retrieve an offer for Progrès insurance within 1-2 seconds via SMS, allowing Helsana to:

- Increase the visibility of its brand against the intensive activities of competitors
- Emphasize the innovative image of the Progrès brand
- Leverage the situational context once the customer is focused on the offer by other marketing activities (18,000 requested premiums resulted in 10,000 calls into the customer call center and 3,500 contract closures in 2003)

For further details about this case, see Reichold, Schierholz, Kolbe, and Brenner (2003) and Reichold, Schierholz, Kolbe, and Brenner (2004).

Gossard G4Me: Mobile Marketing

Company Background

Gossard (http://www.gossard.co.uk) is a manufacturer of lingerie products. Gossard's strategy focuses on *a strong, intimate bond* to customers in the market for string thongs by shifting the brand image towards the self-image of their primary target group: young, modern women and on *a premium, luxury product.*

Challenge

- To position the Gossard brand in a market for products with a strong personal bond requires an in-depth knowledge about customers.
- This conflicts with the goal of a non-invasive, opt-in, and privacy-preserving marketing campaign.

Solution

By launching marketing campaigns and providing giveaways such as coupons to respondents via SMS Gossard could build up a database with high-quality and detailed information about customers in their target group. Thus, Gossard has succeeded in:

- Gaining an in-depth understanding about their customers' desires
- Building a means to address customers for personal products on a very personal channel/medium such as a cellular phone
- Boosting the affectivity of traditional marketing campaigns such as TV spots with an interactive element, reaching an eight months sales target in just eight weeks

For further details about this case, see Lerner and Frank (2004, pp. 72-73).

Cologne Public Transport Authority: Mobile Sales

Company Background

The Cologne public transport authority is a publicly owned organization which runs the public transportation (buses, local trains, underground trains) in the agglomeration of Cologne, Germany and the surrounding suburbs. With the liberation of the market for public transport as required by

the European Union, it will have to compete for the contract, thus it has launched a campaign to improve their image as an *innovative service provider* with *a high customer service level.*

Challenge

- Apart from tickets being sold on subscription, still many tickets for public transport are sold individually, for example, via ticketing machines at bus stops or in trains.
- The image of the public transport authority has been reported as rather mixed. While the core service of public transport has been accepted as good, the level of customer service, for example, flexibility and innovativeness, have been reported as low.
- Providing high level, individual service to anonymous customers buying at ticketing machines is impossible.

Solution

A mobile ticket has been introduced by which customers can order a ticket for public transportation by simply calling a free 1-800 number. The ticket is delivered as a text message to their mobile within seconds. Customers need to register before they can use the service (except for one free trial ticket per mobile phone number). With this new system, it is possible:

- To allow discounts for customers who repeatedly buy single tickets (e.g., customers who by the third single ticket within one day, receive a full-day pass, and save about 20% of the fare)
- To create customer profiles and individualize services for customers
- To improve the Cologne public transport authority's image as an innovative service provider

eBay in Germany: Mobile Transactions

Company Background

eBay is probably the most well-known online auction platform in the world. Customers range from professional sellers (power sellers) to occasional private sellers to private buyers. eBay's customer value proposition focuses on *global reach, variety of traded items, efficient information services,* and *low trading fees and item prices.* eBay experiences impressive growth rates, in many figures constantly around and above 30%.

Challenge

- eBays biggest challenge currently is to maintain the large growth rates in a more and more saturated market. eBay tries to maintain this growth by acquiring new members (which is hard to achieve in almost saturated markets), activation of passive members and maximizing activity of active members. Currently, all member activity is dependent on the member's access to a Web-enabled PC, since eBay is a typical Web application.

Solution

In order to increase the reach of the platform and the activity of existing members, eBay introduced an SMS-based bidding process in Germany. Members who have placed the highest bid on an item can register for an alerting service, which notifies them when another member has placed a higher bid or when the auction is over and they won. Also, in response to the message about a higher bid, the recipient can place a new bid via SMS. With this new interaction channel, eBay members:

- Receive up-to-date information on auctions they personally participate in wherever they are
- Can respond to higher bids by other members and thus stay active in their auctions even when not in reach of Web-enabled PC

Eneco: Mobile Field Force

Company Background

Eneco (http://www.eneco.nl) is a Dutch energy supplier with about €2 billion annual turnover (2002). The corporate strategy aims to achieve customer loyalty by supplying a *high level of customer service* and a *reliable energy supply at affordable rates (i.e. in this case competitive price).*

Challenge

- Field force agents have no access to crucial information while in the field.
- Customers' issues cannot be resolved immediately due to lack of information.

Solution

Agents are provided with a PDA-based mobile application connecting them to the corporate IT via a mobile middleware module. This improves customer service by:

- Allowing for real-time processing of billing-relevant data collected by agent
- Better coordination of and information supply for agents on-site
- More visited customers per agent and a higher on-site solution rate

For further details about this case, see Lerner and Frank (2004, pp. 18-20).

SOS Médecins: Mobile Field Force

Company Background

SOS Médecins (http://www.sosmédecins.ch) is an initiative of more than 50 doctors in the Geneva region, providing medical treatment at home, in emergencies, and otherwise. SOS Médecins' strategy focuses on *best possible medical service (i.e., in this case product quality)* and *maximum availability.*

Challenge

- Doctors have no direct access to the patient's records, thus the need for time-consuming calls to the central office instead of treating patients
- The scheduling of doctors' routes proves inefficient due to lack of location information
- The travel routes of doctors prove inefficient due to lack of navigational support

Solution

By providing doctors with a PDA- GPRS-based solution, secured via VPN technology, SOS Médecins provides doctors with most current patient records and navigation support as well as optimized the scheduling efficiency and thus achieved to:

- Increase the time each doctor can effectively treat a patient
- Decrease the delay between the patient's call and the doctor's arrival
- Improve medical treatment itself

For further details about this case, see Lerner and Frank (2004, pp. 54-56).

Verizon: Mobile Sales Force

Company Background

Verizon (http://www.verizonwireless.com) is the largest wireless telecommunication provider in the U.S., serving about 39 million customers; generating an annual revenue of $22.5 billion (2003). Verizon's strategy aims to show its ability to *innovate* while providing *the best service* in consulting their business customers according to their particular, individual needs.

Challenge

- In an innovative field like MB, Verizon needs to demonstrate ability to deliver innovative solutions.
- To maintain the solutions' innovativeness Verizon needs to reduce time-to-market for its products and services as much as possible.

Solution

Verizon equipped its own sales force with mobile corporate data access, for example, to its CRM application, and thereby:

- Improved customer service and consulting due to better and more proactive information availability for sales agents
- Improved its visibility as an innovator by demonstrating wireless solution know-how on-site

For further details about this case, see Lerner and Frank (2004, pp. 36-37).

Novartis: Mobile Info Services

Company Background

Novartis (http://www.novartis.ch, http://www.novartis.co.uk) is a Swiss pharmaceutical manufacturer with $24.8 billion annual turnover and $5 billion annual profit (2003). Novartis' Consumer Health business unit positions itself as an innovative company having a *positive impact on people's lives (i.e., customer intimacy),* making *available the right information at the right time (i.e., accessibility).*

Challenge

- Novartis wants its brand to be seen as a partner helping lower the burden of allergies in everyday life.
- The "Aller-eze" product should be seen as the main product in the anti-allergy (especially hay fever) market, that is, customers' creating the association between the two intuitively.

Solution

To introduce the new anti-allergic product "Aller-eze" Novartis' British affiliate launched a mobile marketing initiative. By offering a subscription service providing patients with timely, location-specific allergy warnings and hints for patients, Novartis succeeded in:

- "Aller-eze" being perceived as a partner providing daily support, easing the pain of allergy patients
- Emphasizing the innovative image of Novartis as a whole

For further details about this case, see Lerner and Frank (2004, pp. 74-75).

Lufthansa: Mobile Info Services

Company Background

Lufthansa is one of the largest airlines and a founding member of the star alliance, the largest network of cooperating airlines in the world. It considers itself a *full-range service provider addressing all customer segments* with different product variations. Price-sensitive customers can book *cheap rates with low service level*, business customers can book *flexible rates with high service level* and luxury customers are treated with *exclusive service*. Lufthansa also cultivates an innovative image, for example, by adopting new technologies early.

Challenge

- Events such as delays or cancellations create new information which is viable for customers, which are traveling and therefore mobile by nature.
- Cost pressure in the airline industry in general requires the airline to streamline processes and to raise operational efficiency.
- Procedures such as check-in are time critical, since there is only a short time frame for the handling. Additionally, especially business customers are typically in a hurry and appreciate shorter handling procedures.

Solution

Lufthansa offers multiple mobile services, with different service levels for different customer segments. Services include general information such as timetables (available freely for customers and non-customers), flight-related information such as alerts about delays and gate changes (available to all customers), and mobile check-in service (available to premium customers only). With these services, Lufthansa was able:

- To further differentiate the service levels for different customer segments, for example, allow for a more flexible check-in procedure for premium customers
- Streamline the customer handling procedures since customers are informed about flight-related events proactively instead of having to request or search the information
- To demonstrate its innovation capabilities by leveraging a new technology earlier than most competitors

For further details about this case, see Schierholz, Glissmann, Kolbe, and Brenner (2006).

Lotto NL: Mobile Gaming

Company Background

Lotto NL (www.lotto.nl) is a publicly owned lottery service in the Netherlands. Its strategy is to *cover the entire market* with a mix of service channels, *providing anywhere-anytime access to lottery services.*

Challenge

- A high market saturation in traditional channels requires an exploitation of new market segments/channels to allow for revenue growth.

Solution

By allowing customers to take part in lottery games via SMS Lotto NL:

- Lowers customers' efforts to take part in lottery games and thus engages more customers
- Leverages and enhances the traditional marketing activities such as TV spots by providing instant access to its services

For further details about this case, see Lerner and Frank (2004, pp. 34-35).

STRATEGY'S IMPLICATIONS FOR PROCESS SELECTION AND DESIGN

Looking at the cases of Verizon, Helsana/Progrès, Cologne Public Transportation Authority, Gossard, and Lufthansa, we identified an attribute to include in addition to the ones listed in the framework as introduced in the second section. These cases show that especially new technologies (such as currently mobile technology) can be a means for businesses to demonstrate innovative capabilities. In current literature, we also found approaches, suggesting innovativeness to be an important strategic attribute of market strategy (Kim & Mauborgne, 1999; Micheal, Rochford, & Wotruba, 2003). Thus, we added the attribute of innovativeness which can be supported by the application of MB technologies. This results in the following common strategic focus attributes/goals which make up the strategic framework used for our analysis:

- **Price:** Offering low, transparent, and fair prices compared to the market
- **Customer intimacy:** Offering hassle-free service on a personal level, establishing a one-to-one relationship with customers
- **Product quality:** Offering the best product features in the market
- **Accessibility:** Offering simple, anytime-anywhere-anyhow access to products
- **Innovativeness:** Being perceived as an innovator or early-adopter of new, innovative technologies

Table 3 summarizes the prioritization of these attributes across the analyzed cases; applying the classification from Crawford and Mathews (2001) of primary focus, secondary focus, and no

Table 3. Overview of strategic focus in the analyzed cases

	Price	Intimacy	Product	Accessibility	Innovation
Progrès	●	⦸	⦸	○	●
Gossard G4Me	⦸	○	●	⦸	●
Cologne PTA	⦸	○	⦸	⦸	●
eBay in Germany	●	⦸	⦸	○	●
Eneco	●	○	⦸	⦸	⦸
SOS Médcins	⦸	⦸	○	●	⦸
Verizon	⦸	●	⦸	⦸	○
Novartis	⦸	○	⦸	●	⦸
Lufthansa	⦸	○	⦸	⦸	●
LottoNL	⦸	●	⦸	○	⦸

Legend: ○ = primary focus ● = secondary focus ⦸ = no focus (market average)

focus (i.e., the company pursues market average performance).

Obviously, the strategic framework as introduced in the second section and which has been extended here can be used to classify the mobile initiatives in the analyzed cases. The following common aspects of the selected customer-oriented processes, depending on the strategic orientation, can be observed.

Comparing the correlations between strategic focus and process selections for mobile support in the case studies, the following observations can be made. Companies focusing primarily on price were not found in the sample, companies who focus on price as the second distinguishing attribute have chosen processes where either process steps could be eliminated by removing media breaches (Eneco) or processes where mobile technology allows for better price transparency and the communication of the price on an individual basis (Progrès). Most companies have the focus on customer intimacy. These companies have chosen processes where mobile technology allows for customer support in spontaneous or emergency situations (eBay, Novartis, Lufthansa) or processes where contextualization (mostly personalization) allowed for a convincing one-to-one interaction (Gossard, Cologne PTA). Only SOS Médcins focuses on product quality, their product is a service offering which is improved by better information support to mobile service agents (doctors). Companies who focus on accessibility have used mobile technology to offer their customers a direct and interactive interaction channel, either for requesting/receiving information or individual offers (Progrès, eBay, Novartis) or to order services or products or perform other transactions (eBay, LottoNL). Companies who focus on innovation have chosen processes which have a high external visibility and which occur frequently (Progrès, Cologne PTA, eBay, Verizon, Lufthansa).

Strategic Focus on Price

A strategy focused on competitive and transparent price can be supported, when business processes are mobilized in which information is passed on and the point of creation (PoC) and the point of action (PoA) of the information differ. For example, Eneco could raise operational efficiency, and thus lower operational costs of their field force by supporting them with mobile devices, which were connected and integrated with Eneco's billing system and other IS which provide them with information to help them in solving customer incidents on-site.

Strategic Focus on Customer Intimacy

A strategy focused on customer intimacy and the best customer experience can be supported when business processes are mobilized that support the customer in spontaneous situations or where anonymity can be overcome by using a personalized mobile medium. Here a customer experiences that the company is there to help and provide knowledge for the customer when he/she needs it. Also, the generation of personalized knowledge about the customer can be well supported by mobile solutions, since the devices used in such solutions, such as cellular phones, usually have a strong personal touch. For example, Novartis could support its customers with crucial information, personalized to the location and allergic profile of each individual customer and the Cologne public transport authority could provide personalized discounts on repeated purchases.

Strategic Focus on Accessibility

A strategy focused on accessibility is probably the most obvious one to be supported by mobile technology, even though again the support of physical products seems to be hard. A strategy focusing on accessibility should leverage mobile technology's

potential to extend the communication channels the customer can use to obtain a service from a business to location- and time-independent media such as cellular phones. For example, Lotto NL could extend its reach to occasional gamblers, who were not taking part in the lottery because of the burden of having to obtain a lottery ticket from a store or from an Internet-connected PC. By offering the purchase of lottery tickets via cellular phones, customers can now purchase lottery tickets anytime (i.e., independent from office hours of points of sale), anywhere (i.e., independent of where the next point of sale is located at), and anyhow (i.e., it is the customer's choice via which channel to purchase).

Strategic Focus on Innovativeness

A strategy focusing on the demonstration of innovativeness of a company can be supported by MB at least nowadays, when mobile technology has not yet become a commodity. To support an innovative image of a company the processes obviously have to be externally visible to have an impact on the company's image. For example, Verizon chose the sales agents because they have immediate customer contact and thus can best show Verizon's ability to put innovative products into operational use. Also, addressing young customer market segments can be well supported with innovative marketing based on technologies considered "cool" by these customers as the cases Helsana/Progrès and Gossard illustrate.

Strategic Focus on Product Quality

The analyzed cases indicate that a strategy focused on quality of product is hard to support unless the product is either closely related to mobile technology or is a knowledge-intensive service product. For example, Verizon could support the product quality and how this quality is perceived by customers by showing its products and services on-site via its own sales agents. The knowledge aspect played an important role in the case of SOS Médecins, where the product of medical service has been greatly enhanced by providing the doctor with complete and current knowledge about the visited patient.

OUTLOOK

The MB industry will mature further. On the one hand, the technological evolution will bring about more sophisticated devices and networks, which allow for more sophisticated applications and services. On the other hand, the market will likely experience a shakeout leading to more clearly distinguished roles in the value chain. Currently, especially in Western Europe, many mobile network operators try to control the entire value chain, leaving little room for other partners, claiming large parts of the profit potential and thus rendering the applications of MC and MB relatively unattractive. Price competition and a maturing market will likely cure this phenomenon.

CONCLUSION AND FURTHER RESEARCH

The analysis presented in this chapter shows how the alignment of the use of MB technology with corporate strategy can be achieved, with special respect to business processes in customer interaction. We have identified five different strategic focuses and explain which criteria the processes should fulfill to provide the best support to the corporate strategy when being mobilized, thus promise to realize their full potential (i.e., the best ROI of the related IT investments).

Since the analysis so far is based on 10 cases, which are not representative for a general target audience, the framework should be further validated by further cases studies and quantitative empirical research. Other aspects that should be addressed by further research include a detailed

method for process selection, business process redesign, and technology selection to provide businesses with a structured method on how to achieve best effects with the application of MB technology.

REFERENCES

Ali, S., Torabi, T., & Ali, H. (2006). A case for business process deployment for location aware applications. *International Journal of Computer Science and Network Security, 6*(8a), 118-127.

Amberg, M., Figge, S., & Wehrmann, J. (2003). A cooperation model for personalised and situation dependent services in mobile networks. In A. Olivé, M. Yoshikawa, & E. S. Yu (Eds.), *Advanced conceptual modeling techniques* (2784 ed.). Berlin, Germany: Springer.

Amberg, M., & Remus, U. (2003). Multi-channel commerce: Hybridstrategien und controlling. In W. Uhr, W. Esswein, & E. Schoop (Eds.), *Wirtschaftsinformatik 2003/Band II* (pp. 795-817). Heidelberg, Germany: Physica-Verlag.

Anckar, B., & D'Incau, D. (2002a). Value creation in mobile commerce: Findings from a consumer survey. *Journal of Information Technology Theory and Application, 4*(1), 43-64.

Anckar, B., & D'Incau, D. (2002b, January 7-10). *Value-added services in mobile commerce: An analytical framework and empirical findings from a national consumer survey.* Paper presented at the 35th Annual Hawaii International Conference on System Sciences, Big Island.

Andrews, K. (1969). Toward professionalism in business management. *Harvard Business Review, 47*(2), 49-60.

Ansoff, I. (1965). *Corporate strategy: An analytical approach to business policy for growth and expansion.* New York: McGraw-Hill.

Ansoff, I. (Ed.). (1969). *Business strategy.* Harmondsworth, UK: Penguin.

Anthony, R. N. (1965). *Planning and control systems.* Boston: Harvard University Press.

Bakos, J. Y., & Treacy, M. E. (1986). Information technology and corporate strategy: A research perspective. *MIS Quarterly, 10*(2), 106-120.

Balasubramanian, S., Peterson, R. A., & Jarvenpaa, S. L. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science, 30*(4), 348-362.

Brynjolfsson, E. (1993). The productivity paradox of information technology: Review and assessment. *Communication of the ACM, 12*(36), 66-77.

Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives, 14*(4), 23-48.

Chandler, A. (1962). *Strategy and structure.* Boston: Harvard University Press.

Chandler, A. (1977). *The visible hand.* Boston: Harvard University Press.

Chen, M. (2005). A methodology for building mobile computing applications: Business strategy and technical architecture. *International Journal of Electronic Business, 2*(3), 229-243.

Christensen, C. M. (1997). *The innovator's dilemma: When new technologies cause great firms to fail.* Boston: Harvard Business School Press.

Clarke III, I. (2001). Emerging value propositions for m-commerce. *Journal of Business Strategies, 18*(2), 133-148.

Colgate, M. R., & Danaher, P. J. (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent

execution. *Journal of the Academy of Marketing Science, 28*(3), 375-387.

Crawford, R., & Mathews, R. (2001). *The myth of excellence: Why great companies never try to be the best at everything.* New York: Crown Business.

Cummings, S., & Angwin, D. (2004). The future shape of strategy: Lemmings or chimeras? *Academy of Management Executive, 18*(2), 21-36.

Dickinger, A., Arami, M., & Meyer, D. (2006, January 4-7). *Reconsidering the adoption process: Enjoyment and social norms—Antecedents of Hedonic mobile technology use.* Paper presented at the 39th Hawaii International Conference on System Sciences, Kaua'i.

Dous, M., Salomann, H., Kolbe, L., & Brenner, W. (2004). *CRM—Status quo und zukünftige Entwicklungen*. Switzerland: Universität St. Gallen.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532-550.

Eppler, M., Seifried, P., & Röpnack, A. (1999, April 8). *Improving knowledge intensive processes through an enterprise knowledge medium*. Paper presented at the ACM SIGCPR conference on computer personnel research.

Feldman, S. (2000). Mobile commerce for the masses. *IEEE Internet Computing, 4*(6), 74-75.

Figge, S. (2001, May 3). *Situation dependent m-commerce applications.* Paper presented at the Conference on Telecommunications and Information Markets—COTIM, Providence, RI.

Figge, S. (2002). Die open mobile architecture—Systemumgebung für mobile Dienste der nächsten Generation. *Wirtschaftsinformatik, 44*(4), 375-378.

Figge, S., Schrott, G., Muntermann, J., & Rannenberg, K. (2002, June 16). *EARNING M-ONEY—A situation based approach for mobile business models.* Paper presented at the European Conference on Information Systems, Neapel, Italy.

Fleisch, E. (2001). *Business perspectives on ubiquitous computing.* St. Gallen, Switzerland.

Fleisch, E., & Bechmann, T. (2002). *Ubiquitous computing: Wie "intelligente Dinge" die Assekuranz verändern*. St. Gallen, Switzerland.

Fleisch, E., Mattern, F., & Österle, H. (2002). *Betriebliche anwendungen mobiler technologien: Ubiquitous commerce*. St. Gallen, Switerzland.

Fritsch, L., & Rossnagel, H. (2004). SIM-based mobile electronic signatures: Enabling m-business with certification on demand. *Card Forum International, 8*(1), 38-40.

Funk, J. L. (2003). *Mobile disruption: The technologies and applications driving the obile Internet.* Hoboken, NJ: Jon Wiley & Sons.

Gebauer, J. (2002, June 17). *Assessing the value of emerging technologies: The case of mobile technologies to enhance business to business applications.* Paper presented at the 15th Bled Electronic Commerce Conference, Bled, Slovenia.

Gebauer, J., & Shaw, M. J. (2004). Success factors and impacts of mobile business applications: Results from a mobile e-procurement study. *International Journal of Electronic Commerce, 8*(3), 19-41.

Gebert, H., Geib, M., Kolbe, L. M., & Brenner, W. (2003). Knowledge-enabled customer relationship management. *Journal of Knowledge Management, 7*(5), 107-123.

Geib, M., Reichold, A., Kolbe, L. M., & Brenner, W. (2005, January 3). *Architecture for customer relationship management approaches in financial services,* Big Island, HI.

Han, S.-Y., Cho, M.-K., & Choi, M.-K. (2005, July 11). *Ubitem: A framework for interactive marketing in location-based gaming environment.*

Paper presented at the Proceedings of the Fourth International Conference on Mobile Business (mBusiness), Sydney, Australia.

Hartline, M. D., Maxham, J. G., & McKee, D. O. (2000). Corridors of influence in the dissemination of customer-oriented strategy to customer contact service employees. *Journal of Marketing, 64*(2), 35-50.

Hartmann, S., & Dirksen, V. (2001). Effizienzsteigerungen von unternehmensinternen Prozessen durch die Integration von Komponenten des M-Business. *Information Management & Consulting, 16*(2), 16-19.

Hitt, L., & Brynjolfsson, E. (1994, December). *Creating value and destroying profits? Three measures of information technology's contribution.*

Humpert, F., & Habbel, F.-R. (2002). Mobile Dienste für die Öffentlichkeit. *HMD Praxis der Wirtschaftsinformatik, 226*, 37-43.

Kadyte, V. (2005, July 11). *Process visibility: How mobile technology can enhance business-customer care in the paper industry.* Paper presented at the Proceedings of the Fourth International Conference on Mobile Business (mBusiness), Sydney, Australia.

Kenny, D., & Marshall, J. F. (2000). Contextual marketing—The real business of the Internet. *Harvard Business Review, 78*(6), 119-125.

Kim, W. C., & Mauborgne, R. (1999). Creating new market space. *Harvard Business Review, 77*(1), 83-93.

Kohli, R., & Devaraj, S. (2004). Realizing the business benefits of information technology investments: An organizational process. *Misqe, 3*(2), 53-68.

Kung, H.-Y., Hsu, C.-Y., Lin, M.-H., & Liu, C.-N. (2006). Mobile multimedia medical system: Design and implementation. *International Journal of Mobile Communications, 4*(5), 595-620.

Kunze, C. P., Zaplata, S., & Lamersdorf, W. (2006, June 13). *Mobile process description and execution.* Paper presented at the 6th IFIP WG 6.1 International Conference on Distributed Applications and Interoperable Systems, Bologna, Itlay.

Laukkanen, T. (2005, July 11). *Comparing consumer value creation in Internet and mobile banking.* Paper presented at the Proceedings of the Fourth International Conference on Mobile Business (mBusiness), Sydney, Australia.

Laukkanen, T., & Lauronen, J. (2005). Consumer value creation in mobile banking services. *International Journal of Mobile Communications, 3*(4), 325-338.

Lehner, F. (2003). *Mobile und drahtlose Informationssysteme: Technologien, Anwendungen, Märkte.* Berlin, Germany: Springer.

Lerner, T., & Frank, V. (2004). *Best practices mobile business* (2nd ed.). BusinessVillage.de.

Levitt, T. (1983). After the sale is over... Author(s): Source: ; Sep/Oct83, Vol. 61 Issue 5, p87, 7p. *Harvard Business Review, 61*(5), 87-94.

Liang, T.-P., & Wei, C.-P. (2004). Introduction to the special issue: Mobile commerce applications. *International Journal of Electronic Commerce, 8*(3), 7-17.

Looney, C. A., Jessup, L. M., & Valacich, J. S. (2004). Emerging business models for mobile brokerage services. *Communications of the ACM, 47*(6), 71-77.

Luo, X., & Seyedian, M. (2003). Contextual marketing and customer-orientation strategy for e-commerce: An empirical analysis. *International Journal of Electronic Commerce, 8*(2), 95-118.

Mallat, N., Rossi, M., & Tuunainen, V. K. (2004). Mobile banking services. *Communications of the ACM, 47*(5), 42-46.

Micheal, K., Rochford, L., & Wotruba, T. R. (2003). How new product introductions affect sales management strategy: The impact of type of "newness" of the new product. *Journal of Product Innovation Management, 20*(4), 270-283.

Mintzberg, H. (1980). Structure in 5's: A synthesis of the research on organization design. *Management Science, 26*(3), 322-341.

Mintzberg, H. (1987). The strategy concept II: Another look at why organizations need strategies. *California Management Review, 30*(3), 11-24.

Möhlenbruch, D., & Schmieder, U.-M. (2001). Gestaltungsmöglichkeiten und Entwicklungspotenziale des Mobile Marketing. *HMD Praxis der Wirtschaftsinformatik, 220,* 15-26.

Mylonopoulos, N. A., & Doukidis, G. I. (2003). Mobile business: Technological pluralism, social assimilation, and growth [Special issue]. *Internation Journal of Electronic Commerce, 8*(1), 5-22.

Nah, F. F.-H., Siau, K., & Sheng, H. (2004). *Values of mobile technology in education.* Paper presented at the Tenth Americas Conference on Information Systems, New York.

Nah, F. F.-H., Siau, K., & Sheng, H. (2005). The value of mobile applications: A utility company study. *Communications of the ACM, 48*(2), 85-90.

Paavilainen, J. (2002). *Mobile business strategies: Understanding the technologies and opportunities.* London: Addison-Wesley.

Park, C. (2006). Hedonic and utilitarian values of mobile Internet in Korea. *International Journal of Mobile Communications, 4*(5), 497-508.

Parvatiyar, A., & Sheth, J. N. (2000). The domain and conceptual foundations of relationship marketing. In J. N. Sheth & A. Parvatiyar (Eds.), *Handbook of relationship marketing* (pp. 3-38). Thousand Oaks: Sage.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: Understanding access anytime, anywhere. *Transactions on Computer-Human Interaction, 8*(4), 323-247.

Porter, M. E. (1979). *How competitive forces shape strategy.* Boston: Harvard Business School Press.

Porter, M. E. (1996). What is strategy? *Harvard Business Review, 74*(6), 61-78.

Porter, M. E. (1998). *Competitive advantage: Creating and sustaining superior performance*. New York: Free Press.

Pousttchi, K., Turowski, K., & Weizmann, M. (2003). *Added value-based approach to analyze electronic commerce and mobile commerce business models.* Paper presented at the International Conference of Management and Technology in the New Enterprise, La Habana, Cuba.

Raisinghani, M. (2002). Mobile commerce: Transforming the vision into reality. *Information Resources Management Journal, 15*(2), 3-4.

Rannenberg, K. (2004). Identity management in mobile cellular networks and related applications. *Information Security Technical Report, 9*(1), 77-85.

Reichold, A., Schierholz, R., Kolbe, L. M., & Brenner, W. (2003, September 1). *M-Commerce at Helsana health insurance: Mobile premium calculator.* Paper presented at the Dexa '03, Prag.

Reichold, A., Schierholz, R., Kolbe, L., & Brenner, W. (2004). Mobile-commerce bei der Helsana: Mobile Prämienerstellung. In K. Wilde & H. Hippner (Eds.), *Management von CRM-Projekten*: Gabler.

Reichwald, R. (2002). *Mobile Kommunikation: Wertschöpfung, Technologien, neue Dienste*. Wiesbaden: Gabler.

Reichwald, R., & Meier, R. (2002). Generierung von Kundenwert mit mobilen Diensten. In R. Reichwald (Ed.), *Mobile Kommunikation—Wertschöpfung, Technologien, neue Dienste* (pp. 207-230). Wiesbaden, Germany: Gabler.

Romano, N. C., & Fjermestad, J. (2002). Electronic commerce customer relationship management: An assessment of research. *International Journal of Electronic Commerce, 6*(2), 61-113.

Romano, N. C., & Fjermestad, J. (2003). Electronic commerce customer relationship management: A research agenda. *Information Technology and Management, 4*(2-3), 233-258.

Rust, R. T., Inman, J. J., Jia, J., & Zahorik, A. (1999). What you don't know about customer-perceived quality: The role of customer expectation distributions. *Marketing Science, 18*(1), 77-92.

Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing, 68*(1), 109-127.

Rust, R. T., Zeithaml, V. A., & Lemon, K. N. (2000). *Driving customer equity: How customer lifetime value is reshaping customer strategy.* New York: The Free Press.

Sadeh, N. (2002). *M-Commerce: Technologies, services, and business models.* New York: John Wiley & Sons.

Schierholz, R., Glissmann, S., Kolbe, L. M., & Brenner, W. (2006, July 6). *Mobile systems for customer service differentiation—The case of Lufthansa*. Paper presented at the 10th Pacific Asia Conference on Information Systems, Kuala Lumpur, Malaysia.

Schierholz, R., Kolbe, L. M., & Brenner, W. (in press). Mobile customer relationship management: Foundations, challenges and solutions. *Business Process Management Journal.*

Schulze, J., Thiesse, F., Bach, V., & Österle, H. (2000). Knowledge enabled customer relationship management. In H. Österle, E. Fleisch, & R. Alt (Eds.), *Business networking: Shaping enterprise relationships on the Internet* (pp. 143-160). Berlin, Germany: Springer.

Shani, D., & Chalasani, S. (1992). Exploiting niches using relationship marketing. *The Journal of Consumer Marketing, 9*(3), 33-42.

Sheng, H., Nah, F. F.-H., & Siau, K. (2005). Strategic implications of mobile technology: A case study using value-focused thinking. *The Journal of Strategic Information Systems, 14*(3), 262-190.

Siau, K., Sheng, H., & Nah, F. F.-H. (2004). *Value of mobile commerce to customers.* Paper presented at the Tenth Americas Conference on Information Systems, New York.

Silberer, G., Wohlfahrt, J., & Wilhelm, T. (Eds.). (2001). *Mobile commerce. Grundlagen, Geschäftsmodelle, Erfolgsfaktoren*. Wiesbaden, Germany: Gabler Verlag.

Skelton, G. W., & Chen, L.-d. (2005). Introduction to m-business applications: Value proposition, applications, technologies and challenges. In G. W. Skelton & L.-d. Chen (Eds.), *Mobile commerce application development* (pp. 1-21). Hershey, PA: Idea Group Inc.

Stafford, T. F., & Gilleson, M. L. (2003). Mobile commerce: What it is and what it could be. *Communications of the ACM, 46*(12), 33-34.

Stake, R. E. (1995). *The art of case study research*. London: Sage.

Stender, M., & Ritz, T. (2006). Modeling of B2B mobile commerce processes. *International Journal of Production Economics, 101*(1), 128-139.

Tarasewich, P. (2003). Designing mobile commerce applications. *Communications of the ACM, 46*(12), 57-60.

Titkov, L., Poslad, S., & Tan, J. J. (2006). An integrated approach to user-centered privacy for mobile information services. *Applied Artificial Intelligence, 20*(2-4), 159-178.

Treacy, M., & Wiersema, F. (1994). *The discipline of market leaders.* Reading: Addison-Wesley.

Turowski, K., & Pousttchi, K. (2003). *Mobile commerce: Grundlagen und Techniken*. Berlin, Germany: Springer.

Urbaczewski, A., Valacich, J. S., & Jessup, L. M. (2003). Mobile commerce—Opportunities and challenges. *Communications of the ACM, 46*(12), 31-32.

Van der Heijden, H., & Valiente, P. (2002, June 6-8). *The value of mobility for business process performance: Evidence from Sweden and The Netherlands.* Paper presented at the European Conference on Information Systems, Gdansk, Poland.

Wamser, C. (2003). Die wetbbwerbsstrategischen Stoßrichtungen des Mobile Commerce. In J. Link (Ed.), *Mobile commerce* (pp. 65-93). Berlin, Germany: Springer.

Wamser, C., & Buschmann, D. (2006). *Erfolgsfaktoren des Mobile Business.* Rheinbach: Deutsche Gesellschaft für Management Forschung (DMGF).

Weill, P. (1992). The relationship between investment in information technology and firm performance: A study of the valve manufacturing sector. *Information Systems Research, 3*(4), 307-333.

Weill, P., & Broadbent, M. (1998). *Leveraging the new infrastructure—How market leaders capitalize on information technology.* Boston: Harvard Business School Press.

Weill, P., Subramani, M., & Broadbent, M. (2002). Building IT infrastructure for strategic agility. *MIT Sloan Management Review, 44*(1), 57-65.

Weill, P., & Vitale, M. (2002). What IT infrastructure capabilities are needed to implement e-business models. *MIS Quarterly Executive, 1*(1), 17-34.

Wernerfelt, B. (1984). A resource based view of the firm. *Strategic Management Journal, 5*(2), 171-180.

Wohlfahrt, J. (2001). One-to-one marketing in mobile commerce. *Information Management & Consulting, 16*(2), 49-54.

Wolf, H., & Wang, M. (2005, July 11-13). *A framework with a peer fostering mechanism for mobile P2P game.* Paper presented at the Proceedings of the Fourth International Conference on Mobile Business (mBusiness), Sydney, Australia.

Wong, C. C., & Hiew, P. L. (2005, July 11-13). *Mobile entertainment: Review and refine.* Paper presented at the Proceedings of the Fourth International Conference on Mobile Business (mBusiness), Sydney, Australia.

Yin, R. K. (2002). *Case study research. Design and methods* (Vol. 5, 3rd ed.). London: Sage.

Zobel, J. (2001). *Mobile business und m-commerce—Die Märkte der Zukunft erobern*. München, Germany: Hanser.

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, & J. Day, pp. 88-110, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.10
Universal Approach to Mobile Payments

Stamatis Karnouskos
Fraunhofer Institute FOKUS, Germany

András Vilmos
SafePay Systems, Ltd., Hungary

INTRODUCTION

An old saying coming from the telecom world states that nothing can be really considered as a service unless you are able to charge for it. The last several years have seen a boom in interest in mobile commerce, mainly due to the high penetration rates of mobile phones. Furthermore, there is evident the need for a real-time, open, and trusted payment service that can be used any time, anywhere, and that can handle any transaction in any currency. Such a service would promote not only content creating activities but would empower the electronic and mobile commerce area and kick-start new innovative services. The time is right for such a mobile payment service, because the infrastructure, the business models, and other conditions that favor its existence are realistic and in place (Vilmos & Karnouskos, 2004). Up to now, we have witnessed the rise and fall of several efforts in the area, ranging from realizing simple intangible good purchases, up to interaction with real points of sale (POS) and person-to-person (P2P) transactions. Day by day, new trials are initiated, targeting different sections in the MP area; however, there is still no solution that is open and widely accepted. In this article, we first introduce the reader to the mobile payment area, present the guiding forces behind it, and subsequently examine such an open, secure mobile payment approach that has been successfully designed, implemented, and tested. Furthermore we identify some midterm future trends that we consider will be of high importance to the further development of the area.

BACKGROUND

Payments are the locomotive behind the business domain and heavily depend on trust and security. A global study by Little (2004) estimated that

m-payment transaction revenues would increase from $3.2 billion in 2003, to $11.7 billion in 2005, and to $37.1 billion in 2008 world wide. Mobile payments are seen as the natural evolution of existing e-payment schemes that will complement them (Heng, 2004). The increasingly popular ownership of mobile personal, programmable communication devices worldwide promises an extended use of them in the purchase of goods and services in the years to come (Mobey Forum, 2003). Security in payment transactions and user convenience are the two main motivation reasons for using mobile devices for payments.

The context of mobile payments can be defined as follows: Any payment where a mobile device is used in order to initiate, activate and/or confirm this payment can be considered as a mobile payment. A mobile payment solution can be used in multiple applications and scenarios. The simplest scenario involves only the user, the device and a single payment processor, such as a mobile operator, bank, broker, or an insurance company. The user identifies himself or herself to the mobile device through secure identification mechanisms, including physical possession and password or even via biometric methods; the device then authorizes the transaction to the payment processor for the money transfer. More complex transactions involve at least one additional party, the merchant. In this case, the merchant may be affiliated with a different payment processor; therefore the two payment processors must be able to interoperate.

Based on the amount to be paid we can have different categorization of mobile payments. Generally we have:

- Micropayments: These are the lowest values, typically under $2. Micropayments are expected to boost mobile commerce as well as pay-per-view/click charging schemas.
- Minipayments: These are payments between $2 and $20. This targets the purchase of everyday's small things.
- Macropayments: These payments are typically over $20.

Currently, there are several efforts at the international level to accelerate and solidly support emerging mobile payment solutions. Most of the heavyweight companies that deal with hardware or software products for the mobile market and companies such as the mobile network operators (MNO) and financial service providers try via international fora and consortia to define the guidelines to which such a system should comply. The aim is to produce an approach that is widely acceptable and that would reach a global audience and not address just a specific customer base or isolated scenario. Towards this end, several consortia have aroused such as Simpay (www.simpay.com—ceased operation in summer 2005), Starmap Mobile Alliance, Mobey Forum (www.mobeyforum.org), Mobile Payment Forum (www.mobilepaymentforum.org), Mobile Payment Association (mpa.ami.cz), Paycircle (www.paycircle.org), Mobile electronic Transactions (www.mobiletransaction.org), and so forth. Apart from these "pure" mobile payment consortia, whose work directly affects the mobile payments, there are also other actors that indirectly are evolved with the mobile payment area and come from the financial/banking sector. Karnouskos (2004) provides an overview of these consortia.

For mobile payments to succeed, several requirements need to be addressed. Simplicity and usability largely determines whether users will use a service. This includes not only a user-friendly interface but also the whole range of goods and services one can purchase, the geographical availability of the service, and the level of risk the user is taking while using it. A promising mobile payment service should be offered widely and in a transparent fashion covering the biggest range of mobile payment transactions such as person to person (P2P), business to consumer (B2C), and business to business (B2B), domestic, regional and global coverage, low- and high-value payments. It

should be based on open standards that will allow it to interact with other systems and easily scale. It should also be secure by means of technology and processes, and preferably be built on existing trust relationships. The new systems should be, at the end, more cost effective than the legacy approaches (e.g., the technology used may cost more, but if the fraud is minimized, at the end of the day, it is a cost-saving solution). Furthermore, they should also create new revenue flows or better tackle existing ones in order to justify their existence. Finally, understanding the nature and key rules of each local market as well as providing integration with existing approaches (e.g., reuse existing infrastructure and legacy billing systems) may also lead to its rapid acceptance. It should also be kept in mind that, apart from the technology part, the right legislation framework must be in place and ease approaches, especially when we refer to a global payment service. Experience has shown that even when a common directive exists (for instance within the European Union), its full interoperable implementation at per country level still remains a challenging task (Merry, 2004).

Within the past years, several mobile payment solutions have been developed. Some of them even managed to leave the prototype level and enter the commercial market. A detailed insight on these payment approaches is provided by Henkel (2001), Krueger (2001) and Karnouskos (2004). The mobile payment area is an active one and is rapidly changing. However still existing approaches have done little to fully address all of the requirements needed to establish a global, widely accepted open and secure mobile payment service. For instance regarding security in such services; most MP procedures today use SMS or IVR (interactive voice response) as a method to verify user's identity, methods that have been proven to be insecure. Furthermore, users are usually asked to provide their personal information to a third-party service provider in order for them to be able to register and get the service. Therefore they are asked to place immediate trust of their money and personal data on a previously unknown party. This third party is able to have the complete set of data for any transactions users make, therefore it is able to monitor users' private lives and of course do indirect profiling. It must be kept in mind that user-perceived security (the combination of technical security and trust in the procedures of the approach) is a critical factor (Heng, 2004) that decides on the success or failure of a payment service and therefore it has to be done correctly from day one. Generally existing solutions today are either not trusted, not available to a large enough audience, not speedy enough, not user friendly, not secure enough, tailored for special applications and transaction types, are only available to a limited closed circle of customers and merchants, or have a limited business model. SEMOPS, which we shortly present here, has designed and implemented an approach that realizes a secure, universal, real-time electronic payment service, which effectively covers most of the requirements such a global service poses. To our knowledge past and current mobile payment approaches (Karnouskos, 2004), address only fractions of the mobile payment domain needs, while SEMOPS takes a holistic approach, therefore complementing any existing system.

SEMOPS: A SECURE MOBILE PAYMENT SERVICE

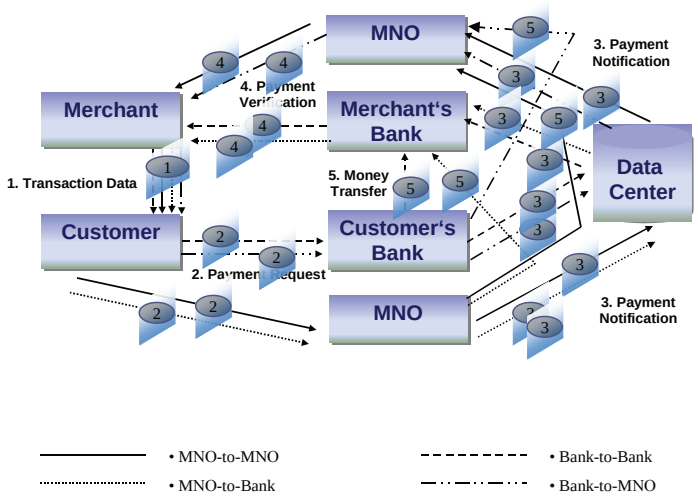
SEMOPS is a mobile payment solution that is capable of supporting micro, mini as well as macro payment transactions. It is a universal solution, being able to function in any channel, including mobile, Internet and POS; it can support any transaction type, including person to person (P2P), business to consumer (B2C), business to business (B2B) and of course person to machine (P2M), with a domestic and/or international geographic coverage.

As in every payment system, SEMOPS is capable of transferring funds from the customer to the merchant or, in more general terms, from the payer to the payee. Typically, this transfer is realized via a payment processor, such as a bank or a mobile operator. The SEMOPS payment solution, however, is novel in that it enables cooperation between different payment processors (e.g., cooperation between banks and mobile operators), in achieving a global, secure, real-time, user-friendly, and profitable mobile payment service that can be used in both electronic and mobile commerce transactions. The payment service designed, developed, and currently in trial within the SEMOPS project establishes a customer-driven transaction flow and follows a simple credit push model. Basic principle of the business model is that it is based on the cooperation of banks and MNOs. This situation has two consequences (a) actors' resources can be combined and (b) revenue has to be shared. This is quite a challenge but SEMOPS proves that this is a win-win situation for all participants.

In Figure 1, one can distinguish the main players and components in a mobile payment scenario. Each user (customer or merchant) interacts with his or her payment processor (e.g., home bank or mobile network operator (MNO)) only. The banks and MNOs can exchange messages between them via the Data Center (DC). We should mention that the legacy systems of the bank and the merchant are integrated in the SEMOPS infrastructure and are used as usual. A typical scenario assumes that:

1. The merchant (generally any real/virtual POS) provides to the customer the necessary transaction details, invoices.

Figure 1. SEMOPS transaction flow



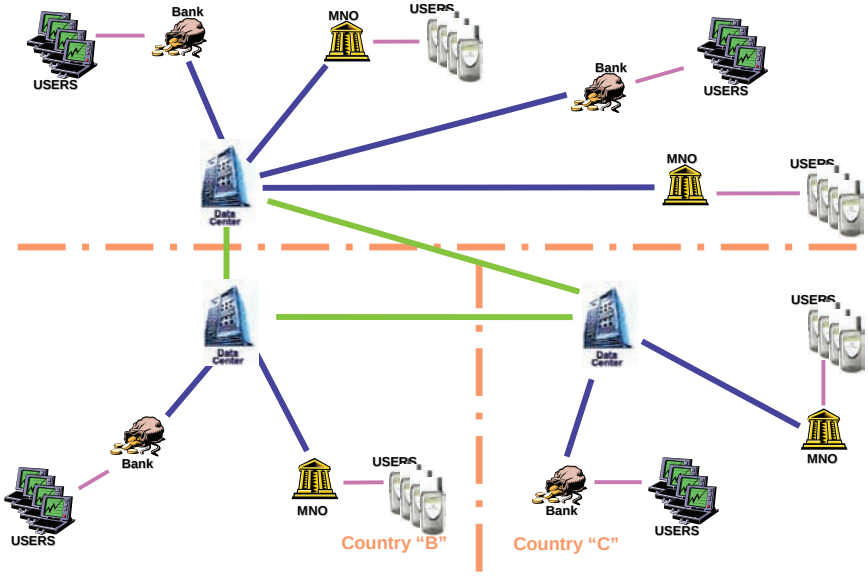
2. The customer receives the transaction data and subsequently initiates the payment request, authorizes it and forwards it to the payment processor (at the customer's bank or MNO).
3. The payment processor identifies the customer, verifies the legitimacy of the payment request, checks the availability of funds and forwards this request to the merchant's payment processor via the DataCenter (DC).
4. The merchant's bank receives the payment notice, identifies the merchant, notifies him or her about the payment being made, or requests from him to confirm or reject the transaction.
5. Once the merchant side confirmation comes, the fund transfer is done and all parties are notified about the successful payment.

There can be different combinations, depending on whether the user (customer or merchant) uses his or her bank or MNO account and whether the merchant accepts the payment in his bank or MNO account. The SEMOPS model (Karnouskos, Vilmos, Hoepner, Ramfos, & Venetakis, 2003) is extensible, therefore any third service provider that can offer the customer an account (e.g., credit card or financial service provider, even a utility company) can also easily slip in the role of the bank. It is however important to note, that although SEMOPS enables any account managers to play the role of a payment processor, the actual participation may be limited by legal constrains.

The new payment solution only has a chance to be accepted on the market if it makes good economic sense for the key players to promote the service. All the features, offered to the end users, the security, the comfort, the wide reach may be in vain if there is no economic incentives for the service providers. However it is obvious also that the service providers alone cannot make a success story of the service if the users are dissatisfied with either the service or the terms of the usage. The SEMOPS approach is based on decentralization. In each country where the service is introduced there is a local entity, the license holder, who organizes the service, contracts with the banks and mobile operators, contracts with the local service providers, ensures that local regulations are complied with, makes sure that the general service requirements are followed.

The flexibility of the model and its capability of integrating quickly new payment processors are critical for its survival. As it can be seen in Figure 2, the customers of any new financial provider that connects to the infrastructure can immediately transact with all other customers of the other providers in a transparent for them way. That will lead to a rapid expansion of the service that can establish it as a global payment service. SEMOPS follows a trust-delegation model. The new customers added do not have to place any trust into the SEMOPS approach itself; they need to trust the service that their banks and MNOs are providing to them (therefore extend the existing trust they already have placed to these institutions). The banks and MNOs are connected via a financial infrastructure with its own rating system and its own trust relationships that exist today. As a result, in a transaction scenario, user A does not have to know personally or trust directly user B to perform the transaction. The SEMOPS approach has several features, including means to secure transactions, notify in real-time its users and protect their privacy by even allowing anonymous payments to be made. Further info on the approach can be found in (Karnouskos, Vilmos, Ramfos, Csik, & Hoepner, 2004). Beyond using existing trust relationships among banks/MNOs and their customers, SEMOPS deploys also state of the art security (digital signatures and encryption) as well as processes that protect the user privacy (Karnouskos, Hondroudaki, Vilmos, & Csik, 2004).

Figure 2. Distributed MP service architecture



FUTURE TRENDS

Currently, almost all existing approaches focus on 2G or 2.75G infrastructures in order to achieve the critical mass once they are commercial. However, the mobile network infrastructure itself is rapidly evolving. The début of UMTS, wireless LAN, WiMAX and other 3G and beyond technologies will provide new capabilities that will free MP from some its current limitations and allow more sophisticated approaches to be developed. Once this infrastructure becomes mainstream, we will witness also solutions that take into account the new security capabilities, which are nonexistent today, offered by such infrastructure for security, privacy and trust management.

The device manufacturers continue to bring on the market mobile phones that have advanced capabilities and host their own execution environment. It is a matter of time until advanced cryptographic services are integrated in these devices that will make possible diverse secure communication and authentication procedures. Mobile public key infrastructure (mPKI), mobile digital signatures, encryption, and biometric authentication are expected to be widely available in the near future. Furthermore Identity Management efforts are ongoing for the Internet community and several standardization consortia such as Liberty Alliance (www.projectliberty.org) work toward federated identity in the virtual world. If such efforts are successful, they will have a catalytic effect on MP domain, as they will provide a homogeneous identity framework capable of bridging universally the real and virtual world.

With the rise of technological approaches, other communication channels will flourish. Today the basic channels of payment services are the SMS, Voice, and lately IrDA and communication over GPRS/EDGE. However, other

innovative approaches seem also promising such as instant messaging (IM) and near field communications (NFC). The IM will not only allow bridging together the Internet and mobile services and payments but will also make trivial P2P payments, where the two or more parties are not in the same physical space (Karnouskos, Arimura, Yokoyama, & Csik, 2005).

Digital Rights Management (DRM) is an integrated complex context covering not only technologies that limit or prohibit the unauthorized copying or distribution of these products but include also laws, contracts and licenses that regulate and restrict the use of such material (Becker, Buhse, Günnewig, & Rump, 2003). As content generated for mobile devices is increasing, mobile DRM systems are expected to play a significant role in the future (Beute, 2005). Standardization initiatives like the Open Mobile Alliance (OMA; www.openmobilealliance.org) work towards developing an advanced mobile DRM standard with the ability to support richer content business models. However rich payment capabilities also need to be in place and the existing MNO billing schemes will not be enough. In the future coupling content management with a global payment capability, preferably real-time (e.g., via instant messaging), will result in a powerful combination, where the mobile user any time anywhere can access legitimately content and instantly pay for it according to his preferences (Karnouskos, 2004).

CONCLUSION

Mobile payment has sparked a lot of interest in research and commerce communities and is viewed as an integral part of our future life. The area is an extremely active one, and rapid commercial evolvement is expected in the short and mid term. The need for a mobile payment service that can address in a global way existing needs is evident, and the first steps have already been done. However, although several mobile payment services have been designed, implemented and even commercialized, up to today there is no such service that can be widely accepted and cover adequately most of the transaction spectrum that we have referred to. For any service to evolve and reach the critical mass, several issues including business as well as technology aspects have to be approached in the right way.

SEMOPS (www.semops.com) presents a promising approach as it integrates state of the art technology, a flexible cooperative business model and builds over trust relationships that exist in the real world today. SEMOPS demonstrated a fully functional service with live users in the premier computer industry event CEBIT 2005 (www.cebit.de). Currently we are in the process of setting several pilots mainly in Europe, but later also in Asia and the United States, while the aim is to make SEMOPS a successful commercial service within the short-term future.

REFERENCES

Becker, E., Buhse, W., Günnewig, D., & Rump, N. (Ed.). (2003). Digital rights management: Technological, economic, legal and political aspects. Lecture Notes in Computer Science (Vol. 2770).

Beute, B. (2004). Mobile DRM––Usability out the door? Telematics and Informatics Journal, 22(1-2) 83-96. Elsevier.

Heng, S. (2004). E-payments: Modern complement to traditional payment systems. Economics: Digital economy and structural change (Deutsche Bank Rep. No. 44). Frankfurt am Main, Germany: Deutsche Bank Research.

Henkel, J. (2001). Mobile payment: The German and European perspective. In G. Silberer (Ed.), Mobile commerce. Wiesbaden, Germany: Gabler.

Karnouskos, S. (2004). Mobile payment: A journey through existing procedures and standardization initiatives. IEEE Communications Surveys & Tutorials, 6(4). Retrieved from http://www.comsoc.org/livepubs/surveys/public/2004/oct/pdf/KARNOUSKOS.pdf

Karnouskos, S., Arimura, T., Yokoyama, S., & Csik, B. (2005). Instant messaging enabled mobile payments. In A. Salkintzis & N. Passas (Eds.), Wireless multimedia: Technologies and applications. John Wiley & Sons.

Karnouskos, S., Hondroudaki, A., Vilmos, A., & Csik, B. (2004, July 12-13). Security, trust and privacy in the secure mobile payment service. Proceedings of the Third International Conference on Mobile Business 2004 (m>Business), New York.

Karnouskos, S., Vilmos, A., Hoepner, P., Ramfos, A., & Venetakis, N. (2003, September 29-October 1). Secure mobile payment—Architecture and business model of SEMOPS. EURESCOM Summit 2003, Evolution of Broadband Service, Satisfying user and market needs. Heidelberg, Germany.

Karnouskos, S., Vilmos, A., Ramfos, A., Csik, B., & Hoepner, P. (2004). SeMoPS: A global secure mobile payment service. In W-.C. Hu, C-.W. Lee, & W. Kou (Eds.), Advances in security and payment methods for mobile commerce (pp. 236-261). Hershey, PA: Idea Group Publishing.

Krueger, M. (2001, August). The future of m-payments—Business options and policy issues. Electronic Payment Systems Observatory (ePSO). Institute for Prospective Technological Studies. Retrievable from http://epso.intrasoft.lu/papers/Backgrnd-2.pdf

Little, A. (2004). Global m-payment report 200—Making m-payments a reality. Retrieved from www.adlittle.com

Merry, P. (2004, July). Mobile transactions in Europe: The challenge of implementation and ramifications of EU directives (Industry survey from the ARC Group). Retrieved from www.arcgroup.com

Mobey Forum. (2003). White paper on mobile financial services. Retrieved from http://www.mobeyforum.org/public/material/

Vilmos, A., & Karnouskos, S. (2004, July 12-13). Towards a global mobile payment service. Proceedings of the Third International Conference on Mobile Business 2004 (m>Business), New York.

KEY TERMS

Authorization: Granting of rights, what includes granting of access based on access rights or privileges. It implies the rights to perform some operation, and that those rights or privileges have been granted to some process, entity, or human agent.

DRM: Digital Rights Management (DRM) is a concept for managing and controlling the access and utilization of digital assets.

Macropayment: These payments are typically over $20.

Micropayment: These are the lowest values, typically under $2. Micropayments are expected to boost mobile commerce as well as pay-per-view/click charging schemas.

Minipayment: These are payments between $2 and $20. This targets the purchase of everyday small things.

Mobile Payment: Any payment where a mobile device is used in order to initiate, activate, and/or confirm this payment can be considered as a mobile payment.

POS: Point of Sale is a location where a transaction occurs. This may be a realPOS (e.g., a checkout counter), or a virtualPOS (e.g., an e-shop in the Internet).

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 1114-1119, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.11
Influence of Mobile Technologies on Global Business Processes in Global Organizations

Dinesh Arunatileka
University of Western Sydney, Australia

Abbass Ghanbary
University of Western Sydney, Australia

Bhuvan Unhelkar
University of Western Sydney, Australia

ABSTRACT

Organizations are globalizing their business primarily due to the communications capabilities offered by Internet technologies. As a result, there are global business processes that span across multiple geographical locations and time zones. The influence of mobility on these global business processes does not appear to have been studied in sufficient detail. Furthermore, mobile technology goes far beyond its ubiquitous use as a mobile phone for voice communication or for the exchange of messages. This chapter discusses and recommends a model for transition and integration of mobility into global business processes. We also envisage the accommodation of mobile Web services in mobile transformations enabling business applications to collaborate regardless of their technological platforms.

INTRODUCTION

Technology has changed the way people and businesses communicate with each other. Information and communications technology (ICT) has made a big leap in communications in the previous decade. With the communications backbone of

the Internet, ICT has influenced the very way of life for many people and organizations. As a result, the method and manner in which the organizations carry out their businesses have also changed. Ubiquity of business processes, as per Arunatileka (2006), is likely to play a major role in future business environments. This ubiquity is a result of organizations starting to collaborate electronically, paving way for collaborative global businesses, resulting in significant strategic advantages to these organizations. These advantages include business growth in the global market, and building strategic alliances and partnerships for business organizations (Arunatileka & Arunatileka, 2003). However, when it comes to extending and applying the concepts of collaboration through time- and location-independent mobile technologies, it appears that organizations are still relatively nascent. The paucity of literature in this regard provided us with the necessary impetus to study mobile technologies from the point of view of their application in global businesses.

This chapter accomplishes our aim of studying and applying mobility to global business processes. This chapter also reports on construction and application of corresponding models to enable mobile transformations (m-transformations). The lead author of this chapter has also validated and applied these models through "action research" at a real global organization. This well-thought-out application of mobility resulted in streamlining and speeding up of the existing processes, as well as exploration and creation of totally new business processes within that global organization.

Mobile Technology

The advancement of mobile technologies has created the opportunity for organizations to adapt this technology in their business processes. Per Schneiderman (2002), faster access to the corporate database and new applications that embody wireless and Internet connectivity are two great advantages that organizations can develop in terms of their business operations. The usage of mobile devices in the modern era is so important that their incorporation in business processes can be classified as one of the crucial factors in the survival and prosperity of a business. Birchler (2004) clearly points out that the exponential growth of the Internet has challenged the prevailing understanding of network organizations and ownership. Therefore mobility, combined with the Internet, provides organizations with a powerful tool to be used strategically for connections in the electronic business world. Deshpande, Murugesan, Unhelkar, and Arunatileka (2004) describe the requirement of delivering the Web in a single composite device; their vision, under the auspices of the Device Independent Web Engineering (DIWE) Group, is to make the Web accessible to "anyone, anywhere, anytime, anyhow."

The use of portable computing through communications devices is forcing the reappraisal of the capabilities and future of wireless. In today's competitive markets, mobile technology is providing person-to-person communication, resulting in a new era of customer relationship management for organizations (Arunatileka & Unhelkar, 2003). Mobility, in the context of businesses, can be understood as the ability of processes to be executed anywhere and at anytime. Mobile technology encompasses the various devices and applications that have been put together to provide organizations and individuals with the ability to conduct businesses as per the DIWE vision mentioned earlier.

This study started with observing daily business activities of a global organization in order to ascertain how m-transformation could enhance their business processes. These processes were then modeled using the activity diagrams of the unified modeling language (www.omg.org). This modeling was followed by "re-engineering" of the business processes based on the theoretical m-transformation models formulated by the researchers. Thereafter, two selected business processes related to "timesheets" were transformed into

mobile-enabled processes. The following sections describe the research problem and the approach taken in solving it in greater detail.

THE RESEARCH PROBLEM

The research problem was formulated to understand whether m-transformation of certain business processes of the company would result in improved efficiency and productivity. The initial effort was to identify such candidate business processes which could undergo m-transformation. Thereafter those processes were to be transformed using the theoretical model of m-transformation. Once the m-transformed processes are integrated into the mainstream of business processes, their functionality was to be further observed in order to achieve a smooth flow of business processes. Advantages and challenges in terms of this m-transformation were also to be studied, in order to improve and validate the model.

There were two processes that were identified as the potential processes for improvement. This identification was done in consultation with the managing director of the global organization (henceforth referred to as "the company"). The two processes initially observed for this research were: the timesheet operation and information gathering process.

The timesheet process was one of the most important processes since it involved the revenue generation for the branch office. The existing process for timesheet operation had a few lapses, which are described under the analysis. Mainly, there was no activity to initiate the timesheet process. The group leaders (GLs) who initiated the timesheets were too busy with their own work to be involved in this administrative process. The account manager (AM) who was the facilitator for this process only could call and remind the GLs regarding the timesheet. However the GLs had to find a time to sit down and send an e-mail to indicate the timings of each of the workers.

Information gathering was a process to gather information for future business. Again the busy GLs were given the task to find out whatever information they could from the project clients. There was no formal process for this either. Therefore there was no proper information kept regarding the current project status as well as other important information regarding future projects, deadlines, and so forth.

These two processes were carefully studied and modeled in order to analyze potential solutions for them to overcome the current lapses. Input from the AM and GLs was also used to verify the existing processes and workflow in order to get a correct picture.

Modeling of Existing Processes

Modeling of existing processes was carried out with use cases and activity diagrams of the unified modeling language (UML, n.d.). Use cases document, in a step-by-step manner, how actors (people and the system) interact with each other. These descriptions of use cases are used to further visually model them with the help of activity diagrams—graphically representing the documentations of the use cases.

Timesheet Operation

There were five use cases identified from the study of the existing timesheet operations. Once the use cases were described, the information therein was converted into activity diagrams. Thereafter mobile technology was introduced in order to convert the existing processes into a new suit of business processes.

Information Gathering

Whatever information is gathered at a branch office currently is collected in a very ad-hoc way. However there is a need for systematic gathering of information since the market is becoming more

competitive and needs to have the constant attention of the marketing staff. Three main types of information required by the management were identified. Apart from the marketing people, the AM in this instance, the best people who could have access to such information were people working on a project such as the GLs. They could get information via the grapevine. However since they were extremely busy with their own expert contributions, a method to get such information was very important. However this area is not discussed in detail in this chapter.

THE COMPANY BACKGROUND

This action research was carried out in a truly global organization, headquartered in India and with operations around the world including in Australia. The company is a provider of professional services to technology organizations. The specific study was in the Sydney, Australia branch office.

The company in Australia is based mainly in Sydney and Melbourne. Its clients mostly include ICT technology providers. There are GLs in each of the projects, and they report to the AM in charge of that project. The AM is overall in charge of the project and reports to the regional director (RD). One AM could handle several projects; a project could be classified as the provision of the company expertise to a client organization in terms of technical expertise, project management, and other related services. One client could have several projects being run in parallel. Each project would have one group leader who is responsible to report to the account manager about the number of hours worked by each of the company employees attached to that project.

The objective of the branch office is to conduct business activities in Australia and New Zealand, and generate revenue for the parent company in India. The company has several areas of core business, namely: billing support systems, networks, mobility services, location-based services, and next-generation networks (NGNs).

Being an overseas company, the branch in Australia must outsource legal expertise in Australia. These outsourced services are in the area of accounting, including taxation, immigration laws, reporting procedures to the ASIC and other such government institutions, and other relevant local laws.

The company has a major challenge in collating the timesheets by its 50-plus consultants working in this region. It takes a while for the accounts manager to collect all the timesheets, collate them, consolidate them, and then send them across to the head office for invoicing to the clients in Australia and New Zealand.

The information gathering process is a proactive process where the GLs who work in various projects are expected to provide feedback for the company with regard to the projects. The information is relevant to current projects as well as projects in the pipeline for the same client. GLs who work within the client companies are considered the best people to gather this type of information for the branch office. The following section briefly describes the theoretical background for action research.

METHODOLOGY: ACTION RESEARCH

This research carried out at the company followed the 'action research' approach in order to validate our theoretical presuppositions. Action research has been defined by Rapoport (1970) as follows:

"Action research aims to contribute both to the practical concerns of people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework."

The action research at the company mainly concentrated on two business processes which were very significant and also had problems right now. These two processes of weekly timesheets

and monthly information sheets were considered for this study. The managing director emphasized the need to streamline these processes to enable getting timesheets to the head office as quickly and efficiently as possible.

Action research was very appropriate for this research study because it:

- Established research method use in the social and medical sciences.
- Increased importance for information systems toward the end of 1990s.
- Varied in form and responded to particular problem domains.

In addition, its most typical form is a participatory method, and the fundamental contention of the action researcher is that complex social processes can be studied best by introducing changes into these processes and observing the effects of these changes.

There is widespread agreement by action research authorities, as indicated by Peters and Robinson (1984), that it has four common characteristics:

- An action and change orientation.
- A problem focus.
- An "organic" process involving systematic and some times iterative stages.
- Collaboration among participants

The key assumptions underlying action research which were also applied in the company setting were that social settings cannot be reduced for study and action brings understanding.

Thus the existing processes were modeled using UML. Thereafter, the models of the existing processes were studied within the context of the m-transformation model in order to ascertain the areas within the process where mobility can be incorporated.

The existing procedure was first modeled drawing use cases in order to identify the interactions between various actors. The use cases were documented and then converted to activity diagrams. There were altogether five use case descriptions and five activity diagrams modeling the existing business process for timesheets. These activity diagrams were then verified by the RD and the AM at the branch office.

The action and change orientation in this instance were the two operations selected. The specific problems were the delays in timesheet operations and the unstructured nature of the current information gathering process. The use of UML for modeling provided the process involving systematic and sometimes iterative stages. The collaboration among participants was provided through various meetings held to verify the existing and proposed processes.

Thus action research offered a very scientific and subtle way of observing changes in an organization and introducing new changes without drastically changing the ongoing operations of the organization. The following section analyzes the current operation and uses the theoretical model for m-transformation to transform these processes.

ANALYSIS

Timesheet Operation

Once the existing timesheet operation was modeled through the activity diagrams, an industry survey was carried out through a literature survey. This survey revealed the existing technology available for projects of this nature. There were also discussions with the account manager, group leader, and director on which processes and areas to be improved. These discussions with the TML stakeholders also revealed various drawbacks with the existing timesheet operation, as follows:

- Timesheets were not submitted on time by the group leaders.

- There was no formal method of triggering this important activity.
- The client project manager was not involved in the current process.
- Discrepancies could take considerable time to be corrected.
- The proper monthly timelines—thus affecting the ERP software cycle (PeopleSoft) at the head office—are missed at times.
- Collection of account receivables goes on overtime due to the lapses in the process.

Information Gathering Process

The information gathering process was triggered by a form which could be sent to the AM by the GL at the end of each month. An short message service (SMS) reminder would be sent by the AM during the first week of the month for such information for the previous month. The information gathering process was fraught with lack of proper structure. This problem was further exacerbated by the fact that the GLs were overloaded with their own consulting work, resulting in lack of time to gather such information. This situation appeared to provide a rich opportunity for application of mobility for its alleviation. As a result, a form was proposed that would tell the GL the exact information that was required from the project. If the GL could observe such information even before the form was due in the month (e.g., hearing about a new project in the pipeline), the GL could immediately convey that to AM via SMS so that AM and the director had time to prepare themselves to talk to the client.

Mobile Transition Roadmap

While the aforementioned two processes were the focus of the study, it was also essential to provide a carefully thought-out approach to transforming these processes to mobile-enabled processes. The researchers created a model of the "mobile transformation roadmap," which provided the necessary and robust theoretical basis for transitioning the processes into mobile-enabled processes. Figure 1 illustrates the crux of this roadmap model. What is shown is a generic model to transform a business organization into a mobile-enabled organization by transitioning its business processes.

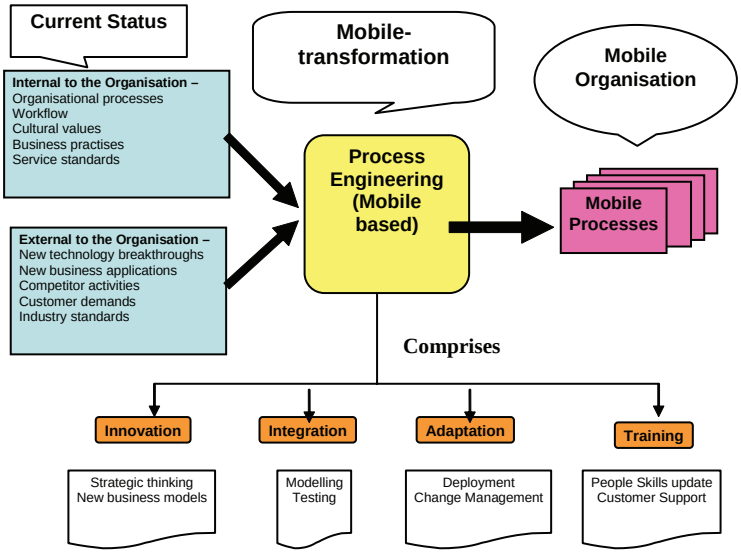
The business process transformation modeled in Figure 1 is further broken down into smaller and manageable steps when it is executed. In the business process transition area, factors internal to the organization are considered. The current operations of the organization are looked at in the wake of customers and employees demanding new mobile technologies as well as easier and efficient work methods.

The technology transition, although a part of the overall business process transition, is focused on the application of emerging technologies, and corresponding new tools and applications that need to be incorporated into the mobile-enabled processes. The incorporation of such new technology in the business adds pressure for the organization to effectuate change in its technology usage. This leads to new thinking and new business processes to get the full effect of emerging technology.

The training of employees and the forming of the new-look organization creates a more customer-oriented work ethic. The transitioned organization is able to follow the well-known objective of being customer centric by making judicious use of mobile technology. As a result, the service standards, delivery periods, response times, and similar attributes change positively in the organization leading to a new business culture. Thus, three perspectives—namely, technology with regard to emerging technologies, methodology with regard to business processes, and sociology with regard to new business culture towards customer orientation—are identified in the transitioned organization.

The m-transformation roadmap points out the four action areas where process engineering

Figure 1. The proposed mobile transition roadmap



should very carefully considered. These are innovation, integration, adaptation, and training. These key areas should be revisited—iteratively, incrementally, and parallel—until the new business processes are satisfactorily transformed. Furthermore, the m-transformation process can also be supported by corresponding computer aided software engineering (CASE) tools that can make the process easier as well as measured.

Innovation tools include brainstorming as well as critical evaluation of the current business processes. Furthermore, theoretical innovation models such as one by Rogers (2003) that include the five stages in the innovation-decision process—namely, knowledge, persuasion, decision, implementation, and confirmation—can also be of immense importance in m-transformation. In our case, they were applied in identification of the processes that had potential for m-transformation.

Integration required us to model the processes undergoing m-transformation to a stage where they could be studied, modified, and made ready for integration with the business itself. This required us to make use of the unified modeling language to model both existing as well as the new business processes that would result from m-transformation.

Adaptation involved the need to 'settle' the new processes in the business environment. While the integration stage required extensive modeling, the adaptation stage had a need to ensure that the freshly modeled processes were adaptable to the environment—including the business, its employees, and customers. This stage of the m-transformation processes was crucial in our exercise as it required us to make the employees aware of the proposed processes as well as putting together service standards and rules.

Training is the last and most crucial stage of the m-transformation process. This stage required us to provide training in the use of the mobile-enabled process to employees as well as customers in order to understand and use the new processes in a smooth manner.

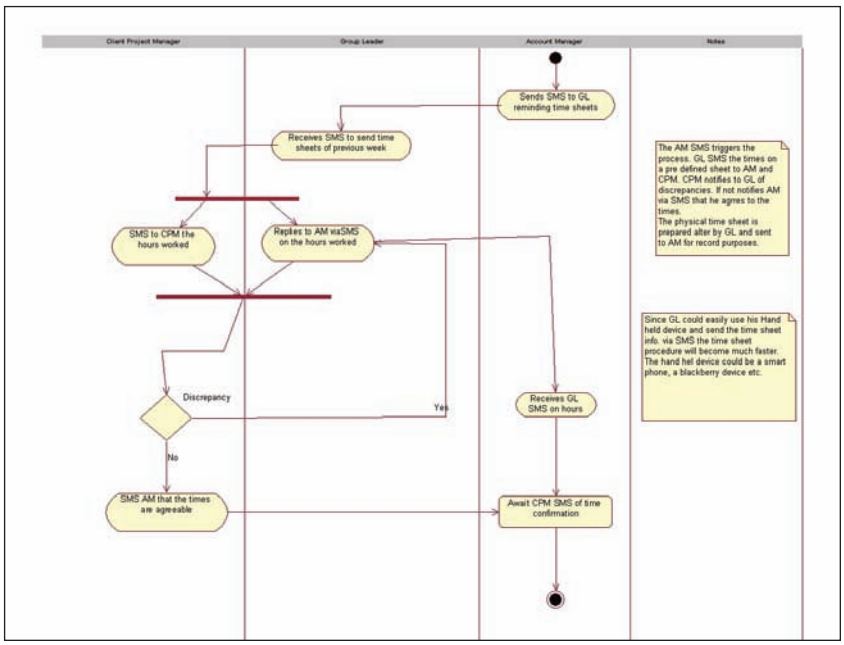
Having looked at the existing processes, new activity diagrams were drawn in order to highlight and rectify the drawbacks. There were four such new activity diagrams highlighting the introduction of mobile technology and the improvement of the business processes which are listed herein.

Proposed Process Activity Diagrams for Timesheet Process

Figure 2 depicts the initial timesheet process where the AM is triggering the process with an SMS message of billable time. The actual timesheet data is delivered by the GL to the AM via SMS. The AM sends back time data to the client project manager via SMS again for time confirmation. Therefore the whole process is based on SMS messages where all the unnecessary delays are reduced and the process is fast tracked since the SMS could be sent anytime anywhere at the real time, with quick reactions, rather than delaying sending the data. Therefore any such delays in getting back to the office to send the e-mail and finding the time to do such action are minimized. See Figure 2 for the origination of a timesheet by the AM and the GL which shows the fundamental change effectuated by incorporation of mobility.

Further, since AM is triggering the timesheet operation by an SMS message in the proposed process, the GL has to reply to the SMS in order to send in the timesheet. Therefore the process is controlled at the branch office, as opposed to the

Figure 2. Proposed activity 1—account manager/group leader pre-timesheet activity



current process where the control is with the GL, who is an employee of the company, but physically working in an outside client organization.

As depicted in Figure 3, the AM awaits the SMS reply from the GL. The difference in the proposed process as compared to the existing process is, the AM does not await the arrival of the actual timesheet, but instead prepares the Excel sheet based on the agreed times transmitted through the SMS message. This process of recording the timings before the paper-based timesheets are received provides crucial advantage to the business at the end of the month. The AM could update any discrepancy every week until the end of the month with the actual timesheet data. However, at the end of the month he will forward the data in the Excel sheet and his last update would be based on an SMS message.

Figure 4 illustrates the process at the head office when it receives the Excel worksheet on work times. Since the AM forwards the Excel sheet with SMS-based data, for the last week of the month, there would be no delays in processing and sending invoice data to the head office.

Therefore every PeopleSoft cycle would be producing a set of invoices. Thus the invoicing operation is streamlined that each PeopleSoft cycle is printing a set of invoices to be sent to customers. Any discrepancies will be dealt with later through a process of credit and debit notes.

It should also be noted that Figure 4 would change very much if the mobile Web service, which is described as a future enhancement later in this chapter, is adopted by the company. The entire activity of manual entry of data into the PeopleSoft system would be taken care of by the mobile Web service.

Figure 5 illustrates the finalizing of the process with CPM that deals with sending the invoices to the accountant for payment. Note that a credit/

Figure 3. Proposed activity 2—timesheet preparation and verification

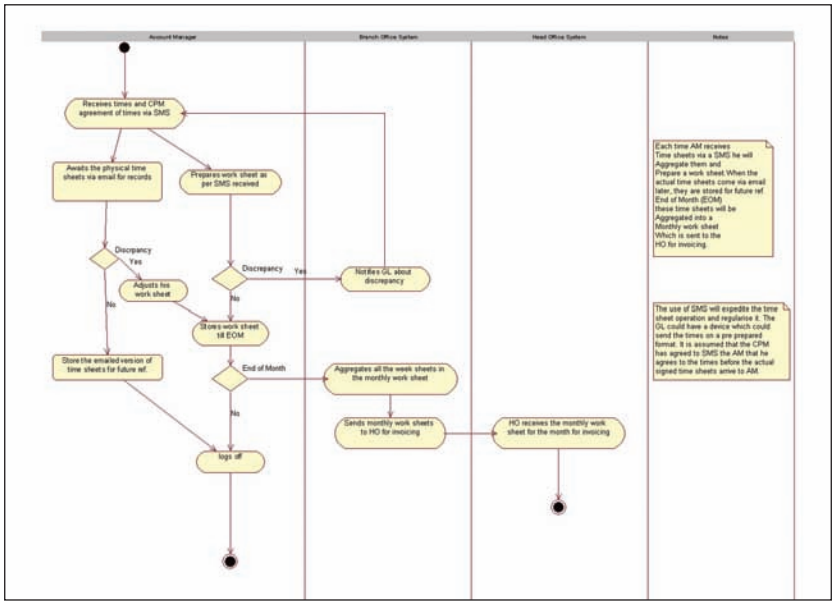


Figure 4. Proposed activity 3—head office invoice operation

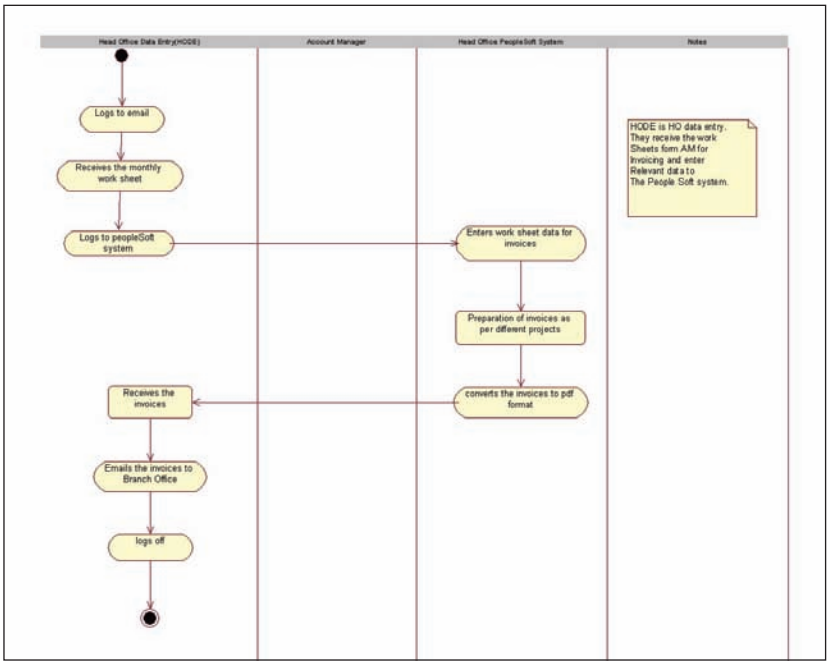
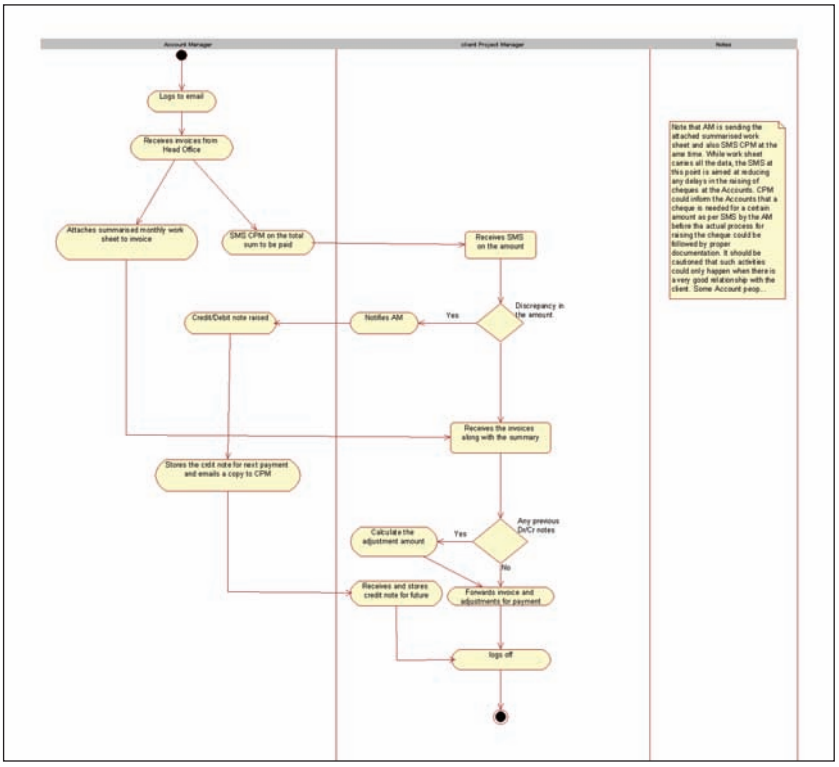


Figure 5. Proposed diagram 4—branch office and project client finalizing of invoices



debit note is provided within the process for any discrepancies for the previous accounting period since the actual invoice is partly based on the SMS data. This ability to provide a credit/debit note provides the necessary flexibility to amend errors that are likely to occur when the process is m-transformed with the use of SMS data.

EFFECT OF MOBILE TRANSFORMATION

Having considered how the selected business processes for the international organization under study would change as a result of m-transformation, we now do a brief comparison of the effect of such transformation on the organization itself. In our case, this comparison will be between the existing timesheet process and corresponding mobile-enabled timesheet process. This comparison showcases the difference between the two processes—without and with mobile technology—especially in the way in which it affects the roles played by people in the organization. There are five activities listed which are depicted in the existing activity diagrams. These activities are then compared with the proposed timesheet operation activity diagrams.

Comparison of the Existing Process with the Proposed Process

In the action research work that was carried out at the company, application of mobile technology for the selected business processes of the organization resulting in new processes was proposed. The transformation of the business processes are guided by the m-transition roadmap, which is a generic model for transformation of business processes with the application of mobile technology. The roadmap looks both inward and external to the organization in deciding what is happening in the organization as well as in the industry sector in which the organization belongs. In this specific case, we looked at how the technology services providers work in a general way when deciding how to improve the business processes of the organization. By doing that, the organization is compared with other similar organizations in the same industry. Therefore it provides a guideline as to what new technology has to be introduced in order to keep pace with the competition.

A further improvement of the timesheet process could be looked at in terms of Web services and mobile Web services. A detailed description on how the business processes could be further improved by use of this emerging technology is discussed in the next section.

EMERGING TECHNOLOGIES

Based on our understanding, the company under study has an excellent opportunity to use emerging technologies to further enhance the processes modeled with mobility. Web services and mobile Web services are such technology which could be used to communicate with the PeopleSoft system directly. This would result in further improving the process by eliminating manual data entry and use the Web service to send the data directly to the PeopleSoft system. Therefore it is useful look at Web services and mobile Web service technology for future expansion.

Web Services (WS)

Web services (WS) is a unit of business, software application, or a system that can be accessed over a network by extensible markup language/simple object application protocol (XML/SOAP) messaging. It is a delivery mechanism that can serve many different consumers on many different platforms at the same time. Stacey and Unhelkar (2004) describe WS technology as an enabler to connect incompatible standalone systems to integrate a complex distributed system in a way that was not possible with previous technologies,

Table 1. The comparison of existing processes with proposed m-enabled processes

Activity	Current Process	Proposed Process
Group Leader (GL) Timesheet Operation	There are no official indications or reminders sent to GL for timesheets.	Account manager (AM) sends official reminder to GL via SMS, which prompts faster action.
Branch Office Timesheet Verification	AM awaits receipt of timesheet of GL to start filling out Excel worksheet to be sent to head office (HO).	AM is prompted by the times via SMS from GL and the confirmation from the client project manager (CPM).
Head Office Invoice Operation	HO awaits the Excel sheet with times for invoice preparations.	Still awaits Excel sheet, but it happens faster since the times are delivered via SMS to AM, vs. the current process where AM has to await for the e-mails.
Branch Office Dispatch of Invoices to Clients	The invoices are occasionally dispatched late due to delays in preparation and collection of timesheet data.	The invoices are dispatched more accurately based on SMS time data. There is provision for credit notes if there are any discrepancies for amounts invoiced, which are carried forward to the next period.
Invoice Operation at the Project Client	The invoices are checked by the CPM for any discrepancies at this point. If the invoice has been delayed due to delayed times sent to head office, this procedure would be around 2-3 months later than the actual month of invoicing.	There is no specific process needed at this point since the invoicing has been done and a credit/debit note could simply be raised to make the adjustment at this point. However, the invoicing would be completed within the next month due to the faster action via SMS messaging.

while Chatterjee and Webber (2004) believe that Web services represent a new architectural paradigm for applications. WS capabilities access other applications via industry standard network, interfaces, and protocols.

Web services are software programs that enable document applications to talk to each other. Taylor (2005) describes that although Web services are centered around the documents, they do not necessarily follow that such a document should be readable by people; this is reflected to be the core goal of WS.

Web services could be defined as open standard-based (XML, SOAP, etc.) Web applications that interact with other Web applications for the purpose of exchanging data. Initially used for the exchange of data on large private enterprise networks, Web services are evolving to include transactions over the public Internet (Lucent, n.d.).

A more specific form of WS using mobile technology is described next which is relevant to the organization under study.

Mobile Web Services (MWS)

A mobile application that is using the WS to transmit its data is classified as MWS. According to Pashtan (2005), mobile terminals and mobile services are an integral part of the extended Web that includes the wireless domains that facilitate automated interoperation between terminal and network services. WS can replace less flexible methods for information exchanging of specific transaction data. WS enables the building of software applications that execute on the Internet and use the same software paradigms that were successfully applied in the development of enterprise application.

According to the Australian Computer Society (ACS, 2005) report on MWS, with Web services, phones now have the potential to actually consume useful services. But before developing a mobile client, you might want to think twice before taking the simple object application protocol/hyper text transfer protocol (SOAP/HTTP) route. First of

all, turning your phone into a SOAP client might have some performance costs related to slow data speeds and processing both HTTP commands and XML. Secondly, most phones do not come with Web services support built in. Finally, you can hide the Web services complexity and leverage existing technologies to make use of their widespread availability. This would require a gateway to sit in between the phone and the Web service to handle the passing and conversion of messages, but you no longer have to worry about client-side performance issues or even deploying a client (ACS, 2005).

Microsoft service providers define MWS as an initiative to create Web services standards that will enable new business opportunities in the personal computer and mobile space, and deliver integrated services across fixed (wired) and wireless networks. Mobile Web services use existing industry standard XML-based Web services architecture to expose mobile network services to the broadest audience of developers (http://www.microsoft.com/). The functionality of MWS is examined in the light of how MWS could enhance the current process enhancing its functionality to talk to the PeopleSoft system directly via Web eliminating the second data entry at the head office, which is happening during the current process.

Business Process Reengineering with the Global Company Perspective

Having discussed WS and MWS, an important question worth considering is how these technologies can be incorporated into the business processes of the company that is undergoing m-transformation. Thus, the discussion on MWS is a part of re-engineering business processes with mobility. "Reengineering," as defined by Hammer and Champy (2001), is a fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in the critical, contemporary measure of performance such as cost, quality, services, and speed. Reengineering a company's business processes ultimately changes practically all aspects of a company such as people, jobs, managers, and values that are linked together.

The global company under study has undergone m-transformation. However, the transformed business processes could be further enhanced by the use of mobile Web services. For instance, the process of sending the timesheet data to the head office for invoicing could be easily done through such mobile Web services. By using mobile Web services, the data could be directly sent using a mobile device. This would cut down on reentering the data into the PeopleSoft system if the MWS could be configured to talk directly to the PeopleSoft system. Such direct integration would make the process faster and also make it less error prone due to reducing of human intervention. Such integration could be used in all the offices of the organization globally, to communicate with the PeopleSoft system sending timesheet data. The security concerns arising could also be addressed through MWS itself by authenticating the invoice printing through a second MWS communication.

As has been mentioned earlier, the main focus of this research is to investigate the business processes to see how they could be made more productive. By reengineering the business processes, they could collaborate and integrate with the business processes of the other organizations, the head office, and branch offices in this context. As stated in ACS (2005), processes have linkage to each other; this research is concentrating on this linkage of the processes to another organization's processes that may not be known to each other. Thus the total global operation of the company could be integrated together using WS/MWS. These processes could help each other, thus the branch offices of the entire organization could automatically collaborate with each other without going through the head office.

CONCLUSION AND FUTURE RESEARCH DIRECTION

This research investigated the current systems at a global organization in order to look at possible solutions to enhance the business processes of the company under study. Selected business processes were studied in order to introduce mobile technology to those processes. This resulted in transformation of existing business processes to mobile-enabled business processes which would be integrated into the organization. The company is implementing this proposal systematically and carefully under the guidelines of the m-transformation roadmap, concurrently looking at adaptation and training issues. Once all the proposed business processes are fully operational, the actual impact of the m-transformation could be identified.

This research has a value to all the stakeholders who took part in it. We, the researchers, benefited by the overall research findings to understand how the organizational processes would change in order to accommodate mobile technology. The company benefited by implementing some outcomes from this research and streamlining its very important business processes. The company could further explore and get far-reaching benefits, if it would invest to implement WS/MWS technology into its business processes, which would benefit the global organization in communicating to the head office directly using MWS/WS, thus improving productivity of the global organization.

REFERENCES

ACS (Australian Computer Society). (2005, September 8). *Web services overview.* Retrieved November 10, 2005, from http:// www.acs.openlab.net.au/content.php?article.131

Alonso, G. (2004). *Web services, concepts, architecture and applications.* Berlin: Springer-Verlag.

Arunatileka, D. (2006). In B. Unhelkar (Ed.), *Mobile business: Technological, methodological and social perspectives.* Hershey, PA: Idea Group.

Arunatileka, D., & Unhelkar, B. (2003). Mobile technologies, providing new possibilities in customer relationship management. In *Proceedings of the 5th International Information Technology Conference, Colombo, Sri Lanka.*

Arunatileka, S., & Arunatileka, D. (2003). E-transformation as a strategic tool for SMEs in developing nations. In *Proceedings of the 1st International Conference on E-Governance, New Delhi, India.*

Birchler, M. (2004). Future of mobile and wireless communications. In P. Smyth (Ed.), *Mobile and wireless communications: Key technologies and future applications.* UK: Institution of Electrics Engineers.

Cabrera, L.F., & Kurt, C. (2005). *Web services architecture and its specifications: Essential for understanding WS.* Redmond, WA: Microsoft Press.

Chatterjee, S., & Webber, J. (2004). *Developing enterprise Web services: An architect's guide.* Englewood Cliffs, NJ: Prentice Hall.

Deshpande, Y., Murugesan, S., Unhelkar, B., & Arunatileka, D. (2004). Methodological considerations and challenges: Moving Web applications from desk-top to diverse mobile devices. In *Proceedings of the Device Independent Web Engineering Conference*, Munich, Germany.

Ghanbary, A. (2006). Evaluation of mobile technologies in the context of their applications, limitations and transformation. In B. Unhelkar (Ed.), *Mobile business: Technological, methodological and social perspectives.* Hershey, PA: Idea Group.

Hammer, M., & Champy, J. (2001). *Reengineering the corporation, a manifesto for business evolution.* UK: Nicholas Brealey.

Lucent. (n.d.). *W-definitions.* Retrieved May 18, 2006, from http://www.lucent.com/search/glossary/w-definitions.html

Marmaridis, I., & Unhelkar, B. (2005). *Proceedings of the 1st MobiComm, Mobile Business Conference, Sydney, Australia.* Retrieved from http://www.mbusiness2005.org/contact.html

Pashtan, A. (2005). *Mobile Web services.* Cambridge: Cambridge University Press.

Peters, M., & Robinson, V. (1984). The origins and status of action research. *Journal of Applied Behavioral Science, 20*(2), 113-124.

Rapoport, R.(1970). Three dilemmas in action research. *Human Relations, 23*(4), 499-513.

Rogers, E.M. (2003). *Diffusion of innovations* (5th ed.). New York: The Free Press.

Schneiderman, R. (2002). *The mobile technology, question and answers.* AMACOM.

Stacey, M., & Unhelkar, B. (2004). Web services in implementation. *Proceedings of the 15th ACIS Conference, Hobart, Australia.*

Taylor, I.J. (2005). *From P2P to Web services peers in a client/server world.* Berlin: Springer-Verlag.

UML. (n.d.). *UML 2.0.* Retrieved May 4, 2006, from http://www.uml.org/#UML2.0

Unhelkar, B. (2005). Transitioning to a mobile enterprise: A three-dimensional framework. *Cutter IT Journal, 18*(8), 5-11.

ADDITIONAL READING

ACMA. (2003, August). *Mobile commerce: Regulatory and policy outlook discussion paper.* Retrieved December 7, 2005, from http://www.acma.gov.au/ACMAINTER.4849984:STANDARD:1004384283:pc=PC_7126

Adam, O., Chikova, P., & Hofer, A. (2005). Managing inter-organizational business processes using an architecture for m-business scenarios. *Proceedings of ICMB 05, Sydney, Australia.*

Alag, H. (2006). Business process mobility. In B. Unhelkar (Ed.), *Mobile business: Technological methodological and social perspectives* (vol. 2, pp. 583-601). Hershey, PA: Idea Group.

Anckar, B., & D'Incau, D. (2002). Value added services in mobile commerce: An analytical framework and empirical findings from a national consumer survey. In *Proceedings of the 35th Hawaii International Conference on System Sciences.*

Archer, N. (2004). The business case for employee mobility support. In *Proceedings of the IADIS International Conference on E-Commerce, Lisbon, Portugal.*

Barjis, J. (2006). Overview and understanding of mobile business in the age of communication. In B. Unhelkar (Ed.), *Mobile business: Technological methodological and social perspectives* (vol. 2, pp. 719-726). Hershey, PA: Idea Group.

Barnes, S.J. (2002). The mobile commerce value chain: Analysis and future development. *International Journal of Information Management, 22*(2), 91-108.

Basole, R.C. (2004). *The value and impact of mobile information and communication technologies.* Atlanta: Georgia Institute of Technology.

Basole, R.C. (2005). Transforming enterprises through mobile applications: A multi-phase framework. In P*roceedings of the 11th America's Conference on Information Systems, Omaha, NE.*

Chan, J.C., & Hoang, D.B. (2005). Novel user-centric model for m-business transformation. In *Proceedings of the International Conference on Mobile Business.*

Cousins, K., & Varshney, U. (2001). A product location framework for mobile commerce environment. In *Proceedings of the Workshop on Mobile Commerce,* co-located with MobiComm2001, Rome.

Di Pietro, R., & Mancini, L.V. (2003). Security and privacy issues of handheld and wearable wireless devices. *Communications of the ACM, 46*(9).

El Kiki, T., & Lawrence, E. (2006). Government as a mobile enterprise: Real-time, ubiquitous government. In *Proceedings of the 3rd ITNG Conference,* Las Vegas, NV.

Er, M., & Kay, R. (2005). Mobile technology adoption for mobile information systems: An activity theory perspective. In *Proceedings of ICMB 05, Sydney, Australia*.

Falcone, F., & Garito, M. (2006). Mobile strategy roadmap. In B. Unhelkar (Ed.), *Handbook of research in mobile business: Technical, methodological, and social perspectives.* Hershey, PA: Idea Group.

Forouzan, B.A. (2004). *Data communications and networking* (3rd ed.). New York: McGraw-Hill.

Gershman, A. (2002). Ubiquitous commerce—always on, always aware, always pro-active. *Proceedings of the 2002 Symposium on Applications and the Internet* (SAINT 2002).

Guizani, M., & Raju, A. (2005). Wireless networks and communications security. In Y. Xiao, J. Li, & Y. Pan (Eds.), *Security and routing in wireless networks* (vol. 3, p. 320). New York: Nova Science.

Hawryszkiewycz, I., & Steele, R. (2005). A framework for integrating mobility into collaborative business processes. In *Proceedings of the International Conference on Mobile Business (ICMB 2005).*

Herzberg, A. (2003). Payments and banking with mobile personal devices. *Communications of the ACM, 46*(5), 53-58.

Hsu, H.Y.S., Burner, G.C., & Kulviwat, S. (2005). Personalization in mobile commerce. *Proceedings of IRMA 2005, San Diego.*

Huber, J.F. (2002). Toward the mobile Internet. *Computer, 35*(10), 100-102.

ITU. (2006a). *The regulatory environment for future mobile multimedia services.* Retrieved September 15, 2006, from http://www.itu.int/ITU-D/ict/partnership/index.html

Jarvenpaa, S.L., Lang, K.R., Takeda, Y., & Tuunainen, K. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41-44.

Kalakota, R., & Robinson, M. (2002). *M-business: The race to mobility.* New York: McGraw-Hill Professional.

Lalis, S., Karypidis, A., & Savidis, A. (2005). Ad-hoc composition in wearable and mobile computing. *Communications of the ACM, 48*(3).

Lee, C. (2006). Mobile CRM: Reaching, acquiring, and retaining mobility customers. In B. Unhelkar (Ed.), *Mobile business: Technical, methodological and social perspectives.* Hershey, PA: Idea Group.

Lee, Y.E., & Benbasat, I. (2003). Interface design for mobile commerce. *Communications of the ACM, 46*(12), 48-52.

Lyytinen, K., & Yoo, Y. (2002). Issues and challenges in ubiquitous computing. *Communications of the ACM, 45*(12), 62-65.

Mallat, N., Rossi, M., & Tuunainen, V.K. (2004). Mobile banking services. *Communications of the ACM, 47*(5), 42-46.

Manecke, N., & Schoensleben, P. (2004). Cost and benefit of Internet-based support of business processes. *International Journal on Production Economics, 87,* 213-229.

McGregor, C., & Morris, B. (2006). A survey of recent research to support remote neonatal

care via mobile devices. *Proceedings of the IMB Conference, Sydney, Australia*.

Patel, A. (2006). Mobile commerce in emerging economies. In B. Unhelkar (Ed.), *Mobile business: Technological, methodological and social perspectives.* Hershey, PA: Idea Group.

Raisinghani, M., & Taylor, D. (2006). Going global: A technology review. In Y.U. Lan (Ed.), *Global integrated supply chain systems.* London: Idea Group.

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Sarker, S., & Wells, J.D. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM, 46*(12), 35-40.

Schwiderski-Grosche, S., & Knospe, H. (2002). Secure mobile commerce. *Electronics & Communication Engineering Journal, 14*, 228-238.

Sheng-Tzong, C., Jian-Pei, L., Jian-Lun, K., & Chia-Mei, C. (2002). A new framework for mobile Web services. *Proceedings of the Symposium on Applications and the Internet* (SAINT 2002), Nara, Japan.

Stafford, T.F., & Gillenson, M.L. (2003). Mobile commerce: What it is and what it could be. *Communications of the ACM, 46*(12), 33-34.

Stanoevska-Slabeva, K. (2003). Towards a reference model for m-commerce applications. *Proceedings of ECIS 2003, Naples, Italy.*

Sun, J. (2003). Information requirement: Elicitation in mobile commerce. *Communications of the ACM, 46*(12), 45-47.

Tarasewich, P. (2003). Designing mobile commerce applications. *Communications of the ACM, 46*(12), 57-60.

Tian, M., Voigt, T., Naumowicz, T., Ritter, H., & Schiller, J. (2004). *Performance considerations for mobile Web services. Computer Communications, 27,* 1097-1105. Retrieved December 3, 2006, from http://www.elsevier.com/locate/comcom

Tsai, H.A.B., & Gururajan, R. (2005). Mobile business: An exploratory study to define a framework for the transformation process. In *Proceedings of the 10th Asia Pacific Decision Sciences Institution (APDSI) Conference, Taipei, Taiwan.*

Urbaczewski, A., Valacich, J.S., Jessup, L.M., & Guest Editors. (2003). Mobile commerce: Opportunities and challenges. *Communications of the ACM, 46*(12), 30-32.

Varshney, U. (2002). Mobile payments. *Computer, 35*(12), 120-121.

Varshney, U. (2004). Vehicular mobile commerce. *Computer, 37*(12), 116-118.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications*, 185-198.

Waegemann, P. (2006). Mobile solutions offer providers flexibility in managing care. *Managed Healthcare Executive, 16*(2), 58.

Wang, Y., Van der Kar, E., Meijer, G., & Hunteler, M. (2005). Improving business processes with mobile workforce solutions. In *Proceedings of the International Conference on Mobile Business, Sydney, Australia.*

This work was previously published in Handbook of Research on Global Information Technology Management in the Digital Economy, edited by M. Raisinghani, pp. 519-534, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.12
Optimal Number of Mobile Service Providers in India:
Trade-Off between Efficiency and Competition

Rohit Prasad
Management Development Institute, India

Varadharajan Sridhar
Management Development Institute, India

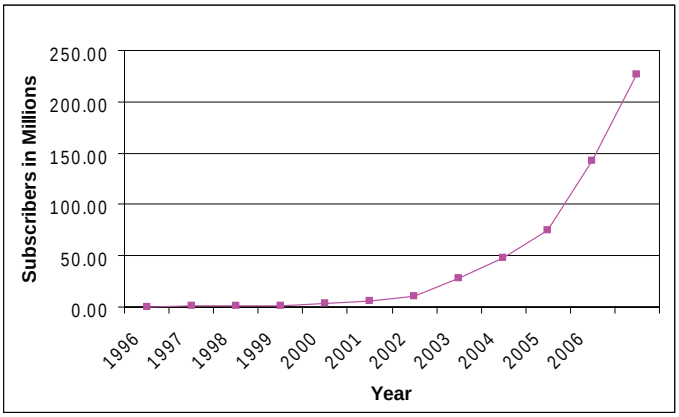
ABSTRACT

With 225 million subscribers, India has the world's third largest mobile subscriber base in the world. The Indian mobile industry is also one of the most competitive in the world with 4-7 operators in each service area. A large number of operators bring competition and its associated benefits such as decrease in price and hence corresponding growth of the market. On the other hand in the presence of economies of scale, too many operators may result in inefficient scales and high unit costs. This article analyses the trade-off between competition and economies of scale by estimating the production function for mobile subscribers and traffic carried. Analysis of panel data reveals the existence of economies of scale in the Indian mobile sector. We then derive an upper bound on the optimal number of operators in each license area and discuss policy implications.

INTRODUCTION

Quick deployment, competition, advancement in technologies, and reduced cost of access have propelled the growth of mobile services in India much like in other emerging countries. The Indian mobile subscriber base continues to grow and has reached about 225 million in December 2007 from about 142 million a year ago. Figure 1 illustrates the exponential growth of mobile services in India. India currently has the world's third largest mobile subscriber base, and is slated to exceed that of the U.S. by the end of this year to become the second largest in the world, next

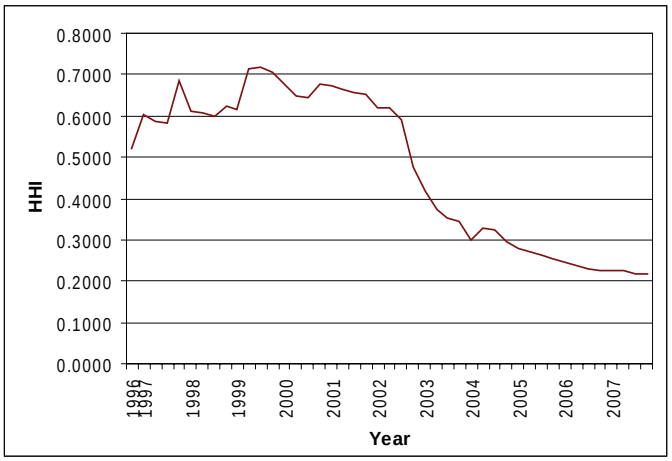
Figure 1. Growth of mobile services in India



only to China. The compounded annual growth rate of the mobile subscriber base has been 84.2% over the last 5 years. Revenue from cellular mobile services touched $12.5 billion for the fiscal year ending March 2007 (Voice & Data, 2007).

The Indian mobile industry is also one of the most competitive in the world. There are as many as 7 mobile operators in certain areas of the country. Figure 2 illustrates the amount of competition and market power as indicated by

Figure 2. HHI over time period across different categories of Service Areas



the Herfindahl Hirschman Index (HHI). (Viz. the lower the HHI, the higher the competition). As indicated, the overall HHI is about 0.2180 (as on December 2007) indicating very low market concentration.

A large number of operators bring competition and its associated benefits such as reduced prices, variety of products and services, and quality. On the other hand, in the presence of economies of scale, too many operators may result in inefficient scales and high unit costs (Nuechterlein & Weiser, 2005).

It is the objective of this article to analyze whether economies of scales are present in the Indian mobile industry using a classical production function approach. In the following section, we present an overview of the mobile industry in India including the method of allocating licenses. Next, we discuss the conceptual framework of economies of scale as applicable to the telecom industry. Thereafter, we discuss our model and illustrate the variables and data that we have used in the study. In the subsequent section, we discuss results of our analysis and finally conclude with limitations of our study and future research directions.

MOBILE SERVICES IN INDIA

As in China and the U.S., mobile operators in India are licensed to operate in designated geographical operating areas, referred to as "Licensing Service Areas (LSAs)." In India, there are 23 LSAs, which are categorized as metros, A, B and C. The categorization is based on the expected revenue potential with category C circles being the lowest in that order.

The licensing process for cellular mobile services started in India in 1992 with a duopoly market structure in each service area. Global Systems for Mobile (GSM) was mandated as the technology to be adopted. Beauty contest and single stage auction procedures were used as the allocation mechanisms for metros and category A, B, C circles, respectively, for the award of license and spectrum. The first digital cellular services started in the metros in 1995 (for details regarding the licensing procedure, the reader is referred to Desai (2006) and Jain (2001)). Subsequently, the third operator license was awarded to the government operator in 2001. The fourth operator license was further issued in 2000 using a three-stage auction procedure.

During this period, the government also liberalized the Basic Telecom Services (BTS) market, which typically provided traditional landline-based Plain Old Telephone Service. Jain and Sridhar (2003) provide details of the basic telecom services operations and its growth in India. In the year 2000, BTS operators approached the government with a proposal that they could provide local access loop at much lower cost using Code Division Multiple Access (CDMA) wireless technology. BTS operators argued that quick deployment of wireless CDMA service provides high spectral efficiency and lower per line cost compared to landline services, and hence is definitely a better alternative compared to wired access loop for certain areas of the country. This increased the competition in mobile services. After a couple of years of litigations between the BTS operators and cellular operators, the Indian government announced Unified Access Service (UAS) Licenses in November 2003 that allowed migration of basic service license holders to provide full mobility-based services with a stipulated entry fee (Sridhar, 2007b).

Today, mobile services in India are split between the GSM and CDMA operators with about 76% being GSM cellular mobile subscribers. Table 1 illustrates the categorization of circles in India and their corresponding subscriber base. In 2007, the Indian telecom regulator—Telecommunications Regulatory Authority of India (TRAI)—recommended no cap be placed on the number of operators for the provisioning of access (mobile, nonmobile) services (TRAI, 2007). The Indian

Table 1. Subscriber base across different services areas (as of December 2007)

Circles	GSM Subscribers	CDMA Subscribers	Total Subscribers
Metro LSA			
Delhi	10,114,314	5,187,769	15,302,083
Mumbai	8,328,136	1,679,099	10,007,235
Chennai	5,205,651	1,176,086	6,381,737
Kolkata	4,481,603	2,374,430	6,856,033
Category A LSA			
Maharashtra	12,800,578	5,300,118	18,100,696
Gujarat	11,959,678	3,174,521	15,134,199
Andhra Pradesh	12,627,111	5,500,706	18,127,817
Karnataka	12,006,799	3,330,097	15,336,896
Tamil Nadu	13,085,080	2,951,284	16,036,364
Category B LSA			
Kerala	7,652,577	2,364,978	10,017,555
Punjab	8,580,276	2,061,759	10,642,035
Haryana	4,186,038	1,527,740	5,713,778
Uttar Pradesh (W)	7,844,066	3,278,305	11,122,371
Uttar Pradesh (E)	11,053,797	3,575,118	14,628,915
Rajasthan	8,822,952	3,039,590	11,862,542
Madhya Pradesh	8,023,536	2,864,050	10,887,586
West Bengal, Andaman & Nicobar	6,311,324	1,580,648	7,891,972
Category C LSA			
Himachal Pradesh	1,696,024	247,305	1,943,329
Bihar	7,025,778	2,458,282	9,484,060
Orissa	3,467,335	790,771	4,258,106
Assam	3,048,446		3,048,446
North East	1,695,945		1,695,945
Jammu & Kashmir	1,868,820	185	1,869,005
Total	171,885,864	54,462,841	226,348,705

government also announced in 2007 that spectrum up to 25 MHz will be made available from the Department of Defence for commercial mobile services. The government soon was flooded with a large number of applications for mobile service licenses and the associated spectrum (refer to Sridhar, 2007c, for details).

In this context, it is important to analyze the presence of economies of scale in the mobile services market in India and accordingly calculate the optimal number of operators who can efficiently provide service.

ECONOMIES OF SCALE

The commonly held view is that competition is the most effective market structure to ensure low prices and high quality. However, in industries such as fixed line telecom services and electricity distribution, economies of scale and scope are large enough to warrant low levels of competition, even monopolies, to minimize unit costs. Telecommunication carriers face huge initial costs, including, for example, laying down copper lines from the Central Office to each subscriber location in case of basic fixed line services; constructing cell sites and Base Transceiver Stations (BTSs) in case of mobile services; and laying optic fiber cables to interconnect their access networks to backbone networks. These costs are both *fixed*, in that the operator must incur them up front before it can provide any volume of service, and *sunk*, in that, once made, the investment cannot be put to some other use. In contrast, the marginal cost of providing services to each additional customer, once the network is operational, is often less. Given the enormous fixed costs and negligible marginal costs, the carrier's long-run average costs within the defined geographical area may well decline with every increase in the size of the network. In other words, it is often cheaper for an operator to provide services to the one-millionth customer than to the one-thousandth customer (for details on economies of scale and scope in telecommunications, reader is referred to Nuechterlein & Weiser, 2005).

The presence of economies of scale poses a ticklish question for votaries of competition. Allowing a few firms to dominate a market would lead to greater efficiency in production, but at the risk of increased mark-ups over marginal cost on account of market power. In this article, we first establish the presence of economies of scale in the Indian mobile industry and then estimate the number of operators.

THE MODEL

Production functions of various forms have been applied in telecommunications (see Bloch, Madden, & Savage, 2001; Eldor, Sudit, & Vindod, 1979; Fishelson, 1977; Pentzaropoulos & Geikos, 2002; Sudit, 1973; Vinod, 1972). However, there are criticisms of the validity of some of the complex non-homogenous functional forms including Constant Elasticity of Substitution (CES) and corresponding translog functions (Bloch et al., 2001; Fishelson, 1977). We apply the classical production function of Cobb-Douglas form used by many researchers to determine the presence of economies of scale in the Indian mobile industry.

We estimate a production function of the Cobb-Douglas form for mobile services using a panel of data collected over 7 years across all the 23 LSAs for different GSM operators [1] providing services in their respective circles (Prasad & Sridhar, 2007a). A function of this form was used by Roller and Waverman (2001) and Sridhar and Sridhar (2007) in supply side estimation of telecom services growth. The production function is specified as follows:

$$X = Ay^{\beta}z^{\gamma} \qquad (1)$$

In the above equation, the dependent variable X refers to mobile subscriber base or the traffic carried on the network. The two important factor inputs considered as the independent variables include: (i) allocated amount of spectrum that provides the required channel capacity for traffic (y) and (ii) deployed mobile infrastructure such as Base Transceiver Stations (BTS) which provides connectivity to mobile handsets (z).

Taking log on both sides, (1) can be expressed as:

$$\ln X = \ln A + \beta \ln y + \gamma \ln z \quad (2)$$

We estimate (2) in two regression equations: (i) using mobile subscriber base and (ii) the Minutes of Usage (MoU) which is representative of the traffic carried on the network as the dependent variable. We describe below the variables in detail.

Subscriber Base and Minutes of Usage

Operator-wise subscriber base data was used in the model across all the 23 LSAs. Figure 3 illustrates the growth of mobile subscriber base over different categories of LSAs over time.

Although subscriber base is a noncontroversial, transparent metric to determine growth of mobile services, it is argued that traffic generated may be a better measure (Sridhar, 2006a). We calculate the traffic generated using average Minutes of Usage (MoU) per month per subscriber. It is pointed out in Sridhar (2007a) that MoU in India is one of the highest in the world compared to matured markets such as Finland. The MoU of prepaid subscribers are relatively less than that of post-paid. However, we take blended MoU for calculating the average traffic generated on the mobile network. Figure 4 illustrates the trend in MoU for mobile services in India.

Although there is a general increase in MoU, there is a decrease in rate of growth of MoU from 2006 onward. This is due to an increase in less intensive and less Average Revenue per User (ARPU) users subscribing to mobile services.

Radio Frequency Spectrum

Mobile operators are allotted certain bandwidth of radio frequency spectrum for offering mobile

Figure 3. Growth of subscribers across different categories of LSAs over time

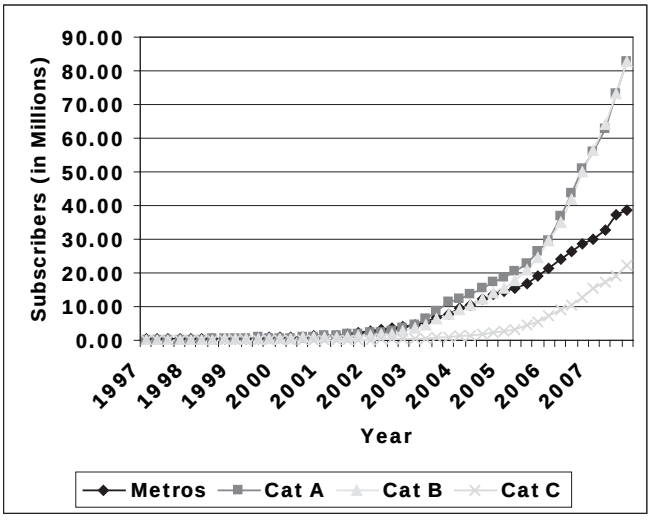
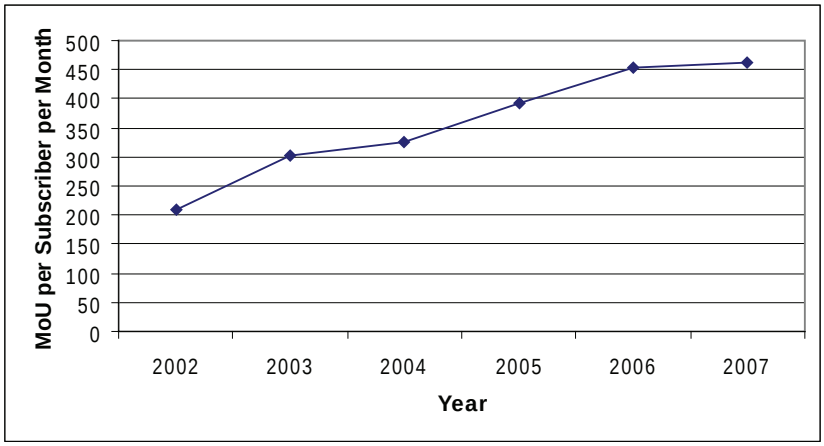


Figure 4. Trends in minutes of usage of mobile services



services using mechanisms such as auction and beauty contest. The management of radio frequency spectrum is important to meet the following objectives: a) granting of exclusive rights to spectrum, (b) ensuring efficient use of it, and c) promoting competition in services (Sridhar, 2006c). The costs experienced by the service provider may depend directly on how much spectrum is available and hence is an important factor input to the provisioning of mobile services (Hills & Yeh, 1999). Hills and Yeh (1999) point out that per subscriber investment falls with increasing available spectrum. However, they point out to diminishing returns as more spectrum is added. The allocation procedure has affected the amount of spectrum available for each operator over the time period studied. Hence, a brief word on the allocation rules follows.

Spectrum Allocation for Mobile Services in India

Before the introduction of mobile services, spectrum was entirely in the control of India's Department of Defense. In 1995, 4.4 MHz of spectrum each was auctioned to two operators in the 900 MHz band to provide GSM-based services. Spectrum has been made available as and when defense vacated spectrum. In 2000, the government operators were also given the third operator license and start-up spectrum for their GSM services.

In 2001, when the fourth operator was introduced, start-up spectrum of 4.4MHz was allotted in the 1800 GSM band through auction. In 2003, when CDMA operators were allowed, a start-up spectrum of 2.5 MHz was given in the 800 MHz band to each operator. Further allocation of spectrum was based on network rollout and subscriber base of the operators. In 2005, the regulator TRAI came out with its spectrum policy recommendations in which subscriber-based norms (viz. minimum subscriber base required for additional carriers) were introduced for additional allocation of spectrum to incumbents. The existing subscriber base criteria for additional spectrum allocation are illustrated in Table 2[2].

Table 2. Spectrum allocation criteria for mobile services in India(Source: TRAI, 2007)

LSA	2 × 6.2 MHz	2 × 8 MHz	2 × 10 MHz	2 × 12.4 MHz	2 × 15 MHz
Delhi/Mumbai	0.3	0.6	1.0	1.6	2.1
Chennai/ Kolkata	0.2	0.4	0.6	1.0	1.3
A	0.4	0.8	1.4	2.0	2.6
B	0.3	0.6	1.0	1.6	2.1
C	0.2	0.4	0.6	0.9	1.2

Minimum CDMA Subscriber Base (in Millions) Criteria

LSA	2 × 3.75 MHz	2 × 5 MHz	2 × 6.25 MHz	2 × 7.5 MHz
Delhi/Mumbai	0.3	1.0	1.6	2.1
Chennai/ Kolkata	0.2	0.6	1.0	1.3
A	0.4	1.2	2.0	2.6
B	0.3	1.0	1.6	2.1
C	0.15	0.5	0.9	1.2

Minimum GSM Subscriber Base (in Millions) Criteria

Mobile Network Infrastructure

Access to mobile services is provided through Base Transceiver Stations (BTSs) located in each cell site. The efficiency of a given amount of spectrum can be enhanced by creating smaller cells (viz. by reducing intersite distance) and increasing reuse of frequency. This method increases the number of BTSs in the network. Hence, for a given amount of spectrum, the number of BTS is a factor input that can affect the traffic density carried in the area or the number of subscribers supported. However, increase in BTS also increases the capital expenditure for the operators.

TRAI recommended infrastructure sharing as a way to reduce duplication in infrastructure and expedite the penetration of mobile services, especially in rural and remote areas of the country (TRAI, 2005). However, until recently both active and passive infrastructure could not be shared by the operators as the policy favoured deployment of telecom infrastructure in the country. Only recently, the Indian mobile operators were allowed to share the passive infrastructure (viz. tower space, electricity supply to the BTS). Hence, number of BTS in an LSA is considered as another important factor input in our model.

Although in earlier studies by Roller and Waverman (2001) and Sridhar and Sridhar (2007), "telecom investment" was considered as one of the important supply side input to the provisioning of telecom services, we use number of BTSs as indicated above as the factor input. Moreover, LSA data on telecom investment is not available. It is expected that larger the number of cell sites or BTS, larger is the investment and hence its

effect on mobile services growth. Figure 5 gives the growth of BTS across different categories of LSAs. As can be seen from the figure, the growth pattern for BTSs closely resembles growth of mobile subscriber base.

Table 3 gives the regression results for all the circles as well as individually for metros, category A, Category B and Category C circles, respectively. The regression results in all cases clearly indicate the existence of economies of scale and scope in the mobile industry across all categories of circles (i.e., $\beta + \gamma \succ 1$). This means that unit cost decreases with increasing scale. Hence, larger firms will be able to operate more efficiently using the factor inputs.

The above finding may seem to be counter intuitive in view of the fact the Indian operators continue to sustain profit levels despite falling usage charges and increased market fragmentation. The paradox is resolved by examining the rapid increase in the size of the market. Figure 6 shows the increase in subscribers per operator over 1996-2006 period for service areas, which have been witnessing the lowest HHI.

The graph clearly indicates that the operators are leveraging on economies of scale to remain profitable. This is especially true of large operators, which are becoming more profitable every year with profit margins of more than 40%.

UPPER BOUND ON THE OPTIMAL NUMBER OF OPERATORS

Assuming the standard "U" shaped average cost function, each firm *j* minimizes the average cost by producing c_j units of the product/service. Given an estimate of the total demand *T* in the market, we can then determine the optimal number of such operators $n = T/c_j$. However, because the Indian mobile market is still evolving, it is difficult to empirically determine c_j where the operator's cost function is minimized. Hence, instead of

Figure 5. Growth of BTSs in India across different categories of LSAs

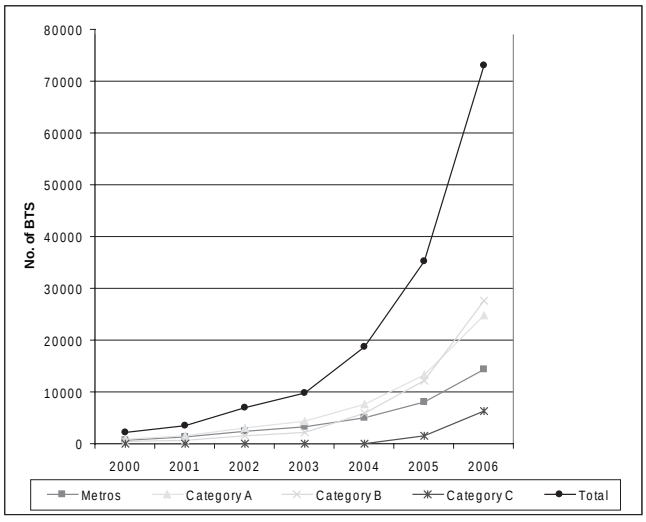
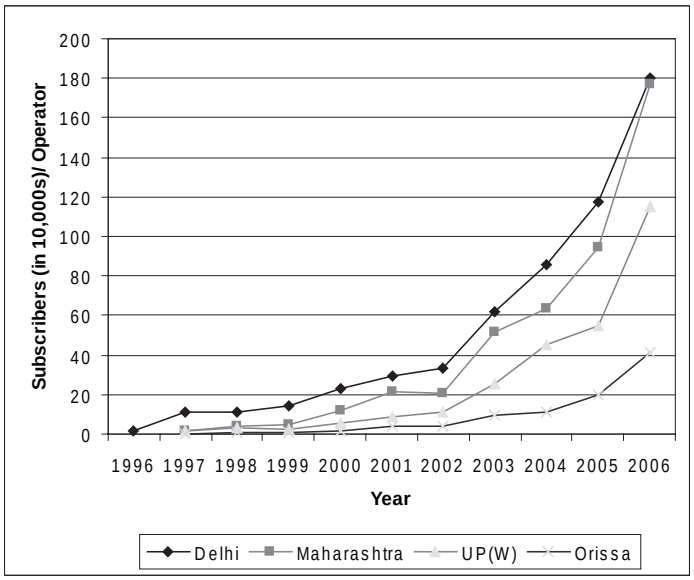


Table 3. Regressions results

Regression Results (for All Circles)				
Independent Variable	Dependent Variable: Subscribers (in Millions): N=387; Adj R^2=0.887; F=1512.93; p<0.00		Dependent Variable: Traffic (in Millions Minutes of Usage)>: N=307; Adj R^2=0.851; F=871.284; p<0.00	
	Coefficients	T (Sig)	Coefficients	T (Sig)
Constant	-8.097	-56.380 (0.000)	-3.049	-14.716 (0.000)
Spectrum	1.526	14.485 (0.000)	1.632	13.294 (0.000)
BTS	0.753	30.731 (0.000)	0.859	25.697 (0.000)
Regression Results (for Metros)				
Independent Variable	Dependent Variable: Subscribers (in Millions): N=88; Adj R^2=0.930; F=576.723; p<0.00		Dependent Variable: Traffic (in Millions Minutes of Usage): N=64; Adj R^2=0.901; F=290.914; p<0.00	
	Coefficients	T (Sig)	Coefficients	T (Sig)
Constant	-7.562	-36.524 (0.000)	-3.394	-9.010 (0.000)
Spectrum	1.668	9.548 (0.000)	2.174	9.971 (0.000)
BTS	0.611	11.961 (0.000)	0.710	11.065 (0.000)
Regression Results (for Category A circles)				
Independent Variable	Dependent Variable: Subscribers (in Millions): N=117; Adj R^2=0.902; F=534.344; p<0.00		Dependent Variable: Traffic (in Millions Minutes of Usage): N=88; Adj R^2=0.855; F=257.641; p<0.00	
	Coefficients	T (Sig)	Coefficients	T (Sig)
Constant	-8.437	-34.547 (0.000)	-3.891	-9.938 (0.000)
Spectrum	1.717	8.123 (0.000)	2.027	8.245 (0.000)
BTS	0.752	14.206 (0.000)	0.868	11.174 (0.000)

Regression Results (for Category B circles)				
Independent Variable	Dependent Variable: Subscribers (in Millions): N=114; Adj R^2=0.877; F=491.770; p<0.00		Dependent Variable: Traffic (in Millions Minutes of Usage): N=113; Adj R^2=0.875; F=395.705; p<0.00	
	Coefficients	T (Sig)	Coefficients	T (Sig)
Constant	-8.502	-27.271 (0.000)	-3.891	-10.632 (0.000)
Spectrum	1.884	7.693 (0.000)	2.018	8.018 (0.000)
BTS	0.719	17.209 (0.000)	0.889	17.310 (0.000)
Regression Results (for Category C circles)				
Independent Variable	Dependent Variable: Subscribers (in Millions): N=43; Adj R^2=0.808; F=89.177.770; p<0.00		Dependent Variable: Traffic (in Millions Minutes of Usage): N=40; Adj R^2=0.764; F=65.871; p<0.00	
	Coefficients	T (Sig)	Coefficients	T (Sig)
Constant	-8.337	-12.740 (0.000)	-2.397	-3.306 (0.002)
Spectrum	1.334	3.493 (0.000)	1.439	3.634 (0.001)
BTS	0.872	11.494 (0.000)	0.868	9.755 (0.000)

Figure 6. Increase on subscriber base in select service areas



calculating the optimal number of operators in the market, we determine a conservative upper bound on the optimal number of operators.

We assume that the economies of scale would cease to exist beyond the range of subscriber minutes observed in the sample. The highest subscriber minutes observed were C (=2,234 million subscriber minutes) in our sample. We hence assume that the unit costs would be minimized at C in every circle. Using equation (3) we calculate UB_i, the upper bound on the number of operators at service area i for the years 2007 and 2010. In equation (3), T_i is the total traffic in service area i which is calculated from total subscriber base (including both GSM and CDMA subscribers) in each service area i and the average MoU[2].

$$UB_i = T_i / C \qquad (3)$$

Results are presented in Table 4. The table indicates that in 2007, there is no LSA in which the upper bound on the optimal number of operators is greater than three. In all LSAs the optimal number of operators is significantly less than the actual number, which is 6-7 on average.

It is to be noted that if our assumption does not hold good and that increasing returns to scale persist beyond the range of our sample, then the average cost function will have a minima at a capacity higher than C (=2,234 million subscriber minutes), thus lowering the optimal number of operators. Hence, our estimate is a conservative upper bound on the optimal number of operators. The number of operators in most of the LSAs in 2007 was higher than even this upper bound.

However, as is seen in Figure 1, the mobile subscriber growth continues to grow, thus increasing T_i which might result in larger values of the optimal number of operators in the future. We calculate the optimal number of operators for the year 2010 based on the target of 500 million mobile subscribers set by the regulator and the

Table 4. Upper bound on optimal number of operators in 2007

LSA	Subscriber Base (in Millions as of Sep. 2007)	Existing Number of Operators	Upper Bound on Optimal Number
Metros			
Delhi	13.9585	6	3
Mumbai	11.2403	6	2
Chennai	5.6824	6	1
Kolkata	6.1996	6	1
Category A			
Maharashtra	15.8119	6	3
Gujarat	13.6236	6	3
Andhra Pradesh	16.2266	6	3
Karnataka	13.7761	6	3
Tamil Nadu	13.6789	6	3
Category B			
Kerala	9.1175	6	2
Punjab	9.7207	7	2
Haryana	5.3041	6	1
Uttar Pradesh (West)	9.8635	6	2
Uttar Pradesh (East)	12.8670	6	3
Rajasthan	10.4260	7	2
Madhya Pradesh	9.1448	6	2
West Bengal, Andaman & Nicobar	6.7733	6	1
Category C			
Himachal Pradesh	1.6544	7	1
Bihar	8.0538	6	2
Orissa	3.6719	6	1
Assam	2.5982	4	1
North East	1.4775	4	1
Jammu & Kashmir	1.6732	4	1
	202.5436		

ministry (TRAI, 2007). We first forecast the subscriber base in service area *i* which is part of category *k* as follows:

$$N_{i,2010} = M_{ik,2007} \times MD_{k,2010} \times Pop_{k,2010} \quad (4)$$

where $N_{i,2010}$ is the projected number of subscribers in service area *i* in year 2010; $M_{ik,2007}$ is the market share of service area *i* in the corresponding category *k* in year 2007; $MD_{k,2010}$ is the assumed mobile density per 100 population (100%, 50%, 50% and 40% in metros, category A, B and C service areas, respectively) in category *k* in year 2010; and $Pop_{k,2010}$ is the projected population of category *k* in year 2010[3]. The assumed mobile density $MD_{k,2010}$ satisfies the constraint that the total number of subscribers across all categories equals the number projected by the regulator and the ministry. Using $N_{i,2010}$, we then calculate the upper bound on the optimal number of operators using equation (3). Results are presented in Table 5. In 2010, the upper bound on the optimal number varies from 2 to 9.

POLICY IMPLICATIONS

Mergers and Acquisitions

The calculation of the optimal number of operators should not be viewed as a recommendation to place a cap on the number of operators. Here the amount of spectrum available and the minimum spectrum required for services should be the only consideration for restricting entry. If there are too many operators then the market will eventually, through a process of mergers and acquisition, converge on the optimal number. It is pointed out in Grover and Saeed (2003) that one of the main reasons for mergers, acquisitions and partnerships in telecom market is for attaining economies of scale and scope. Sridhar (2006b) illustrates many inter-LSA acquisitions by large mobile operators in the country for the same reason. Hence, the regulator's main concern should be to put in place the appropriate mergers and acquisitions guidelines.

In a "contestable market" even the specification of a floor on the number of operators is not necessary, as the threat of entry is sufficient to eliminate market power. However, after elaborating on the concept of contestability we will argue that the Indian mobile industry is not perfectly contestable. Hence, the specification of a floor becomes necessary, and it is here that our calculation of the upper bound on the optimal number of operators has important policy implications.

Contestable Market

Traditionally, measures such as the HHI are used by regulators to monitor competition in the marketplace. The U.S. Department of Justice, for example, uses a high HHI as a measure for raising antitrust concerns during mergers and acquisitions (Prasad & Sridhar, 2007b).

The theory of price competition by Bertrand (1883) notes that the presence of even two firms is sufficient to prevent the exercise of market power, thus generating a perfectly competitive behavior. Baumol (1982) espoused "contestability" as a broader benchmark of competition than market power as measured by the HHI. A "contestable" market, is one in which entry and exit of firms are free. In this case, the incumbent(s) face a real threat of rapid entry and exit of firms if they make supernormal profits. Hence, a contestable market never offers more than a normal rate of profit and is characterised by the absence of any sort of inefficiency in production.

As mentioned earlier, in a perfectly contestable market even the stipulation of a floor on the number of operators is not necessary. However, the telecom market in India is not contestable for the following reasons (TRAI, 2007):

a. Entry of new operators is difficult, as they have to get license and the scarce spectrum before starting their operations;
b. Exit of the operators is difficult as they have rollout obligations. For example, each UASL operator should cover up to 90% of metro service area within 1 year of the effective date of licensing. There are similar roll out

Table 5. Projection of upper bound on optimal number of operators in 2010

LSA	**Projections for 2010**	
	Projected Subscriber Base (in Millions)	**Estimated Upper Bound on Optimal Number of Operators**
Metros		
Delhi	17.0000	4
Mumbai	20.0000	4
Chennai	8.0000	2
Kolkata	15.0000	3
Category A		
Maharashtra	21.1280	4
Gujarat	19.0138	4
Andhra Pradesh	22.2521	5
Karnataka	19.0950	4
Tamil Nadu	18.5110	4
Category B		
Kerala	29.5892	6
Punjab	32.8379	7
Haryana	18.2486	4
Uttar Pradesh (West)	32.7392	7
Uttar Pradesh (East)	41.2319	9
Rajasthan	33.5115	7
Madhya Pradesh	29.9194	6
West Bengal, Andaman & Nicobar	21.9223	5
Category C		
Himachal Pradesh	8.7896	2
Bihar	40.6363	8
Orissa	19.0463	4
Assam	14.6843	3
North East	7.8920	2
Jammu & Kashmir	8.9514	2
	500.0000	

obligations for other categories of LSAs as well; and

c. The entrants have to incur large capital expenditure, as towers, electronics and net-

works need to be put in place before starting services.

In this context, the regulatory approach should be to promote contestability wherever possible, without excessively curtailing the scale of operation, as a low HHI may mean high unit costs dues to diseconomies.

Measures available to the regulator for increasing contestability include:

a. Introduction of Mobile Virtual Network Operators (MVNOs) who lease radio access from the existing mobile network operators and provide mobile voice and data services; and
b. Introduction of mobile number portability, which allows subscribers to shift to different operators at ease.

The regulator can also conduct significant market power assessment as done in countries in EU using benchmarks from circles where the market size allows a low HHI and imposing obligations on operators as appropriate (Sridhar, 2007a).

However, given that these measures will not make the market perfectly contestable, attention also needs to be paid to ensure participation of a minimum number of players in the market. Here, based on our analysis of the production function, we recommend that the regulator needs to specify a floor which is not greater than the upper bound on the optimal number calculated for the year 2010. Note that in no circle is the upper bound less than two. Hence, the floor recommended by us is compatible with a minimum level of competition.

In contrast to our approach, the regulator in India recommended the following (TRAI, 2007):

a. The minimum number of wireless access service providers in each service area should be four;
b. The market share of merged entity in the mobile access market should be capped at 40%; and
c. The cross equity holding in the same circle should be capped at 20%.

The above guidelines seem to be pro-competition. However, they will result in a high cost structure for the Indian mobile services industry and may not result in reduced price for the subscriber (Prasad & Sridhar, 2007b). Allowing efficient scale of production along with promoting contestability as suggested earlier would be the ideal policy to serve the best interests of the consumer.

CONCLUSION

The existing literature while addressing competition and growth of mobile services, does not analyze in depth the economies of scale and industry efficiency. This is one of the first such attempts to analyze the trade-offs between low market power and economies of scale for sustained growth of mobile services in the country. Our analysis of the data on mobile services in India indicates the existence of economies of scale in this sector. We also calculate the upper bound on the optimal number of operators in each License Service Area so that policies that make appropriate trade-offs between competition and efficiency can be formulated.

Note that there are other factor inputs such as labor that are relevant (Sung & Gort, 2000). In the absence of data on employment at circle level, we crosschecked the conclusion of increasing returns using available data on operating expense per subscriber. Data shows significant negative correlation between operating expense per subscriber and number of subscribers. This implies that our estimation of economies of scale is relatively conservative.

Because of limitations on availability we could not include CDMA data in our data set. Although CDMA service contributes to less than 25% of the market share, technology efficiencies of CDMA service is different from that of GSM service. Extension of this work could include the CDMA data as well as for accurate identification of economies of scale in the entire mobile industry.

An interesting extension of this work could be applied to the mobile industry in China where only two operators (viz. China Mobile and China Unicom) provide services in 31 LSAs. Sridhar (2007b) contrasts the market structure between China and India and points out that despite having an HHI of more than 0.5, Chinese mobile market continues to grow with affordable prices.

REFERENCES

Baumol, W. (1982). Contestable markets: An uprising in the theory of industry structure. *The American Economic Review,* 1-15.

Bertrand, J. (1883). Theorie Mathematique de la Richesse Sociale, *Journal des Savants*, 499-508.

Bloch, H., Madden, G., & Savage, S. (2001). Economies of scale and scope in Australian telecommunication. *Review of Industrial Organization, 18*(2), 219-227.

Department of Telecommunications (DoT). (2000). *Guidelines for the issue of license for cellular mobile telephone service.* Retrieved May 9, 2008, from http://www.dotindia.com

Department of Telecommunications (DoT). (2005). *Guidelines for unified access service license*. Retrieved May 9, 2008, from http://www.dotindia.com

Desai, A. (2006). *India's telecommunications industry: History, analysis, diagnosis*. New Delhi, India: Sage.

Eldor, D., Sudit, E.F., & Vindod, H.D. (1979). Telecommunications, a CES production function: A reply. *Applied Economics, 11*, 133-138.

Fishelson, G. (1977). Telecommunications, a CES production function. *Applied Economic, 9*, 9-18.

Grover, V., & Saeed, K. (2003). The telecommunication industry revisited. *Communications of the ACM, 46*(7), 119-125.

Hills, A., & Yeh, H.Y. (1999). Spectrum use and carrier costs: A critical trade-off. *Telecommunications Policy, 23*, 569-584.

Jain, R.S. (2001). Spectrum auctions in India: Lessons from experience. *Telecommunications Policy, 25*, 671-688.

Jain, P., & Sridhar, V. (2003). Analysis of competition and market structure of basic telecommunication services in India. *Communications & Strategies, 52*, 271-293.

Nuechterlein, J., & Weiser, P. (2005). *Digital crossroads*. Cambridge, MA: MIT Press.

Pentzaropoulos, G.C., & Giokas, D.I. (2002). Comparing the operational efficiency of the main European telecommunications organizations: A quantitative analysis. *Telecommunications Policy, 26*, 595-606.

Prasad, R., & Sridhar, V. (2007a, September 12). Mobile competition: Fewer the better. *Economic Times.* Retrieved May 9, 2008, from http://www.economictimes.com

Prasad, R., & Sridhar, V. (2007b, November 2). For a contestable mobile market. *Economic Times*. Retrieved May 9, 2008, from http://www.economictimes.com

Roller, R.L., & Waverman. (2001). Telecommunications infrastructure and economic development: A simultaneous approach. *American Economic Review, 91*(4), 909-923.

Sridhar, V. (2006a, January 3). Link spectrum allocation to actual traffic generation. *The Financial*

Express. Retrieved May 9, 2008, from http://www.financialexpress.com

Sridhar, V. (2006b, September 12). Mega telecom partnerships. *Financial Express*. Retrieved May 9, 2008, from http://www.financialexpress.com

Sridhar, V. (2006c, September 29). To allocate spectrum, study real estate. *Financial Express*. Retrieved May 9, 2008, from http://www.financialexpress.com

Sridhar, V. (2007a). *Analyzing the future growth of mobile communication services in developing countries such as India: Lessons from Finland*. MDI Working Paper (No. 004). Gurgaon, India: Management Development Institute.

Sridhar, V. (2007b). Analyzing the growth of mobile telecommunication services across service areas of China and India. In R. Garg & M. Jaiswal (Eds.), *Bridging digital divide* (pp. 46-65). New Delhi, India: Macmillan Advanced Research Series.

Sridhar, V. (2007c, November 6). How should spectrum be allocated. *Economic Times*. Retrieved May 9, 2008, from http://www.economictimes.com

Sridhar, K., & Sridhar, V. (2007). Telecommunications infrastructure and economic growth: Evidence from developing countries. *Applied Econometrics and International Development, 7*(2), 37-60.

Sudit, E.F. (1973). Additive non-homogeneous production functions in telecommunications. *The Bell Journal of Economics and Management Science, 4*, 499-514.

Sung, N., & Gort, (2000). Economies of scale and natural monopoly in the U.S. local telephone industry. *The Review of Economics and Statistics, 82*(4), 694-697.

Telecommunications Regulatory Authority of India (TRAI). (2005). *Recommendations on the growth of telecom services in rural India*. Retrieved May 9, 2008, from http://www.trai.gov.in

Telecommunications Regulatory Authority of India (TRAI). (2007). *Recommendations on review of license terms and conditions and capping of number of access providers*. Retrieved May 9, 2008, from http://www.trai.gov.in

Vinod, H.D. (1972). Non-homogenous production functions and applications to telecommunications. *The Bell Journal of Economics and Management Science, 3*, 531-543.

Voice & Data. (2007). *Cellular: Gaining strength*. Retrieved May 9, 2008, from http://www.voicendata.com

ENDNOTES

1 Due to lack of data availability for CDMA operations, we restrict our analysis for the GSM services. We assume that the production function we estimate for GSM services that accounts for more than 75% of the market equally holds good for the CDMA operations as well.

2 Note that C is calculated based on GSM subscriber base and MoU data. However, while calculating the upper bound on the number of operators, we use the total subscriber base (including both GSM and CDMA) assuming that CDMA technologies exhibit increasing returns over the same range. We round off the results of equation (3) as appropriate.

3 For example, if the population for metros is expected to be 80 million in 2010, and the market density for metro category is assumed to be 100% in 2010 and if Delhi has currently got 30% of the total number of subscribers present in metros, then the number of subscribers in Delhi will be 24 million in 2010.

This work was previously published in International Journal of Business Data Communications and Networking, Vol. 4, Issue 3, edited by J. Gutierrez, pp. 69-88, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 6.13
Evolution of Telecommunications and Mobile Communications in India:
A Synthesis in the Transition from Electronic to Mobile Business

Chandana Unnithan
Deakin University, Australia

Bardo Fraunholz
Deakin University, Australia

ABSTRACT

Electronic business is a concept that has been adopted by businesses all over the world. The developing world takes it as a viable economic opportunity to catch up with other economies. A significant underlying factor in this development is the evolution of telecommunication infrastructure, especially in developing economies. In this chapter, we have synthesized this critical evolution in India. In the process, we found that there is a second layer of evolution into mobile communications and subsequently mobile business, which is gaining momentum in India. We conclude with an outlook for the future for these developments.

INTRODUCTION

The electronic business revolution, an idea that caught the imagination of many businesses, governments and individuals across the world during the second millennium has now become a reality (Chen, 2001). Many economies across the globe have transitioned their business processes and service delivery into the electronic mode, ushering

in the digital or rather electronic business era. It is now widely accepted by policy makers, enterprises and societies that information communication technologies are at the centre of an economic and social transformation that is affecting all countries (UNCTAD, 2003). Lately, the technological advances in mobile communications, which form part of the information communications infrastructure, has caught the attention of many an economy. While it is a natural transition for developing nations, to use mobile technologies to facilitate electronic businesses and progress from electronic to mobile business, developing nations find potential in the low-cost, convenient infrastructure it offers.

Among the world's population, more than 80 percent live in developing countries where socio-economic progress continues to be slow, due to a variety of reasons such as poor infrastructure, low education, etc. (Splettstoesser, 2002). In this context, India is a geographically disparate developing economy, with a population of over 1.2 billion people spread over 35 states, speaking different languages, relatively unequal in the distribution of wealth, education and progress, is finding the transition of its domestic economy into the digital world rather challenging. To begin with, there was regulation not permitting foreign direct investment in the country for decades until the 1990s. However, during the latter half of 1990s, with the deregulated telecommunications industry opening up to private competition, combined with the federal regulations permitting foreign direct investment into the economy, the domestic economy began its transition. This transition has been enabled by the government and the growing middle-class information technology professionals, who found potential for progress through information communications technologies. Lately, it has been realized that mobile communications provide a low-cost infrastructure which provides for economic progress within the economy by enabling electronic business processes beginning with electronic government delivery to revenue generating models for network operators.

In this chapter, we aim at capturing the transition of India into the digital world by closely examining the key influencers, i.e., the evolution of telecommunications and mobile communications. In the process, we have also touched upon other influencers in the evolution, such as the effect of information technology and some government initiatives. This chapter is meant to inform academia, policy makers as well as all concerned forums involved in developing nations. This chapter offers an example for other developing nations who wish to exploit the power of telecommunications and especially mobile communications, in their economic progress as well as their transition into electronic business processes.

BACKGROUND

Examining the impact of globalization and the Internet on developing nations, Kshetri (2001) contends that the main factors that lead to the explosion of e-commerce include the development of better and faster computer technology, the creation of more user-friendly software, people's trust in electronic transactions, and low costs. However, some factors such as cultural beliefs, lack of computer literacy, technological infrastructure, and government policies are a major deterrent for the spread of e-commerce in developing nations. The UNCTAD (2003) report, which analyzed developing nations, confers with this view, suggesting that developing nations have manifold challenges such as coping with new technologies as well as exploiting their full potential, and managing embedded logical relationships with the developing world in their transformation to knowledge societies. This section leads into various subsections that briefly examines and traces the evolution of India in this context.

India, being a developing nation, has many interesting challenges. The main issues have been

the availability of bandwidth and power across the nation (Unnithan, 2002). Although the situation has improved over the past five years, as of March 2003, there were only 13 million Internet users in India out of a population count of over 1.2 billion (ITU, 2003). It is commonly believed that these issues impact computer literacy and obstruct technology absorption into the society. Paradoxically, India is one of the largest exporters of software, human capital and information technology enabled products in the world, which sets it apart from other developing nations (NASSCOM, 2002). Therefore, it can be established that computer literacy and availability of technologies does not have a direct correlation in India.

If the impact of government policy is examined, over the fragmented population, there is a visibly growing middle-class layer of information technology professionals, or knowledge workers, who seem to be driving the technology-enabled processes within the economy by influencing government policies (Unnithan, 2002). A small example is that these professionals, who now seem to be the growing working population, are impatient to stand in a queue to pay something as mundane as a utility bill, which usually requires a day's leave. Over the past decade, they have influenced the government to facilitate technology-enabled processes. The economy seems to be in the midst of a subtle social revolution which is pushing it into the digital world (Unnithan, 2002).

As the UNCTAD (2003) report rightly suggests, government intervention is necessary for universal access to facilities and to provide a climate for foreign direct investment within the country, which in turn helps e-commerce growth. The form of political system and legal laws adopted by people around the world may also contribute to a low enrollment in global e-commerce (Kshetri, 2001). Countries with authoritarian governments may interfere with their constituents' freedom of speech and association. Such governments can also impose great trade tariffs on business conducted via the Internet to discourage and control global e-trade. In the following subsection we briefly trace the impact and evolution of information technology sector and policies that impacted on telecommunications reform, which the government had a significant role to play, as a preamble to our major thrust in the chapter.

TRANSITION AND IMPACT OF INFORMATION COMMUNICATION TECHNOLOGY REVOLUTION

The Information Communication Technology revolution in India had its beginnings in 1975, when the government of India strategically decided to take effective steps for the development of information systems and the utilization of information resources (Moni & Vijayaditya, 2002). The federal government of India, with a view to informatics-led development, decided to introduce decision-support systems within government ministries and departments, mainly to facilitate planning/implementation of socio-economic programs during the fifth planning period. The National Informatics Centre under the Electronic Commission/ Department of Electronics was the outcome of this view and was assisted by the United Nations Development Program (UNDP).

The 1990s were a period of rapid development in the technology-based industries, and de-regulation of markets following the removal of protection by the government lead to the growth of entrepreneurial activity. These developments were supported by the growing levels of expertise in information technology; venture capitalism and increasing amounts of foreign investment (Reddy, 2000). Toward the end of 1990s, with the opening up of the economy, deregulation and privatization, India became a favored destination for software development due to cheap labor and highly skilled manpower. Investment in knowledge-based industries was expected to

boost India's dominance in the next millennium (Ministry of Technology, 1999).

The formation of the National Association of Software Service Companies in the early 1990s reflects on India's strength in this sector (Nasscom, 2002). This association along with the government has been a catalyst in forming the "software backbone" of India. The Indian IT industry has grown from USD 1.73 billion in 1995 to USD 13.5 billion in 2002. In terms of GDP, the figures have risen from 0.59 percent to 2.87 percent. The export orientation of the information technology industry in India is very evident from these figures. In 1999, taking in the all around socio-economic growth with the emerging digital economy, the government created a new Ministry of Technology by merging the Department of Electronics and the National Informatics Centre with the Electronics and Software Export Promotion Council (Nasscom, 2002). The Ministry of Technology envisioned Internet-based information facilitation for the common public by various government agencies at all levels to be made available by 2005, the establishment of 100 million Internet connections and one million Information kiosks (i.e., one to two connections per village) by 2008 with private sector and unorganized sector participation, promotion of Indian language content over the Internet, re-engineering of government processes leading to e-governance and launching of a mass campaign on IT awareness (Ministry of Technology, 2000).

The Centre for Development of Advanced Computing (C-DAC) was one of the pioneering ventures of the Government of India in the early 1990s to promote research in the area of advanced computing. Towards the end of 1990s, this organization became a catalyst in developing multilingual software that facilitated e-governance India-wide (C-DAC, 2002). India's software and services industry grossed annual revenue of USD 8.26 billion during 2000-01, with its export market taking USD 6.2 billion. One out of every four global software giants outsourced their mission critical software requirements to India in 2000-01. The sector also accounted for almost 2 percent of the country's GDP (Nasscom, 2002). Parallel to these developments was the Internet growth in terms of subscribers projected to touch 1.5 million in 2002. Within India, the Central/State administrations, insurance companies, financial institutions, defence segment, the public tax system, ports, customs, telecommunications and educational institutions rapidly adopted e-governance, thus boosting domestic software revenues (Nasscom, 2002).

The Central Board of Direct Taxes, Ministry of Finance, Government of India issued a notification in September 2000 listing IT-enabled services exempted from income tax including, back office operations, call centres, data processing, engineering and design, geographic information system services, human resource services, insurance claim processing, legal databases, medical transcription, payroll, remote maintenance, revenue accounting, support centres and Web site services. The provision of including other services in this list is progressively being implemented (Nasscom, 2002). The Information Technology Act of 2000 heralded a new Cyber Law regime in the country. Nasscom is committed to catalyze Internet proliferation in the country, the ultimate goal to get 2Mbps of bandwidth for every adult citizen by 2005.

According to Nasscom (2003), India currently offers a strong value proposition of all IT-enabled services due to its abundance of skilled English-speaking manpower, which rates high in the area of qualifications, capabilities, quality of work and ethics. This places India ahead of competitors such as Singapore, Hong Kong, China, Philippines, Mexico, Ireland, Australia and Holland, among others (Nasscom, 2003). Nasscom is working with international certification agencies to set standards and India has been found to be uniquely capable for setting, measuring and monitoring quality targets. When compared to their western counterparts, the number of transactions per hour for back of-

fice processing has achieved higher productivity levels. With its unique geographic positioning that makes it possible to offer 24x7 services and reduction in turnaround times by leveraging time differences is yet another strong point. The regulatory environment, specially relating to ICTs is highly progressive and most of the policy recommendations made to the government have been accepted and acted upon. Incentives such as income tax holiday until 2010 have been provided for the export of IT-enabled services. The Government of India has announced a special policy for call centers. Many state governments in India are offering incentives and infrastructure for setting up IT-enabled services (Nasscom, 2003).

In the next section, we examine the evolution of the telecommunications sector, which is considered as one of the key influencers in the transition of India into the digital world.

EVOLUTION AND IMPACT OF TELECOMMUNICATIONS

Jain (2001) argues that many developing countries have noted the constraint of a state monopoly in telecommunications as standing in the way of spurring internal growth and competing in an increasingly global economy. Even though it is a century since telecommunications emerged, developing nations such as India do not share the benefits of a universally distributed telecommunications service (Maxwell, 2000). Historically, until the mid-1990s, India was still struggling with electronic business activity, especially for the domestic market. However, several developments constituted a sudden, although modest, surge in this activity after 1996 (Nasscom, 2003). Liberalization of the telecommunications sector and the opening up of the economy to foreign direct investment constituted the major facilitators. Before the early 1990s, telecommunications in India was a state-owned and bureaucracy laden. For example, to obtain a telephone connection, there was a high fee and a few years of waiting period. In a geographically disparate terrain, laying down infrastructure across the nation was and is still a challenging dream (Unnithan, 2002).

Indian telecommunications sector was wholly under government ownership until 1984 (Jain, 2001; India Infoline, 2001; Sinha, 1997, Dhar, 2000). The Post and Telegraph was separated from the sector in 1985 to form the Department of Telecommunications or DOT. Subsequently, DOT set up two public sector corporations, Mahanagar Telephone Nigam Limited (MTNL) and Videsh Sanchar Nigam Limited (VSNL), to allow greater autonomy and flexibility. While MTNL took over the operation, maintenance and development of telecommunications services in the metropolitan areas of Mumbai and New Delhi, VSNL was set up to plan, operate, develop and accelerate international telecommunications services in India (Jain, 2001; India Infoline, 2001; Sinha, 1997). MTNL enjoyed a monopoly position in the two metropolitan cities, until recently, but VSNL was given a monopoly over all international access to India through its gateways (India Infoline, 2001).

The Telecommunications Commission with representatives from many government departments, including electronics and finance, was a result of government realization that regulation of the Telecommunications sector remained with the DOT (Jain, 2001; Dhar, 2001). In 1997, a separate regulatory body, the Telecommunications Regulatory Authority of India (TRAI) was formed by an act of Parliament, with the main function of finalizing toll rates and settling disputes between the main players (India Infoline, 2001; Bagchi, 2001). Following the National Telecommunications Policy of 1994, the government announced private participation in basic and cellular services. The country was divided into 20 "circles" and one private operator was allowed to compete with the DOT in each of these circles. However, DOT was to give the licenses to operators with a fee, driving out competition with heavy license fees and tariffs (Bagchi, 2001). This led to the announce-

ment of the National Telecommunications Policy of 1999, taking into account the convergence and existing anomalies in the sector (Bagchi, 2001). As a result of this policy, 70 ISPs became operational in India. The government also encouraged several ISPs to set up international gateways to the Internet, bypassing the VSNL monopolized gateway (Bagchi, 2001). In addition to opening up international telephony, the government also decided to end VSNL's monopoly, two years before the WTO-set deadline of 2004.

The Centre for Development of Telematics (CDOT) in India was perhaps one of the earliest government initiatives to research and develop technology suited for the Indian climate (Jain, 2001). CDOT was able to champion the idea of technology for the masses, with rural automatic exchanges designed specifically for Indian climatic conditions. Many regional areas including villages, small towns and B class cities were connected and public telephone booths became part of Indian society. An indigenously developed technology, adaptable for Indian conditions, was successful and, by end of the year 2000, 10 million of the 20 million lines installed in India were using CDOT exchanges (Jhunjunwala, 2001).

Several technological changes made it imperative for the government to view IT, telecommunications and broadcasting legislation in a coherent and convergent manner, which led to the drafting of the Information, Communications and Entertainment Bill (Jain, 2001). The Communication Bill 2000 had the objective of facilitating the development of a national infrastructure for an informed society, establishing a licensing framework for carriage and content of information in the converging areas of television, broadcasting, data communications, multimedia and other technologies (Bagchi, 2001). This convergence bill along with the Information Technology Act of 2000 clearly indicate that India's government is moving towards a single communication network catering to all types of technologies including the Internet, Datacom, Telecommunications, Wireless, Fixed, Mobile, Cellular, Satellite Communications and e-commerce (Jain, 2001).

India Infoline (2001) pointed out that the country had an approximate tele-density of only 2 fixed lines per 100 persons (India Infoline, 2001). However, telephone penetration is not dependant on phone ownership. As in many developed countries, private space in houses is not abundant, and phones tend to be shared (Unnithan, 2002). In many interior areas, public call offices or telephone booths tend to be used (Jain, 2001). The socio-economic changes within the country spurred by the Internet have seen the emergence of cyber cafes and computer institutes all over the country. Interestingly, this development has been accentuated by the growing need for technology education within the country, essentially facilitated by the software industry. Although telephones may not have reached every home, cyber cafes are in great demand (Unnithan, 2002).

Convergence of ICTs, telecommunications, broadcasting and entertainment toppled most of the old value chains, bringing forth yet another revolution within India (Moni & Vijayaditya, 2002). Studies have shown that a large populace of television users would embrace the Internet, video-on-demand and greater interaction with content, but may be diffident about buying or using a personal computer. India has the highest cable penetration percentage of 46.8 percent among low telephone penetration countries (Nagaraj, 2001). There is a drive for cable modems, particularly by work-at-home households and Internet users. Satellite is another broadband access technology, which has immense potential for a country as large as India. Technological developments now permit the network used to carry broadcasting signals to the customer premises to be used for carrying telecommunications and data. India already has had a critical mass addicted to television and cable channels and this has fuelled the growth of cable Internet (Unnithan, 2002).

Indian society has had limited resources to absorb unprofitable innovations and the majority of the population has historically responded only with caution and economic necessity. Interestingly, the growing upper-middle class, characterized by the computer professionals, is increasingly driving the diffusion of technological innovation. Many of these professionals are non-residents willing to invest in innovative telecommunications ventures, as they are seen as progressive icons for the economy. The impact of global trends, technological innovations, and a growing generation of technically skilled youth who are driven by rational views, moving away from the older generation with their nationalist attitudes, making further modification of attitudes and actions inevitable (Jeevan, 2000).

Together with Nasscom, the government of India is now committed to push electronic commerce in India, as reflected in the announcement of the Information Technology Act of 2000, the announcement of ISP policy for the entry of private Internet service providers in 1998, permission grants to private ISPs to set up international gateways, permission of Internet access through cable TV infrastructure, initiation of a National Internet backbone, announcements of national long distance service beyond the service area to private operators, complete non-monopolization of undersea fibre connectivity for ISPs in 2000, free right of way facility with no charge to access providers to lay fibre optic networks along national, state highways; interconnectivity of government and closed user networks, and establishment of public tele-info centres (PTIC) with multimedia capabilities (Nasscom, 2002, 2003).

After having examined the telecommunications sector, we now propose to examine the mobile communications sector, which is fast becoming a key enabler in facilitating electronic business process delivery.

EVOLUTION AND IMPACT OF MOBILE COMMUNICATIONS

Towards the turn of the century, the government of India recognized the key role of telecommunications in its developing economy and decided to invest significantly in the mobile communications sector (COAI, 2002, 2003). Mobile services were introduced initially as a duopoly under a fixed-license regime for a period of 10 years. With liberalisation in the telecommunications sector, the country was divided into four metropolitan cities and 19 telecommunications circles which were then roughly analogous with the states of India, and licenses awarded to private operators, bringing in competition. Cellular licenses were awarded to the private sector, first in the metropolitan cities of Delhi, Mumbai, Kolkata and Chennai in 1994 and then in the 19 telecommunications circles in 1995 (COAI, 2002, 2003).

The initial response of the private sector was very encouraging with the attractiveness of the Indian market — the low teledensity, the high latent demand and a burgeoning middle class — brought in some of the largest global telecommunications players, foreign institutional investors and the major Indian industrial houses to invest. Annual foreign investment in telecommunications increased steadily from an insignificant INR 20.6 Million in 1993 to INR 17,756.4 Million in 1998. However, the attractiveness of the Indian market did not last for very long, as by 1997-98 the private cellular operators were confronted with a series of problems that threatened their very viability and survival. As a result of this, Foreign Direct Investment inflow into telecommunications dropped sharply, declining by almost 90 percent to INR 2126.7 Million in 1999 (COAI, 2002, 2003). As private-sector participation preceded the set up of regulatory authority and tariff rebalancing, licenses were auctioned at exorbitant amounts, leading to high cost structure and unaffordable tariffs. Therefore, for the common public, although mobile telephony was a convenient faster option,

as against a fixed phone, the unaffordable tariffs did not help the situation.

According to the COAI (2003), one of the key factors of this critical state was the manner in which liberalization was undertaken. Usually, deregulation is preceded by tariff rebalancing, institution of a strong and independent regulator and then private sector participation is invited. In India, private sector participation was invited in 1992, the Regulatory Authority was set up in 1997 and the tariff rebalancing exercise commenced in 1999 and is still far from complete. The regulatory authority had considerable ambiguity on its powers, which resulted in virtually each and every order of the authority being challenged by the licensor. In addition, consumer benefit was the least priority by the government and the sector was a key revenue generator for the government. Although the National Telecommunications Policy of 1994 identified the primary objective as affordable cellular services, this was almost disregarded during implementation. Licenses were granted through an auction process to an enthusiastic private sector deluded by the huge potential of the Indian market and lured into bidding exorbitant sums of money for cellular licenses. These huge license fees resulted in a high cost structure leading to unaffordable tariffs and lower growth of the market. Subsequently, the cellular industry was on the verge of bankruptcy by the end of 1998 (COAI, 2002, 2003).

Under these circumstances, the government introduced a new policy called NTP 99 and the amendment of the Telecommunications Regulatory Authority Act in January 2000. The policy replaced the high–cost, fixed licensing regime with a lower cost licensing structure through revenue sharing, providing a greater degree of competition and flexibility in the choice of technologies. Existing private cellular operators migrated to the new telecommunications policy regime beginning in August of 1999. Cellular tariffs have dropped by over 90 percent since May of 1999. The average airtime tariff in 2001 was prevailing around INR2 per minute as against the peak ceiling tariff of INR16.80 per minute when NTP 99 was announced. There was also a significant drop in the mobile phone costs with cellular handsets costing around INR 30,000 or US$645 in the initial days to INR2000 or US$42 (COAI, 2002, 2003).

More specifically, as the government rationalized levies, resulting in high turnover, and cellular operators were able to venture more into cities and towns. Parallel to this development, the operators are able to offer services to consumers on a contract or plan basis, subsidizing the cost of phones. Consumers may be offered a certain tariff for buying the phone over a period of 12 months on a contract, where they also are bound to the operator for that period for services provision. Thus, on a plan the consumer may be charged as low as USD42 for buying the phone, over a period of 12 months, which may be the term of the contract. The government is promoting the mobile sector as it generates revenue for the exchequer, but also reduces the costs of infrastructure rollout especially when connecting remote villages. Low tariffs, along with price wars by cellular operators are supported by massive consumer demand, especially the youth in metropolitan cities (Fraunholz & Unnithan, 2004a).

By the end of 2002, the mobile subscriptions surpassed fixed-line networks. As against owning a PC and getting an Internet connection at home, or using an inconvenient option of a public Internet booth without privacy, ownership of a mobile phone constituted to this surge. Another enabler of mobile telephony success is touted to be the Short Messaging Service, which has facilitated not only the growth of a new culture but also many business models that support electronic business. Although there is modest growth in the area of electronic commerce as such, there seems to be an interesting trend toward developing mobile commerce. Many vertical industry segments, especially the banking industry, have introduced mobile commerce successfully. The

transition into mobile commerce is supported by the introduction of 3G networks in 92 cities in 2003 (COAI, 2003).

On the technological frontier, the Indian Government, when considering the introduction of cellular services into the country, made a landmark decision to introduce the GSM standard, thus avoiding adolescent technologies and standards. Although cellular licenses were made technology neutral in September 1999, all the private operators are presently offering only GSM-based mobile services. In July 2001, cellular licenses were awarded and all of the new licensees have opted for the GSM standard to offer their mobile services (COAI, 2002). According to Gartner (Indiantelecomnews, 2003a). CDMA technology is particularly attractive to India, as the point-to-point concept of communication within specific circles is an important factor for India. The Indian market had clearly defined points of usage within a telecommunications circle where CDMA is likely to work better as opposed to GSM or unlimited mobility. IDC forecasts (Indiantelecomnews, 2003b) that the cheap CDMA connections are likely to affect GSM operators, with up to 20 percent of subscribers willing to try cheaper CDMA services. Beginning with 3G advanced wireless services, the Reliance India Mobile service marked the first CDMA2000 1X nationwide commercial launch in India, bringing advanced wireless data and voice services to 92 cities (3g, 2003) in May 2003. However, in India, the CDMA is being adopted as a platform for launching 3G but it still is expected to co-exist with current GSM and future 3GSM services.

On another note, the unique nature of multiple network operators licensed within each circle, leading to co-operation and competition — or co-opetition (Xu et al., 2003) is the way SMS operates within India. The term co-opetition describes competing businesses cooperating to create and enlarge the market rather than competing to divide (Brandenburger & Nalebuff, 1996). The success of SMS in India has been a stimulant for network operators who seem to be optimistic about the forthcoming MMS to go along the same success route. The vast geographic terrain also offers opportunities in niche segments such as farming or agro sectors in central and north western India, trade sector in Gujarat, IT professionals in the southern region, industrialists in the central region, literary communities of eastern region and so forth. Each of these communities along with the growing youth population in metros offer significantly different opportunities for the providers as they increasingly demand value-added services tailored to their needs, whether they be different languages, script, content and so forth (Fraunholz & Unnithan, 2004a).

The infrastructure problems associated with the geographic region motivates the population — due to the convenience it offers — as well as the government authorities to promote the mobile sector. It not only generates revenue for the exchequer, but also reduces the costs of infrastructure rollout especially when connecting remote villages. With stimulation from the government, and population demand, cellular network providers seem optimistic about their future growth in India. However, the licenses issued clearly indicate some significant players who hold the major market shares, whether it is through subsidiaries or sister organizations. The industry itself is showing the signs of becoming an oligopoly in the future (Fraunholz & Unnithan, 2004a).

Figures from the Cellular Operators Association of India or COAI showed that the industry had 5.725 million subscribers, up from 3.27 million at the end of January 2001 and 5.48 million subscribers at the end of year 2001. The data showed that the industry added 246,281 users in January 2002 alone, led by the four main city markets of Bombay, New Delhi, Madras and Calcutta, which together added 93,070 customers. The overall Indian mobile subscribers jumped to more than 12 million in first quarter of 2003 (Fraunholz and Unnithan, 2004a,b).

A recent Reuter (2002) report claimed that the number of mobile phone subscribers in India is likely to rise to 120 million by 2008 because it has the cheapest call rates in the world. India's US$5.0-billion mobile phone sector, billed as one of the fastest growing markets globally in this decade, has eight million users spread across some 1,500 cities and more than 60,000 villages. The main driver for a more than 100 percent growth each year in the past six years has been falling tariffs in a sector where a dozen money-losing firms have launched a fierce price war to grab market share. The eight-year-old sector has the lowest rate of USD16 a month for a 300-minute talk time plan compared with other developing nations such as USD21 in China and USD77 in Brazil (Reuters, 2002; Rediff.com, 2002). However, mobile operators pay between 8.0 to 12 percent of their revenues as license fee compared with no license fee in China. Data released by the Cellular Operators Association of India showed that the industry is expected to be one of the world's fastest growing markets this decade (Rediff.com, 2002).

As of late 2002, there were 24 companies and 42 networks on air all over India (Ramachandran, 2002). For the first time, over a three-month period from April to June 2002, the cellular subscriptions went up by 960,000 as against the fixed lines increase of 300,000 over the same period (COAI, 2003). Low tariffs, along with price wars by cellular operators are supported by massive consumer demand, i.e., the youth and businessmen in metropolitan cities as well as the relatively new and upcoming households in rural areas (Reuters, 2003). The tight monopoly control over telecommunications and aggressive efforts to curtail competition had led into slow growth of the Internet. This in turn led to the boom of entrepreneurs in India.

India has a low teledensity of 4.5 percent compared with a global average of more than 15 percent (Reuters, 2003). The number of households in the rural areas is expected to grow to 360 million by 2010, making them an attractive audience. On the other hand, the thickly populated urban city areas are less motivated to get a fixed-line network. To explain the cause of this de-motivation, an example would be the thickly populated metropolitan Mumbai, where every suburb is connected with metro railway lines lined by illegal slums (Fraunholz and Unnithan, 2004a). To get a fixed-line cable network that runs across these slums into the 12th level apartment, in itself is a significant feat, as it would require cutting through many bureaucratic angles including the people employed for installation (Fraunholz & Unnithan, 2004a).

On the other hand, once a fixed line is in place, the fear of a slum dweller cutting into the line, resulting in massive bills every month — which cannot be traced — often deters households from connecting fixed lines. On yet another note, due to the small limited spaces available within households and the relative lack of privacy often drives people into public booths — which are perceived to have more privacy — especially for the youth who would like to be away from the earshot of family members (Fraunholz and Unnithan, 2004a). The mobile phone, although starting off as a high cost affair, is now a lure for those who seek privacy, relative safety from slum dwellers and freedom from bureaucratic tangles. Most firms expect the market for mobile services to grow by between 10-14 million new subscribers in 2003 (CellularOnline, 2002).

The affordability of mobile phones has become more and more possible in India, with the government cutting down on the levies. In addition, the network operators have opened up the avenue of subsidizing the phone, through call plans, which has almost brought the cost of the phone down to INR2000 or USD43. In spite of a heady growth in the cellular services market, following the subsidies offered through call plans, the legal market for cellular handsets has remained very small (Indiainfoline, 2002). A large percentage of the handsets sold in the country are through the unauthorized or the grey channel, which includes

smuggled handsets, parallel import and handsets brought by people travelling abroad.

The share of the unauthorized market in the overall market has shot up from 74 percent in the year 2000 to 86 percent in the first half of 2001 and an estimated 89 percent in second half of 2001, according to an IDC report on the Cellular Handsets Market in India (Fraunholz and Unnithan, 2004a). This increase can be attributed to the price differential between handsets bought from legal and grey channels. The difference in price is at least 25 percent to 35 percent and arises due to the high level of duties like customs or sales tax paid by the vendors selling in the legal market. In a metropolitan city like Mumbai, there is also additional taxes, such as Octroi (a levy on the basis of getting into the metro area). Yet another reason for this flourishing grey market is that the handset vendors do not provide extensive after sales support because of the small size of the actual legal cellular market and therefore absence of economies of scale. Without any supporting infrastructure, buyers do not feel the need to go on a call plan — to buy a phone — when the same phone is available with much cheaper rates from the grey market. There is no incentive for buyers in real terms (Indiainfoline, 2002).

Mobile communications has re-invented the role of fishing captains into logistics and supply chain managers (Karkera, 2002). For example, the fishing industry in Kerala, a southwestern state of India, generates USD600 million in a year in revenues. During the day prices vary throughout the day at 17 landing ports around the main port of Cochin. Currently, 8,000 fishing boats carry mobile phones, to locate the best offers before landing in the port, saving expensive fuel by calling in carrier boats that take the catch to the shore. In addition, the agents, handlers and middlemen also carry mobiles to get their best deals. Two competitive firms are offering services to these "communities of interest" (Keen and Mackintosh, 2001). The boom of young IT professionals carrying PDAs and mobiles and also the growing concept of mobile workers in densely populated metropolitan cities of Mumbai, where commuting otherwise takes hours, are becoming increasingly commonplace.

Short messaging services has brought mobile communications to Indian life, whether it is student or executive and urban or rural life (Thomas, 2002). India is an economy widespread geographically, but prefers the closeness in society. The cost-effectiveness and convenience of the mobile combined by this new SMS and lately multimedia messaging services or MMS, is becoming increasingly common. The growth of mobile workers, especially in the IT area, within metropolitan cities and with increasing demands on their time have further added to the increasing popularity of SMS. A fair example of an office executive, stuck in a traffic jam before a presentation, sending an SMS or even connecting through a PDA to send the agenda or presentation through is becoming part of daily life (Fraunholz & Unnithan, 2004b).

Mobile Youth (2002) claimed that SMS is creating a revolution in India with an estimated 60 messages sent per phone per day from India's 8 million mobile phones owners in early 2002. On special days such as national festivals, SMS traffic increases to clog most networks. SMS was reported to be four times more than normal as people sent festival greetings on Diwali, as this is the most economical, convenient and instant mode of communication (Chatterjee, 2002). There seems to be a 500 percent jump over normal usage during this major festival season equivalent to New Year's Eve (MobileYouth, 2002). In a joint initiative, Ericsson Mobile and Bharti Telecom — the network provider who holds major market share within India — worked together to develop an SMS-based service for school children (Ericsson, 2001). Four to five million school children obtained their test results by sending their identification numbers and receiving their results in SMS form. Not only did they receive their overall percentage, but also individual subject marks. Following this

project, Bharti's (which has 25 percent market share) SMS traffic grew more than 100 percent (Ericsson, 2001).

Pereira (2002) reported that on the 13th of May 2002, the capital city of New Delhi was introduced to traffic police SMS. The service was aimed at providing aid in answering the average queries of a motorist as well as to help the traffic police operating the field. Commuters can get information on traffic blockages and diversions while investigative journalists can acquire information on accidents or prosecutions immediately through this service. A vehicle being towed away is immediately notified to the owner. News sites in India such as Mid-day and the India Times have expanded their SMS alert options. Major portals such as Yahoo and Rediff have launched SMS services for instant messaging via gateways on their Web sites. In the southern state of Kerala, fishermen use their mobiles to send SMS messages to their partners on the shore about their catch, so that the price can be fixed and faster transactions can be made. For many families living apart in the large geographic region, SMS is a lifeline to keep in touch (Thomas, 2002).

The southern Indian city of Chennai outstripped all the other metropolitan cities in SMS usage with 80 percent of the mobile owners sending SMSs (Times, 2003). In an interesting IDC survey, women were found to be SMS'ing more than their male counterparts — averaging 4.2 messages per day within the same period, in Chennai, which reported maximum SMS usage. Interestingly, human rights activists in India have condemned the diffusion of SMS especially among the youth as a cause of breaking up relationships (Mobileyouth, 2002). For example, a typical 'U4Me' message was cited to have sparked marital discord ending in a divorce. However, the growth of SMS seems undeterred with operators clocking a staggering nine million short messages in one single festival day, in the capital city of New Delhi alone. Many celebrities now provide mobile numbers to fans as SMSs do not intrude their privacy (Thomas, 2002). A vital aspect behind the success of SMS is that the costs range from INR 2 or USD 0.042 in some circles to 50 Paisa or USD 0.010 and free in others. The income from SMS is currently 10 percent of the total revenue for many network providers (Thomas, 2002). The affordability and the trendy, cool aura that it provides to the youth seem to be the key factor in SMS success in India. From the corporate point of view, at least in the metro cities, SMS is becoming vital for communications amongst traffic congestions and traveling (De, 2001).

Evidently, government support and stimulation has progressed the telecommunications reform. Over the 1990s, growth of professional youth has supported the massive demand of mobile communication services. Combined with these factors were the after effects of deregulation that stimulated the previously stagnant or rather inflexible telecommunication sector. With digital transmission becoming increasingly popular in India during late 1980s, the next evolution into mobile phones was slow but gradual over the period. However, initially the non-affordability of mobile handsets itself was impeding the growth of this sector. With the new telecom policies that subsidized the levy on handsets, and parallel growth of the grey market, once the influx of foreign goods became open after 1995, handsets became more affordable (Fraunholz & Unnithan, 2004a).

Businesses as well as individuals have now become common users of mobile phones, due to the convenience and cost-effectiveness that it offers in the Indian context, especially compared to the hassles of obtaining a fixed-line phone. With the blessing of government subsidies, increased demand and subsidized handset levies, network service providers (or cellular network providers as they are known in India), have launched into a fierce price war, especially targeting the youth market (Fraunholz and Unnithan, 2004b). In the end, the network providers who will offer the most affordable as well as innovative services will drive

the mobile communications market. The uptake of SMS/MMS as a value-added service seems to be driving a mini social revolution within India. It has in a way provided the youth with "affordable freedom" within a restrictive society. It has to be noted that a fixed-line telephone is not considered "private" as it exists within a closely condensed household, where people can eavesdrop. The administrative hassles of obtaining a fixed phone, as compared to a mobile phone, are making individuals as well as businesses opt for mobile communications (Fraunholz & Unnithan, 2004b).

In turn, SMS is an attractive business model for network providers to extend into mobile commerce. While a combination of SMS and voice transmission is able to help governance, service provisioning such as traffic information, it will be a while before mobile commerce can take over from electronic business over the Internet. India still has a long way to go to reach the critical mass or saturation regarding mobile phones. 3G with CDMA standard is becoming a beacon of hope for rural and geographically spread out areas, as well as affordability (3G, 2003). However, the CDMA standard will co-exist with the established GSM networks for a long time to come. If the 3G is absorbed as fast as it is perceived to be, and holds up to its promises within India, it may be the perfect launching platform for many mobile commerce models such as location-based services, which would be very lucrative in the disparate terrain (Fraunholz and Unnithan, 2004a).

As it can be seen from this section, mobile communications have not only facilitated a low cost infrastructure, which has overtaken fixed-line subscriptions, but also have facilitated revenue generating models within the domestic economy. With rapid absorption of mobile technologies, the day is not too distant when electronic business in India may transition into mobile business. The next section is a brief appraisal of the electronic business processes within the economy.

ELECTRONIC BUSINESS IN INDIA

Indian government had recognized the potential savings for the exchequer by introducing digital service delivery (Nasscom, 2003). It would cut down transaction costs of governance, thereby stretching the taxes paid by the average citizen to provide more services across the economy. Therefore, a major effort was made to introduce digital governance into the country with the Central and State administrations, customs, ports, the public tax system and education system pioneering the venture. A number of state governments implemented e-commerce initiatives aimed at cost effectively taking various facilities to citizens. Innovations in the area of land records, taxation, procurement, etc., were witnessed in the sector, with the Internet pervading significant government transactions (Nasscom, 2002). The government of India issued guidelines that 2-3 percent of every ministry or department plan budget was to be utilized in achieving digital governance using IT (Raje, 1999). As pointed out, many state governments have taken initiatives to provide "one-stop shops" to deliver a host of services to citizens such as domicile certificates, driving licenses, property tax payments, electricity and water bills, etc. In parallel, to achieve mass customization, the government of India decided to set up a National Institute of Smart Government as a tripartite venture between government, business and community (Raje, 1999).

With the increasing recognition that information technology is catalyzing economic activity and efficient governance, Indian government has made significant investments in the sector. One of the interesting challenges for the government was to implement a common e-governance thread in a geographically dispersed, demographically multilingual India. Out of the 1.2 billion population, 95 percent (950 million) speak or practice 18 officially recognized languages. For the Centre for Development of Advanced Computing (C-DAC) this presented an opportunity. An initiative in

developing Indian language tools with natural language processing, in evolving script and font standards through GIST technology was pioneered by C-DAC, with directives from the government (C-DAC, 2002). Some of the successfully commissioned initiatives include:

- **Hospital Management System:** implemented to improve healthcare services for patients in speciality and government hospitals across India.
- **State of Maharashtra:** Public Works Department with 250 state-wide offices was networked, the GIS-based land management was implemented providing Web-based access to land data covering allotment, transfer, mortgage, surrender, etc., of the industrial development units. Archives Computerization was deployed for the Department of Archives and Octroi (a type of tax) collection was computerised and networked.
- **Stamp registration in Maharashtra and UP States:** Online property registration, valuation and report generation across 366 offices at various state administrative units, reducing time and increasing revenue.
- **Karnataka State:** Major functions of property tax valuation/collection, issue and record of death/birth certificates, water supply billing, consumer complaints and internal MIS functions were computerised to provide improved citizen services.
- **Andhra Pradesh:** Implemented a data warehouse of land and population data of 60 million people to enable well informed, timely policy decisions by government officials across various departments.

Compaq India established a memorandum of understanding with Electronics Development Centre of India (ER & DCI) in the year 2000 to initiate e-governance in NOIDA city and extend it subsequently to various states. The project was to smart link/interface between citizens and to develop a system that automates rural development, arms and licenses, regional transport offices, land records, citizen databases, electricity board payments, and set up GIS (Compaq, 2000). This project brought Internet/intranet infrastructure up to section officers level, IT empowerment of officers and officials through training, IT-enabled services including government G2G, G2B, G2C portals and development of BPR methodology for electronic services delivery, among other initiatives (Moni & Vijayaditya, 2002).

In December 2001, the entire state of Gujarat was networked up to the small taluk level of government. Everything from collecting the posts to disposal of files is computerised and the GSWAN has come into existence. The Concept Centre of Electronic Governance set up by the Indian Institute of Management in Ahmedabad was able to identify worthwhile applications and disseminate knowledge for successful implementation of e-governance applications amongst bureaucracy and other stakeholders. E-governance has been able to bridge the digital divide in this state with a 50 percent literacy level and more than half of the population living in rural areas. The priority services include pension processing and ration cards (CIOL, 2002). As Zdnetindia (2002) reports, Tamil Nadu State has a comprehensive state government information site with application forms in English, Tamil, comprehensive land records, a pilot project of utility bill payment over the Internet, tele-medicine projected proposal, application software for regional transport offices, registrar office and major intended IT projects for high court and police departments. It is evident that similar ventures in other states are ongoing here as well.

The most important initiative of the Karnataka government is the lodging of taxes online. Computerised land records — the Bhoomi Scheme — and registered transfer certificates are on the anvil. In addition, the policing system, forestry, agriculture and regional transport system is being computerised. The state also focuses on education

with YUVA, a program aimed at underprivileged youth, women and families with low income (Banerjee, 2001). In Madhya Pradesh, 5,500 centres for computer literacy were announced with a program called Headstart. The state is advocating e-governance through education. In addition, the commercial tax department, registration department, treasuries and agriculture marketing departments were computerised. The state is promoting a "build-operate-transfer" system for smart cards in the transport department for registering vehicles and issuing driving licenses (Singh, 2002).

The Indian government has been taking key initiatives over the past few years to create an environment conducive for e-commerce activity and some of them include the Information Technology Act 2000, which brought in the cyber-law regime in the country, entry of private Internet service providers in 1998, granting permission of Internet access through Cable Television networks, the establishment of Public TeleInfo Centers (PTIC) with multimedia capabilities and allowing 100 percent foreign direct investment in B2B ventures. The nationwide Internet backbone has also been initiated. Whilst the federal government has laid out several such initiatives, there has also been support from state-level ministries initiating various developmental initiatives for public welfare and for promoting business, especially Small and Medium Size businesses (Ministry of Technology, 2003).

Despite these, there is a modest e-commerce activity estimated to be around USD 300 million in the year 2002 (Nasscom, 2002). The Business-to-Commerce spending in India is estimated to have grown in the year 2002, with the travel sector accounting for 23 percent of transactions. Business-to-Business e-commerce implementation was low except in certain vertical sectors such as automobiles, banking and finance (Nasscom, 2003). Experts have argued that the low cost of personal computers, a growing installed base for Internet use, and an increasingly competitive Internet Service Provider (ISP) market will help fuel e-commerce growth. Dataquest, an Indian computer journal, has found that the rise of Indian Internet subscribers will ultimately depend on the proliferation of network computers and Internet cable (Gartner, 2003). Cyber cafes will also continue to provide low-cost access.

Currently, the lion's share of current e-commerce revenue is generated from an ever-expanding business-to-consumer (B2C) rather than business-to-business (B2B) market. As in the United States, B2C transactions have taken the form of online purchases of music, books, discounted airline tickets, and educational resources. In a recent McKinsey-Naccson report, it was estimated that some 80 percent of e-commerce in India over the next few years could be B2B if the correct environment were developed. The B2B market is expected to increase following greater investment in the Indian telecommunications infrastructure, once intellectual property rights and legal protections for commerce over the Internet are addressed. There are still enormous challenges facing e-commerce sites in India. The relatively small credit card population and lack of uniform credit agencies create a variety of payment challenges unknown in the United States. Increased distribution of online purchases could be complicated by India's complex postal system and an uncertain regulatory environment. Nonetheless, everyone from Yahoo, Microsoft, and IBM to local carpet vendors, hotels, and some 300 Indian ISP's are trying to claim a slice of the rapidly emerging Indian e-commerce market (Nasscom, 2003).

As Kripalani and Clifford (2000) have rightly commented, India has always had enormous potential, but a difficult time living up to it. Corrupt governments, outbursts of Xenophobia or communal violence have affected its confident progress, not to mention the disparate socio-economic, cultural and geographic spread. Successive governments have made great strides to reduce stifling, socialist-era regulations, but the shadow

still exists as many politicians still are reluctant to relinquish power. For e-commerce to surge, there is a need for ample telecommunications capacity, computers, and electricity. Even with progressive reforms, India still needs to attain the critical mass with telephony — wired or wireless. Power shortages are chronic while access to PC/Internet continues to be low. However, on the positive side, India has free media and democracy which many developing economies lack. And the country is proactive to solve its shortcomings. In telecommunications, the government is dismantling curbs on foreign investment and competition. Consortiums are building fiber-optic networks. Satellite communications and TV set-top boxes are expected to help bring the Internet to households that still lack phone lines. Wireless telephony has gained momentum (Kripalani & Clifford, 2000).

MOBILE BUSINESS

According to Nasscom's Strategic Review 2003, in the year 2002, the m-commerce market was in the region of USD50 billion worldwide (Nasscom, 2003). While the US and European markets are expected to dominate forecasted revenues until 2005, Asia-Pacific and the rest of the world are expected to account for 40 percent of the estimated $225 billion m-commerce market by 2005. However, the m-commerce market in India has not seen as much growth as was expected. Experts opine that it is still in a very nascent stage and will take time to reach the maturity level to match EU and the US standards.

According to Anil Lekhi, VP–IT for Spice Telecom in Punjab (Expresscomputer, 2003), the company that witnessed hardly any m-commerce transactions a year ago, is now doing business of around INR 25,000 to 30,000 daily through m-commerce. The average value of these transactions varies from INR 1.50 to 10. Presently almost 9 percent of Spice's overall revenues come from value-added transactions of which m-commerce is a part. The optimistic country operations manager of Motorola PC division commented on m-commerce potential in India based on the ability to address nearly 50 million subscribers by 2005 (Expresscomputer, 2003).

Indian cellular operators today are under tremendous pressure to sustain and grow their Average Revenue Per User or ARPU. Fierce competition among operators has consistently driven down tariffs, reducing revenue from the voice-based operations of the wireless networks. Operators today are providing value-added services to sustain and grow their ARPU (Fraunholz & Unnithan, 2004a). Optimism prevails that with subscribers already checking movie schedules and airline/railway booking status among other things using the mobile handset, it is only a matter of time before they start booking tickets as well. The advent of wireless Internet in the form of GPRS and CDMA 1x is expected to further boost m-commerce in India (Expresscomputer, 2003).

The two types of m-commerce transactions are low-value and high-value. Low-value transactions usually imply music downloads, logo downloads, picture downloads, ring tone downloads, etc., some banking, value-added services like news, stock alerts and services like m-coupons and wallets. On the other hand, there are high-value transactions, which involve credit and debit card transactions, point-of-sale terminals, going to the merchant location and paying through the handset. There are only low-value transactions in India at present. Major advances in m-commerce are not going to happen until higher-bandwidth networks are deployed and wireless service providers cooperate with each other instead of pushing competing standards (Expresscomputer, 2003).

Yet another major reason is the lack of commerce-capable cellular networks, which can route real-time transactions over the cellular network to a remote payment gateway and guarantee security over the transactions (Alexander, 2003). Operators need to instill confidence among mobile users to start taking buying decisions on deals initiated

from mobile phones and turning the mobile into a wireless debit card. The different parties involved in the entire m-commerce value chain are wireless infrastructure providers, wireless service providers, certifying authorities, applications/software providers, equipment manufacturers, credit card companies and banks. And the absence of proper coordination between them will hamper growth (Alexander, 2003). Also all players in the value chain, from biometrics to SIM card providers, cellular operators, network providers, application developers, banking to semiconductor companies have to coordinate.

However, on the progressive side, in their rollouts the operators are actively deploying next generation wireless Internet technologies to facilitate data services. The cellular subscriber base is growing phenomenally and reaching respectable levels of adoption. So, the stage seems to be set for m-commerce and the industry is already seeing the first signs of evolution in value-added services that operators have started providing. Quick adoption is imminent, keeping in mind the ARPU decline from voice (Expresscomputer, 2003). If statistics are anything to go by, the SMS rage will drive m-commerce in India. According to Merrill Lynch, SMS could bring in as much as $75.6 million of revenues for Indian GSM operators by the year 2005. The stock-brokering firm predicts that by the end of 2003, close to 700 billion application-driven SMS would be sent from mobile phones, which would be almost half of the total SMS traffic. During 2000-2003, while peer-to-peer messaging has been growing at a the annual growth rate of 46 percent, the application-driven SMS traffic has been growing by 204 percent during the same period (Expresscomputer, 2003).

On the other hand, the software services sector, which has been a catalyst in Indian electronic business is aiming at competing in emerging markets of wireless communications and mobile commerce through client software development and embedded systems design (Shankar, 2004). As the growth rate of mobile phones has already outnumbered the growth of fixed-line phones in India, with the development of a secure, easy-to-use method for paying over a mobile is devised, m-commerce will become a reality in India.

CONCLUSION

Over the past decade, progress is visible from the Indian perspective as regards government initiatives in promoting Information Communication Technologies infrastructure, telecommunications and especially mobile communications that supports electronic/mobile business. Despite this, in a geographically diverse nation, with different grades of socio-economic progress, density of population and attitudes, it is still a challenge to catch up with developed nations as regards electronic business. Although mobile communications seem to be a boon for an economy with infrastructure issues for telecommunications, it will be a while before the critical mass is achieved for mobile communications, Internet and overall technology access.

The overall attitude in the economy is now driven by growing middle-class professionals, who form the majority of the population. Policies seem to be based on the needs of these professionals and a proactive government, which seem to realize the potential of ICTs and electronic/mobile business for the socio-economic progress within the economy. The use of wireless technologies has become increasingly common especially in densely populated urban markets. This in turn offers potential for lucrative business models such as location-based services which drives mobile business.

Outlook

India's ruling government has made information technology the cornerstone of its political agenda of generating high economic growth while surrendering little sovereignty to multinationals

(Kripalani and Clifford, 2000). The hope is to spread information communication access to unify a nation otherwise divided by cultural and economic disparities. Many remote villages in India are now connected to the Internet. From craftspeople to daily farmers, rural Indians have begun using mobile communications or mobile Internet facilities to sell goods and monitor prices. It seems to be a novel, but very effective, approach for a progressive resurgent India. This chapter has provided a synthesis of many interesting factors, mainly in the telecommunications area that influenced the growth of electronic business and its transition into mobile business within India. It is expected to inform researchers, academia, policy makers, and all players who are involved in electronic business development in developing nations.

REFERENCES

3g. (2003). 3G Launches in India. *3G Newsletter.* Retrieved June 30, 2003, from *http://www.3g.co.uk/PR/May2003/5369.htm*

Alexander, G.C. (2003). It is time to shop through cellphones. Retrieved from *http://www.rediff.com/money/2003/dec/22betterlife.htm*

Bagchi. (2000, December). Telecommunications reform and the state in India: The contradiction of private control and government corporation. Center for Advanced Study of India, CASI Occasional Paper #13. Retrieved from *http://www.sas.upenn.edu/casi/reports/Bagchipaper120000.pdf.*

Brandenburger, M. A., & Nalebuff, J. B. (1996). *Co-opetition.* New York: Doubleday.

CellularOnline. (2002). India's mobile industry grows 75% in January. Retrieved from *http://www.cellular.co.za/news_2002/021502-india_growth_75%25.htm*

Chatterjee, S. (2002). 'Shrt n swt. C u 2nite,' with SMS, love goes mobile at touch of a button. *News India times Online.* Retrieved from *http://www.nesindia-times.com/2002/02/22/sp-valentines-day-sms.html*

Chen, S. (2001). *Strategic management of e-Business.* London: John Wiley.

COAI. (2002, August 23). COAI News Bulletin 20. Retrieved from *http://www.coai.com/docs/nb20.pdf*

COAI. (2003). Cellular operators association India statistics. Retrieved from *http://www.coai.com/stats.2003.q1.htm*

De, R. (2001). SMS from Yahoo: Net profits not a myth. *Express Computer.* Retrieved from *http://www.expresscomputeronline.com/20011231/cover1.shtml*

Dhar, S. (2001). *Indian telecommunications liberalisation and development* (a report). Essar Comvision Limited, India.

Ericsson. (2001). Ericsson mobility world India and Bharti India's largest private communication services provider launch SMS project for students. *Ericsson Online.* Retrieved from *http://www.ericsson.com/mobility world/sub/articles/success_stories/india_bharti_launch_sms_project_for_students?PU*

Expresscomputer. (2003). Dial M for m-commerce. Retrieved from *http://www.express computeronline.com/20030818/indtrend1.shtml*

Fraunholz, B., & Unnithan, C. (2004a). Critical success factors in mobile communications: A comparative roadmap for India and Germany. *International Journal of Mobile Communications, 2,* N1.

Fraunholz, B., & Unnithan, C. (2004b). SMS growth and diffusion: A preliminary investigation of three economies. *Proceedings of ISoneworld 2004,* Las Vegas.

Gartner. (2003). Gartner Press Room, Quick Statistics, Mobile Phones. Retrieved June 30, 2003, from *http://www.dataquest.com/press_gartner/quickstats/phone.html*

India Infoline (2001). Telecommunications. Indian Telecommunications Industries Sector Report. Retrieved from *http://www.indiainfoline.com/sect/tesp/ch01.html*

India Infoline. (2002, August 29). 85% of the cellular phones come from the grey market. Retrieved from *http://www.indiainfoline.com/nevi/cell.html*

Indiatelecommunicationsnews. (2003a). CDMA is better for India – Gartner, India Telecommunications News. Retrieved from *http://www.indiatelecommunications news.com/technology.htm*

Indiatelecommunicationsnews. (2003b). GSM way ahead-IDC, India Telecommunications News. Retrieved from *http://www.indiatelecommunicationsnews.com/technology.htm*

Jain, R. (2001). The telecommunications sector. India Infrastructure Report 2001, IIM Ahmedabad. Retrieved from *http://www.iimahd.ernet.in/ctps/iir8.pdf*

Jhunjunwala, & Ramamurthy, B. (2001). *Enabling telecommunications and Internet connectivity in small towns and rural India*. India Infrastructure Report 2001.

Karkera. (2002). Of fishermen, mobile phones and changing lifestyles. Rediff.com special report/George Type. Retrieved from *http://www.rediff.com/news/2002/aug/13spec.htm*

Keen, P.G.W., & Mackintosh, R. (2001). *Freedom economy: Gaining the mCommerce edge in the era of the wireless Internet*. CA: Osborne/McGRawHill.

Kripalani, M., & Clifford, M.L. (2000, February 21). Information technology is lifting the economy, and the politicians are backing it. *Business Week, Asian Edition*. Retrieved from *http://egov.mit.gov.in/*

Kshetri, N. B. (2001). Determinants of the locus of global e-commerce. *ElectronicMarkets*, *11*(4), 250-257.

Ministry of Technology. (1999). *Annual Report of 1999-2000*. Ministry of Information Technology, Government of India.

Mobile Youth. (2002). Indian youth love affair with SMS condemned by cultural activists. Mobile Youth Online. Retrieved from *http://www.mobileyouth.org/news/mobileyouth629.html*

Moni, M., & Vijayaditya, N. (2002). Convergence and eGovernance: National informatics centre – An active catalyst and facilitator in India. Retrieved from *http://waterinfo.nic.in/news/egover_convergence.html*

Nair, S. (2002). *Governance and public management*. Strategy Paper.

Nasscom. (2002). IT industry. Retrieved from *http://www.nasscom.org/it_industry.asp*

Pereira, M. (2002). Traffic cops and SMS. *Online NIC*. Retrieved from *http://www.delhitrafficpolice.nic.in/art3.htm*

Ramachandran, T.V. (2002). The Indian cellular mobile sector - Activities and concerns. *Proceedings of the June 13th Conference in Bangkok*. Retrieved from *http://www.itu.int/ITU-D/pdf/4597-11.1-en.pdf*

Rediff.com. (2002). Indian mobile sector grows 4.8% in July. *Rediff.com report*. Retrieved from *http://www.rediff.com/money/2002/aug/09cell.htm*

Reuters. (2002). Indian mobile users to touch 120 million by 2008. *Telephony*. Retrieved from *http://031102.coverstory.telephonyonline.com/microsites/newsarticle. asp?mode=print&newsarticleid=2652994&releaseid=&srid=10750&magazineid=7&siteid=3*

Reuters. (2003). Indian rural market has huge potential – Telecommunications sector. *Reuters Online.* Retrieved from *URL:http://in.tech.yahoo.com/030306/137/21u37.html*

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier.* UK: Addison-Wesley.

Shankar, J. (2004). mcommerce new mantra for Indian software firms. Retrieved from *http://www.islamonline.net/iol-english/dowalia/techng-2000-june-21/techng3.asp*

Singh, D. (2002). Keynote address by Chief Minister of Madhya Pradesh. *Proceedings of the Second Roundtable on IT in Governance,* March 12, Hyatt Regency, New Delhi.

Sinha, S. (1997, July-September). The risks of financing telecommunications projects. Indian Institute of Management, Ahmedabad, Vikalpa. *The Journal of Decision Makers, 22*(3), 1- 15.

Splettstoesser, D., & Kimaro, F. (2000). Benefits of IT-based decision-making in developing countries. *The Electronic Journal on Information Systems in Developing Countries, 3,* 1-12.

Thomas, K. S. (2002). r u hooked? Communications. The Week. Retrieved from *http://www.the-week.com/22feb03/life9.htm*

Times. (2003). Chennai beats other cities in SMS usage. *The Economic Times.* Retrieved from *http://economictimes.indiatimes.com/cms.dll/html/comp/articleshow? artid=14872448*

Unnithan, C. (2002). eGovernance in India – Initiatives and drivers – A preliminary investigation. *Proceedings of 2nd European Conference on eGovernment,* October 1-2, Oxford University, UK

Xu, H., Teo, H.H., & Wang, H. (2003). Foundations of SMS commerce success: Lessons from SMS messaging and co-opetition. *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS 36),* January.

This work was previously published in Electronic Business in Developing Countries: Opportunities and Challenges, edited by S. Kamel, pp. 170-192, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 6.14
Linking Businesses for Competitive Advantage:
A Mobile Agent–Based Approach

Tong-Seng Quah
Nanyang Technological University, Republic of Singapore

Chye-Huang Leow
Singapore Polytechnic, Republic of Singapore

ABSTRACT

In a highly competitive business environment, every organization is trying to achieve more using fewer resources. This is particularly true in this Internet era, where many businesses are moving from "brick-and-mortar"-based operation towards having at least an Internet presence, where e-commerce is fast gaining acceptance. Recent advances in mobile agent technology promise yet another powerful tool to gain competitive advantage—to deliver cost-effective services through utilizing Internet infrastructure. Such a development helps both individuals and organizations achieve higher productivity at lower cost. In this chapter, the authors describe an intelligent mobile agent-based system that links hotels and restaurants to provide gourmet goers with a convenient way of searching for their choice restaurants. The system sends off intelligent mobile agents to automatically roam the Internet, gather the relevant information about food and services from participating restaurants, and provide the most optimized selection as suggestions to help the users make their meals decision. This greatly reduces information overload for the users. Participating business establishments also benefit, through increased business.

INTRODUCTION

Agent-based systems have gained prominence over the last few years. One of the most interesting categories of agents is mobile agents (Lange & Oshima, 1998). Unlike static agents, which are restricted to operate within a single machine or address space, mobile agents have the ability to migrate over the network, execute tasks at each

location and potentially interact with other agents that cross their paths. Advantages of mobile agents include their ability to reduce network usage, increase asynchrony between clients and servers, add client-specified functionality to servers and introduce concurrency. These features help lower computing costs of modern businesses as well as better manage network traffic, as illustrated below.

Many online business transactions involve processes that require extensive database searches and matches. For example, users of an online bookstore are likely to view various catalogs, matching descriptions with preferences they have in mind before deciding which books to purchase. As such, information search and filtering applications often download and process large amounts of server-resident information and generate comparatively small amounts of result data. The scenario is greatly different with a mobile agent-based system, where mobile agents move to and execute on server machines and access server data without using the network, reducing bandwidth requirements. Many of today's applications involve repeated client-server interactions, which require either maintaining a network connection over an extended period or making several separate requests. If mobile agents are used instead, the client system does not have to maintain a network connection when its agents access and process information. This permits increased asynchrony between the client and server. This feature is especially useful for mobile computers (such as laptops and PDAs), which typically have low-bandwidth, unreliable connections to the network and are often switched off to save power consumption. Also, the repeated client-server interactions are reduced to two agent-transfer operations, reducing the frequency of network usage, as well.

An example of a user-level application would be an electronic marketplace. Vendors can set up online shops, with products, services or information for sale. A customer's agent would carry a shopping list along with a set of preferences, visit various sellers, find the best deal based on user preferences and purchase the product using digital forms of cash. An added advantage of such a system is that businesses may also be linked up to form a chain, such that mobile agents may move between stores within a business chain to make their purchases. Such a setup will enhance the competitive advantages of participating online stores.

Apart from mobility, a mobile agent-based system will need mechanisms for restricted resource access, secure electronic commerce, protection of agent data, robustness and user control over roaming agents. These will be discussed in later sections.

MOBILE AGENTS: ISSUES AND DEVELOPMENTS

Agents-Enabled Electronic Commerce

Mobile agents offer a number of useful possibilities:

- The agent can express the application-level protocol required to perform a transaction. This includes dialogs on choices and options, configurations, availability, delivery methods and opportunities for setting up, as well as complete and accurate capture of information required by the vendor in a particular format. Mobile agents technology is a plausible method for vendors to distribute the client end of a transaction protocol in a device-independent way.
- Alternatively, the mobile agent may be able to present the consumer's desire as a query to a number of potential vendors to determine degree of match, price, availability and so forth.

- The agent may also be able to consult a "consumer guide" or other advisor before making a purchase.
- The agent can provide a secure vehicle for the transaction, providing bilateral authentication and privacy.
- The agent can provide a transaction currency for settlement. The agent's account is presumably reconciled periodically against "real" money.

To facilitate the development of mobile agents distributed applications, and to overcome some problems and issues that arise from this approach, some requirements must be addressed. Systems that support the use of the mobile agent paradigm have to provide a basic set of services and characteristics. These will be discussed later.

Achievable Competitive Advantages Using Mobile Agent-Based E-Commerce Platforms

While many potential competitive advantages can be achieved using a mobile agent-based e-commerce system, the following are being highlighted:

1. **Efficient supply chain management:** Enterprises may link up to provide a wider range of products and services to customers. This will likely attract a larger customer base, and benefits all parties involved. For example, by linking up a hotel server with restaurant chain servers, guests in the hotel get the impression of a wider range of cutlery service available. The restaurants, on the other hand, make their presence noticed and are likely to get more business. All these can be achieved by the mobile agent applications "weaving" through the servers to retrieve and recommend cutlery establishments that match the hotel guests' preferences.
2. **Effective inventory control:** Many businesses have stores and shop-fronts at multiple locations. To minimize overstocking of inventories and tie up precious cash flow, many businesses keep their inventory low. However, this risks loss of sales when a customer wants goods that are out of stock at a certain branch store. The situation can be saved if such businesses link up their store-front computers using a mobile agent-based system. Such a system will enable a shopkeeper to find the availability of certain stock that matches a customer's request, thus capturing sales instead of letting a customer walk out of the shop and be disappointed.
3. **Powerful information searches for decision making:** Accessibility to information are crucial for important decision making such as loan approval—especially if the quantum is big. Using this example, credit providers may join a bureau which captures the 'worthiness' of private individuals. An agent-based system may be deployed to consolidate the credit situations of a loan applicant with various banks by utilizing the restricted access rights to the bureau controlled databases. This will help the loan approving officers to make informed decision and hence reduces bad debt for the lenders.

Agent Mobility

The primary identifying characteristics of mobile agents is their ability to autonomously migrate from host to host. Thus, support for agent mobility is a fundamental requirement of the agent infrastructure. An agent can request its host server to transport it to some remote destination. The agent server must then deactivate the agent, capture its state and transmit it to the server at the remote host. The destination server must restore the agent state and reactivate it at the remote host, thus completing the migration.

The state of an agent includes all its data, as well as the execution state of its thread. At the lowest level, this is represented by its execution context and call-stack. If this can be captured and transmitted along with the agent, the destination server can reactivate the thread at precisely the point where it requested the migration. An alternative is to capture the execution state at a higher level, in terms of application-defined agent data. The agent code can then direct the control flow appropriately when the state is restored at the destination.

Security Issues

The introduction of mobile agent code in a network raises several security issues. In an open network, such as the Internet, it is entirely possible that the agent and server belong to different administration domains. In such cases, they will have much lower levels of mutual trust. Servers are exposed to the risk of system penetration by malicious agents, analogous to viruses and Trojan horses. Security-related requirements are discussed in the following sections.

Privacy and Integrity

Agents carry their own code and data with them as they traverse the network. Parts of their state may be sensitive and need to be kept secret when the agent travels on the network. For example, a shopper agent may carry its owner's credit card number or personal preferences. The agent transport protocol needs to provide privacy to prevent eavesdroppers from acquiring sensitive information. Also, an agent may not trust all servers equally. We need a mechanism to selectively reveal different portions of the agent state to different servers. For example, a shopping agent may solicit quotations from various vendors. To ensure fairness, one vendor's quotation must not be readable or modifiable by others.

A security breach could result in the modification of the agent's code as it traverses the network. We need some means of verifying that an agent's code is unaltered during transit across a distrusted network or after visiting a distrusted server. An agent's state typically needs to be updated during its journey so it can collect information from servers. While we cannot assume that all servers visited are benign, we can provide mechanisms that allow such tampering to be detected.

Cryptographic mechanisms can be used to provide a secure communication facility, which an agent can use to communicate with its home site, or servers can use to transport agents safely across distrusted networks. Selective revealing of state can be accomplished by encrypting different parts of the state with different public keys belonging to the servers allowed to access those parts of the state. Mechanisms such as seals can be used to detect any tampering of agent code.

Authentication

When an agent attempts to transport itself to a remote server, the server needs to ascertain the identity of the agent's owner to decide what rights and privileges the agent will be given in the server's environment. A vendor's server needs to know the visiting agent's identity to determine which user to charge for service rendered. Conversely, when an agent migrates to a server, it needs some assurance of the identity of the server itself before it reveals any of its sensitive data to the server. Digital signature systems have been used to develop mutual authentication schemes. To verify signatures, agents and servers need to reliably know the signing entity's public key. This requires a key certification infrastructure. Public keys certified by trusted agencies can be posted in network-wide directories that can be accessed by agents and servers.

Authorization and Access Control

Authorization is the granting of specific resource access rights to specific principles (such as owners of agents). Some principals are more trusted than others, and thus, their agents can be granted less-restrictive access. This involves specifying policies for granting access to resources based either on identities of principals, their roles in an organization or their security classification.

Metering and Charging Mechanisms

When agents travel on a network, they consume resources, such as CPU time, disk space and so forth at different servers. These servers may legitimately expect to be reimbursed monetarily for providing such resources. Also, agents may access value-added services—information and so forth—provided by other agents, which could also expect payment in return. For example, in a marketplace, users can send agents to conduct purchases on their behalf. Thus, mechanisms are needed so that an agent can carry digital cash and use it to pay for resources used by it. Operating system-level support may be needed for metering of resource usage, such as the CPU time used by an agent or the amount of disk space needed during its visit.

Agent Monitoring and Control

An agent's parent application may need to monitor the agent's status while it executes on a remote host. If exceptions or errors occur during the agent's execution, the application may need to terminate the agent. This involves tracking the current location of the agent and requesting its host server to kill it.

Similarly, the agent owner may simply recall its agent back to its home site and allow it to continue executing there. This is equivalent to forcing the agent to execute a migrate call to its home site. The owner can use an event mechanism to signal the agent or raise an exception remotely. The agent's event/exception handler can respond by migrating home. This capability of remotely terminating and recalling agents raises security issues—only an agent's owner should have the authority to terminate it. Thus, some authentication functions need to be built into these primitives; that is, the system must ensure that the entity attempting to control the agent is indeed its owner, or has been authorized by the owner to do so.

COMPARATIVE STUDY OF E-COMMERCE REQUIREMENTS, AGLETS, AND HP WEB SERVICES

Finally, a set of e-commerce requirements will be defined to analyse Aglets' and HP Web Services' capabilities to fulfill them. The e-commerce requirements range from simple information exchange and bulk data transfer to secure firewall traversal, close collaboration and dynamic relationship requirements. It will be shown where each technology has its advantages and domains. This comparison also shows how the combination of both technologies can provide combined advantages and strengths.

Information Exchange in E-Commerce, Aglets, and HP Web Services

Many of today's e-commerce applications include complex business processes with a large number of concurrent tasks. These tasks may persist for a long duration; they may require long waiting times and could be nested within other tasks. Additionally, they are highly asynchronous, expose continues changes and may configure on the fly.

Thus, any flat conversation management, like message exchange, lacks the scalability for handling and tracking such sizable applications. Unfortunately, message exchange is the way

Aglets interact. These messages always follow the same basic scheme. They are composed of a "message type" in form of a string and a "message content", which can be any type of object. However, they do not support the demands of modern e-commerce.

Any more complex transactions in Aglets are usually implemented through a centralized scheduling architecture, where one Aglet host serves a coordination unit and does the scheduling, monitoring and execution control. This may work well within one single enterprise, but causes serious problems for inter-enterprise transactions.

HP Web Services, on the other hand, evolved from the Distributed Computing paradigm, which is primarily involved in handling such transactions. The e-brokering system was added on top of that, and it closely follows the e-commerce model. Business tasks are modeled as services and can be composed through other lower-level nested services. A typical complex HP Web Services request is broken down into simpler requests. The set of service providers for each of these simple requests is then dynamically discovered. Subsequently, the best match is invoked, and its execution mediated. This model used by HP Web Services fits exactly into the demands of e-commerce.

Bulk Data Transfer in E-Commerce, Aglets, and HP Web Services

As personalized, continuously running and semi-autonomous entities, Aglets can be used to mediate users and servers to automate a number of time-consuming tasks in e-commerce. However, again, Aglets communicate via message exchange, which may not be suitable for bulk data exchange. Routing and caching a large amount of data imposes a considerable burden for Aglets. For example, moving data between an operational database and a data warehouse via an Aglet is very unlikely.

HP Web Services can provide asynchronous and synchronous communication in the same environment. Bulk data transfer is an easy task for HP Web Services, as well as for other distributed computing environments, like CORBA and RMI. It fits closely into distributed computing and is a direct extension from Networking Transport Protocols (like TCP/IP).

Extensible Mark-up Language (XML) as Joint Communication Language in Aglets and HP Web Services

In today's technical world, many different domain specific ontologies (Hewlett Packard, 2000e) are used. Ontology refers to the common vocabulary and agreed semantics specific for a subject domain. Both HP Web Services and Aglets mainly focus on establishment of collaboration, mediation and providing services. They thereby aim at generic solutions to be applied across many different sectors of businesses. However, a banking institution may use an entirely different ontology than a CD retailer.

Currently, XML is in the process of solving this problem. Through the use of Document Type Definition (DTD), each sector can create its own semantic that fits individual needs and yet remains generally usable across sector boundaries. The power of XML and its role in e-commerce have been widely recognized. Consequently, HP Web Services provides support for XML in its Application Programming Interface (API).

The software developed during this project enables communication between Aglets and HP Web Services. The software can receive and send Aglet messages as well as deploy HP Web Services. And it exports all these functionalities in the form of handy modules, to be configured together to fit individual needs. Furthermore, reuse was one of the major design considerations for this project. The software could be easily extended with additional modules to implement

a proxy between the Aglet world, HP Web Services and the Internet. A DTD-based interpreter should closely fulfill these requirements. This would enable document-driven Aglet cooperation. Moreover, it would allow Aglets to share ontology (Hewlett Packard, 2000e) for multiple or even dynamic domains. In this way, the cooperation of dynamic Aglets would support *plug-and-play commerce*—mediating businesses that are built on one another's service. Aglets would acquire some of the key functionalities of HP Web Services.

Firewalls in Aglets and HP Web Services

Internet-based e-commerce involves multiple enterprises separated by firewalls. *Intra-enterprise* process management differs from *inter-enterprise* process management significantly. Different enterprises are not only separated by firewalls, but also have self-interests and individual data sharing scopes. When they are involved in a business process, they are unlikely to trust and rely on a centralized workflow server. Rather, they need support for peer-to-peer interactions. This has become the major impendence for using the conventional centralized workflow systems for inter-enterprise e-commerce automation.

One difficulty for the Aglet technology to fit into this picture consists in the limitation of its coordination model. HP Web Services, on the other hand, has Firewall Traversal as one of its standard services. Since HP Web Services has its roots in distributed operating systems research, it also has an integrated support for fine-grained access control. The HP Web Services Engine can be inserted at multiple points in the chain between clients and remote services. These remote services will act and look just like a local service, since the HP Web Services Engine acts like a kernel. Thus, the administrator can see and control access to services inside his network and firewall traversal is supported.

Collaboration in E-Commerce, Aglets and HP Web Services

An e-commerce scenario typically involves the following activities: identifying requirements, brokering products, brokering vendors, negotiating deals, or making purchase and payment transactions. Today, these activities are initiated and executed by humans.

Using Aglets or, in general, Mobile Agents technology, to support e-commerce automation is a promising direction. Aglets could be personalized, continuously running and semi-autonomous, driven by a set of beliefs, desires and intentions (BDI). They could be used to *mediate* users and servers to automate a number of the most time-consuming tasks in e-commerce with enhanced parallelism.

HP Web Services was primarily designed for enabling the creation of dynamic, Internet-based business relationships through the ad hoc discovery and interaction of e-services. E-services include applications, computing resources, business processes and information, delivered securely over the Internet. The HP Web Services Framework Specification (SFS) defines standard business interactions and conventions as XML documents that allow e-services to dynamically discover and negotiate with each other and compose themselves into more complex services.

Dynamics in E-Commerce, Aglets and HP Web Services

E-commerce applications operate in a distributed computing environment, involving multiple parties with dynamic availability and a large number of heterogeneous information sources with evolving contents. Dynamic relationships among a large number of autonomous service requesters, brokers and providers is common. A business partnership (e.g., between suppliers, resellers, brokers and customers) is often created dynamically and maintained only for the required duration, such

as a single transaction. E-commerce activities typically rely on distributed and autonomous tasks for dealing with such operational dynamics. Thus, e-commerce is a *plug-and-play* environment. Services need to be provided on demand. To support such dynamics, an e-commerce infrastructure must support the cooperation of loosely coupled e-business systems.

Aglets with predefined functions but without the ability to modify their behavior dynamically may be too limited for mediating e-commerce applications properly. They cannot switch roles or adjust their behavior to participate in dynamically formed partnerships. For Aglets to participate in such relationships, a complex and sophisticated *dynamic behavior modification* infrastructure has to be developed on top of the standard Aglet services.

Turning Aglet cooperation from conversation level to process level could be a solution. In general, businesses collaborate following certain rules, such as, "if you send me a price request then I will send you a quote"; and "if the quote I sent you is acceptable, then you will send me an order." These rules include sequences of steps to form a business process. Such business collaboration usually involves multiple Aglets, each responsible for managing or performing certain tasks that contribute to the process. Adding inter-enterprise cooperative process management capability into agent-based systems is critical for these business collaborations.

HP Web Services started from the beginning with the vision of dynamic brokering. Again, SFS allows e-services to dynamically discover and negotiate with each other and compose themselves into more complex services. This creates an open-service model, allowing all kinds of digital functionality to be delivered through a common set of APIs. SFS presents a uniform service abstraction and mediated access. New service types and semantics can be dynamically modeled using the common service representation of an HP Web Services resource. However, that requires all parties to comply with HP Web Services' service representation.

ARCHITECTURE OF THE CROSS-PLATFORM BRIDGE

As already mentioned, the whole software system exports utility methods for collaboration between HP Web Services' and Aglets' applications. All functionalities can be accessed independently and are designed in a highly modular way. The important implication of the bridge is that it provides a means for mobile agent applications to port from one platform to another, thus enlarging potential applications.

The whole software can be partitioned into three subsystems, each operating in different environments:

HP Web Services Client Software

The HP Web Services Client subsystem operates in a pure HP Web Services environment. Only the standard HP Web Services components and configurations are needed in the same way, as they are required for HP Web Services legacy applications. The environment includes a Client HP Web Services Core, where the HP Web Services Client software connects to and runs on top of it. The main purpose of this system is to provide access for external HP Web Services legacy applications to the collaboration functionalities provided by the Bridge Manager system.

Aglet Client Software

The Aglet Client software is the Aglet-side correspondent to the HP Web Services Client software. It operates from inside a Client Tahiti Server and exports a graphical user interface, as well as software interfaces to Aglet legacy applications. Again, only minimal Aglet configuration is required. This system mainly serves as a gateway for

Figure 1. HP Web Services client class diagram

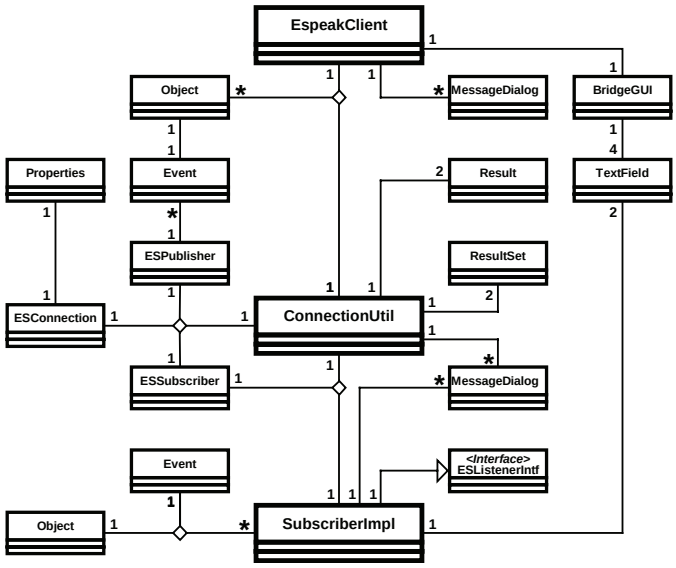
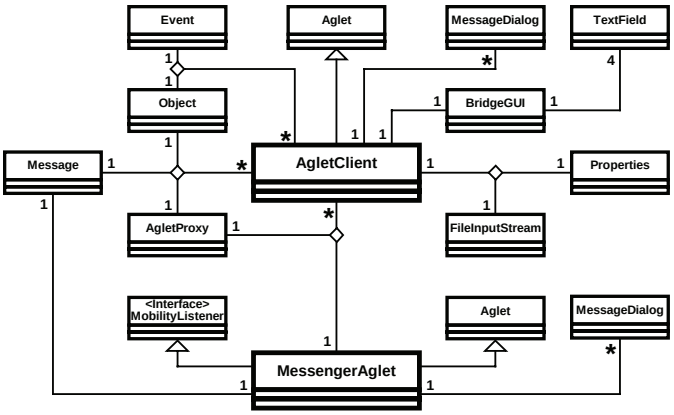


Figure 2. Aglet client class diagram



Aglets to deploy the Bridge Manager's collaboration services, similarly as the HP Web Services Client software does for HP Web Services.

Bridge Manager System

The Bridge Manager consists of both a Bridge HP Web Services Core and Bridge Tahiti Server. This combined unit allows collaboration between any HP Web Services and Aglet environment. The operating system hosting this entity needs to have both the configuration required for the HP Web Services environment and the configuration required for the Tahiti Server and Aglets environment. The HP Web Services side of the Bridge Manager mainly handles intercommunication and collaboration with HP Web Services legacy applications, whereas the Tahiti Server side attends to Aglets legacy applications.

There can be many instances of HP Web Services Client systems and Aglet Client systems. Their number is mainly limited by hardware

Figure 3. Bridge manager class diagram

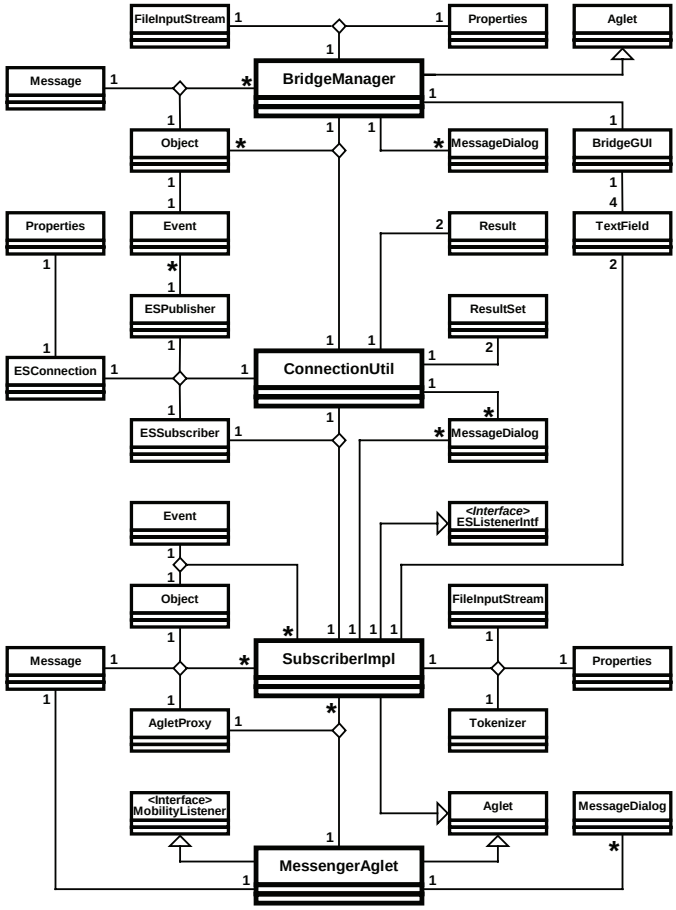
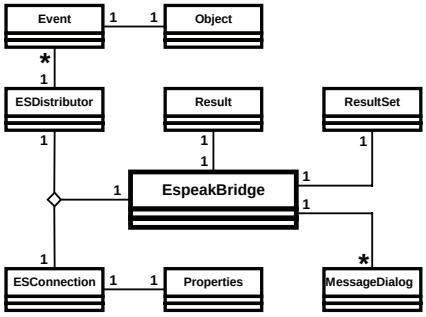


Figure 4. Event distributor class diagram



resources and network conditions. Additionally, there can also be many instances of Bridge Manager systems. They can be hosted on the same physical machine or in a distributed environment. Collaboration between all those entities will still be maintained, and they can form different communities of collaboration. Multiple memberships in different communities are also possible as dynamic entries and leaves.

DESIGN PRINCIPLES FOR THE BRIDGE BETWEEN AGLETS AND HP WEB SERVICES

The following sections elaborate on methodologies and general design principles that have to be deployed throughout the course of the design. These mainly describe fundamental design techniques necessary for developing HP Web Services software as well as Aglet software.

Choosing the HP Web Services Interface

Two interface options are available with HP Web Services:

J-ESI

This interface to HP Web Services is based on Java and allows interaction with the HP Web Services core or HP Web Services through APIs (Dantas et al., 2002).

WebAccess

This interface to HP Web Services is based on XML and enables interaction with the HP Web Services core or HP Web Services through standard Web browsers by returning HTML or XML documents (Glitho, Olougouna, & Pierre, 2002).

The Java model is oriented towards traditional API interfaces. Services are described by having an API or set of APIs. The client can make calls to discover services, retrieve a stub object and then invoke the services. These are typically synchronous methods, with calls to methods producing results, which the client will wait on.

The XML model, on the other hand, is a fundamentally asynchronous, document-based interface. Services are described not by a set of APIs but by a schema, which describes a set of XML documents that those services can understand. To find a service, a document defining the query for services is sent to WebAccess, which will then return a document describing these services, which fit the query criteria.

Computational Services

Computational services fit well with the API-style (Java) model. For instance, the contributed service of the Virtual File System is based on the Java model and exposes a core set of functional methods (Read, Write, Open, Close), which can be invoked by a client.

The API model typically assumes knowledge of the exact interface at programming time; usually through importing the Interface Definition Language (IDL) definitions at compile time to generate the stubs needed. This means that the interface must remain unchanged though the life of that version of the client. If the interface changes or is extended, the clients must be recompiled to handle or take advantage of the changes.

Business Services

Informational-, business- or broker-type services fit well with the document mode.

The client can discover changes or extensions in the document model when he or she downloads the schema (DTD). On one hand, the document model requires some additional effort in parsing the schema and handling different formats for documents; but on the other hand, this allows greater flexibility for the client software, since it is possible to handle a wider range of changes with recompiling.

Overview of Creating E-Services

The procedure of building and deploying an e-service with J-ESI (Dantas et al., 2002) involves three main steps:

- Specification of the contract (interface) for the service.
- Writing the implementation code for the interface.
- Deploying the service.

One of the first steps is to create the contracts (interfaces) for the services. E-services can be built with any programming language, but the interfaces used with J-ESI (Java HP Web Services Interface (Java-API)) have to conform to the HP Web Services IDL. The development stage involves writing the implementations for the interface. To accelerate development time, one may choose to deploy existing services (HP Web Services's standard services or third-party e-services) as components or convert legacy applications into e-services.

The next step is to specify relevant attributes and use the vocabulary service provided by the HP Web Services engine to describe the e-service. The e-services are registered through HP Web Services elements that are connected to a service engine and can be discovered across multiple groups of HP Web Services Cores within a community. Interactions between e-services are mediated by the HP Web Services infrastructure.

The Service Contract

This service contract (interface) is defined as an HP Web Services IDL, which is similar to the Java-RMI IDL and must have a ".esidl" extension for the IDL compiler to recognize it as an HP Web Services IDL file. It would have the following structure:

```
public interface SomeServiceIntf {
   public <returnType> firstMethod (<type> inParam,
      ...);
    public <returnType> secondMethod (<type> in-
      Param, ...);
}
```

The IDL file needs to be compiled using the HP Web Services IDL compiler:

```
java net.espeak.util.esidl.IDLCompiler
SomeServiceIntf.esidl.
```

The IDL compiler generates the following files, which are used by J-ESI (Dantas et al., 2002):

```
SomeServiceIntf.java
SomeServiceStub.java
SomeServiceIntfMessageRegistry.java
```

The file SomeServiceIntf.java is a copy of SomeServiceIntf.esidl, with minor changes to make it an HP Web Services interface. SomeServiceStub.java is the stub class that the service finder returns to the client when it discovers the look-up service. For every method defined in the interface, the stub class contains the code to create messages, marshal parameters and send it to the service provider. The SomeServiceIntf MessageRegistry.java is used by J-ESI to register the object types.

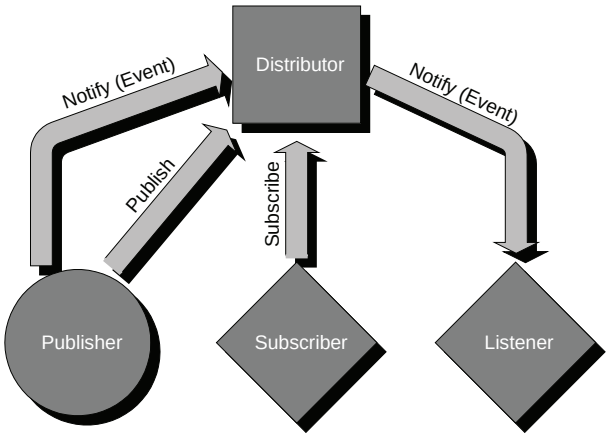
As mentioned in the previous chapter, one may choose to accelerate development by making use of HP Web Services build-in features. The J-ESI interface already provides functionalities to implement the publisher subscriber model. For this particular problem of collaboration utility methods, the authors' recommendation is to deploy J-ESI's net.espeak.jesi.event.ESListenerIntf class.

The next section describes the design details of HP Web Services's Event Service, an extensible service targeted at loosely coupled, distributed applications. Events provide a publish-subscribe mechanism for communication built on top of HP Web Services messaging.

The Event Model

HP Web Services supports an extended form of the familiar publisher-subscriber event model. There are four logical entities in the HP Web Services

Figure 5. HP Web Services Event Model



Event Model, whose interactions are illustrated in Figure 5. These entities are Publisher, Listener, Distributor and Subscriber.

A Publisher is an entity that generates an Event notification message. The recipient of an Event notification is called a Listener. A Distributor is an extension of a Listener. It receives Events and forwards them to other Listeners. A Subscriber is an entity that registers interest in a particular Event with a Distributor and designates the Listener to whom Events are sent. The Subscriber and Listener are typically the same physical entity. Similarly, it is fairly typical for a Publisher to act as a Distributor of its own Events.

In J-ESI (Dantas et al., 2002) these entities have the following representation:

- Event: net.espeak.infra.cci.events.Event
- Distributor: net.espeak.jesi.event.ESDistributor
- Listener: net.espeak.jesi.event.ESListenerIntf
- Publisher: net.espeak.jesi.event.ESPublisher
- Subscriber: net.espeak.jesi.event.ESSubscriber

The Core itself is an example of an Event Publisher. It sends Events to a trusted Client, called the Core Distributor, to signal state changes, such as a change in a Service's attributes. The Core Distributor may then distribute these Events to interested Clients that have appropriate authority.

Figure 6 illustrates a typical Event notification process where the Subscriber and Listener have been folded into a single Client.

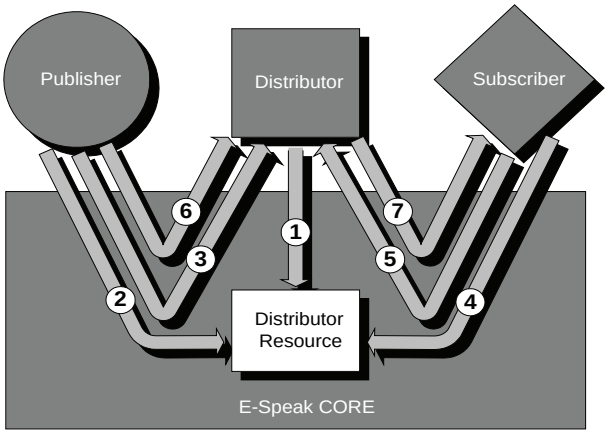
The following numbers in the figure represent these steps in the process:

1. The Distributor registers with the Core.
2. The Publisher discovers the Distributor.
3. The Publisher sends a Publish request to the Distributor describing the Events it will be generating.
4. The Subscriber discovers the Distributor.
5. The Subscriber sends a Subscribe request to a Distributor, describing the Events in which it is interested.
6. The Publisher sends the Event to the Distributor using a notify message.
7. The Distributor forwards the Event to the Subscriber (also using a notify request).

Communication in Aglets

The principal way for Aglets to communicate is by message passing. Inter-Aglet messaging is based on a simple event scheme that requires an Aglet

Figure 6. HP Web Services Event notification process



to implement handlers for the kinds of messages it is supposed to understand. These *"kinds of messages"* in Aglets are the direct correspondent to the *"event types"* in HP Web Services. The message-handling method is not directly called to send a message to an Aglet. Instead, a '`sendMessage`' method on a proxy is invoked. This proxy serves as a message gateway for the Aglet. One of the benefits of using a proxy is that it provides a location-independent interface for sending messages to Aglets. In other words, it does not matter whether a remote proxy (a proxy on a remote Aglet) or a local proxy is used to send a message; the interface remains the same.

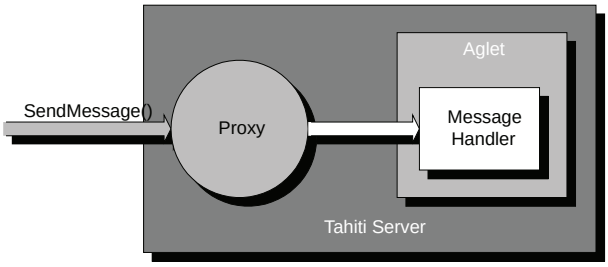
The '`sendMessage`' method of the proxy takes a message object as an argument and sends the message to the Aglet for which the proxy is acting as a gateway. The method may return an object in reply to the message.

The '`handleMessage`' method of the Aglet class is one of the key methods that one has to override. It enables the Aglet to respond to messages sent to it. Typically, the implementation of this method will consist of a switch statement that tests for the kind of the incoming message. The handler is supposed to return a Boolean value, whether the message kind was understood and handled. Figure 7 illustrates the message passing mechanism in Aglets.

Messages in Aglets are objects. A message object is characterized by its *kind*. This property is used to distinguish messages from each other. The `com.ibm.aglet.Message` class supports a range of constructor that all have *kind* as a mandatory argument. Message objects also contain an optional argument field for data associated with a particular message. The argument field can be either atomic (String, int, etc.) or tabular (Hash table). The many message constructors represent shortcuts for the initialization of the argument field. After an Aglet message handler receives messages, it will determine the kind of messages and then retrieve a possible argument from the message.

The Message class also provides methods for handling non-atomic arguments. The reason is that messages often need to carry multiple arguments to the receivers. Such arguments are most effectively handled as key-value pairs. The "`setArg`" and "`getArg`" methods are convenient for organizing multiple arguments into a table. Another group of methods in the Message class enables direct replies to incoming messages. The message handler thereby uses the incoming message object to deliver a reply.

Figure 7. Aglet message passing mechanism



As mentioned in this chapter, the proxy object plays a fundamental role in Aglet messaging. The next chapter, therefore, will introduce the Aglet proxy and describe the rationale behind this element of the Aglet API.

Aglet Collaboration through Proxies

An Aglet is fundamentally a mobile event and message handler. Associated with each Aglet is a proxy object that serves several purposes. Two of its most important roles are (1) as a shield to avoid uncontrolled access to the Aglet's public methods, and (2) as a convenient handle for a local, remote or deactivated Aglet.

When an Aglet is created (`AgletContext.createAglet()`), it is automatically associated with a proxy object that is returned to the application. The application should then use this proxy to control the Aglet. Unless the Aglet gives away an object reference to itself, it is impossible for the application or any other Aglet to access any of the public methods and fields in the Aglet.

It should never be necessary to operate directly on the Aglet itself. The application can control the Aglet through the proxy's methods: `clone()`, `dispose()`, `dispatch()`, `deactivate()` and `activate()`.

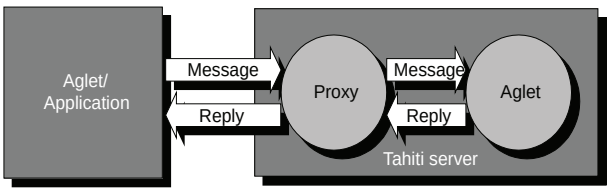
The `dispatch()` method will return a new proxy that gives control of the remote Aglet. It is here that location transparency comes in. The proxy returned by the `dispatch()` is identical to the proxy to a local Aglet, but referred to as remote proxy. As a consequence of this architecture, an Aglet can have only one local proxy, but multiple remote proxies. Figure 8 shows the relationship between an Aglet and its proxy.

AN IMPLEMENTATION EXAMPLE

Context

Every organization is trying to do more with less. This is particularly true of today's highly competitive global business environment. Many organizations are starting to turn towards technology to advance their competitive edge.

Figure 8. Relationship between Aglet and proxy



Mobile agent technology is a logical step in the evolution of e-commerce. This software technology enables a paradigm shift for the Internet from a "do-it-yourself" model to a "do-it-for-me" model.

Agent technology promises to deliver cost effective Inernet-specific labor. At the low end, agents can roam the Internet while we sleep and present us with documents of interest first thing in the morning. More sophisticated Internet-based agents can perform statistical mapping to find those elusive documents, songs or movies. In the near future, they may represent us on the Internet, negotiating and purchasing things we want.

Objective

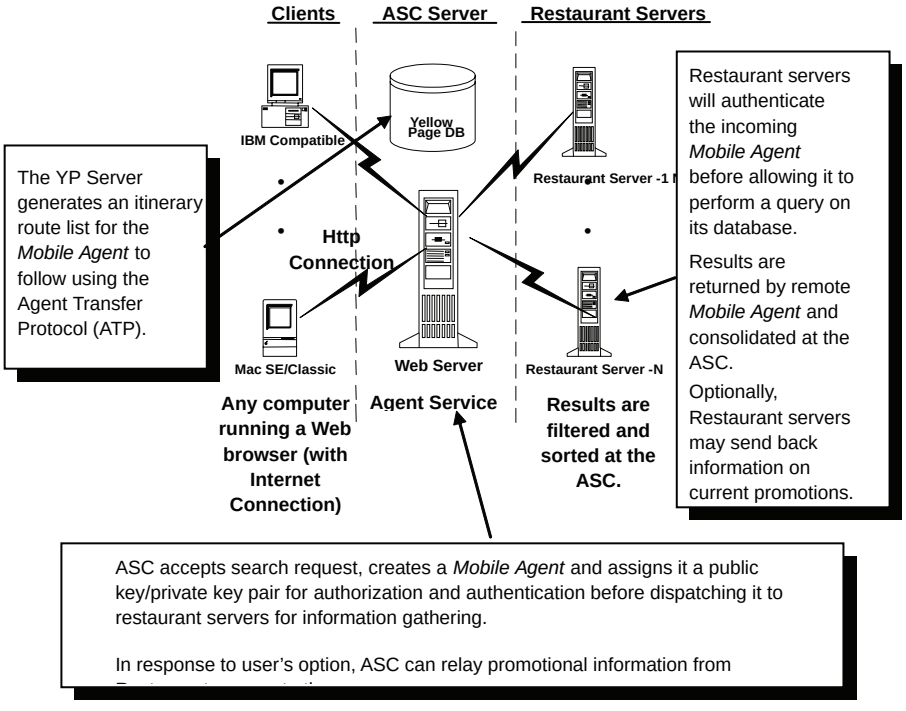
A project was started with the aim of building an intelligent software agent-based framework for Internet information gathering. A system for recommending local restaurants to hotel guests is built to demonstrate the framework. The intelligent software agent can automatically roam the Internet, gather the relevant information about some food and services from online restaurants and provide the most optimized selection as a suggestion to help the user make his decision. In addition, the restaurant servers can "push" promotional information to the customers. The system also provides security features to protect information and communications between the participating host servers.

An Overview of the System Structure

The mobile agent framework consists predominantly of Client, Agent Service Center (ASC) and the Restaurant Server Platform (Figure 9).

The hotel guest will be able to access the online restaurant recommendation system through a Web browser in his or her hotel room. The search request(s) made by the guest will be registered

Figure 9. Architecture framework for online restaurant guide system



with the ASC. The ASC will then process each request and generate a list of online restaurant sites that will likely to provide the pertinent information (i.e., food dishes) requested by the guest. The Yellow Page (YP) server, which provides a database of such online sites, helps to facilitate the compilation and generation of a list of such relevant online sites. This list of online sites will constitute the itinerary list that the Mobile Agent (MA) will have to visit. As mentioned, another main functionality of the ASC server would also be to generate a mobile agent that will begin traveling to the online restaurant sites to gather data on behalf of the user.

The mobile agent will abide by the generated itinerary list as it travels from one online restaurant server to another to complete its search for food and restaurant information. Upon arrival at each online site, the mobile agent will ask the restaurant server to search for the food based on the user's search requirements. In addition, the restaurant server can retrieve its promotional information and push it to the ASC server, which in turn displays the promotional information to the users.

Design Subsystems

The design subsystems are depicted in Figures 10 and 11.

ASC

The ASC subsystem accepts a search request from the Web Server and dispatches a Mobile Agent to Restaurant Servers for information gathering. The information gathered by the Mobile Agent will be consolidated at the ASC, which will then forward the findings to the Web Server and be displayed on the hotel guest's PC.

Within the Agent Service Subsystem exist the Agent Management subsystem and User Management subsystem, which provide the overall functionality of the parent subsystem. The Agent Management subsystem manages the creation of agents and provides search results to users. The User Management subsystem manages the creation of users (at the arrival of hotel guests) and maintains a user database.

Figure 10. Design subsystem

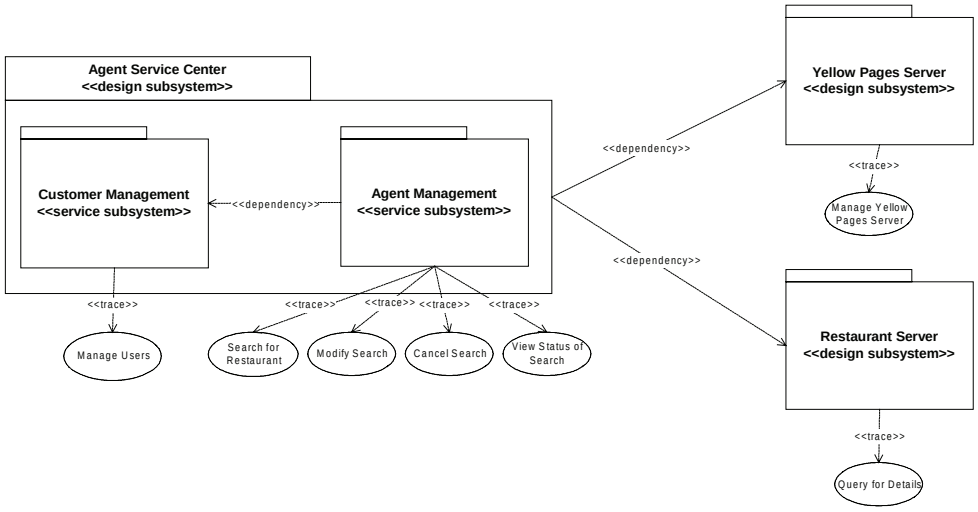
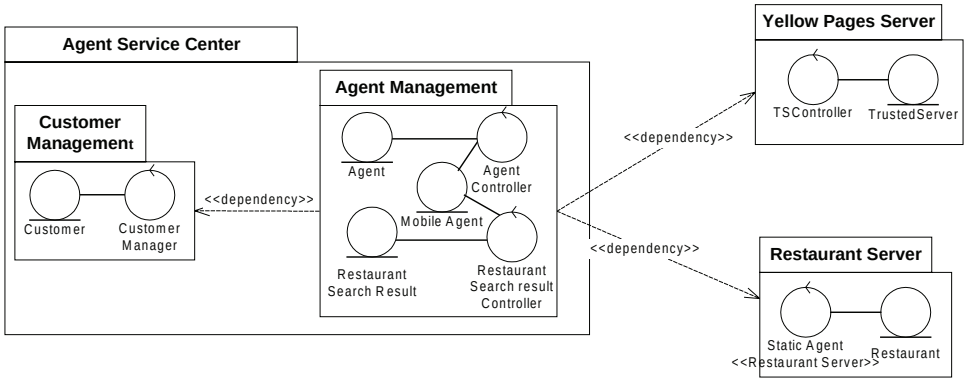


Figure 11. ASC



YP Server

The YP Server provides a list of online restaurant servers where the Mobile Agent can be dispatched. It consists of a database that provides the route list.

Restaurant Server

The Restaurant Server is normally provided by each participating restaurant to host the Mobile Agents. The server provides all the information related to the search request submitted by the customer. The dispatched Mobile Agent from the ASC will roam the Restaurant Servers according to the order as inscribed on the route list.

Key Services Identified

As a first version of the system, the following services have been identified.

Search for Restaurant

The user will select the "search for restaurant" option the main Web page. This will take him or her to a new Web page that will form the interface for specifying new search parameters. The user will be prompted to specify the search criteria. The system will generate a public/private key pair and route list for the newly created agent. The agent will then examine the route list and move to the first destination. The destination server will authenticate the mobile agent before allowing the parameters to be passed to the static local search agent. The search handler will start searching for stipulated requirements. The results will be filtered before passing them on to the mobile agent for return to update the search result database.

Modify Search

The user will select the "modify search for restaurant" option in the main Web page. This will take him or her to a new Web page that will form the interface for specifying new search parameters. The user will be prompted to specify the new search criteria. The Web page will pass this information to the ASC. The host will retract the existing mobile agent, following which, the system will generate a public/private key pair and a route list for a newly created agent. The agent will then examine the route list and move to the first destination.

Cancel Search

The user will select the "cancel search for restaurant" option in the main Web page. This will take him or her to a new Web page that will form

the interface for canceling the search. The system will show a list of active user's agents. The user then chooses the agents to be canceled. The user will be prompted for confirmation of the cancellation. On confirmation from the user, the host will track the location of the selected agent and retract it.

View Status of Search

The user selects an agent from a list of mobile agents using the system Web page. Upon selecting an agent, the user views the search status of selected agent by invoking the "view search" option. The agent controller coordinates the get search status event by asking the static agent to get the search results from the restaurant search results database, which resides in the ASC server. The results will then be displayed on the user screen.

Main Parameters of the System

The basic aim of the system is to allow the customer to search for restaurants in town that have an Internet presence using the mobile agent system. To achieve this, the customer will provide the following pieces of information to the system:

- **Ambience:** User can specify "air-conditioned," "non air-conditioned," "pool side" and so forth.
- **Average Price Rating of Meals:** User can choose a rating on how much he or she is willing to pay for a meal.
- **Location:** User can specify the region in the country where he or she wishes to have the meal; for example, Orchard, Marina Bay, City Hall and so forth.
- **Restaurant Specialty:** Whether the user has preference for any particular kind of food; for example, Italian, Mexican, Continental, Chinese and so forth.
- **Name of Dish:** Specify the name of any particular dish the user is looking for. He or she can choose the dish based on a textual description provided. Photos images will be shown when available.

Based on these inputs from the user, the mobile agent will roam from one server to another (each server being hosted by a restaurant and providing information about the restaurant) looking for restaurants that match the user's requirement. After the results are consolidated, they will be displayed on the user's PC. If the number of results obtained from the search is large—for example, more than 20—the system will then inform the user that the search has led to many results and will give him or her the option of either seeing all of them or redefining the search criteria.

Flow of Events

The following describe the process a user of the system has to go through:

1. User has been authenticated by the system as he logs in.
2. User invokes the systems by entering the search information needed to aid the search for restaurants or food before a new mobile agent is created and sent into the network. Parameters that the user has to provide were described earlier.
3. A new search record is created in the agent-track list and a mobile agent is created and dispatched to the restaurant server to search for the required information.
4. The search result is retrieved by the mobile agent from the restaurant server and is returned to ASC for filtering and updating of the search results workspace. The mobile agent then moves on to the next server for information gathering.

4a. At times, certain restaurants may have some promotional dishes on offer. When the mo-

bile agent visits those restaurant servers, it will leave the Internet address of its origin with the servers, so that the latter may send promotional information to the ASC.

5. Once the mobile agent has visited the last restaurant server on its itinerary, it will signal its status to the ASC, which will then present the consolidated search results to the user.

Key User Interfaces of the System

1. New search (creating new mobile agent)
2. Displaying results of search request
3. Announcing the availability of special promotional dishes
4. Displaying details of promotional dishes

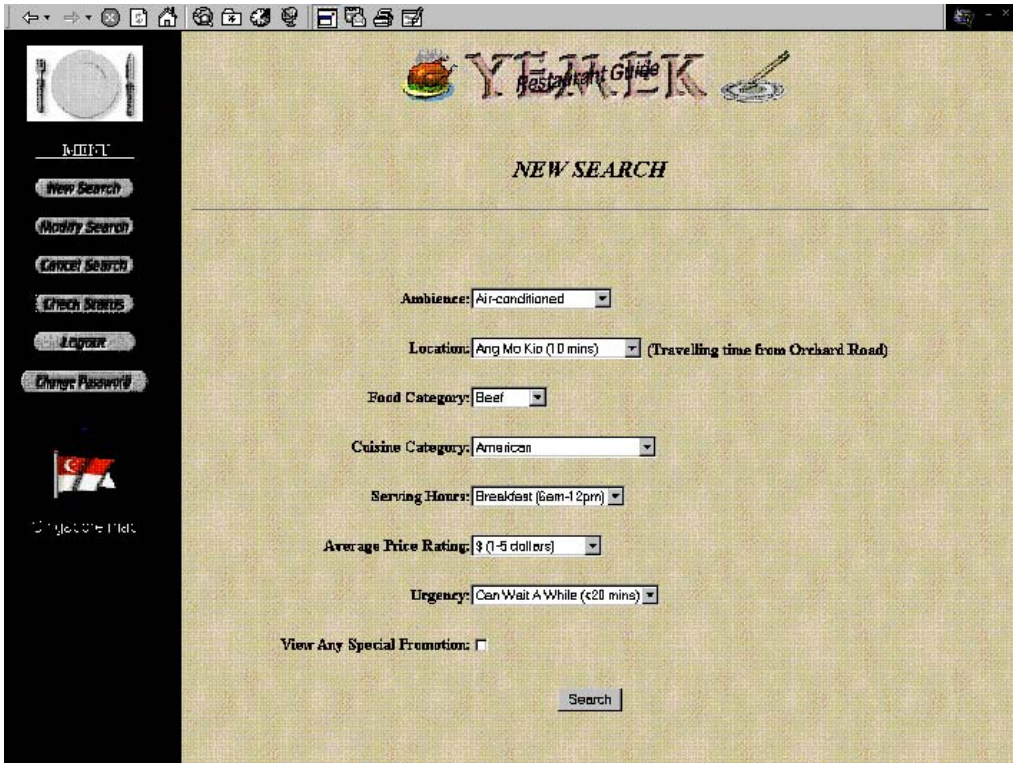
Summary of System Features

A mobile agent system has been developed to perform a restaurant food search for customers from restaurants with a Web presence. Also, details on promotional offers are also "pushed" to the user. While not described in this chapter, the security of the system and integrity of the data are ensured by means of cryptography and digital signature schemes. The system provides a user-friendly environment for easy usage.

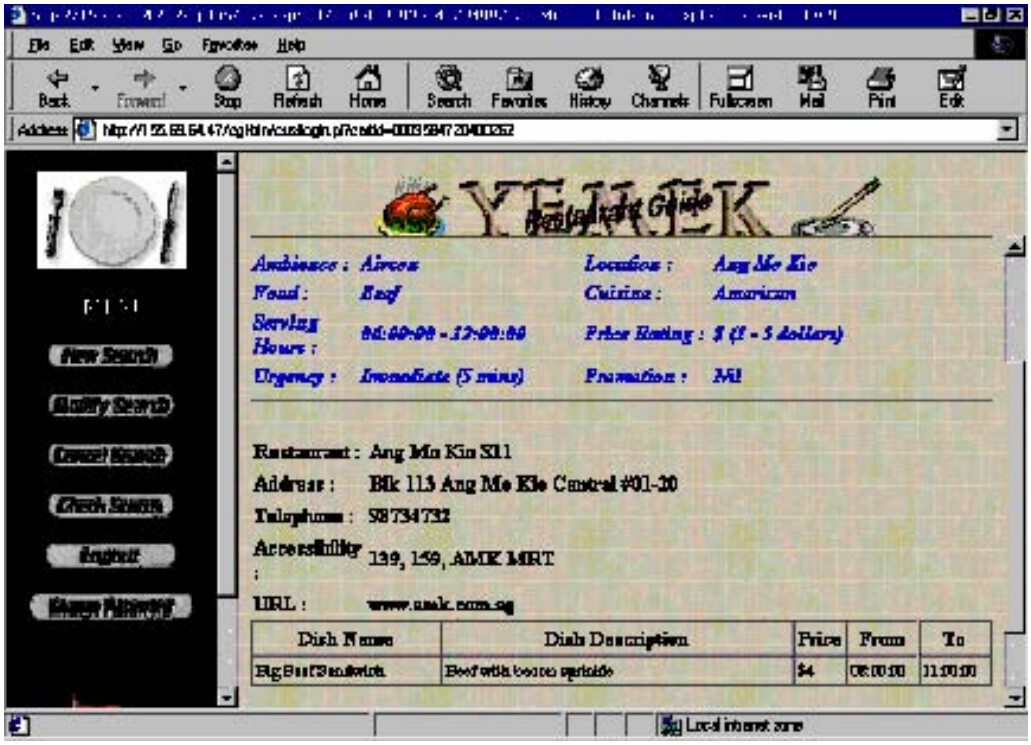
FUTURE ENHANCEMENTS

Besides fine-tuning the system to run in both Aglets and HP Web Services environments, we are also working on the following areas: We are extending the ASC functions to handle reservation

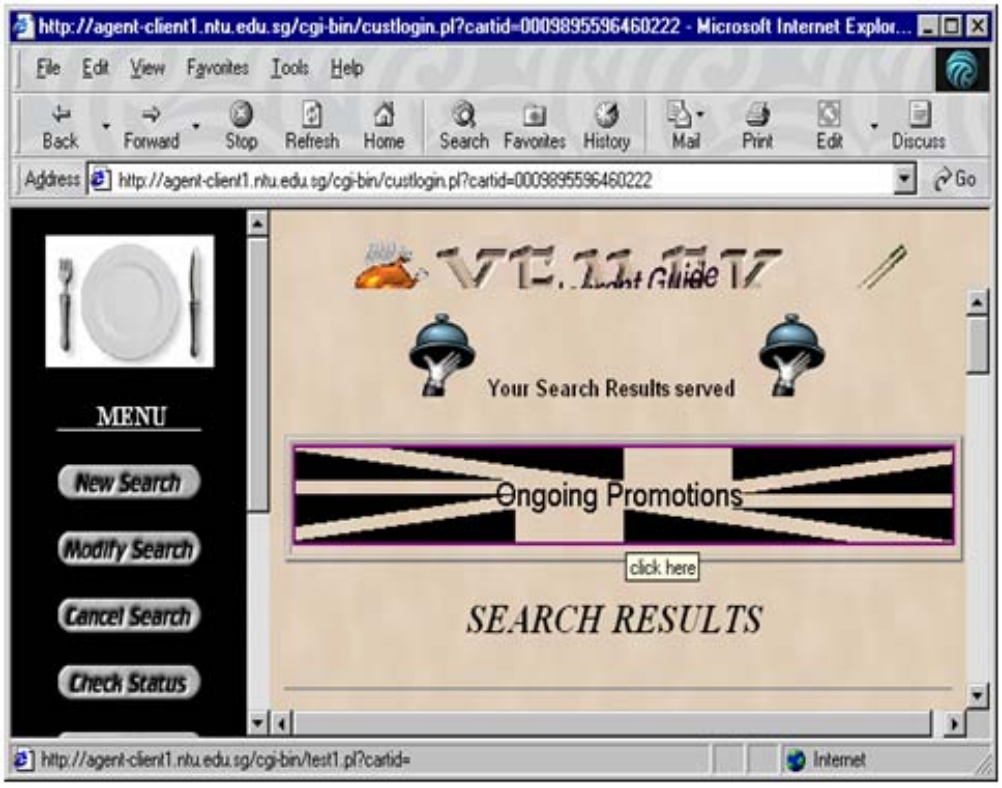
1. New search (creating new mobile agent)



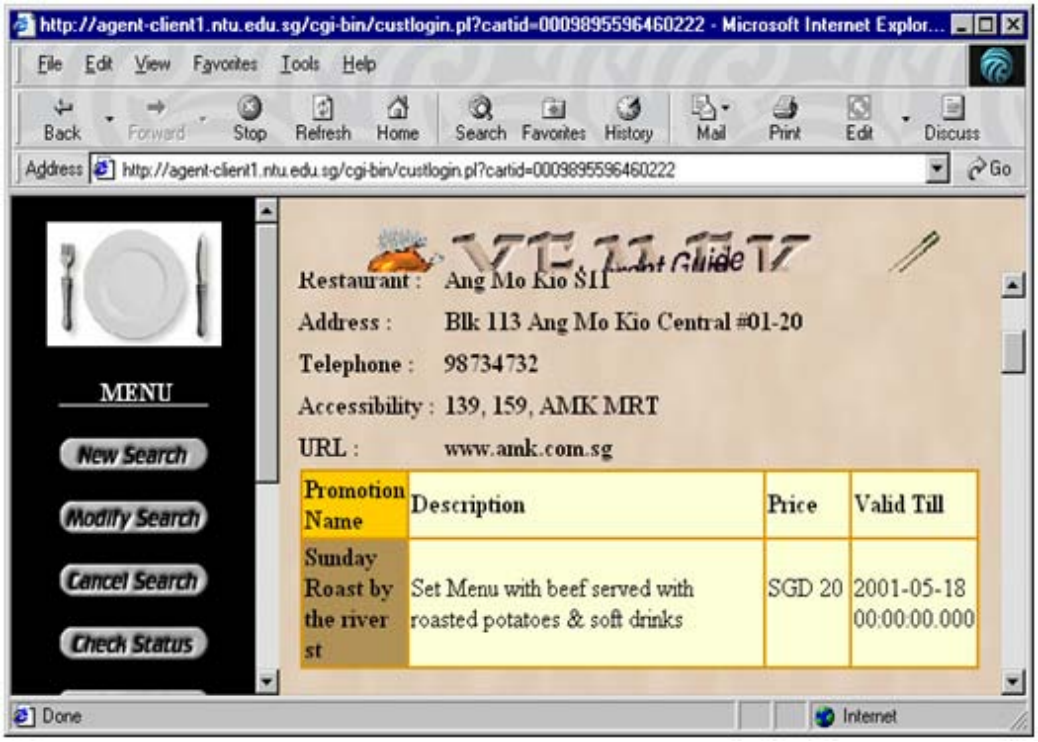
2. Displaying results of search request



3. Announcing the availability of special promotional dishes



4. Displaying details of promotional dishes



requests from the user. This will require further mobile agent activity such that the mobile agent will interact with the restaurant reservation system to place a booking. This will naturally involve payment options that must be provided for the user to pay for his meals. To push the technology further, we also are exploring the possibility of allowing autonomous negotiation by the mobile agent. Basically, mobile agents representing the users and the restaurant servers will meet at some cyberspace negotiation room to transact their requests for the respective hosts they are representing (Quah & Goh, 2002).

CONCLUSION

The above sections and application example have demonstrated the feasibility of creating a bridge to link Aglet and HP Web Services into a virtually common platform for recreating mobile agent applications. Such a mobile agent-based e-commerce system can indeed offer competitive advantages to businesses and help manage information flows to strategically link enterprises. Main advantages achieved through such a system are:

1. Lower cost of operation. This is mainly achieved through autonomous processes of the mobile agents.
2. Efficient supply-chain management. In the application example, hotels and restaurants have achieved a win-win collaboration through linking their services into a seamless system that provides added value for their common customers—hotel guests.
3. Conveniences to customers. This will likely increase patronage and thus improve revenue in-flow and enlarge market share.

In conclusion, it can be envisaged that the future of multi-platforms mobile agent-based systems is bright, and the number of potential applications is enormous. Such systems are likely to bring forth competitive advantages to business enterprises.

REFERENCES

Baek, J-W., Yeo, J-H., & Yeom, H-Y. (2002). Agent chaining: An approach to dynamic mobile agent planning. *Proceedings of the 22nd International Conference on Distributed Computing Systems* (pp. 522-529).

Binder, W. (2002). Using mobile agents for software distribution and maintenance: Autonomous stations capable of securely executing dynamically uploaded applications. *Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid* (pp. 352-353).

Chavez, A., & Maes, P. (1996). Kasbah: An agent marketplace for buying and selling goods. *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology,* London (pp. 75-90).

Cheng, S-T., Liu, J-P., Kao, J-L., & Chen, C-M. (2002). A new framework for mobile Web services. *Proceedings of the Symposium on Applications and the Internet (SAINT) Workshops* (pp. 218-222).

Cockayne, W.R., & Zyda, M. (1998). *Mobile agents.* Greenwich, CT: Manning Publications Co.

Dantas, M.A.R., Lopes, M.A.R., & Ramos, T.G. (2002). An enhanced scheduling approach in a distributed parallel environment using mobile agents. *Proceedings of the 16th Annual International Symposium on High Performance Computing Systems and Applications* (pp. 166-170).

Desic, S., & Huljenic, D. (2002). Agents based load balancing with component distribution capability. *Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid* (pp. 354-358).

Glitho, R.H., & Magedanz, T. (2002). Applicability of mobile agents to telecommunications. *IEEE Network, 16*(3).

Glitho, R.H., Olougouna, E., & Pierre, S. (2002). Mobile agents and their use for information retrieval: A brief overview and an elaborate case study. *IEEE Network, 16*(1), 34-41.

Goldschmidt, B., Laszlo, Z., Doller, M., & Kosch, H. (2002). Mobile agents in a distributed heterogeneous database system. *Proceedings of the 10th Euromicro Workshop on Parallel, Distributed and Network-based Processing* (pp. 123-128).

Gunter, M., & Braun, T. (2004). Internet service monitoring with mobile agents. *IEEE Network, 16*(3), 22-29.

Hewlett Packard. (2002). *HP Web Services platform*. Retrieved August 13, 2002, from www.hpmiddleware.com/SaISAPI.dll/SaServletEngine.class/products/hp_web_services/default.jsp

Hewlett Packard. (2000a). *Ten ways to think HP Web Services.* HP Web Services Documentation. Retrieved August 13, 2002, from www.HPWeb Services.net/library/pdfs/ThinkEspeak.pdf

Hewlett Packard. (2000b). *System management.* Retrieved August 13, 2002, from www.HP Web Services.net/library/pdfs/SysMgmt.pdf

Hewlett Packard. (2000c). *Architecture specification.* Retrieved August 13, 2002, from www. HP Web Services.net/library/pdfs/HP Web ServicesArch.pdf

Hewlett Packard. (2000d). JESI programmers guide. Retrieved August 13, 2002, from www. HP Web Services.net/library/pdfs/Jesi-Pgm-Guide.pdf

Hewlett Packard. (2000e). Service framework guide.

Hewlett Packard. (2000f). Contributed services. Retrieved August 13, 2002, from www.HP Web Services.net/library/pdfs/ContributedServices.pdf

Hewlett Packard. (2000g). WebAccess programmers guide. Retrieved August 13, 2002, from www.HP Web Services.net/library/pdfs/Webaccess-PgmGuide.pdf

IBM. (2002). Aglets software development kit 2. Retreived October 15, 2002, from www.trl.ibm.com/aglets/index_e.htm

Johansen, D., Lauvset, K.J.& Marzullo, K. (2002). An extensible software architecture for mobile components. *Proceedings of the 9th Annual IEEE International Conference and Workshop on Engineering of Computer-Based Systems* (pp. 231-237).

Komiya, T., Ohsida, H., & Takizawa, M. (2002). Mobile agent model for distributed systems. *Proceedings of the 22nd International Conference on Distributed Computing Systems* (pp. 131-136).

Lange, D.B., & Oshima, M. (1998). *Programming and deploying Java mobile agents with Aglets.* Boston: Addison-Wesley.

Liotta, A., Pavlou, G., & Knight, G. (2004). Exploiting agent mobility for large-scale network monitoring. *IEEE Network, 16*(3), 7-15.

Manvi, S.S., & Venkataram, P. (2002). Adaptive bandwidth reservation scheme for multimedia traffic using mobile agents. *Proceedings of the 5th IEEE International Conference on High Speed Networks and Multimedia Communications,* 370-374.

Marques, P., Fonseca, R., Simoes, P., Silva, L., & Silva, J.G. (2002). A component-based approach for integrating mobile agents into the existing Web infrastructure. *Proceedings of the Symposium on Applications and the Internet (SAINT) Workshops* (pp. 100-108).

Moukas, A., Guttman, R., Zacharia, G., & Maes, P. (1998). Agent-mediated electronic commerce. *MIT Media Laboratories.* Retrieved August 16, 2002, from http://ecommerce. media.mit.edu

Nakazawa Aoki, S., & Tokuda, J.H. (2002). Autonomous and asynchronous operation of networked appliances with mobile Agent. *Proceedings of Distributed Computing Systems Workshops* (pp. 743-748).

Object Management Group. (2000). Component object request broker architecture. Retrieved August 17, 2002, from www.omg.org

Papavassilio, S., Puliafito, A., Tomarchio, O., & Ye, J. (2003). Mobile agent-based approach for efficient network management and resource allocation: Framework and applications. *IEEE Journal on Selected Areas in Communications, 20*(4), 858-872.

Pleischm, S., & Schiper, A. (2002). Non-blocking transactional mobile agent execution. *Proceedings of the 22nd International Conference on Distributed Computing Systems* (pp. 402-403).

Puliafito, A., & Tomarchio, O. (2002). Design and development of a practical security model for a mobile agent system. *Proceedings of the 7th International Symposium on Computers and Communications* (pp. 477-483).

Quah, T.S., & Goh, L.Y. (2002). *E-negotiation by mobile agent* (technical report). Singapore, Nanyang Technological University.

Quah, T.S., & Schmid, A. (2001, November 8-11). Contrast, comparison and integration of Aglets and HP Web Services in e-commerce. *Proceedings of the E-Commerce Research Conference,* Dallas, TX.

Rusinikiewicz, M., Klas, W., Tesch, T., Wasch, J., & Muth, P. (1995). Towards a cooperative trans-

action model—The cooperative activity model. *Proceedings of the 21st International Conference on Very Large Databases,* Zurich, Switzerland (pp. 194-205).

Samaras, G., Spyrou, C., & Pitoura, E. (2002). View generator: A mobile agent based system for the creation and maintenance of Web views. *Proceedings of the 7th International Symposium on Computers and Communications* (pp. 761-767).

Scarpa, M., Zaia, V.M., & Puliafito, A. (2002). From client/server to mobile agents: An in-depth analysis of the related performance aspects. *Proceedings of the 7th International Symposium on Computers and Communications* (pp. 768-773).

Schmid, A. (2001). *Mapping of Aglets into HP Web Services* (masters dissertation). Singapore, Nanyang Technological University.

Schmidt, D.C. (2000). *Distributed object computing with COBRA middleware.* Washington University. Retrieved October 11, 2002, from www.cs.wustl.edu/~schmidt/cobra.html

Sun Microsystems. (2000). Java Remote Method Invocation (RMI). Retrieved August 13, 2002, from http://java.sun.com/products/jdk/rmi/

Sun Microsystems. (2000). Jini connection technology. Retrieved August 13, 2002, from www.sun.com/jini/

Urakami, M., Sigeyasu, T., & Matsuno, H. (2002). Performance evaluation of mobile agent on its living time and target existing rates in servers. *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops* (pp. 341-346).

World Wide Web Consortium (W3C). (2000). Extensible Mark-up Language (XML). Retrieved September 21, 2002, from www.w3.org/XML/

This work was previously published in IT-Enabled Strategic Management: Increasing Returns for the Organization, edited by B. Walters and Z. Tang, pp. 160-190, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 6.15
Integrating Mobile Technologies in Enterprise Architecture with a Focus on Global Supply Chain Management Systems

Bhuvan Unhelkar
University of Western Sydney, Australia

Ming-Chien Wu
University of Western Sydney, Australia

Abbass Ghanbary
University of Western Sydney, Australia

ABSTRACT

This chapter investigates opportunities to integrate mobile technologies within an organization's enterprise architecture (EA),
with an emphasis on supply chain management (SCM) systems. These SCM systems exist within the overall EA of the business. SCM systems are further influenced by the increasing modern-day need for information and communications technologies (ICTs) within a business, to bring together all of its disparate applications. The resultant enterprise application integration (EAI) also stands to benefit immensely from the incorporation of mobile technologies within it. Traditionally, supply chain management systems have involved management of the flows of material, information, and finances in a complex web of networks that include suppliers, manufacturers, distributors, retailers, and customers. Thus, these traditional supply chain management systems have a great need for integration under the umbrella of EAI. Mobile technologies can provide time and location independence to these EAIs in terms of

information in the supply chain systems, creating the possibility of multiple business processes that traverse diverse geographical regions. This chapter, based on the research conducted by the authors at the University of Western Sydney, discusses the opportunities that arise in supply chain management systems due to the time and location independence offered by mobility, and the resultant advantages and limitations of such integration to the business.

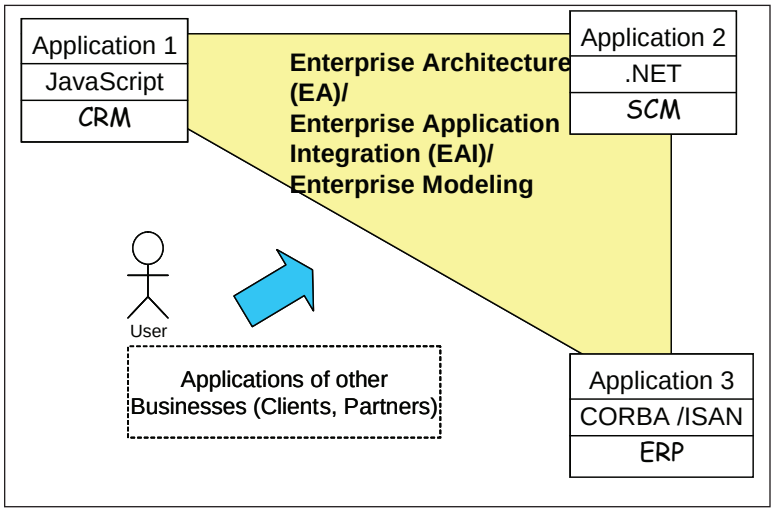
INTRODUCTION

A business enterprise uses a suite of different software applications to fulfill its various activities. These systems include supply chain management (SCM) systems, customer relationship management (CRM) systems, enterprise resource planning (ERP) systems, business intelligence (BI) systems, and other supporting financial and business systems. These enterprise systems do not operate in isolation. In fact, each of these systems depends on other systems, as well as large amounts of data in the background, to fulfill their own requirements. Specifically, supply chain management systems involve management of the flows of materials, information, and finance in a complex web of networks that include suppliers, manufacturers, distributors, retailers, and customers. The complexity of an SCM system requires it, as per Poirier (1999), to offer the right combination of data, products, and services to customers at the right time, right place, and right price. With rapidly increasing Internet access and business-to-business (B2B) connectivity, users of SCM are able to get their information needs easily—leading to what can be called electronic supply chain management (E-SCM) systems. E-SCM (Internet-based) systems are integrated together with all other enterprise applications, resulting in a comprehensive enterprise architecture (EA). Such an EA delivers the company a competitive advantage by opening up opportunities to streamline processes, reduce costs, increase customer satisfaction, and enable thorough strategic planning (Unhelkar & Lan, 2006). In today's modern business environment, it is important to further extend the advantages by incorporating wireless technologies and handheld devices in the organization's overall enterprise architecture. As Barnes (2002) mentions, the impact of wireless telecommunication on the Internet has taken a new turn. We use the mobile technology application for communications, working, banking, and shopping. The "time and location" independence provided by mobile technologies leads us into the era of mobile supply chain management (M-SCM) systems. It is important to understand these M-SCM systems within the context of the overall enterprise architecture. This chapter starts with a brief review of enterprise architecture and the issues related to enterprise application integration (EAI). This is followed by an understanding of the traditional SCM systems, together with the study of mobile technologies and applications. The chapter then describes the details of E-SCM and M-SCM. Finally, an outline of a model for integration of mobile technologies with SCM processes is then presented, together with its advantages and limitations.

ENTERPRISE ARCHITECTURE

An enterprise architecture represents the enterprise's key business, information, application, and technology strategies, and their impact on business functions and processes. EA consists of four key components: enterprise business architecture (EBA), enterprise information architecture (EIA), enterprise solution architecture (ESA), and enterprise technology architecture (ETA). The overall EA comprises software systems that may have been created using different programming languages and databases, and may be operating on different technology platforms. Figure 1 presents how EA is composed of different enterprise

Figure 1. Enterprise application integration composed of different enterprise systems



systems. However, Figure 1 also shows that users of the system want to see a unified view of the EA. This need for a unified view requires the enterprise to bring these various applications together, in an integrated fashion, resulting in enterprise application integration.

An enterprise business architecture that defines the enterprise business model and process cycles and timing also shows what functions should be integrated into the system. The enterprise information architecture focuses on which data and the corresponding data model should be integrated into the system. The enterprise solution architecture, also referred to as an application portfolio, is the collection of information systems supporting the EBA, which helps the user to easily understand and use the interface and components. Enterprise technology architecture is a consistent set of ICT standards which use infrastructures to support the EBA, EIA, and ESA. The infrastructures span across various different technical domain architectures, and include databases, applications, devices, middleware, networks, platforms, security, enterprise service buses, hosting, WLAN, LAN, Internet connection, operation system, servers, systems management, and so on (Pulkkinen, 2006).

Enterprise application integration maintains data integration and process integration across multiple systems, and provides real-time information access among systems. EAI not only links applications together, but also provides more effective and efficient business processes to the enterprise. There are numerous technologies that can be used for enterprise application integration, such as bus/hub, application connectivity, data format and transformation, integration modules, support for transactions, enterprise portal, Web service, and also service-oriented architecture (Finkelstein, 2006). Most importantly, however, the technologies of Web services build on extensible markup language (XML), Web services description language (WSDL), and universal description, discovery and integration (UDDI), which provide an excellent basis for integrating the applications of the enterprise—particularly when they are on

separate platforms.

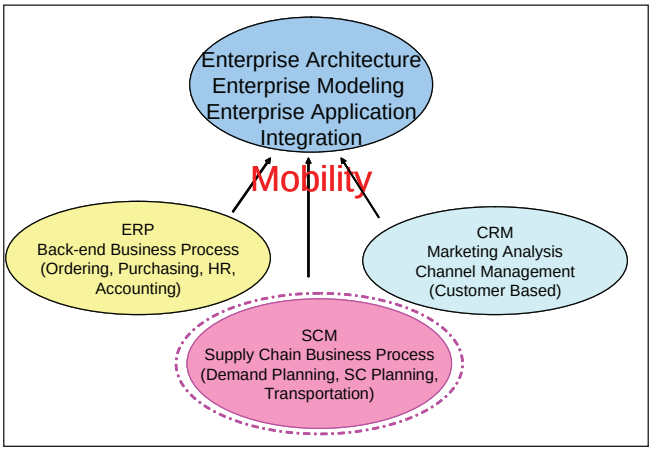
Linthicum (2000) declared that EAI enables the original chaotic enterprise processes to reach a semblance of order after the integration is achieved, resulting in increased efficiency on process flows, data integration, and data transportation. Furthermore, EAI also makes a more effective extension of enterprise processes. Irani, Themistocleous, and Love (2002) divided application integration into three categories: intra-organization, hybrid, and inter-organization. Intra-organizational application integrates packaged systems, custom applications, and EPR systems, and there are no transactions between external users or partners. Hybrid application integrates business-to–consumer (B2C) applications with IT infrastructure. Inter-organizational application integrates all processes between extended enterprises, such as the supply chain, and also can be a transaction between virtual enterprises, for example, e-procurement.

Further to the above discussion, in order to bring about a successful EAI, it is also important to create a model of such integration—called an enterprise model (EM). An EM would describe the objectives pursued by an enterprise and, as per Brathaug and Evjen (1996), focus on four aspects: process, product, organization, and systems. Doumeingts, Ducq, and Kleinhans (2000) defined enterprise modeling as the representation of enterprise activities at a global and a detailed level, using functions and processes in understanding its running. A good EM would take into account not only the technical aspects but also business, social, and human aspects of the enterprise. Such a comprehensive EM will also make it easier for the incorporation of mobility.

Kamogawa and Okada (2004) pointed out that integration of enterprise systems focuses on integrating collaboration agreements, collaboration profiles, business scenario integration, business process integration, and messaging technology. The three major applications—CRM, SCM, and ERP—are all shown in Figure 2. Figure 2 also shows how these applications are integrated through the enterprise model. Our aim is to further apply mobility to the enterprise model. However, in this chapter the discussion on the application of mobility to EM will be restricted to its application to SCM systems.

EA/EM/EAI all integrate enterprise applica-

Figure 2. Enterprise model (based on Kamogawa & Okada, 2004)



tions. Such integration cannot only enable the enterprise to present a unified view of the system to its suppliers and clients, but also reduces errors and improves quality by reducing or even eliminating duplication of data entry. Enterprise application integration brings about not only internal integration, but through extension also offers much more efficiency to its external suppliers, customers, and other trading partners over the Internet. Thus, providing mobility to EAI, and especially the SCM system, will connect existing and new systems to enable collaborative operation within the entire organization in real time—providing new and improved services without location and time limitations.

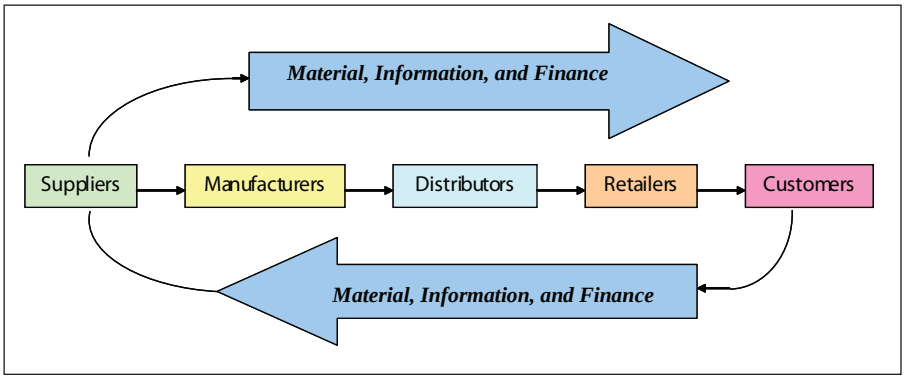
EXISTING MODEL OF SUPPLY CHAIN MANAGEMENT

In order to successfully apply mobility to SCM, we consider SCM within the overall context of EM. Traditional supply chain management involves the flows of material, information, and finance in a network, including suppliers, manufacturers, distributors, retailers, and customers. This is shown in Figure 3, where the various parties are connected to each other through the ability to supply goods, provide information about the goods, and deal with the financial aspects of those supplies.

The traditional supply chain is a push model. As outlined by Lee (2000), material flow includes both physical products flowing from the suppliers to customers through the chain and reverse flows via product returns, servicing, recycling, and disposal. Information flows involve order transmission and delivery status. Financial flows include credit terms, payment schedules, consignment, and title ownership arrangements. These flows cut across multiple functions and areas within a company and across companies (and sometimes industries); this leads to big challenges in terms of both technologies and business. Organization and integration of these flows within and across companies are important for effective supply chain management.

Supply chains can exist in both manufacturing and service organizations. In manufacturing organizations, they are mostly concerned with the flow of products and information between supply chain member organizations—procurement of materials, transformation of materials into finished products, and distribution of those products to end customers. However, in service organizations, supply chains

Figure 3. The supply chain flows (based on Lee, 2000)



can focus on "values added" to the services being offered. Overall, the current information-driven, integrated supply chains enable organizations to reduce inventory and costs, to add product value, to extend resources, to increase speed time to market, and to retain customers (Burt, Dobler, & Starling, 2003).

There are complex relationships in supply chains, such as multiple suppliers serving multiple customers, or a supplier who may be a customer or even a competitor in different parts of the chain. This complexity is the reason why we refer to supply chains as "supply networks" or "supply webs." Because of the network complexity, correct quality and time transfer of information can be extremely difficult. In particular, the multiple layers in a supply chain have the potential to misrepresent order information. This misrepresentation can lead to numerous confusion and errors, such as excessive inventory, inactive capacity, high manufacturing and transportation costs, and increasingly dissatisfied customers. Achieving supply chain efficiency requires exact and timely information. And the longer and more complex the supply chain is, the greater the requirement is, according to Dong (2001), for the supply chain to have precision in terms of the quality and timing of information.

SCM systems primarily include the requirement for integration architecture; the varied objectives of all the participants in a supply chain; mutual collaborative functioning of interfacing systems; and the varied levels of functionalities required for customers, retailers, suppliers, and manufacturers. A comprehensive supply chain management system can be an integration of a customer relationship management system, supplier relationship management system, order/purchasing system, delivery products management (logistics management) system, as well as time management systems.

The goal of supply chain management is to reduce inventories by optimizing material and information flow without sacrificing service level. SCMs have a responsibility to maintain sufficient inventory levels to satisfy the demands. Furthermore, increasingly, industries with short inventory and product cycles, such as high-tech and customer electronics, are highly reliant on SCM to provide them with the ability to interact with their numerous suppliers and retailers. SCM also allows high-quality customer service by delivering the right products to the right place at the right time.

MOBILE (EMERGING) TECHNOLOGIES

Mobile technologies can be considered as one set of the significant emerging technologies (as per Unhelkar, 2005) that have the potential to influence supply chains. Wireless technologies encompass any aspect of communication that is achieved without land-based or wired mechanisms. Thus, mobility, in a strict sense, is a subset of wireless technologies. This is because there can be some wireless communications that need not be mobile (for example, transmissions from a wireless radio tower or between two stationary servers). We consider a range of wireless technologies in this section that are likely to influence the various business processes of an organization. Later, we will discuss the specific mobile technologies (such as RFID) from the point of view of their usage in M-SCM.

Wireless Technologies

Wireless technology refers to technology without wires and phone lines that uses a multiplicity of devices for communications (IBM, n.d.). The term "wireless technology" can also be used to describe modern wireless connections such as those in cellular networks and wireless broadband Internet. In modern usage, wireless is a method of communication that uses low-powered radio waves to transmit data between the mobile terminals (Elliott & Phillips, 2003). The terminals, such as mobile phones, iPods, personal digital assistants (PDAs), global positioning systems

(GPSs), watches, email-only devices, handheld computers, and "wearable" technology, are carried by individuals and are far more "personal" than mere desktop PCs. The latest and important wireless technologies that require a brief discussion in this section are:

- "3G" mobile network
- Mobile satellite network
- Infrared
- Bluetooth
- Wireless
- Local area network
- WiMAX
- Radio frequency identification

Third-Generation Mobile Network

The development of 3G-related technologies has overcome the limitation of the previous generation of mobile technologies by allowing higher transmission rates and more complex e-commerce interactions (Barnes, 2002). Kuo and Yu (2005) and Huber (2002) list three 3G standards, including wideband code division multiple access (WCDMA), code division multiple access 2000 (CDMA2000), and time division–synchronized code division multiple access (TD-SCDMA), approved by the International Telecommunication Union (ITU).

W-CDMA is the most popular 3G mobile network which is capable of transferring multimedia between terminals; it is the technology behind the 3G universal mobile telecommunications system (UMTS) standard, combined with the 2G global system for mobile communications (GSM) standard, which is mainly dominated by European and Japanese firms.

Due to the promotion of the GSM organization and the 60% popularity usage of the 2G system in the global market, CDMA2000 gained the attention of many companies, especially U.S. and Korean firms that mainly support it. One of the advantages of the CDMA2000 system is the upgradeability of the narrowband CDMA system, so the user does not need to change his or her mobile device—just upgrade his or her user plan.

The TD-SCDMA includes three main key technologies: (1) TDMA/TDD principle, (2) smart transmitter and receiver, and (3) joint detection/terminal synchronization. It is mainly supported by China's Datang Telecom, which advocates its low-cost infrastructure.

Mobile Satellite Networks

Mobile satellite networks represent the convergence of the latest mobile technologies with space technologies. Satellites are operated at microwave radio frequencies in various bands, which are allocated by the ITU (2001). Olla (2005) declared that integrating space technology into mobile communications offers two main advantages. The first advantage is in providing access to voice and data service anywhere in the world—of which the current popular application is Internet phone (Voice over IP–VoIP). The second advantage is the exact positioning of useful location-sensitive information used for direction-finding-based and map-reading-based services, the current popular application of which is a car GPS. These applications are becoming commonplace, with Fitch (2004) pointing out that the technique for interfacing satellite links to global networks is well developed, including methods to overcome timing problems.

Infrared

Infrared (IR) technology provides directional electromagnetic radiation for "point-to-point" communication within short range. The radiation wavelength of IR communication is approximately between 750 nm and 1 millimeter. IR data transmission is a mobile application for short-range communication between a computer terminal and mobile device, such as a PDA or a mobile phone. Infrared communications are useful for indoor use

in areas of high population density. IR does not transmit through physical barriers such as a wall, and so it does not interfere with other devices in the vicinity. Infrared transmission is, therefore, the most common way for remote controllers to control physical machines. Furthermore, infrared lasers are used to provide the light for optical fiber communications systems; they are the best choice for standard silica fibers, as using infrared lasers can be a cheaper way to install a communications link in an urban area (Okuhata, Uno, Kumatani, Shirakawa, & Chiba, 1997).

Bluetooth

Bluetooth is a short-range radio technology developed to connect devices without wires. It is an effective technology for a new generation of Internet-capable mobile terminals. It enables numerous innovative services and applications, which function regardless of the mobile operator. The most important solution enabled by Bluetooth technology is synchronization between a PC server and one or more other mobile terminals. Synchronization has been particularly successful in cooperative applications, providing access to SCM systems (Paavilainen, 2001). Buttery and Sago (2004) describe the Bluetooth application as being built into more and more mobile telephones, allowing some very interesting m-commerce opportunities to be created. As people currently carry mobile phones with Bluetooth technology, these technologies can be used for making payments and related service concepts through simple downloads on their mobile devices. Retailers might also be able to provide samples of products to download via a Bluetooth link located close to the actual item, potentially resulting in better customer service and an enriched shopping experience. Bluetooth can operate up to 10 meters (eventually up to 100 meters in future versions). Since Bluetooth technology is a radio transmission, it does not need line-of-sight with another Bluetooth-enabled device to communicate (Schneiderman, 2002). Once Bluetooth technology is in place, one can envisage consumers walking around and giving out messages wirelessly via Bluetooth in order to buy items from vending machines, or buying low-value tickets, or even making small-value "cashless" purchases, such as newspapers.

Wireless Local Area Network (WLAN)

WLAN technology is closer to the fundamental principle of the Internet, wherein anybody can establish an individual network as long as it follows the general intranet guidelines. The wireless links would provide a network connection to all users in the surrounding area, ranging from a single room to an entire campus. The backbone of such a WLAN network may still use cables, with one or more wireless access points connecting the wireless users to the wired network. Currently, laptop computers and some PDA devices can be attached to a WLAN network using a compact flash (CF) or a Personal Computer Memory Card International Association (PCMCIA) card. In the future, PDAs and mobile phones might support multiple network technologies. WLAN is expected to continue to be an important form of connection in many business areas. The market is expected to grow as the benefits of WLAN are recognized (Paavilainen, 2001; Burness, Higgins, Sago, & Thorpe, 2004).

WiMAX

WiMAX is defined as Worldwide Interoperability for Microwave Access by the WiMAX Forum. The forum describes WiMAX as "a standards-based technology enabling the delivery of last mile wireless broadband access as an alternative to cable and DSL." The forum also states that it "will be incorporated in notebook computers and PDAs by 2007, allowing for urban areas and cities to become 'metro zones' for portable outdoor broadband wireless access" (WiMAX Forum,

2006). WiMAX delivers 72 Mbps over 30 miles point-to-point and four miles non-line-of-sight (NLOS) (Ohrtman, 2005). Its purpose is to ensure that broadband wireless radios manufactured for customer use interoperate from retailer to retailer. The main advantages of the WiMAX standard are to enable the implementation of advanced radio features in a standardized approach, and provide people in a city with online access via their mobile devices.

Radio Frequency Identification

RFID is an emerging technology that has been increasingly used in logistics and supply chain management in recent years. RFID technology can identify, sort, and control the product and information flow, all through a supply chain. Today, RFID is a standard technology that uses radio waves to automatically identify people or objects. There are several methods of identification, the most common of which use RFID tags and readers.

Ngai, Cheng, Au, and Lai (2005) proposed that RFID is made up of two components: the transponder, which is located on the object to be identified; and the reader, which, depending upon the design and the technology used, may be a read or write/read device.

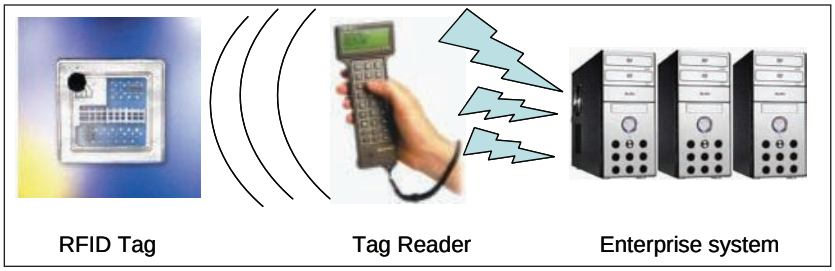
Roberts (2006) states that an RFID system will typically comprise the following three components, as shown in Figure 4:

- An RFID device (tag), which is a unique identifier for an object or person.
- A tag reader with an antenna and transceiver.
- A host system or connection to an enterprise system.

As Figure 4 shows, firstly we incorporate data inside an RFID tag. When the tag goes through a tag reader, the information inside the tag will automatically transfer to the host system. The host system is stored in a data center. After the data center analyzes and organizes the RFID tag information in the host system, specific useful tag information will be sent to a different enterprise SCM system.

Using an RFID system in the supply chain has been demonstrated by Asif and Mandviwalla (2005). Firstly, the SCM system constructs the item "where and when" during processing. When the items leave the manufactory and arrive at the place where they are to be read by the readers, the same information will be transferred directly to the distributor. The items are quickly sent to the correct trucks. As these items arrive at the retail outlet, they are read by the receiving RFID readers, and the retail outlet's inventories are updated automatically. Since the shelves at this outlet also have their own readers, they can directly increase replacement orders. However, using RFID technol-

Figure 4. Important parts of an RFID system (based on Roberts, 2006)



ogy in the SCM system, the items' quality can be automatically updated by the RFID reader sending into the SCM system. This provides highly location-based tracking, reduces the cost and human-error risks, and also improves the effectiveness and efficiency.

EPCglobal, a development of the earlier Auto-ID Center, is one of the two primary RFID standards setting groups. It proposed an Internet-based supply chain model that is aimed at improving supply chain end-to-end efficiency. A key component of the EPCglobal model is the Electronic Product Code or EPC. The manufacturer adds an RFID tag to every item of its product line. Each tag contains a unique EPC, which is a 96-bit code that uniquely identifies objects (items, cases, pallets, locations) in the supply chain (EPCglobal, 2005).

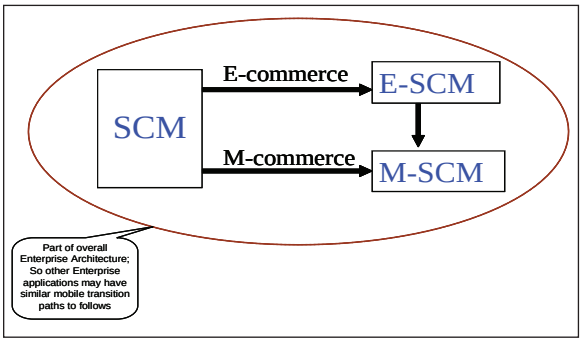
APPLICATION OF MOBILE TECHNOLOGY IN SUPPLY CHAIN MANAGEMENT

Based on the literature surveyed so far, there appears to be a gap between what is considered traditional SCM, electronic SCM, and the potential offered by mobile SCM. The major gap in the current literature in the SCM domain appears to be a lack of discussion between the two types of commerce and, in particular, the value derived from the ability of mobility to provide time and location independence. As shown in Figure 5, SCM can be extended to the e-commerce environment and at the same time it can

Table 1. Mobile technology application comparison table

Mobile Technology	Functions and Applications
3G Mobile Network	• Application: mobile phone device • Higher transmission rate • Popular used and high marketing acceptance
Mobile Satellite	• Application: GPS device and Internet phone (Voice over IP–VoIP) • Space technology • Direction finding and map reading
Infrared	• Application: remote controller • Communication in short distance • Low cost
Bluetooth	• Application: Bluetooth device • Synchronization • Transfers data between a PC server and one or more other mobile device(s)
WiMAX	• Wireless online in urban areas by using mobile device or any computer
WLAN (Wireless Local Area Network)	• Wireless link PC or mobile device network connection in particular surrounding area
RFID (Radio Frequency Identification)	• Application: RFID tag and reader • Product tracking and controlling by automatically updating the RFID tag location through RFID reader

Figure 5. SCM to E-SCM and M-SCM



also be extended to M-SCM. However, there is also a potential, as again shown in Figure 5, for E-SCM to be extended to M-SCM. These are some of the important aspects of SCM systems studied in this chapter.

E-Supply Chain Management

E-commerce deals with a combination of hardware technologies, software applications, and changes to business processes and appropriate customer strategies. E-SCM can enable customers to use electronic connections to obtain the information and associated services from the organization's supply chain system. The objective of E-SCM is to understand customer demographics, purchasing patterns, inventories, orders, and order fulfillments, in order to enable customer satisfaction and creation of new business opportunities (Arunatileka & Unhelkar, 2003). E-commerce provides the basis for much more efficient supply chains that can benefit both customers and manufacturers. This is because e-commerce, through connectivity, brings together various parties involved in commercial transactions. In today's environment, customers are less forgiving of poor customer service and more demanding of customized products or services. As the competition continues to introduce new offerings modified to the special needs of different market sectors, companies have to respond by offering similar custom-made and highly personalized offerings. The ensuing production of various goods and services for multiple countries, customer sections, and distribution exits creates major challenges in forecast, inventory management, production planning, and after-sales service support. Internet-based E-SCM systems bring the companies a competitive advantage by opening up opportunities to streamline processes, reduce costs, increase customer satisfaction, and make possible thorough planning abilities (Unhelkar & Lan, 2006).

The supply-chain e-business model creates a virtual value chain, and information flows across the supply chain. All members of the supply chain have strong electronic systems, and the information sharing to the customer is very effective in the ordering process, product delivery, and other SCM issues (Arunatileka & Arunatileka, 2003).

The e-supply chain is a pull model; in a pull-based supply chain, production and distribution are demand driven so that they are coordinated with true customer demand rather than forecasted demand. This is enabled by fast information flow mechanisms to transfer customer demands to the various supply chain participants. This leads to a significant reduction in inventory costs and enhanced ability to control materials when comparing this to the equivalent push-based system (Levi, Kaminsky, & Levi, 2003).

M-Supply Chain Management

In the 21st century, we are in the era of wireless and handheld technologies, and the impact of the Internet and wireless telecommunication has taken a new turn (Barnes, 2002). Mobile technologies are at the core of the communication revolution. They have increased commercial efforts from the removal of physical connectivity for people, processes, and businesses, resulting in a signifi-

cant impact on communication. Therefore, mobile devices can also be used to optimize the flow of information and materials. An increased number of mobile workers and time sensitivity drive companies towards advanced mobile solutions.

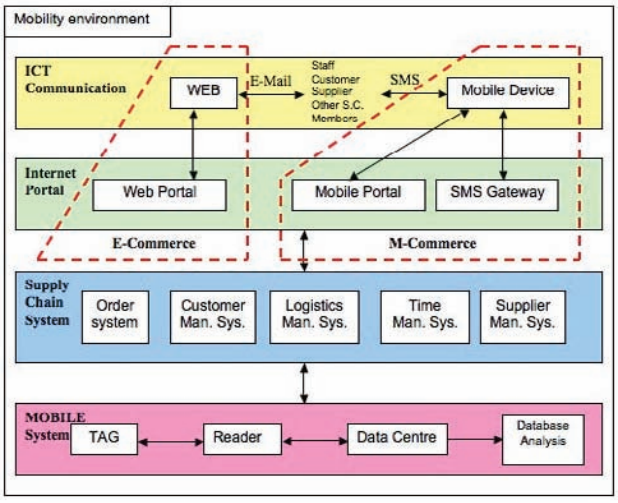
Paavilainen (2001) highlights that the solutions of supply chain management systems are highly time sensitive. The requirement of the time sensitiveness is that SCM systems must have the ability to transact products as close to real time as possible—opening up opportunities for the application of mobile technologies. By incorporating mobility in the SCM system processes, monitoring and receiving of immediate messages from the market can be improved. M-supply chain management focuses on the shortened cycle time from making an order to the fulfillment of that order—which, in most cases, would be delivery of the product to the customer. With mobility, response and confirmation time are much quicker than with the use of the standard Internet connections. Automatic data-reading and updating can be accessed from any mobile devices without the restriction of time and place. Traffic control systems (GPRSs) can be used to deliver products in the supply chain, and also automatically detect the products and report cars' and trucks' locations.

Eng (2005) declares that three main concerns of M-SCM are:

1. a place for efficient distribution of products and services,
2. timing for meeting customer demand and managing logistics, and
3. service quality for responsiveness and customer satisfaction.

In Figure 6, the M-SCM architecture is divided into four parts—the RFID system, SCM system, e-commerce, and m-commerce—and as proposed by us is an extension to the original ideas of Ngai et al. (2005). They state that RFID technology can be integrated with wireless networks to

Figure 6. Architecture of M-SCM systems (this model is proposed by these researchers as an extension to the original ideas of Ngai et al., 2005)



send information to a supply chain management system through a portal to staff, customers, and partners. We define that M-SCM refers to integrating the RFID system, e-commerce, and m-commerce through the Internet portal and ICT communication into a supply chain management system in a mobile environment. Moreover, we declare the RFID system's contents and supply chain management system's contents, and how e-commerce and m-commerce will integrate the SCM through the portal.

The RFID system typically integrates RFID tags, readers, and the host system that will store the received tag information in the data center. After the database analysis process inside the data center, information will be transferred from the RFID tag of each product to each company's enterprise SCM system.

The comprehensive supply chain management system can be an integration of a customer relationship management system, supplier relationship management system, order/purchasing (financial) system, delivery products (logistics) management system, and time management system.

The e-commerce and m-commerce parts show how the communication can occur among customers, staff of enterprises, suppliers, and other supply chain management members to the SCM system. People can use e-mail via the Web to an Internet Web portal to access the SCM system to get information or conduct e-commerce. In addition, they also can use SMS via a mobile device to a mobile portal or SMS gateway to access the SCM system to get information or carry out m-commerce.

The aforementioned model can also be used to understand how existing supply chain processes transfer to incorporate mobility in them. These can be called "m-transformation of SCM" processes. When a company already has the supply chain management system in its enterprise architecture, we can help it add a mobile system and set up the Internet portal with ICT communication devices to upgrade the SCM to M-SCM. In M-SCM, the cooperation of real-time processes by using mobile technology applications provides a stable workflow of up-to-date information from both inside and outside the company. Mobile supply chain applications can also allow users to request information and conduct "ubiquity"—whatever they want, whenever they want it, wherever they want it, and how they want it.

INVESTIGATION INTO ADVANTAGE AND LIMITATION OF MOBILITY IN SCM

The traditional supply chain is a push model. However, the E-SCM and the M-SCM models that we are discussing in this chapter can be considered as pull models. Table 2 depicts the comparison of characteristics between the push and pull models of the supply chain, based on various factors.

From Table 2, Levi et al. (2003) are quite clear in stating that traditional SCM focuses on reducing the cost and recording the location of products into the system. The information flow is based on a push model. As such, the traditional SCM process undergoes a long cycle, and the process time cannot be easily estimated. E-SCM focuses on customer service, and the process is totally based on responding to customers' orders and requirements. Every member of the supply chain can access the information flow from the E-SCM system at any time if they want to track it. Thus, E-SCM can be said to follow the "pull model" of information; it shortens the product cycle and lowers complexity compared with the traditional SCM (push model). The M-SCM can also be said to follow the "pull model," wherein information is extracted from the system when desired by the user.

Both E-SCM and M-SCM provide 24-hour information access, accurate and fast billing, less paperwork/fewer duplicated processes, on-site technical support, and trouble-shooting

Table 2. Push and pull model of SCM (Levi et al., 2003, p. 127)

Portion	Push	Pull
Objective	Minimize cost	Maximize service level
Complexity	High	Low
Focus	Resource allocation	Responsiveness
Lead Time	Long	Short
Processes	Supply chain planning	Order fulfillment

databases. However, in E-SCM, the users of the supply chain can only access the information of SCM when they connect to the Internet. It is limited by the location, in that members still need to sit inside the office to plug in their computers. M-SCM members can access the system at any time and anywhere by using mobile tools with a satellite connection. This brings huge benefits to people with a more convenient and comfortable environment to increase efficiency in their part of the value chain.

Following, we would like to list the advantages and limitations of mobility in SCM.

Advantages

The M-SCM system provides real-time data that can be accessed after logging into the system (Eng, 2005). Batten and Savage (2006) highlight that M-SCM will eliminate considerable duplication of data entry through simplified automated order placement, order status inquiries, delivery shipment, and invoicing. Mobility goes through data entry when it is created to reduce paperwork, document tracking, and human error. It also brings security control. The M-SCM system shortens the organization planning and production cycles, establishes one central data repository for the entire organization, and facilitates enhanced communications through all supply chain members communicating with each other by mobility (Paavilainen, 2001). Moreover, the M-SCM allows all users to request information, place orders for whatever they want, whenever they want it, wherever they want it, and how they want it.

The investigation of mobile technology application into the supply chain systems will bring benefits to all members of the supply chain, which includes suppliers, manufacturers, distributors, retailers, and customers.

Suppliers: The M-SCM system reduces paperwork, the number of administrative employees, and inventory costs, so the company reduces the cost of the products and brings in more customers as well as retaining its present customers. It also brings increased financial incomes to the suppliers (Unhelkar & Lan, 2006).

The RFID system stores the product contents when these have been acquired. When the product is about to expire, all details are contained in each RFID tag for each item. This can help suppliers to quickly obtain the information they need, simply by scanning the RFID tag once only. Also, this can help the suppliers to forecast the schedule arrangements to process the materials and prevent materials from expiring.

Manufacturers: The M-SCM systems can provide the organization with shorter planning and production cycles, and establish one central data repository for the entire organization. Then manufacturers will know the amounts they need to produce, and save money on the products that nobody buys.

RFID has been used in manufacturing to identify items or groups of items, express production procedures, and ensure the correct product quality. It helps that the right materials arrive at the right place, and nothing will be lost (Duckworth, 2004).

Distributors: The M-SCM system ensures quicker time-to-market for the firm's products, provides the retailer with enough stock, and also reduces excessive stocks in the distribution center. A traffic control system can be used as a GPRS to deliver goods to customers/retailers, and also automatically detect control and report locations of cars and trucks.

RFID identifies the distributed items stored inside the containers or trucks, and helps the delivery process to be 100% correct to deliver the right items to the right place at the right time. It has been reported that the RFID program can estimate the need for physical cycle counting, saving companies hundreds of employee hours and days of down time (Mullen & Moore, 2006).

Retailers: No data need to be re-entered into the M-SCM system through the simplified automatic order placement, order status inquiries, delivery shipment, and invoicing. The RFID system processes automatic data-reading and reporting to the supply chain system of the products' location (Wyld, 2006).

The RFID system is also very useful in maintaining enough stock levels in retailers (Karayannis, 2004). It has the ability to automatically record sales, check inventories, and refill or order additional stock in order to reduce the amount of inventory. Some food or items with limited time span can be notified by the system to guarantee retailers having fresh products on the shelves all the time. In addition, it has been proven to reduce theft from the retailers' shelves during the product delivery process in the supply chain. Loss avoidance directly benefits the retailers by minimizing costs and improving selections (Mullen & Moore, 2006).

Customers: The M-SCM system improves customer service substantially by efficient distribution of products and service, timing for meeting customer demands and logistics managing, and service quality for responsiveness and customer satisfaction. It improves the firm's ability to attract new customers and also retain its original customers (Unhelkar & Lan, 2006).

Customers can access information about products via mobile gadgets, and also place/check their orders and pay their bills in the M-SCM system by mobility (Mei, 2004). The M-SCM also allows customers to enquire for information and purchase any products at any time, anywhere, by any method.

Limitations

Supply Chain Integration and strategic partnering: The limitation that has existed in traditional supply chain integration is "reliance and trust" among the partners (other supply chain members) in this mechanism. This is because when the supply chain system integrates all companies in the supply chain, the members need to consider how much information they should announce to other members. Moreover, the more information a company can share, the greater the efficiency and effectiveness of the system. Nevertheless, the company members of the supply chain depend on other members' reliance and trust to share the information. So, this is the critical limitation of the original supply chain integration.

Cost of M-SCM system and facilities implementation: When the companies in the supply chain want to install M-SCM systems in their companies, the cost of systems and facilities during the establishment period is an essential consideration. They need to prepare the project fee at the beginning, but this is not a small amount, and it is the reason why the companies in the supply chain always find it difficult to make decisions to implement the M-SCM system.

Different countries develop mobility at different levels—the Internet speed, WLAN population: As we mentioned earlier, the companies of a supply chain may not be only in one country. Different countries may develop mobile technology at different levels. Some countries are not able to provide wireless local area networks, and can only provide dial-up Internet speeds. In other words, there are many limitations for those companies that want to use the M-SCM systems in the whole supply chain.

Security and privacy issues: Sheng (2006) points out that the issue of security and privacy is indeed a great concern, especially when the members of a supply chain make payments by the m-payment system. The qualifications of any system to provide secure data transfer are regarded as an important standard for both existing and potential users of m-payment systems. The personal privacy of customer information is a major concern to the customer who is deciding whether to use the system or not.

Companies need to change their business process: When the members of a supply chain decide to install the M-SCM systems in their supply chain management system, the current business processes may not be available or suitable to use all the time. The system development team needs to go through the original business processes to modify the new business processes suitable for the M-SCM system users. Therefore, after the system implementation in the companies, the system team needs to prepare a training class for employees regarding the business processes change and the new system utilization in the future.

BENEFITS OF MOBILITY TO THE ENTERPRISE ARCHITECTURE

Enterprise architecture represents a technology-business philosophy that provides the basis for cooperation between various systems of the organization that could be inside or outside the organizational boundary. EA also facilitates the ability to share data and information with business partners by enabling their applications to "talk" with each other. Linthicum (2000) pointed out that many organizations would like to build their entire systems by using the emerging technologies of today, of which mobile technology is a crucial part. Using mobile devices in enterprise modeling can help real-time information access among systems, in production planning and control, inbound and outbound logistics, material flows, monitoring functions, and performance measurements (Rolstadas & Andersen, 2000).

According to Ghanbary (2006), by correct application of mobile technologies into the business processes, the business enterprises are likely to gain advantages such as increased profits, satisfied customers, and greater customer loyalty. These customer-related advantages will accrue only when the organization investigates its customer behavior in the context of the mobile environment.

It is very important to identify that not many organizations have existing mobile solutions. Umar (2005) states that next-generation enterprises (NGEs) rely on automation, mobility, real-time business activity monitoring, agility, and self-service over widely distributed opera-

tions to conduct business. Mobility is one of the most invigorating features, having an enormous impact on how communication is evolving into the future.

Enterprise application integration is a relevant approach to integrating core business processes and data processing in the organization. Lee, Siau, and Hong (2003) state that EAI automates the integration process with less effort. EAI is a business computing term for plans, methods, and tools aimed at modernizing, consolidating, and coordinating the overall computer functionality in an enterprise. With new achievements in information technologies, companies are vulnerable if they do not respond to technologies such as mobile technology in a fast and appropriate manner.

BENEFITS OF MOBILITY TO THE GLOBAL ENTERPRISE

In this globalization era, many enterprises in a supply chain are located in different countries. Enterprises in some countries can provide low labor costs, and some in different countries may have low material costs, or others in different countries may provide professional skills or ideas about product design. However, all enterprises want to sell their product globally. The resultant ability of businesses and customers to connect to each other ubiquitously—independent of time and location—is the core driver of this change (Unhelkar, 2005). It leads the supply chain management to global supply chain management. Mobile technologies are thus a key influence in any efforts towards the globalization of business (Unhelkar, 2004). The processes of such m-transformation can lead an existing business into the mobile business via the adoption of suitable processes and technologies that enable mobility and pervasiveness (Marmaridis & Unhelkar, 2005).

M-SCM can further enhance the global SCM by reducing timing and cost, increasing correct delivery and customer satisfaction, and allowing global enterprises to conduct their business at any time and anywhere. Long (2003) pointed out that international logistics management focuses on international ship delivery schedule management, time, place, and product quality management.

An M-SCM system covers from planning, purchase, and production, to delivery to the customer. Mobile technology raises global enterprises to a much higher level of efficiency and effectiveness. Global enterprises can conduct their business at any time and anywhere, and provide high-quality products at low cost, and also support customer service 24 hours a day, seven days a week, by using an M-SCM system.

COLLABORATIVE SUPPLY CHAIN MANAGEMENT IN AN ENTERPRISE ARCHITECTURE

Electronic collaboration has been studied and experimented on by many studies. In electronic collaboration, there has been ample focus on the effects of dynamic environment and the rapidly evolving technology on organizations. Undoubtedly, these changes cause organizations to restructure and introduce a new suite of business processes to enable them to collaborate with the business processes of other organizations.

There are some critical technological issues that could cause drawbacks in collaboration across multiple SCM organizations. These issues could be classified as collaborations between different platforms, managing technology and maintenance (hardware/software). Web services technology is the solution for the collaboration of applications on different platforms, also independent of their different environments.

The electronic collaboration of the business processes of different SCM organizations is causing many practical issues. These issues are as follows:

- The excess inventory and inefficiencies in the supply chain while different organizations are involved.
- Requesting information by understanding the specific organization's capability to handle the request.
- Collaboration can reduce waste in the supply chain, but can also increase market sensitivity and increase customer expectation.
- Customer satisfaction, which is directly related to the previous issue, as customers expect more.
- Competition among all members of the partnership.

As such, the SCM must also address the following problems:

- **Distribution network configuration:** Number and location of suppliers, production facilities, distribution centers, warehouses, and customers.
- **Distribution strategy:** Centralized vs. decentralized, direct shipment, cross-docking, pull or push strategies, third-party logistics.
- **Information:** Integrate systems and processes through the supply chain to share valuable information, including demand signals, forecasts, inventory, and transportation.
- **Inventory management:** Quantity and location of inventory including raw materials, work-in-process, and finished goods.

Collaboration should be taking place in order to make sure that all parties involved in the collaboration are satisfied. In the past, collaboration was inadequate, with retailers hesitant to share information with others; however, the technology is capable of providing more support for the collaboration. Based on Horvath (2001), collaboration requires individual participants to adopt simplified, standardized solutions based on common architectures and data models. The time to market is critical, and participants will have to forego the luxuries of customization and modification that characterized the proprietary infrastructures of the past.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The chapter has introduced enterprise application integration as a part of a comprehensive enterprise model, followed by a discussion of supply chain management systems in the context of the EM. The traditional supply chain management system and investigation of the opportunities to integrate specific mobile technologies (such as 3G mobile network, mobile satellite network, Infrared, Bluetooth, WiMAX, and WLAN) with supply chain management systems have also been discussed. This chapter has further considered the advantages and limitations of such integration. The discussions in this chapter are important for understanding how and where mobile technologies fit into the overall concept of the enterprise architecture. The discussion on the collaborative nature of Web services and the ability of supply chain systems to capitalize on the connectivity of Web services is also important for globalization and appears in the global supply chain management system.

The aim of this discussion is to provide a solid theoretical basis for future research direction by the authors in the area of mobility and its incorporation in an organization's systems and architecture. Therefore, this chapter is a door opening to further research in the areas of mobile technologies in "enterprise architecture" and "SCM systems." The authors have provided more details of mobility SCM systems—actually enterprise architecture based on contributing the integration of all the enterprise information systems. Our research is opening opportunities

to future research in the area of investigating how mobile technologies influence integrating other systems of enterprise architecture such as enterprise resource planning, customer relationship management, customer order control and planning, material requirement planning, financial accounting, and so on. The team members of our group (MIRAG of AeIMS of UWS) are also investigating mobility influences on business process reengineering, Web services, and project planning. In addition, we still investigate our research into how and what should be included on contributing the comprehensive mobility enterprise architecture (M-EA) model.

REFERENCES

Arunatileka, S., & Arunatileka, D. (2003). E-transformation as a strategic tool for SMEs in developing nations. In *Proceedings of the International Conference on E-Government 2003,* New Delhi, India.

Arunatileka, D., & Unhelkar, B. (2003). Mobile technologies, providing new possibilities in customer relationship management. In *Proceedings of the 5th International Information Technology Conference,* Colombo, Sri Lanka.

Asif, Z., & Mandviwalla, M. (2005). Integrating the supply chain with RFID: An in-depth technical and business analysis. *Communications of the Association for Information Systems, 15,* 393-427.

Barnes, S.J. (2002). The mobile commerce value chain: Analysis and future development. *International Journal of Information Management, 22*(2), 91-108.

Batten, L.M., & Savage, R. (2006). Information sharing in supply chain systems. In Y.U. Lan (Ed.), *Global integrated supply chain systems* (ch. 5). London: Idea Group.

Brathaug, T.A., & Evjen, T.A. (1996). *Enterprise modeling.* Trondheim: SINTEF.

Burness, L., Higgins, D., Sago, A., & Thorpe, P. (2004). Wireless LANs□present and future. In *Mobile and wireless communications: Key technologies and future application* (ch. 3). British Telecommunications.

Burt, D.N., Dobler, D.W., & Starling, S.L. (2003). *World class supply management: The key to supply chain management* (7th ed.). Boston: McGraw-Hill/Irwin.

Buttery, S., & Sago, A. (2004). Future application of Bluetooth. In *Mobile and wireless communications: Key technologies and future application* (ch. 4). British Telecommunications.

Carayannis, J.P. (2004, July). *RFID-enabled supply chain replenishment.* Cambridge: MIT.

Dong, M. (2001). *Process modeling, performance analysis and configuration simulation in integrated supply chain network design*. Faculty of the Virginia Polytechnic Institute and State University, USA.

Doumeingts, G., Ducq, Y., & Kleinhans, S. (2000, August 21-25). Enterprise modeling techniques in year 2000. *Proceedings of ITBM 2000, IFIP 16th World Computer Congress,* Beijing, China.

Duckworth, D.A. (2004). *Potential for utilization of RFID in the semiconductor manufacturing intermediate supply chain.* Cambridge: MIT.

Elliott, G., & Phillips, N. (2003). *Mobile commerce and wireless computing systems.* Boston: Addison-Wesley.

Eng, T.Y. (2005). *Mobile supply chain management: Challenges for implementation*. Elsevier.

EPCglobal. (2005). *Homepage*. Retrieved March 21, 2006, from http://www.epcglobalinc.org/

Finkelstein, C. (2006). *Enterprise architecture for integration: Rapid delivery methods and*

technologies. Artech House.

Fitch, M. (2004). The use of satellite for multimedia communications. In *Mobile and wireless communications: Key technologies and future application* (ch.10). British Telecommunications.

Ghanbary, A. (2006). Evaluation of mobile technologies in the context of their applications, limitations and transformation. In B. Unhelkar (Ed.), *Mobile business: Technological, methodological and social perspectives.* Hershey, PA: Idea Group.

Horvath, L. (2001). *Collaboration: The key to value creation in supply chain management*. Retrieved October 12, 2006, from http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/1770060501.html

Huber, J.F. (2002). Towards the mobile Internet. *Communications of the ACM,* (October).

IBM. (n.d.) *IBM new to wireless technology.* Retrieved February 22, 2006, from http://www-128.ibm.com/developerworks/wireless/newto/#1

Irani, Z., Themistocleous, M., & Love, P.E.D. (2002). The impact of enterprise application integration on information system lifecycles. *Information and Management, 41,* 177-187.

ITU Radio Regulation. (2001). *Frequency allocation* (vol. 1, article 5). ITU.

Kamogawa, T., & Okada, H. (2004). Issues of e-business implementation from enterprise architecture viewpoint. *Proceedings of the 2004 International Symposium on Applications and the Internet Workshops* (SAINTW'04).

Kuo, Y.F., & Yu, C.W. (2005). *3G telecommunication operators' challenges and roles: A perspective of mobile commerce value chain.* Elsevier.

Lee, H.L. (2000, January). *Supply chain management review.* School of Engineering, Stanford University, USA.

Lee, J., Siau, K., & Hong, S. (2003). Enterprise integration with ERP and EAI. *Communications of the ACM, 46*(2).

Levi, D.S, Kaminsky, P., & Levi, E.S. (2003). *Designing and managing the supply chain: Concepts, strategies and case studies* (2nd ed.). New York: McGraw-Hill.

Linthicum, D.S. (2000). *Enterprise application integration.* Boston: Addison-Wesley.

Long, D. (2003). *International logistics global supply chain management*. Boston: Kluwer Academic.

Marmaridis, I., & Unhelkar, B. (2005). Challenges in mobile transformations: A requirements modeling perspective for small and medium enterprises (SMEs). *Proceedings of the M-Business International Conference*, Sydney, Australia.

Mei, Q.R. (2004). *RFID impact supply chain: Innovation in demand planning and customer fulfillment.* Cambridge: MIT.

Mullen, D., & Moore, B. (2006). Automatic identification and data collection. In *RFID: Applications, security, and privacy* (ch. 1). Boston: Addison-Wesley Pearson Education.

Ngai, E.W.T., Cheng, T.C.E., Au, S., & Lai, K.H. (2005). *Mobile commerce integrated with RFID technology in a container depot.* Elsevier.

Ohrtman, F. (2005). *WiMAX handbook: Building 802.16 wireless networks.* New York: McGraw-Hill.

Okuhata, H., Uno, H., Kumatani, K., Shirakawa, I., & Chiba, T. (1997). 4MBPS infrared wireless link dedicated to mobile computing. *IEEE,* 463-467.

Olla, P. (2005). *Incorporating commercial space technology into mobile services: Developing innovative business models.* Hershey, PA: Idea

Group.

Paavilainen, J. (2001). *Mobile business strategies: Understanding the technologies and opportunities.* Wireless Press.

Poirier, C.C. (1999). *Advanced supply chain management: How to build a sustained competitive advantage* (1st ed.). Berrett-Koehler.

Pulkkinen, M. (2006). Systemic management of architectural decisions in enterprise architecture planning. Four dimensions and three abstraction levels. *Proceedings of the 39th Hawaii International Conference on System Sciences.*

Roberts, C.M. (2006). Radio frequency identification. *Computers Security, 25,* 18-26.

Rolstadas, A., & Andersen, B. (2000). *Enterprise modeling improving global industrial competitiveness.* Kluwer Academic.

Scheniderman, R. (2002). *The mobile technology question and answer book.* Amacom.

Sheng, M.L. (2006). Global integrated supply chain implementation: The challenges of e-procurement. In Y.U. Lan (Ed.), *Global integrated supply chain systems* (ch. 6). London: Idea Group.

Umar, A. (2005). IT infrastructure to enable next generation enterprises. *Information Systems Frontiers, 7*(3).

Unhelkar, B. (2004). Globalization with mobility. *Proceedings of ADCOM 2004, the 12th International Conference on Advanced Computing and Communications,* Ahmedabad, India.

Unhelkar, B. (2005). Transitioning to a mobile enterprise: A three-dimensional framework. *Cutter IT Journal, 18*(8).

Unhelkar, B., & Lan, Y.C. (2006). A methodology for developing an integrated supply chain management system. In Y.U. Lan (Ed.), *Global integrated supply chain systems* (ch. 1). London: Idea Group.

WiMAX Forum. (2006). *Frequently asked questions.* Retrieved from http://www.wimaxforum.org/technology/faq

Wyld, D.C. (2006). The next big RFID application: Correctly steering two billion bags a year through today's less-than-friendly skies. In *Handbook of research in mobile business: Technical, methodological and social perspectives* (ch. 54). Hershey, PA: Idea Group.

ADDITIONAL READING

Anckar, B., & D'Incau, D. (2002). Value added services in mobile commerce: An analytical framework and empirical findings from a national consumer survey. *Proceedings of the 35th Hawaii International Conference on System Sciences.*

Basole, R.C. (2004). *The value and impact of mobile information and communication technologies.* Atlanta: Georgia Institute of Technology.

Basole, R.C. (2005). Transforming enterprises through mobile applications: A multi-phase framework. *Proceedings of the 11th America's Conference on Information Systems,* Omaha, NE.

Bernard, H.B. (1999). *Constructing blueprints for enterprise IT architectures.* Wiley Computer.

Carbine, J.A. (2004). *IT architecture toolkit.* Englewood Cliffs, NJ: Prentice Hall.

Chopra, S., & Meindl, P. (2007). *Supply chain management: Strategy, planning & operation* (3rd ed.). Englewood Cliffs, NJ: Pernson Prentice-Hall.

Cook, M.A. (1996). *Building enterprise information architectures reengineering information systems.* Englewood Cliffs, NJ: Prentice Hall.

Cummins, F.A. (2002). *Enterprise integration:*

An architecture for enterprise application and systems integration. Wiley Computing.

Dimitris, C.N. (2001). *Integrating ERP, CRM, SCM, and smart materials.* Boca Raton.

Eckfeldt, B. (2005). What does RFID do for the consumer? *Communications of the ACM, 48*(9).

Er, M., & Kay, R. (2005). Mobile technology adoption for mobile information systems: An activity theory perspective. *Proceedings of ICMB'05*, Sydney, Australia.

Garfinkel, S., & Rosenberge, B. (2006). *RFID: Applications, security, and privacy.* Boston: Addison-Wesley Pearson Education.

Gattorna, J.L., & Walters, D.W. (1996). *Managing the supply chain: A strategy perspective.* Palgrave.

Gershman, A. (2002). Ubiquitous commerce—always on, always aware, always pro-active. *Proceedings of the 2002 Symposium on Applications and the Internet.*

Guitton, A. (2004). *The value of RFID in transportation: from greater operational efficiency to collaboration transportation management.* Cambridge: MIT.

Hammer, M., & Champy, J. (2001). *Reengineering the corporation: A manifesto for business revolution.* London: Nicholas Brealey.

Hawryszkiewycz, I., & Steele, R. (2005). A framework for integrating mobility into collaborative business processes. *Proceedings of the International Conference on Mobile Business.*

Hoque, F. (2000). *E-enterprise: Business models, architecture, and components.* Cambridge: Cambridge University Press.

Jarvenpaa, S.L., Lang, K.R., Takeda, Y., & Tuunainen, K. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41-44.

Kalakota, R., & Robinson, M. (2001). *E-business 2.0: Roadmap for success.* Boston: Addison-Wesley.

Knolmayer, G., Mertens, P., & Zeier, A. (2002). *Supply chain management based on SAP systems.* Springer.

Kou, W., & Yesha, Y. (Eds.). (2006). *Enabling technologies for wireless e-business.* Berlin/Heidelberg: Springer-Verlag.

Lan, Y., & Unhelkar, B. (2005). *Global enterprise transitions: Managing the process.* Hershey, PA: Idea Group.

Lyytinen, K., & Yoo, Y. (2002). Issues and challenges in ubiquitous computing. *Communications of the ACM, 45*(12), 62-65.

May, P. (2001). *Mobile commerce: Opportunities, applications, and technologies of wireless business.* Cambridge: Cambridge University Press.

Medvidovic, N., Mikic-Rakic, M., Mehta, N.R., & Malek, S. (2003). Software architectural support for handheld computing. *Computer, 36*(9), 66-73.

Murugesan, S., & Unhelkar, B. (2004). A road map for successful ICT innovation: Turning great ideas into successful implementations. *Cutter IT Journal, 17*(11), 5-12.

Myerson, J.M. (2005). *RFID in the supply chain: A guide to selection and implementation*. CRC Press.

Passerini, K., & Patten, K. (2005). Preparing IT organizations for the mobile revolution. *Cutter IT Journal, 18*(8), 19-27.

Ptak, C.A. (2000). *ERP: Tools, techniques, and applications for integrating the supply chain.* Boca Raton, FL: St. Lucie Press.

Raisinghani, M., & Taylor, D. (2006). Going

global: A technology review. In Y.U. Lan (Ed.), *Global integrated supply chain systems*. London: Idea Group.

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Sandoe, K., Cornitt, G., & Boykin, R. (2001). *Enterprise integration*. New York: John Wiley & Sons.

Senn, J.A. (2000). The emergence of m-commerce. *Computer, 33*(12), 148-150.

Sharif, A.M., Elliman, T., Love, P.E.D., & Badii, A. (2004). Integrating the IS with the enterprise: Key EAI research challenges. *Journal of Enterprise Information Management, 17*(2), 164-170.

Spewak, S.H., & Hill, S.C. (1992). *Enterprise architecture planning: Developing a bluepoint for data, applications, and technology*. New York: John Wiley & Sons.

Sun, J. (2003). Information requirement: Elicitation in mobile commerce. *Communications of the ACM, 46*(12), 45-47.

Unhelkar, B. (2004). Paradigm shift in the process of electronic globalization of businesses resulting from the impact of Web services based technologies. *Proceedings of IRMA 2004*.

Unhelkar, B. (Ed.). (2006). *Handbook of research in mobile business: Technical, methodological and social perspectives*. Hershey, PA: Idea Group Reference.

Urbaczewski, A., Valacich, J.S., Jessup, L.M., & Guest Editors. (2003). Mobile commerce: Opportunities and challenges. *Communications of the ACM, 46*(12), 30-32.

Varshney, U. (2000). Recent advances in wireless networking. *Computer, 33*(6), 100-103.

Varshney, U. (2003). The status and future of 802.11-based WLANs. *Computer, 36*(6), 102-105.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications*, 185-198.

Vaughan-Nichols, S.J. (2003). Mobile IPv6 and the future of wireless Internet access. *Computer, 36*(2), 18-20.

Wang, Y., Van der Kar, E., Meijer, G., & Hunteler, M. (2005). Improving business processes with mobile workforce solutions. *Proceedings of the International Conference on Mobile Business*, Sydney, Australia.

Weilenmann, A. (2003). *Doing mobility*. Goteborg University, Sweden.

Wisner, J.D., Leong, G.K., & Tank, C. (2005). *Principles of supply chain management: A balanced approach*. Thomson South-Western.

This work was previously published in Handbook of Research on Global Information Technology Management in the Digital Economy, edited by M. Raisinghani, pp. 499-518, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.16
Mobile Business Process Reengineering:
How to Measure the Input of Mobile Applications to Business Processes in European Hospitals

Dieter Hertweck
University for Applied Sciences Heilbronn, Germany

Asarnusch Rashid
Research Center for Information Technology Karlsruhe, Germany

ABSTRACT

There is an ongoing debate about the value of mobile applications for the optimization of business processes in European hospitals. Thus finding satisfying methods to measure the profitability of mobile applications seems to be of great importance. Prior research had its focus mainly on general value dimensions concerning the medical sector or the usability and design aspects of hospital information systems. Conterminous to that, the authors chose a strictly process-oriented approach. They modeled the requirements of future mobile systems as an output of a profitability analysis based on activity-based costing. The cost savings defined as the difference between former and future business processes were used as an incoming payment for an ROI analysis. In a nutshell, the authors present a case study that highlights the value of their analyzing method as well as the enormous benefit of mobile applications in the area of food and medical supply processes in German hospitals.

INTRODUCTION: INCREASED DEMAND FOR EFFICIENT PROCESSES IN HEALTH CARE

Apart from the long-term decline of the population, a great challenge in the contemporary discussion turns out to be the increasing aging of the

population in European industrial societies. This raises various difficulties for our welfare systems and reveals the necessity of long-term adjustment to this development. Aging describes the process of composition of the population shifting for the benefit of elderly people.

Thus, the decisive item is not the increasing number of the elderly but rather their increasing proportion of the population. For example as latest simulations for the development of the German population (Statistisches Bundesamt, 2003) reveal, the proportion of 65-year-old and older people will rise from 17.1 % today to 29.6 % in 2050. At the same time the percentage of geriatric people (80 years and older) will increase to 12 % which means a triplication.

This development causes serious problems in welfare and tax systems that are based on the income of a workforce. Less young people have to pay the pensions and health care of the elderly.

Furthermore, the productivity of our highly automated industry leaves an increasing number of people unemployed. So the real challenges of over aged European industrial societies will be to enhance the productivity of the existing education and health care systems.

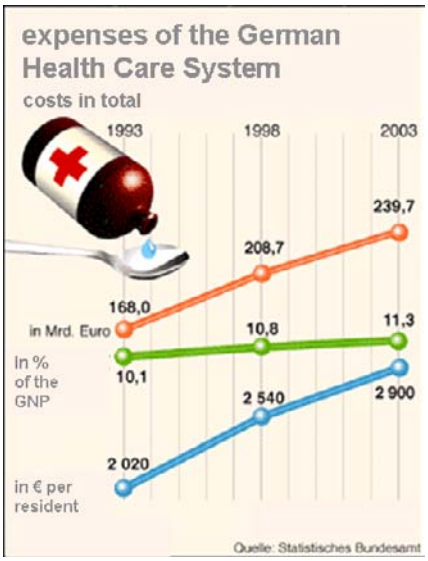
And as productivity is defined as the relationship between output and input factors, there was an intensive discussion going on during the last 2 years about the input factor dimension. Even though the German health care system was able to perform quite well the last decades, from an input point of view the costs and resources to maintain the system were increasing dramatically (see Figure 1).

This development is not typical for the German health care system only; you will find similar developments in all Organsiation for Economic Co-operation and Development (OECD) countries around the world as published at the OECD fact book (OECD, 2006).

One major initiative to stop this cost explosion was the release of a law for financing the clinical sector in 2002. According to this law, a hospital does not get paid for the duration a patient is being treated, but for the respective type of disease. The treatment of every illness is linked to a fixed price—documented in the Diagnosis Related Group (DRG)—the hospital is then paid by the health insurance companies. This system results in the effect that a hospital can only earn good money by improving the business processes in the treatment of the patient, without loosing quality. Since this time on, a big competition between hospitals and different clinical departments to enhance there business process productivity has taken place. Best practices achieving business processes improvements are defined as clinical pathways.

As a parallel to other new deregulation decisions invented by the government, new types of market players like, for example, the Rhön Clinical enterprise emerged. They act as business redeveloper, buying unproductive hospitals

Figure 1. Expenses of the German health care system



and now, by standardizing and optimizing their processes in relation to given DRGs to transfer them to profitable businesses.

This development has only just begun, but the trust in the potential of the business process optimizer is still unbowed, if you take a look to the 3 year curves of their stocks (see Figure 2).

Further potentials in optimizing business processes in the clinical sector are dependent on several issues like:

- Deployment of best practices
- Availability of new technologies that enable high qualified staff to perform the processes in a new way
- Permanent will to benchmark and improve existing business processes

If you look on the infrastructural dimensions of business processes in hospitals today, you can still find a lot of weaknesses linked to the traditional clinical organization, like:

- A very low degree of computerization for support and administrational processes, deeply based in history where traditional administration leaders of the hospitals were not very powerful in comparison to their medical colleagues
- A very poor degree of existing IS integration between different hospital departments, which still causes a lot of medical problems and costs
- A very traditional hierarchical organization with powerful medicinal stuff, but seldom trained in economy
- A very restrictive type of regulation based on the influence of different interest groups and their negotiations in the past

On the other hand, good medical work needs a lot of different information just in time and close to the location, where the patient gets their treatment. From this point of view, one major success factor in business process improvement in the future will be based on technologies that are supposed to be able to transport a big variety of information from different data sources to the medical employee while they are treating their patients, analyzing their physical conditions, or

Figure 2. Chart of the Rhön clinical enterprise

support them with meals and medicine. This is a major reason why the elaborated use of mobile applications in hospitals will be essential for their future success in business

The focus of this paper will be showing you the economic potential of mobile applications, especially their capacity in leveraging the performance of supporting processes like meal and drug supply and how to measure these benefits.

BACKGROUND: THE ECONOMIC POTENTIAL OF MOBILE APPLICATIONS IN A CLINICAL ENVIRONMENT (RELATED WORK)

As mentioned before, it is perhaps, supposed to be a question of organizational survival if the hospitals in European countries are able to decrease there costs far to a level of where they are today. We also argue that only the invention of new, innovatively designed business processes will lead to this target.

But to what degree will and can mobile applications be part of value-added business processes? And how can we measure this value?

If we look at the related literature, we will find two different kinds of studies.

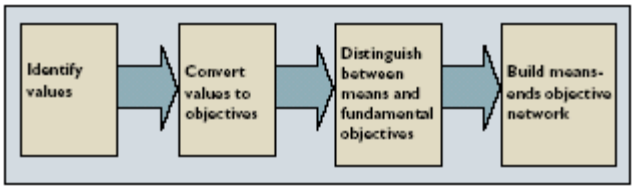
Studies that Deliver Theoretical Basics and Frameworks to the Topic of Value Generation of Mobile Applications

A good example for this kind of research was published by Nah, Siau, and Sheng (2005). Following their definition, value can be defined as: "the principles for evaluating the consequences of action, inaction, or decision" (p.85). The output of their study was a procedure modeled in Figure 3, which was meant to help creating a means-ends objective network including a distinction between basic and fundamental objects as shown in Figure 4.

The methodological way to gain these networks was based on interviews between researcher and employee in the field. A major result can be seen in Figure 4. The strength of this kind of research surely is the identification of relevant factors, which reveals the possibility of enhancing productivity of mobile applications for the whole enterprise.

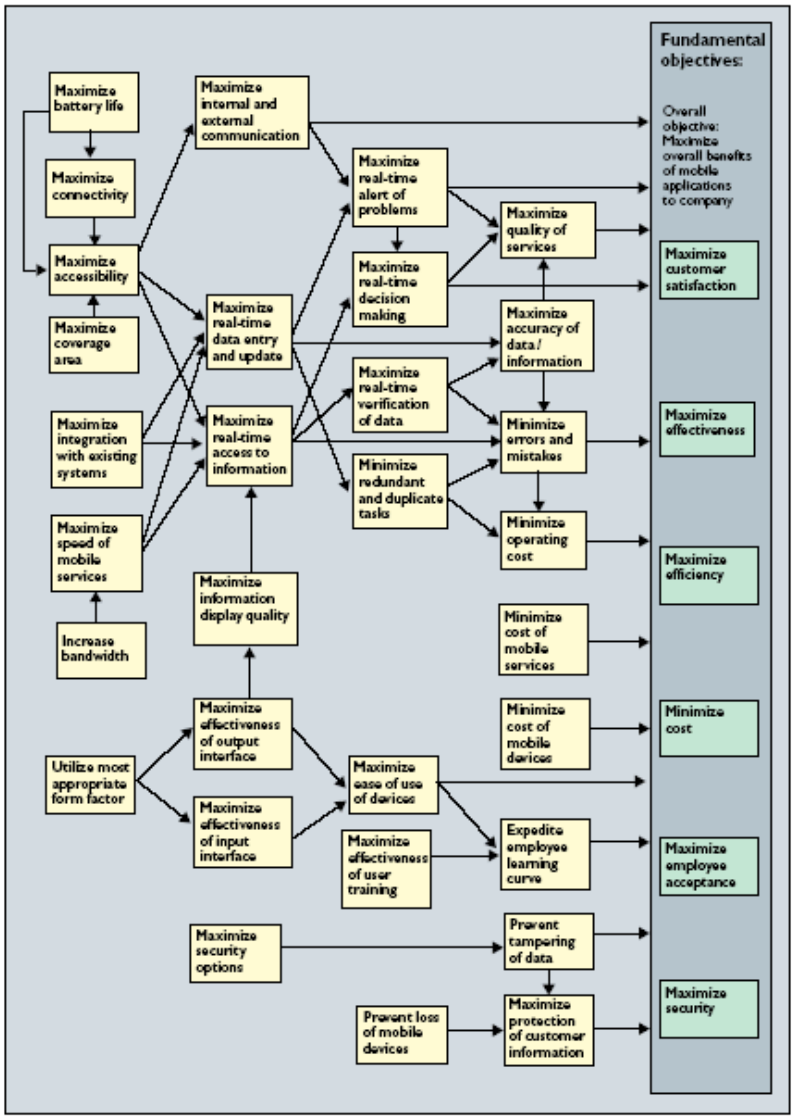
On the other hand, their weakness can be seen in the fact that efficiency of an application system can only be measured in relationship to a supported business process, its core activities, and the output produced. A second weakness of

Figure 3. Procedures of value-focused thinking



their approach was more methodologically. The productivity model of the resulting means-ends objective network of the application system might be based on subjective cognitions, given by the interviewee during the interview situation, instead of time and cost savings measured in a real-world user scenario. A measurement of saved costs between the initial business processes and the redesigned one is not addressed.

Figure 4. The means-ends objective network



Studies Focused on the Business-Oriented Design of Mobile Applications

This approach focuses on the question, what are the basic requirements or architectural patterns that drives the value of a mobile application system in relationship to supported business processes. In addition we can divide this related work in studies that start from an existing business process that has to be improved, and into studies that derive mobile application systems from target processes. There are various studies in the computer supported cooperative work (CSCW) area, which try to enhance the performance of existing business processes by supporting them with fitting mobile system architectures.

A typical research work in this area was published by Shiffman et al. (1999). He shows the benefits of different services (see Figure 5) delivered by a mobile application to a screening workflow for asthma deceases in child care. They described what kind of information management services should be delivered by a pen-based mobile application system to support decision-making workflows in a hospital. Although they get very deep into the interrelationships between system functionalities and the quality of business process support, they did not measure the resulting influence of the system in cost, time, and quality.

Another study from a more transaction network design perspective was published by Morton and Bukhres (1997). It focuses on the improvement of existing transactions in an ambulance scenario. Morton and Bukhres developed a network architecture that enables a hospital to use mobile applications successfully in the ambulatory service. A major challenge in this scenario is the fact, that a mobile host, for example, in an ambulance car could not be online all the time. So they developed an architectural solution, based on a so called base station agent (BSA) that is responsible for the monitoring of transaction of the mobile host during its execution. This architectural innovation developed from existing business processes by interviewing experts and measuring transaction

Figure 5. Delivered services by a mobile application (Shiffman 1999)

Delivered Service	Service Description
Recommendation	the determination of appropriate, guideline-specified activities that should occur under specific clinical circumstances
Documentation	the collection, recording, and storage of observations, assessments, and interventions related to clinical care
Explanation	the provision of background information on decision variables and guideline-specified actions (e.g., definitions, measures of quality or cost) and the rationale that supports guideline recommendations, including evidence and literature citations
Presentation	the creation of useful output from internal data stores
Registration	the recording and storage of administrative and demographic data to uniquely identify the patient, provider(s) and encounter
Communication	the transmission and receipt of electronic messages between the clinician and other information providers
Calculation	the manipulation of numeric and/or temporal data to derive required information
Aggregation	the derivation of population based information from individual patient data

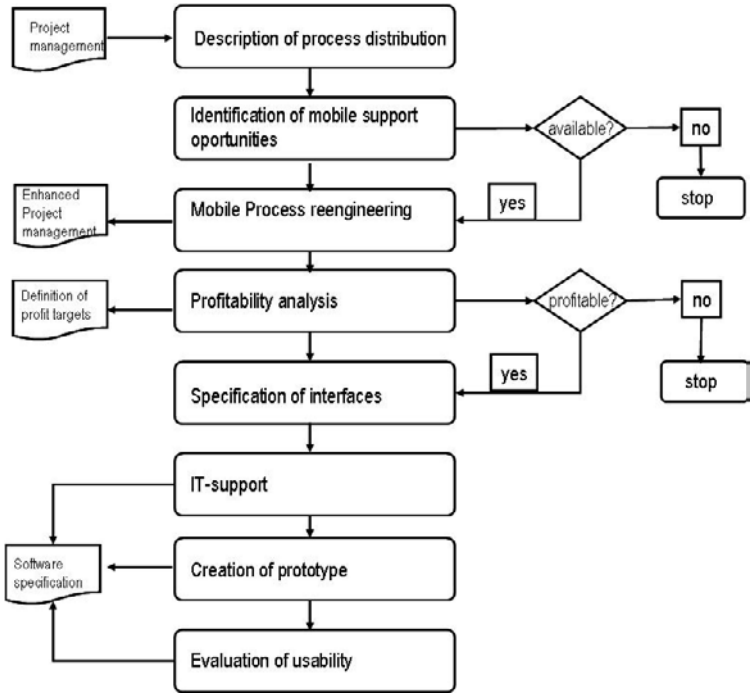
time, which enables the ambulance to use mobile applications for accelerating their transaction times—a very successes critical benchmark in the area of lifesaving.

Different to Shiffman (1999), and Morton and Bukhres (1997), Wang, Van de Kar, and Meijer (2005) published a study that focuses on a mobile application system design, derived from a future target process. Their design methods are based on the collaborative business engineering (CBE) method from Hengst and De Vreede (2004), and tested for the design of a mobile online information system based on a PDA that supports the conductor on the railway delivering actual information and services to railway customers on the train. The output of their kind of research was a very good system design, derived from the requirements of the new, targeted business processes. Although their research design delivers the opportunity to do a business process benchmark between the performance of the current and the target process, an economically motivated measurement of the improvements was not operated in that research study either.

Conclusion: The Integrated Approach

Recurring to related researches as mentioned before, we were looking for a third way that started from an approach as published by Wang et al. (2005) for the necessary system design but also was combined it with a classical activity-based costing analysis. The aim of our approach was to gain data, which enabled us to measure the efficiency of the new designed and mobile

Figure 6. Major steps of mobile process landscaping (MPL)



application supported business processes. We found a suitable method named mobile process landscaping (MPL) developed by Köhler and Gruhn (2004) that supported our research in the mentioned way. MPL includes eight steps shown in Figure 6.

However, based on our long-years experience in the area of business process modeling at the Research Centre for Information Technologies (FZI) and a conclusion drawn in the unpublished doctoral thesis of Högler[1] (2006) we improved the MLP method and adapted it to the mobile business process development and profit analysis of a hospital nursing 300,000 patients a year.

We executed the analysis in the five major steps described in Figure 7:

Step 1: Identification of highly valuable business processes in the business process map of the hospital (see Figure 8).

Highly valuable processes should have a significant cost-saving potential as well as a high degree of distributed activities, so that mobile application support may deliver a remarkable benefit. In our case these processes were easily to be identified, as they were delivered from external service providers, too. Benchmarks for meal and medical supply process performance were far ahead from the performance delivered in the observed hospital.

After quick-win business processes in the area of medical and meal supply had been identified as valuable, and the strategic decision made by the management to still internally deliver these services in the future, a second step in research had to be done—the analysis of the mobile potential of the identified supporting processes as well as the reengineering potential on an activity level (see Figure 7).

Step 2: Identification support opportunities on an activity level and the reengineering potential

The analysis of mobile potentials can be operated in two ways:

a. The first one will be *the identification of activities in an existing process, which could possibly be supported by mobile applications (improvement approach).* Typical indicators for supportable activities are:

Figure 7. FZI mobile process reengineering approach

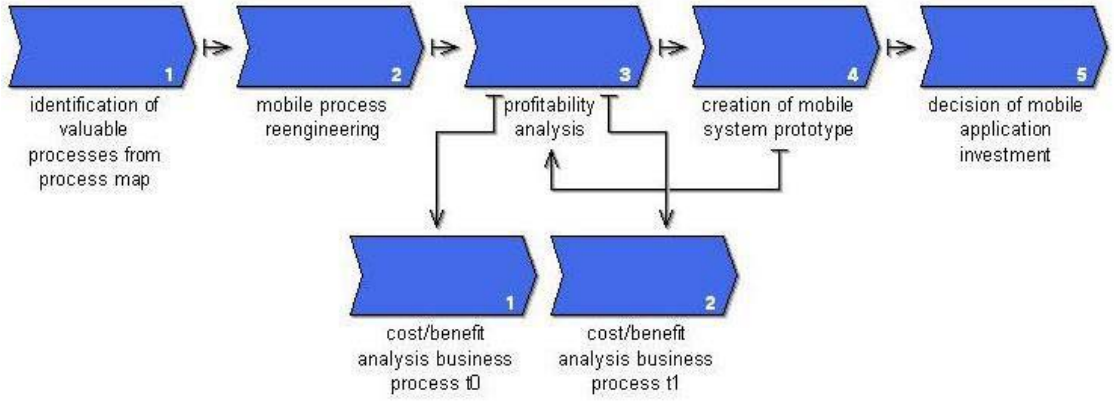


Figure 8. Business process map of the hospital

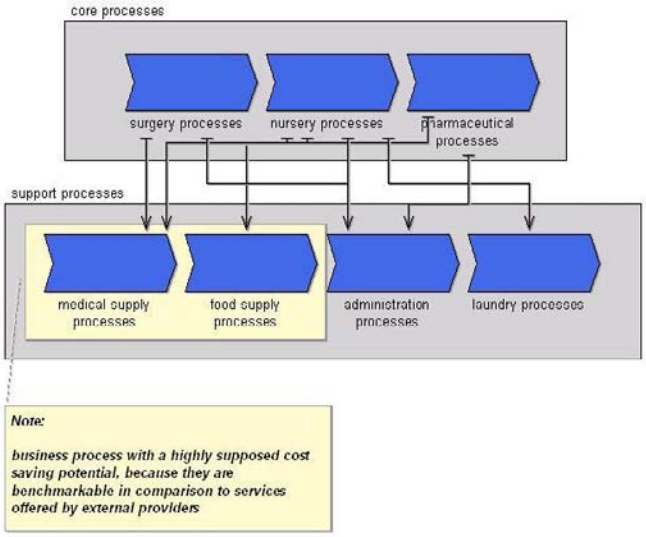
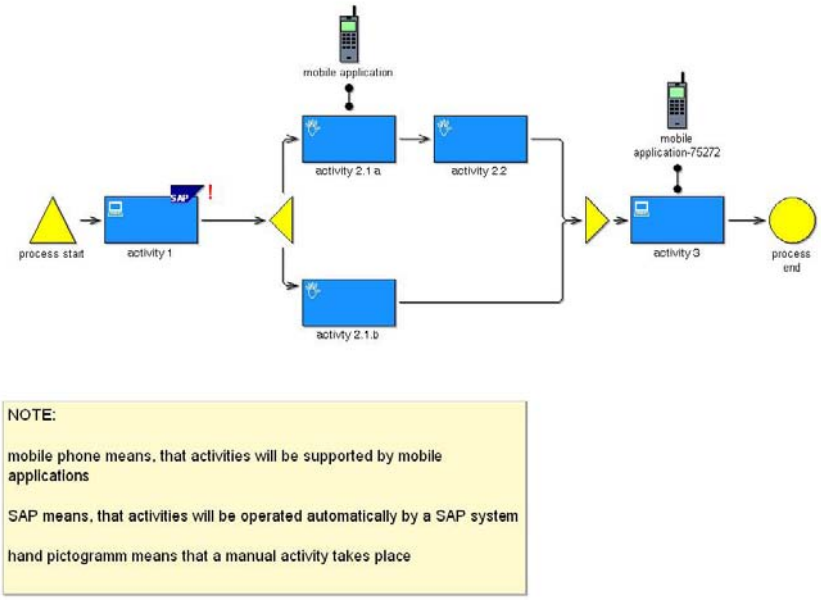


Figure 9. Example of a business process with reengineering potential



- the necessity of having access to actual information at a point work that is not yet connected to a network (information quality and delivery function), and
- the necessity to improve existing data input procedures by delivering high qualitative information back to an integrated hospital information system electronically instead of manually (service quality function).

b. The second one will be *to reach predefined outputs of a needed service with totally new, less cost-intensive mobile business processes (reengineering approach).*

Especially the design of totally new business processes supported by mobile technologies in regard to their later implementation efforts, costs, and maintenance quality is a central research task the FZI has been working on for years now.

The application of steps 3 to 5 of the FZI Mobile Business Process Reengineering Approach (MBPRA), shown in Figure 7, will be described in the case study in chapter 3. It includes the operation of a profitability analysis, and the design of a mobile system prototype fitting into the reengineered business processes. Once the prototype will be able to reach serious savings of activity-based costs, the investment decision for the implementation of the whole system has to be made.

A major focus in our case study will be regarding the question of data quality, which is necessary for modeling and simulating business processes in a way that allows the delivery of detailed information about their cost structure and process performance.

CASE STUDY: MOBILE TERMINALS FOR DRUG AND MEAL SUPPLY

This study was carried out from January until June 2004 at FZI by order of Karlsruhe City Hospital, the largest hospital under public law in the vicinity of Karlsruhe with more than 1,500 beds and 14 clinical departments, handling approximately 300,000 patients per year. The departments are allocated campus-like to numerous buildings.

Figure 10. Comparison between conventional and focused ethnography

Conventional Ethnography	**Focused Ethnography**
long-term field visits	short-term field visits
Experientially intensive	data/analysis intensity
time extensity	time intensity
Writing	Recording
solitary data collection and analysis	data session groups
Open	Focused
social fields	communicative activities
participant role	field- observer role
insider knowledge	background knowledge
subjective understanding	Conservation
Notes	notes and transcripts
Coding	coding and sequential analysis

In this study the hospital's processes of drug supply in the pharmacy of the hospital as well as meal supply in the kitchen were examined. Key stakeholders were nurses, physicians, pharmacists, assistant medical technicians (MTAs), assistant pharmaceutical technicians (PTAs), cooks, cooking assistants, diabetes consultants, haulage service, and the controlling and executive committee.

Used Methodology

As every model of reality is as good as its input, our profitability analysis of mobile business processes was highly dependent on the data quality. As long as time for delivery or processing of activities was supposed to be measured or informal stocks at the wards to be detected, it does make sense to use a mixture of traditional ethnographic and focused ethnographic approach (see Figure 10) as claimed by Knoblauch (2005). In our role as researchers we were involved as participants in the working process, which is not the case in focused ethnography, where the researcher normally acts as an observer. But all the other techniques we used were based on the ethnographic approach of Knoblauch.

The challenge was to gather high qualitative data and transfer it to linear cost models, like Lazarsfeld, Jahoda, and Zeisel (1933) practiced it in their early studies. We used the following data gathering methods:

- **Participating observation:** we worked together with nurses and a druggist as well as kitchen and health care staff for 8 days and night shifts to get a deep understanding of the business processes including the used resources and materials.
- **Inventories:** Inventories of the formal and informal medical stocks at the drug store and the wards
- **Inside interviews:** 34 non-closed expert interviews with different process owners in the hospital (nurses, pharmacist, IT operators, IT management, IS consultants, cooks, diet cooks, garbage man)
- **Outside interviews:** 4 non-closed expert interviews with experts from other hospitals that had just invented mobile systems in the same processes to validate their own gathered data.

Initial Situation/Preface

Previous to this study the hospital was faced with the choice of an investment and integration of mobile devices on wards and pharmacy. These processes were chosen from the process map, because they promised a high potential for optimization. In 2003 the project team "Mobile Computing" was founded by the head of the pharmacy, the chief information officer, and the chief executive officer. In a pilot project they started to develop and test a mobile system in the pharmacy and on two selected stations (oncology and nephrology, each with more than 20 beds) in cooperation with a software company. The aim of the Mobile Computing project was improving the drug supply by reducing administrative activities just as avoiding sources of error by the dint of mobile terminals and their integration into the supply workflow. At the time the study was performed, the Mobile Computing project was in the productive test phase. In case of a positive result, it was planned to roll-out the system on all stations. The study's objective therefore was to provide a basis of decision making about the roll-out by dint of an economic analysis.

In order to examine the portability of the mobile system the study was made up of two scenarios. The first scenario dealt with the processes of drug supply, where, within the scope of the Mobile Computing project, the possibility of comparing past and new situations in a real environment was given. In the scenario two processes of the meal supply were chosen to demonstrate the ability of the mobile system being extendable. In contrast

to the scenario, one of the analyses was limited to the investigation of a business process without any mobile devices due to the fact that there was no test implementation of a mobile system for the meal supply. The analysis of the future processes (with mobile devices) had to be designed with experiences gained from interviews of other hospitals using mobile devices for meal supply already.

By implementing the mobile system, traditional paper-based processes were supposed to be adopted and optimized where exchanging information should be converted from paper-based to electronic-based processes as far as possible. Improvements in workflow by automating administrative jobs (e.g., validity check, sorting of forms, calculating order quantity, etc.) and minimizing sources of error were expected.

In scope of the drug supply the orders were previously taken at the station's PC or at the pharmaceutical rack using PDAs and sent to the enterprise resource planning (ERP) system and the pharmacy's IS via docking station. In the pharmacy, the orders could be processed directly on the PDA by a wireless LAN (WLAN) without the necessity of using paper-based media.

At the meals supply process the paper-based order form was displaced by the digital entry via mobile devices. If necessary, the chosen data could be linked to the patient's incompatibilities or objections. The software was to check up on all data on their plausibility automatically and to send alerts in the case of incomplete or false orders. In the kitchen, all steps for receiving and handling meal orders were automated. Based on the demand of ingredients, the kitchen's staff sent all orders to the supplier. Additionally, the patient nameplates were printed and cut to mark the meal tablets.

Objectives

Of highest interest in this study was the economic potential that could be gained by implementing mobile terminals in a hospital whereat the profitability of the mobile system for the chosen processes for the supply of drugs and meals described previously had to be measured. For this reason processes of the past and future situation were to be documented and compared in order to ascertain possible profit and loss.

Another target was the development of a methodology for the evaluation of economic potentials achieved by the integration of mobile devices into clinical and economical pathways as well as for its testing. For this purpose the general conditions for the use of mobile devices in hospitals had to be considered and the applicability of different kinds of mobile devices to be detected. Another question to be answered was where and how mobile terminals could be integrated in existing business processes. The analysis of advantages and risks that could occur during the roll-out of the mobile system was targeted as well.

Central questions that had to be addressed in the study were:

- How could costs and efficiency of supply processes be measured (with and without the support of mobile devices)?
- What advantages and risks could be expected during the roll-out of mobile systems?
- What profit and loss could result from implementing the mobile system?

Scenario 1: Drug Supply

In the hospital all medicaments like drugs, infusions, and so forth were delivered by the hospital's own in-house pharmacy. The pharmacy is in charge of ordering, producing (in special cases), storing, and delivering drugs to the hospital stations. Pharmacist, MTA, PTA, and warehouseman have to work together closely. The pharmacist is responsible for quality control and consulting services in case of questions about drugs (e.g., compatibility, unlicensed drugs, etc.) for the hospital's physicians and nurses. Furthermore he/she has to determine the demand for drugs and

in special cases produce special pharmaceuticals and infusion bags. Usually, wards order drugs every day.

Former Situation in Drug Supply

The processes of the former situation in the hospital's drug supply are presented in Figure 11. After the prescription had been inserted into the patient record by the responsible doctor, nurses ordered lacking medicaments by an order form. Before that, they had to walk along the medicine shelves on their wards and write down name and amount of the needed medicaments on a notepad. Afterwards they went over their notes and transferred them on the order forms mentioned previously. Due to the numerous classifications of medications (usual medication, infusion, dispensing, cytostatica, anesthetic, etc.) there were eight different types of forms available. The order form had to be signed by the responsible doctor and transported to the pharmacy finally by being thrown in the pharmacy's letter box, faxed, or in urgent cases by using pneumatic post.

In the pharmacy, the pharmacist first sorted the incoming order forms according to their type of form and to their posting station and checked the details on the order forms. Detected discrepancies had to be straightened out by telephone. In case of missing compulsory data (e.g., signature of the doctor) order forms were to be sent back to the station. In case of valid details drugs were collected and put in boxes together. Before delivery the drugs were registered in the ERP system of the pharmacy by scanning the barcode. The so called "Hol- und Bringdienst" ("catch and delivery

Figure 11. Processes of the drug supply in the hospital (former situation)

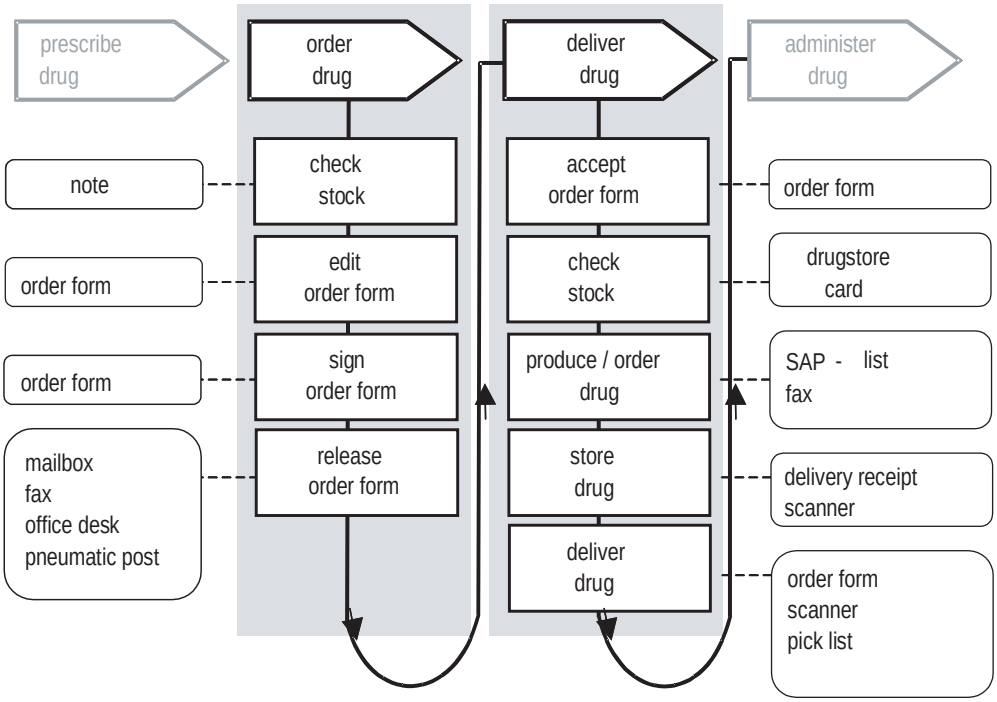


Figure 12. Cost of material for drug supply (traditional supply system)

	amount	unit price	cost of operation/per year
drug documentation form	14064	0,0636 €	900 €
archiving costs (drug store)	15 m²	72 €	1.600 €
drug expiry costs	1	4.770 €	4.770 €

Figure 13. Personnel cost of drug supply (traditional supply system)

Drug order (each station)	**duration [min.]**	**factor**	**costs [per day]**
Fill out order form	6,36	1	3,08 €
Subscription by doctor	1,79	1	0,86 €
Walking to the mail box	2,18	1	1,05 €
Prepare pneumatic post	1,07	1	0,52 €
Edit returning form	0,45	1	0,22 €
Total costs of drug order (per station)			**5,73 €**
Drug delivery (pharmacy)	**duration [min.]**	**factor**	**costs [per day]**
Empty mail box and pneumatic post	64,11	1	51,73 €
Sort order forms	17,14	1	6,77 €
Correct orders, making inquiry calls	31,57	1	19,82 €
Store pick lists	9,00	1	3,80 €
Writing and attaching notes	2,39	1	1,50 €
Automated order	8,50	1	3,59 €
Restructure drug cards	12,00	1	5,06 €
Accounting of narcotics	0,50	6,31	2,63 €
Check of the narcotic accounts	30,00	1/30	0,83 €
Accounting of the sales to clinical employee	60,00	1/7	3,62 €
Data entry of Zytostatika-in PC	90,00	1	75,00 €
Zytostatika-data exchange with ERP-System	45,00	2/30	2,68 €
Register orders for wholesaler	14,00	1	11,67 €
Dictate wholesaler orders to secretary	20,00	1	16,67 €
Empty stock from employee drug sale	5,63	1/7	0,34 €
Additional efforts of employee drug sale	30,00	1/7	1,81 €
Preparation of order forms	2,03	1	0,80 €
Register industrial orders	10,79	1	4,56 €
Archiving of clipboard	27,22	1	11,49 €
Managing call backs of single charges	46,21	1/7	5,50 €
Total costs of drug delivery (pharmacy)			**229,85 €**

service") performed the delivery to the stations. Additionally, nurses are able to fetch ordered drugs at the pharmacy counter.

Information was transmitted by paper-based information media like order forms and notepads as well as by fax, letter box, pneumatic post, telephone, and the staff itself (nurses, pharmacist).

Weak points in the former situation could be identified in several processes. Inefficient activities were the sorting of the order forms, queries by telephone because of non-explicit or uncompleted details, the recording of the anesthetic, output, and sale to hospital staff member as well as orderings from drug maker and wholesaler.

Using pneumatic post (ca. 40-60 times per day) required several time-consuming activities such as loading and unloading the sleeves. Furthermore the pneumatic post system often revealed little reliability and additionally caused high costs for maintenance (see Figure 12 and 13).

The material costs contain those for paper and printing, order forms, occupancy costs for archival storage of order forms, and of thrown away drugs because of the drugs were out of date/expired. The term factor represents the frequency of the respective activities per day. Costs can be calculated by multiplying labor costs by factor.

Present Scenario with Mobile Devices in the Drug Supply

In the present scenario processes of ordering drugs (see Figure 14 and 15) are supported by PDA (personal digital assistant) and by a Web-based order entry system. Nurses now walk along the drugs shelves on their station and fill out the order form

Figure 14. Processes of the drug supply in the hospital (present situation)

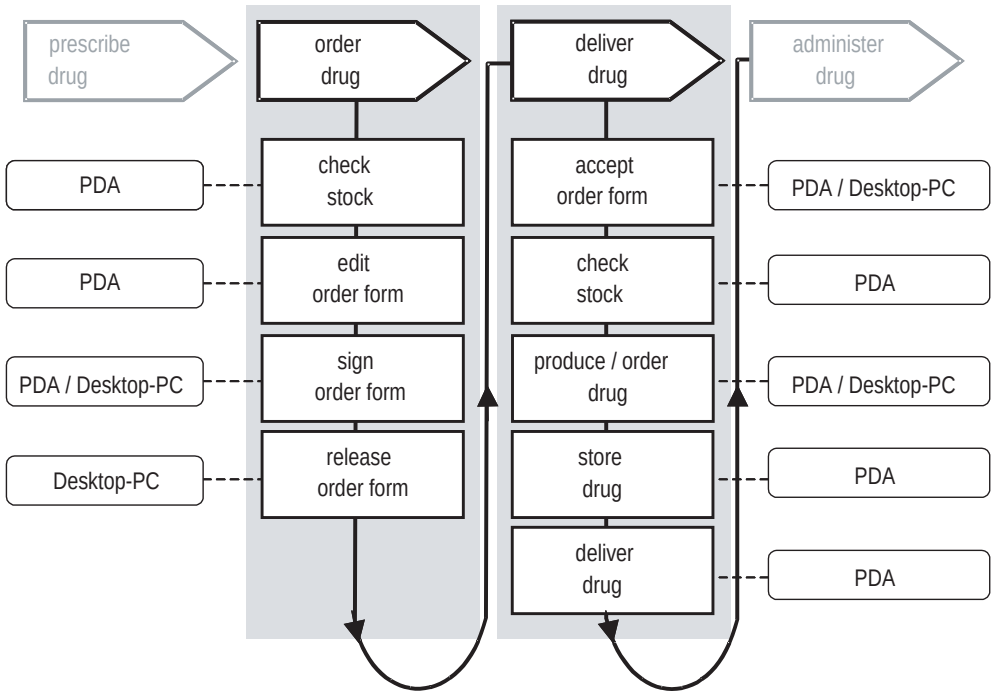


Figure 15. Cost of personnel for drug supply (present situation)

	duration [min.]	factor	costs [per day]
Drug order (each station)			
Fill out order form	2,00	1	0,97 €
Synchronize PDA / Scanner	1	1	0,48 €
Confirrm order	2,00	1	1,16 €
Subscription by doctor	1	1	0,82 €
Edit returning form	0,07	1	0,03 €
Total costs of drug order (per station)			**2,98 €**
Drug delivery (pharmacy)			
Print out pick-list	1	1	0,40 €
Send messages (e.g. quries, information)	0,60	1	0,31 €
Automated order	1,43	1	0,60 €
Accounting of narcotics	1	1	0,83 €
Check of the narcotic accounts	1	1	0,42 €
Data entry of Zytostatika-in PC	10,00	1	8,33 €
Register orders for wholesaler	4,00	1	3,33 €
Register industrial orders	9,25	1	3,90 €
Archiving of clipboard	5,00	1	2,11 €
Managing call backs of single charges	8,00	1/7	0,95 €
Empty stock from employee drug sale	3,00	1/7	0,18 €
Synchonize PDA	1	5	2,56 €
Total costs of drug delivery (pharmacy)			**23,94 €**

directly on a PDA without any use of notepads. With the barcode scanner extension nurses just need to scan the barcode of the drug's package and choose the amount by pull-down menu. In order to save time, the barcodes labels itself are stuck on the shelves' front side, so that there is no more need of taking the drug packages out. Via docking station at the station's computer the order forms are directly transmitted to the ERP system of the pharmacy. A Web front end enables nurses to check, correct, and activate the order. The responsible doctor merely has to check the Web-based order form and sign with his/her personal password to ratify its validity.

The team of the pharmacy now gains access to the order form in an electronic way. By computer and PDA they can view and work on all order forms on the display. Due to the pharmacy's WLAN connectivity the staff has ubiquitous access without any need to sort the order forms. The staff can walk along the shelves of the pharmacy and collect all ordered drugs. By scanning the barcode on the drug package the delivery items are automatically registered.

Review of the Process Improvements in the Drug Supply

The mobile system offered numerous forms of relief and advantages. It is possible to remedy the deficiencies of the former situation without IT support and to avoid disadvantages of station-

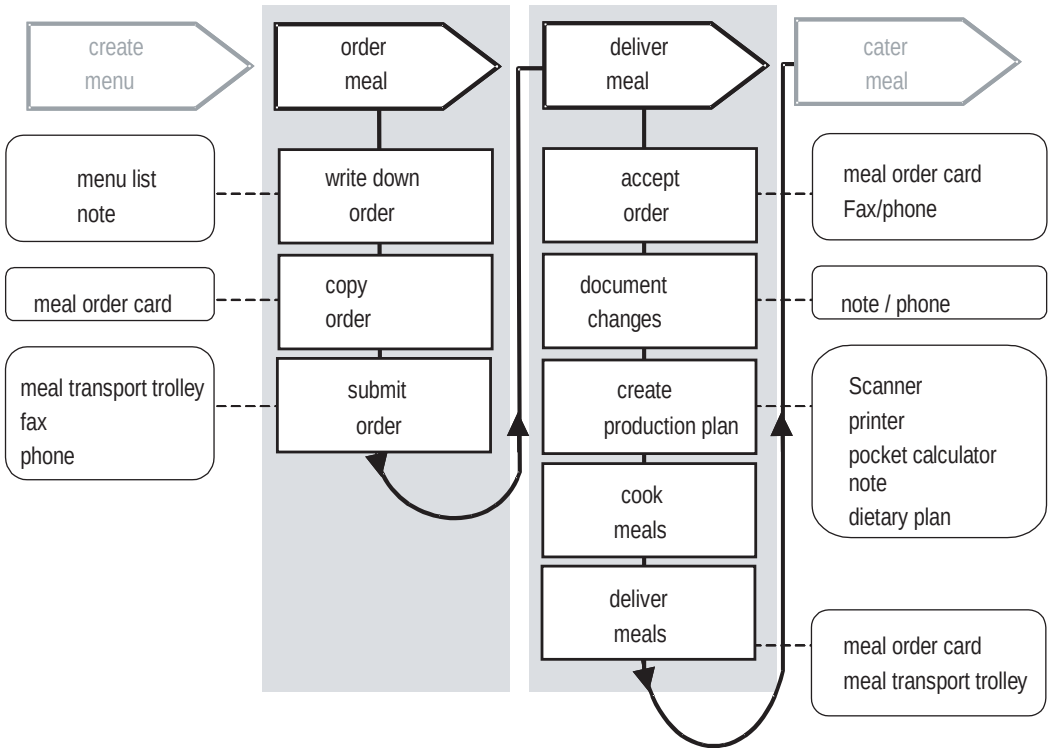
ary computer systems. Communication costs (telephone conversations) between pharmacy and station could be limited by eliminating unreadable handwriting and incomplete details. In the pharmacy numerous activities could be automated or usefully supported.

Because of several circumstances the field of application was limited in some issues. For example, ordering and delivering anesthetic had to be excluded from the mobile system, as dealing with anesthetic had to use paper-based documentation by statute. Another area where the mobile system was discussed to be applied was the pharmacy's stock control. But as the use of drugs could not be registered exactly (e.g., returned drug packages, which were partially used and no more needed), improvement of stock control by the mobile system was not possible either, nor was it to be expected that the amount of drugs that had to be disposed for exceed of expiration date could obviously be reduced.

Scenario 2: Meal Supply

In the hospital there exists one kitchen that supplies personnel, patients, and their dependants with meals. Patients get meals three times per day. One day before each meal, for example, lunch, nurses on the stations ask patients about their wishes for lunch the next day and transmit the orders to the kitchen. Thus, the kitchen's staff is able to plan and organize every meal up to 24 hours before.

Figure 16. Processes of the meal supply in the hospital (former situation)



Former Situation in the Meal Supply

Without the implementation of a mobile system, the supply of meals is a slow process (see Figure 16), afflicted with different kinds of errors. Orders are taken on the basis of menus and forwarded to the kitchen via meal vouchers. In the past these vouchers had been sorted, checked superficially, and afterwards were read in with a special PC-linked scanner. The associated software counted the orders per station and furthermore checked the vouchers' plausibility. Having gathered all orders the production schedule for the next day was printed and handed out to the cooks.

Future Scenario with Mobile Devices in the Meal Supply

On the basis of a new mobile system, the traditional paper-based processes can be adopted and improved (see Figure 17). The orders will be taken at the station's PC or directly at the patient via PDA. The name of the patient is chosen out of a patient administration system, linked to the chosen meal and via PC or docking station sent to the ERP system and the kitchen. If necessary, the chosen data can be linked to the patient's incompatibilities or objections. The software will check all data on their plausibility automatically and send alerts in case of incomplete or false orders. Additionally, it will check the patient's administration software in order to verify which patients have left the hospital or have changed the station.

Review of the Process Improvements in the Meal Supply

These meal vouchers as well as false orders and insufficient deliveries emerge as the most important cost drivers within the old system. For example, it takes up a great deal of time to complete the meal vouchers, to arrange and to correct them as well as to import the vouchers into the system; errors emerging are redundant or needless orders that can not be cancelled. Further cost drivers are

Figure 17. Processes of the meal supply with the mobile system (future situation)

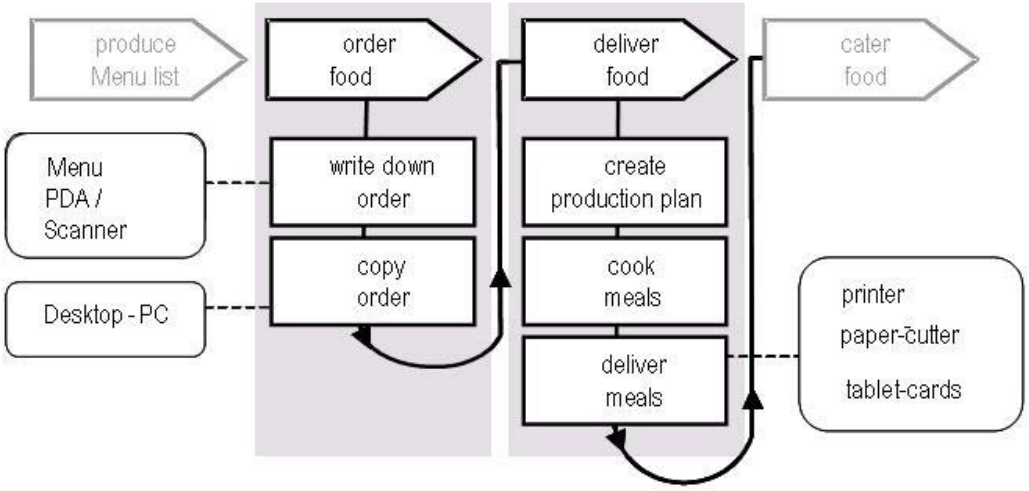


Figure 18. Material costs for meal supply (traditional scenario)

Invest (indispensable)	**amount**	**unit price**	**Total costs**
Software-Update (license paper based system)	1	3.100 €	3.100 €
Software-Update (installation)	1	5.200 €	5.200 €
Software-Update (training costs)	1	1700 €	1700 €
Additional operation costs (per year)	**amount**	**unit price**	**Total costs**
Software-material costs	1.336.746,00	0,018 €	23.968 €
Software-maintenance contract costs	1	911 €	911 €
Waste of meals	29.975	4 €	120.000 €

Figure 19. Cost of personnel for meal supply (traditional supply system)

Meal supply process (per station)	**duration [min.]**	**factor**	**costs [per day]**
Produce collection of different menu	10,00	1/7	0,69 €
Discuss menu with involved employee	5,00	1/7	0,35 €
Prepare meal-voucher	10	3	14,52 €
Sort meal-voucher	1	3	1,45 €
Inform kitchen	2,14	1,00	1,04 €
Total costs of meal order process (per station)			**18,04 €**
Meal delivery process (kitchen)	**duration [min.]**	**factor**	**costs [per day]**
Retrieve meal-voucher from stock	3	3	3,10 €
Sort meal-voucher	120	3	123,86 €
Correct meal-voucher	10	3	10,32 €
Prepare inward meal -voucher for ambulant patients	5	1	1,72 €
Prepare meal -voucher for staff.	5	1	1,72 €
Count token coins of staff	1	1	0,34 €
Operate inward statistics	20	1	6,88 €
Operate discharge statistics	20	1	6,88 €
Operate statistics for additional delivered fruits	2	1	0,69 €
Transport meal -voucher to administration office	3	3	3,10 €
Start meal software	1	3	1,03 €
Scan paper-based meal-voucher	35	3	36,13 €
Print out production plan	2	3	2,06 €
Save data	7	1	2,41 €
Total costs meal delivery (kitchen)			**200,24 €**
Meal delivery process (diet kitchen)	**duration [min.]**	**factor**	**costs [per day]**
Check and correct special diet meal-voucher	45	1	19,86 €
Calculate portions for diet	30	1	13,24 €
Calculate ingredients	30	1	13,24 €
Label plates for special diets	30	1	13,24 €
Total costs meal delivery (diet kitchen)			**59,57 €**

coordinative telephone calls between kitchen and hospital stations that have to be ascribed to the relocation or early release of patients, for example. 80-100 telephone calls per day are necessary due to short-term changes of orders.

These cost drivers add up to approximately 720,000 € each year for ordering activities between the kitchen and all stations of the hospital. This sum includes material costs like order vouchers and wasted meals due to mistakes as well as personnel costs for placing and taking the daily order.

With the help of the mobile system described previously, many of the former cost drivers in meal supply could be eliminated. Having implemented the mobile system, further costs for materials will only arise from the production of tablet cards and non-preventable false deliveries of meals. Non-preventable situations are, for example, when it is not foreseeable after which meal a patient will be released the next day or whether a meal can be taken after a surgery.

In Figure 18 and Figure 19 the costs accruing in the kitchen due to the old IS are listed.

The costs for materials compound of actualizing the software and the contract for maintenance of costs for paper and those for spare meals which can be ascribed either to preventable or non-preventable wrong orders.

Aside from retrenching working expenses, the mobile system also provides an obvious surplus in quality. In contrast to the former system, the patients' incompatibilities like, for example, allergies or possible objections can now be considered with much less effort. Figure 20 shows the personnel costs of the meal supply supported by the mobile system.

Economic Results

Process improvements were performed iteratively in several steps. During the process improvement, ideas and visions of staff members were considered

Figure 20. Cost of personnel for meal supply (with the mobile system)

	duration [min.]	factor	costs [per day]
Meal supply process (per station)			
Discuss menue with involved employee	2,50	1/7	0,17 €
Synchronize PDA	1,00	3,00	1,45 €
Confirm meal order	2,00	1,00	0,97 €
Total costs of meal order process (per station)			**2,59 €**
Meal delivery process (kitchen)			
Count token coins of staff	2,00	1,00	0,69 €
Print out production plan	1,00	3,00	1,03 €
Print out tablet cards	1,00	3,00	1,03 €
Cut tablet cards	10,00	3,00	10,32 €
Total costs meal delivery (kitchen)			**13,07 €**
Meal delivery process (diet kitchen)			
Calculate portions for diet	5,00	1,00	2,21 €
Calculate ingredients	1,00	1,00	0,44 €
Total costs meal delivery (diet kitchen)			**2,65 €**

Figure 21. Formula of the net present value according to Grob (1999)

$$NPV = \left(\sum_{t=1}^{n} \frac{Cash\ Flow_t}{(1 + Discount\ Rate)^t}\right) - Initial\ Investment$$

and the experience gained in cooperative hospitals as well as in Mobile Computing projects in other lines of business were taken into account. Approved methodologies and rules for process improvement by Greiling and Hofstetter (2002, p. 96 et sqq.) were included. The profitability was calculated by the *net present value* (NPV) method (see Figure 21), where n means the time horizon of the project and t numeralizes the years. The term *discount rate* refers to a percentage used to calculate the NPV and reflects the time value of money at an actual average between 3% and 6%.

By dint of the methodology current and target processes of the traditional and the mobile system were modeled and documented, their costs and use in terms of activity-based costing were

Figure 22. Cost of material for mobile system in drug and meal supply

	amount	unit price	Total costs
Invest			
PDA Symbol SPT1846	5	1.200 €	6.000 €
PDA Symbol SPT1550	181	500 €	91.000 €
Printer (kitchen, tablet cards)	2	5.000 €	10.000 €
Paper cutter (kitchen)	2	5.000 €	10.000 €
Printer (pharmacy)	3	750 €	2250 €
WLAN access points	2	200 €	400 €
Software development (ward)	1	39.000 €	39.000 €
Software development (kitchen)	1	32.000 €	32.000 €
Software development (in general)	1	1.000 €	1.000 €
Software development (sales tax.)	1	16.000 €	16.000 €
Course of training (pharmacy)	1	5.000 €	5.000 €
Course of training (station)	108	310 €	33.000 €
Course of training (kitchen)	1	200 €	200 €
Installation (pharmacy)	5	10 €	50 €
Installation (ward)	108	4 €	400 €
Installation (kitchen)	5	80 €	400 €
Additional running costs (per year)			
Costs of lost/ broken PDA (Symbol SPT1846)	0,5	1.200 €	600 €
Costs of lost/ broken PDA (Symbol SPT1550)	20	500 €	10.000 €
Costs of paper (pharmacy)	14.000	0,004 €	50 €
Costs of paper (kitchen)	400.000	0,004 €	1.600 €
Costs of hiring server (electronic data processing center)	1	10.000 €	10.000 €
Costs of data traffic (electronic data processing center)	108.000 MB	0,03 €	3.240 €

determined and compared within the scope of a profitability analysis.

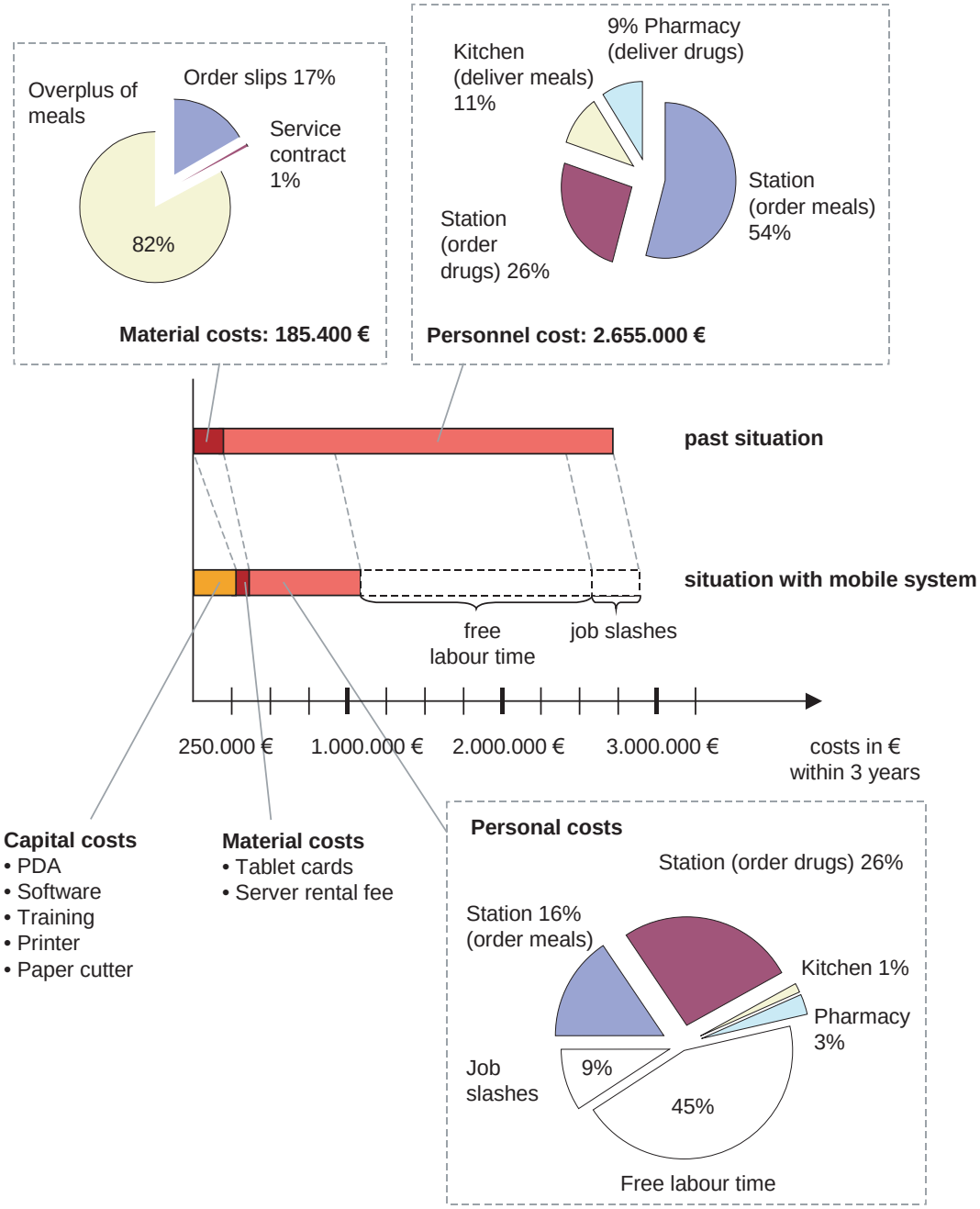
The costs for the optimized drug and meal supply are scheduled in Figure 22. Just as in the drug supply, the one-of expenses for the mobile system comprise of costs for investment and the non-recurring operating expenses. There is also a strong emphasis on staff's training activities to guarantee a maximum acceptance by the staff and a smooth IS implementation.

The investment into the mobile system and the scanner alternative gets profitable at least within one year (see Figure 23 and Figure 24), based on the assumption that operation expenses are lowered strongly and risks narrowed down. The discount rate is chosen on 10% by an initial

Figure 23. Value calculation with the net present value method

Invest (Total)	**274.750 €**
PDA Symbol SPT1846	6.000 €
PDA Symbol SPT1550	91.000 €
Printer (kitchen)	10.000 €
Paper-cutter	10.000 €
Printer (pharmacy)	2.250 €
WLAN access points	400 €
Software development	116.000 €
Training and installation costs	39.100 €
Single savings	**10.000 €**
Additional operation costs (per year)	**25.250 €**
Savings per year (material costs)	
Meal-voucher and maintenance	24.900 €
Waste of meals	30.000 €
Drug order forms	900 €
Costs for needed archiving space	1.000 €
Savings per year (personnel costs)	
½ x diet cook	21.000 €
3x cook assistants	99.000 €
Labor time (general)	591.000 €
Total savings per year (material and personel costs only)	**176.800 €**
Total savings per year (with additional savings of labour time)	**767.800 €**
Period	3 years
Bank rate	10 %
„net present value" (NPV) (without process time)	**108.700 €**
„net present value" (NPV) (with process time savings)	**1.581.000 €**

Figure 24. Schematic diagram of the profitability analysis



investment of 200,000 €. In contrast the payback period is reached within 3 years time, if the implementation of the mobile system is not followed by an appropriate reengineering of the relevant processes. These facts can be ascribed to the integration of autonomous subsystems by mobile devices as well as to the removal of errors. Besides, parallel processes result in the decrease of labor time. Additionally, the mobile devices enable the controlling department to gather more and reliable data for the measurement of costs.

CONCLUSION

In comparison to the big amount of money, the European telecommunication industry has invested in Universal Mobile Telecommunications System (UMTS) and third generation (3G) mobile infrastructure; there still is a very slow adoption of mobile application in different domains, which might result from a lot of different reasons. Two major reasons surely apply: (1) lack of methods to demonstrate and measure the value creation potential of mobile business applications, and (2) the lack of potentially best practices and use cases in different domains.

If we go back to methodologies for measuring the value of mobile-supported business processes, we might be able to demonstrate that it is possible to show the benefits by business process modeling, design, and simulation methods, followed by an activity-based costing and traditional investment appraisal.

However, to do this in a proper way, there was a lot of effort necessary, for example, ethnographic studies and field work for the time of about 2 weeks, participating in the drug and meal supply processes, and recording and analyzing data.

On the other hand the quality of the resulting process model with time and cost data is strongly dependent on the quality of input. But if we look at the investment done in total and the potential savings, the time for research was worth its money.

The process-oriented research approach derived from the MPL method was able to deliver useful results. It was a further enrichment in comparison to research methods that focus on mobile system design (Wang et al., 2005) only, or on abstract means-ends objective networks that describe general value dimensions of mobile applications (Nah et al., 2004).

The business processes that had been chosen from the process map of the hospital turned out to be very fertile for optimization with mobile applications. An ROI of an investment of 200,000 € in mobile systems could be paid back in a time frame of 1 (progressive calculation) or maximum 3 years (conservative calculation).

Hospitals stand to benefit in different ways from implementing a mobile system: There is a high ROI and efficiency gain caused by the use of mobile terminals. The restructuring of processes can reduce running time and thus the workload of health care and administrative personnel. By minimizing the number of errors that do occur during the recording and editing of orders, expenses for the drug and meal supply processes can be reduced, too. Information processing can be automated by the use of mobile electronic systems, thus the waste of material, labor time, and storage space for files (actually an important expense factor) can be significantly decreased.

Another remarkable improvement caused by the new mobile system is the extraction of very detailed information that can be used by the controlling department for the measurement of process performance, process costs, and occurring errors. Furthermore, by setting data in relation to patients, applications can be developed to share patient-oriented information including high traceability and a high transparency in supply processes. For example in drug supply a mobile system can provide new possibilities for enhancing the medication: The hospital's pharmacy now plans to enhance the existing mobile system to a unit dose system, in which every patient gets their individual medication, beginning at the bed-side

prescription by PDAs to the patient-related packaging and ending at the patient-related billing.

Nevertheless, economic potentials can be reduced significantly by the wrong choice of mobile terminals. This fact was revealed by the analysis of the barcode scanner alternative. Though the same processes are supported, the scanner alternative proved to be a less appropriate solution. Although barcode scanner costs half the price of PDAs, they cannot verify the input quality of data, nor display inconsistency checks or warnings. Furthermore, they are not able to alert staff members in case of patients' incompatibilities against drugs or meals. All input can only be checked at the wards' PC. This means, that in case of occurring problems, data has to be recollected again. Comparing the drug supply process scenario operated by the PDA or scanner supported alternative, it becomes clear that barcode scanners lead to additional expenses and to an increasing personnel workload.

In the context of the drug order and supply processes it has not become clear yet if the implementation of intelligent mobile devices can result in measurable advantages in comparison to the implementation of a scanner-supported scenario. This phenomenon is based in the short spatial distance between the ward PCs and the drug cupboards.

But it would make no sense to operate these processes without the PDA system, as it has no measurable negative effects and all different processes can be supported by one system. This lowers the break in barriers for the staff members, who have to be trained only in one system with only one kind of mobile device.

If we go further in the business process map of the hospital, we surely will find other business processes that could be supported by the available mobile system, like, for example, the coordination process for the hospitals laundry.

Once the system's use is established among staff members, the next step might be to support medical business processes, too, if it ends up in a higher efficiency. But before doing that, business process performance analysis will be essential. Nowadays there already exists mobile applications for medical business processes, like decision support systems, e-learning, and telemedical communication systems. One of the prominent mobile systems in Germany is the Stroke Angel system (Holtmann, Rashid, Weinhardt, Gräfe, & Griewing, 2006) in which paramedics use PDAs in emergency ambulance vehicles in case of a stroke to send information about their patient to the targeted hospital. Medicals in this hospital can begin their analysis and make preliminary preparations.. Furthermore, paramedics fill out checklists by PDAs that enable hospitals to make a more precise analysis and collect data for improving the emergency management. Up to present there exists no business process performance analysis, neither of the rescue service nor of the emergency units, in hospitals.

A lot of German hospitals are still not able to tell you how much a treatment belonging to a given and predefined DRG costs. Modeling the existent business processes and enriching them with process costs and time data will be a major first step of improvement—the utilization of mobile applications for performance leverage the second.

Otherwise a hospital might go bankrupt by delivering services in a high quality but to very uncompetitive costs. For example, if a hospital is very famous for its kidney surgery and is planning to enhance it without knowing the cost/benefit structure it is possible to cause serious trouble if the profit contribution is negative, as happened to a competitor of the clinic we had research on.

However, to introduce mobile business process modeling, simulation, and redesign together with activity-based costing will be a major contribution of European hospitals to reduce the costs of their health care systems. Supporting or redesigning distributed clinical processes with mobile applications will be substantial items. This should be an inspiring signal for the telecommunication companies in OECD countries to become an informed partner for the health care sector

with fitting mobile information applications and consultancy services.

REFERENCES

Get Process AG. (2006). *Income process designer.* Retrieved May 25, 2006, from http://www.get-process.de

Greiling, M., & Hofstetter, J. (2002). Patientenbehandlungspfade optimieren—Prozessmanagement im Krankenhaus. Baumann Fachverlag, Kulmbach (Germany).

Grob, L. (1999). *Einführung in die Investitionsrechnung* (3rd ed.). München, Germany: Vahlen.

Hengst, M. den, & De Vreede, G. J. (2004). Collaborative business engineering: A decade of lessons from the field. *Journal of Management Information Systems, 20*(4), 87-115.

Högler, T. (2006). Framework für eine holistische Wirtschaftlichkeitsanalyse mobiler Systeme. In *Proceedings of the MKWI Multikonferenz Wirtschaftsinformatik 2006,* Universität Passau, Germany.

Holtmann, C., Rashid, A., Weinhardt, C., Gräfe, A., & Griewing, B. (2006). Time is brain—Analyse der Rettungskette im Schlaganfall. In *Proceedings of the 5th Workshop of the GMDS Workgroup Mobiles Computing in der Medizin,* Frankfurt, Germany: Shaker Verlag.

Köhler, A., & Gruhn, V. (2004, February 2-3). Mobile process landscaping am Beispiel von Vertriebsprozessen in der Assekuranz. Mobile economy: Transaktionen, Prozesse, Anwendungen und Dienste. In *Proceedings of the 4th Workshop Mobile Commerce,* Universität Augsburg, Germany.

Knoblauch, H. (2005). Focused ethnography. *Qualitative Social Research, 6*(3). Retrieved May 24, 2006, from http://www.qualitative-research.net/fqs-texte/3-05/05-3-44-e.htm

Lazarsfeld, P., Jahoda, M., & Zeisel, H. (1933). *Die Arbeitslosen von Marienthal. Ein soziographischer Versuch über die Wirkungen langdauernder Arbeitslosigkeit.* Germany: Suhrkamp Leipzig.

Morton, S. & Bukhres, O. (1997). Utilizing mobile computing in the Wishard Memorial Hospital ambulatory service. In B. Bryant, J. Carroll, J. Hightower, and K. M. George, (Eds) *Proceedings of the 1997 ACM Symposium on Applied Computing* (SAC '97) (pp. 287-294). San Jose, California, United States. New York, NY: ACM Press.

Nah, F., Siau, K., & Sheng, S. (2005). The value of mobile applications: A utility company study. *Communications of the ACM, 48*(2), 85-90.

Organisation for Economic Co-operation and Development (OECD). (2006). *OECD fact book: Total and public expenditures in health.* Retrieved May 25, 2006, from http://thesius.sourceoecd.org/vl=5439459/cl=16/nw=1/rpsv/factbook/data/10-01-04-t01.xls

Shiffman, R. N., Karras, B. T., Nath, S., Engles-Horton, L., & Corb, G. J. (1999, August). Pen-based, mobile decision support in healthcare. *SIGBIO Newsl, 19*(2), 5-7.

Statistisches Bundesamt. (2003). *Im Jahr 2050 wird jeder Dritte in Deutschland 60 Jahre oder älter sein.* Retrieved May 25, 2006, from http://www.destatis.de/presse/deutsch/pm2003/p2300022.htm

Wang, Y., Van de Kar, E., & Meijer, G. (2005). Designing mobile solutions for mobile workers: Lessons learned from a case study. In *Proceedings of the 7th international conference on Electronic commerce ICEC'05.*

ENDNOTE

[1] She will publish her PhD thesis including improvements of the MPL method this year. First results had been presented in 2006 at the German Conference of Information Systems (WKWI).

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, and J. Day, pp. 174-198, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.17
Information Delivery for Mobile Business:
Architecture for Accessing Large Documents through Mobile Devices

Christopher C. Yang
Chinese University of Hong Kong, Hong Kong

Fu Lee Wang
City University of Hong Kong, Hong Kong

ABSTRACT

In this information-centric age, an organization needs to access the most update and accurate information for fast decision making. Mobile access to Internet provides convenient and portable access to a huge information space. However, loading and visualizing large documents on mobile devices is impossible due to their natural shortcomings such as screen size and computing power. In this chapter, we introduce the fractal summarization model, based on fractal theory, for document summarization on mobile devices. This model generates a brief skeleton of summary at the first stage, and the details of the summary on different levels of the document are generated on demands from users. Such interactive summarization reduces the computation load, which is ideal for wireless access. On the other hand, the hierarchical display in fractal summarization is more suitable for navigation of a large document and it is ideal for small area display. The automatic summarization together with the three-tier architecture and the information visualization are potential solutions to the existing problems in information delivery to mobile devices for mobile business.

INTRODUCTION

Access to the Internet through mobile devices is growing significantly in recent years. The wireless application protocol (WAP) and wireless markup language (WML) provide the universal open

standard and markup language. Many information-centric applications have been developed for mobile devices (Buyukkokten, Garcia-Molina, Paepcke, & Winograd, 2000; Buyukkokten, Garcia-Molina, & Paepcke, 2001a, 2001b, 2001c; Yang & Wang, 2002, 2003b, 2003c). For example, users can now surf the Web, check e-mail, read news, and quote stock prices, using mobile devices. At present, most mobile applications are customer-centered m-services applications. However, mobile computing should not be limited to user-centered applications only. It should be extended to decision support in an m-commerce organization. With a fast-paced economy, organizations need access to large documents or other information sources for fast decision making. As a result, there is an urgent need of a tool for browsing large documents on mobile devices.

Although the development of wireless mobile devices is fast in recent years, there are many shortcomings associated with these devices, such as screen size, bandwidth, and memory capacity. There are two major categories of wireless mobile devices, namely, WAP-enabled mobile phones and wireless personal digital assistants (PDAs). At present, the typical display size of popular WAP-enabled handsets and PDAs is relatively small in comparison with a standard personal computer. The comparatively limited memory capacity of a mobile device also greatly limits the amount of information that can be stored. A large document cannot be entirely downloaded to the mobile device and presented to the user, as the current bandwidth available for WAP is relatively narrow as compared with the broadband Internet connection for PCs.

Despite their convenience, mobile devices impose many constraints that do not exist on desktop computers. The low bandwidth and small resolution are major shortcomings of mobile devices. Information overloading is a critical problem; advance-searching techniques solve the problem by filtering most of the irrelevant information. However, the precision of most of the commercial search engines is not high. Users may only find a few relevant documents out of a large pool of searching results. Given the large screen and high bandwidth for desktop computing, users may still need to browse the search results one by one and identify the relevant information using desktop computers. However, it is impossible to search and visualize the critical information on a small screen with an intolerable slow downloading speed using mobile devices. Automatic summarization summarizes a document for users to preview its major content. Users may determine if the information fits their needs by reading their summary instead of browsing each whole document one by one. The amount of information displayed and downloading time are significantly reduced.

Traditional automatic summarization does not consider the structure of a document, but considers the document as a sequence of sentences. Most of the traditional summarization systems extracted sentences from the source document and concatenated them together as summary. However, it is believed that the document summarization on mobile devices must make use of a "tree view" (Buyukkokten et al., 2001a, 2001b, 2001c) or "hierarchical display" (Mani, 2001). Similar techniques have been applied to Web browsing (Brown & Weihl, 1996): an outline processor organizes the Web page in a tree structure, and the user clicks the link to expand the subsection and view the detail. Hierarchical display is suitable for navigation of a large document, and it is ideal for small area display. Therefore, a new summarization model with hierarchical display is required for summarization on mobile devices.

Summarization on mobile devices in the context of Web pages has been investigated by Buyukkokten et al. (2000, 2001a, 2001b, 2001c). However, a large document exhibits totally different characteristics from Web pages. A Web page usually contains a small number of sentences that are organized into paragraphs, but a large document contains many more sentences that are organized into a more complex hierarchical structure.

Also, the summarization on a Web page is mainly based on thematic features only (Buyukkokten et al., 2001a). However, it has been proven that other document features play as important a role as the thematic feature (Edmundson, 1969; Kepiec, Pedersen, & Chen, 1995). Therefore, a more advanced summarization model combined with other document features is required for browsing large documents on mobile devices.

In this chapter, we propose the fractal summarization model based on the statistical data and the structure of documents. Thematic feature, location feature, heading feature, and cue features are adopted. Summarization is generated interactively. Experiments have been conducted, and the results show that the fractal summarization outperforms the traditional summarization. In addition, information visualization techniques are presented to reduce the visual loads. Three-tier architecture, which reduces the computing load of the mobile devices, is also discussed. In addition to large documents, there is a lot of other valuable information available on the Internet. For example, a great amount of financial news is generated everyday. Access to the most updated and accurate financial information is important during decision making. We will demonstrate the financial news delivery on mobile devices as an example of information delivery for mobile business.

THREE-TIER ARCHITECTURE

Two-tier architecture is typically utilized for Internet access. The user's PC connects to the Internet directly, and the content loaded will be fed to the Web browser and presented to the user as illustrated in Figure 1.

Due to the information-overloading problem, a summarizer is introduced to summarize a document for users to preview before presenting the whole document. As shown in Figure 2, the content will be first fed to the summarizer after loading to the user's PC. The summarizer connects to the database server when necessary and generates a summary to display on the browser.

The two-tier architecture cannot be applied on mobile devices, since the computing power of mobile devices is insufficient to perform summarization and the network connection of a mobile network does not provide sufficient bandwidth for navigation between the summarizer and other servers.

The three-tier architecture as illustrated in Figure 3 is proposed. A WAP gateway is set up to process the summarization. The WAP gateway connects to the Internet through a broadband network. The wireless mobile devices can conduct interactive navigation with the gateway through

Figure 1. Document browsing on PC

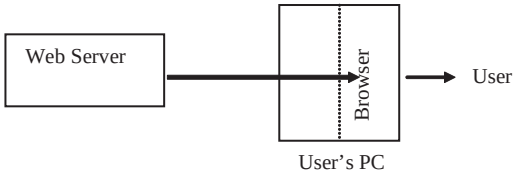


Figure 2. Document browsing with summarizer on PC

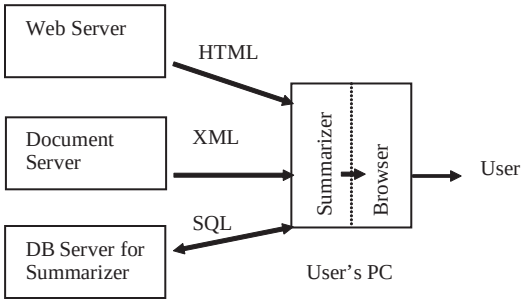
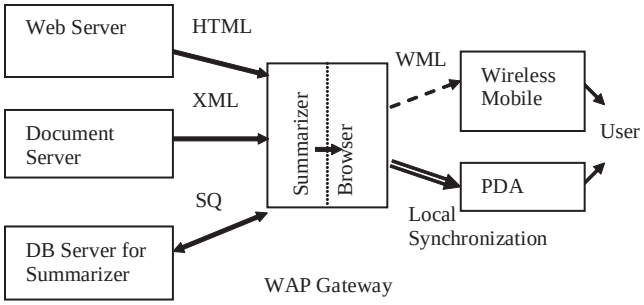


Figure 3. Document browsing with summarizer on WAP



a wireless network to retrieve the summary piece by piece. Alternatively, if the PDA is equipped with more memory, the complete summary can be downloaded to the PDA through local synchronization.

AUTOMATIC SUMMARIZATION

Because there are a lot of shortcomings associated with mobile devices, the Traditional Summarization Model cannot be implemented on a mobile device. A novel summarization model based on hierarchical document structure and fractal theory will be presented.

Traditional Summarization

Traditional automatic text summarization is the selection of sentences from the source document based on their significance to the document (Edmundson, 1969; Luhn, 1958). The selection of sentences is conducted based on the salient features of the document. The thematic, location, heading, and cue features are the most widely used summarization features.

- The thematic feature is first identified by Luhn (1958). Edmundson (1969) proposed to assign the thematic weight to keyword based on term frequency and the sentence weight as the sum of thematic weight of constituent keywords. In information retrieval, absolute term frequency by itself is considered as less useful than term frequency normalized to the document length and term frequency in the collection (Harman, 1992). As a result, the *tfidf* (Term Frequency, Inverse Document Frequency) method is proposed to calculate the thematic weight of keyword (Salton & Buckley, 1988).
- The significance of sentence is indicated by its location (Baxendale, 1958) based on the hypotheses that topic sentences tend to occur at the beginning or end of documents or paragraphs (Edmundson, 1969). Edmondson proposed to assign positive weights to sentences according to their ordinal position in the document—that is, the sentences in the first and last paragraphs and the first and last sentences of the paragraphs. There are several functions proposed to calculate the location weight of sentences. Alternatively,

the preference of sentence location can be stored in a list called Optimum Position Policy, and the sentences will be selected based on their order in the list (Lin & Hovy, 1997).

- The heading feature is proposed based on the hypothesis that the author conceives the heading as circumscribing the subject matter of the document. When the author partitions the document into major sections, he summarizes them by choosing appropriate headings (Edmundson, 1969). The formulation of heading weight is very similar to the thematic feature. A heading glossary is a list consisting of all the words in headings and subheadings. Positive weights are assigned to the heading glossary, where the heading words will be assigned a weight relatively prime to the subheading words. The heading weight of a sentence is calculated by the sum of the heading weight of its constituent words.
- The cue phrase feature is proposed by Edmundson (1969) based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words such as "significant," "impossible," and "hardly." A pre-stored cue dictionary is used to identify the cue phases, which comprises three sub-dictionaries: (i) bonus words, which are positively relevant; (ii) stigma words, which are negatively relevant; and (iii) null words, which are irrelevant. The cue weight of a sentence is calculated by the sum of the cue weight of its constituent words.

Typical summarization systems select a combination of summarization features (Edmundson, 1969; Lin & Hovy, 1997; Luhn, 1958); the total sentence significance score (*SSS*) is calculated as:

$$sss = a_1 \times ss_{thematic} + a_2 \times ss_{location} + a_3 \times ss_{heading} + a_4 \times ss_{cue}$$

where $SS_{thematic}$, $SS_{location}$, $SS_{heading}$, and SS_{cue} are sentence scores based on thematic feature, location feature, heading feature, and cue phrase feature, respectively, and a_1, a_2, a_3, and a_4 are positive integers to adjust the weighting of four summarization features. The sentences with a sentence significant score higher than a threshold are selected as part of the summary. It has been proved that the weighting of different summarization features does not have any substantial effect on the average precision (Lam-Adesina & Jones, 2001). In our experiment, the maximum score of each feature is normalized to one, and the sentence significant score is calculated as the sum of scores of all summarization features without weighting.

Fractal Theory and Fractal View for Controlling Information Displayed

Fractals are mathematical objects that have high degree of redundancy (Mandelbrot, 1983). These objects are made of transformed copies of themselves or part of themselves (see Figure 4). Mandelbrot (1983) was the first person who investigated fractal geometry and developed the fractal theory. In his well-known example, the length of the British coastline depends on measurement scale. The larger the scale is, the smaller the value of the length of the coastline is and the higher the abstraction level is. The British coastline includes bays and peninsulas. Bays include sub-bays, and peninsulas include sub-peninsulas. Using fractals to represent these structures, abstraction of the British coastline can be generated with different abstraction degrees. Fractal theory is grounded in geometry and dimension theory. Fractals are independent of scale and appear equally detailed at any level of magnification. Such property is known as self-similarity. Any portion of a self-similar fractal curve appears identical to the whole curve. If we shrink or enlarge a fractal pattern, its appearance remains unchanged.

Figure 4. Koch curve at different abstraction levels

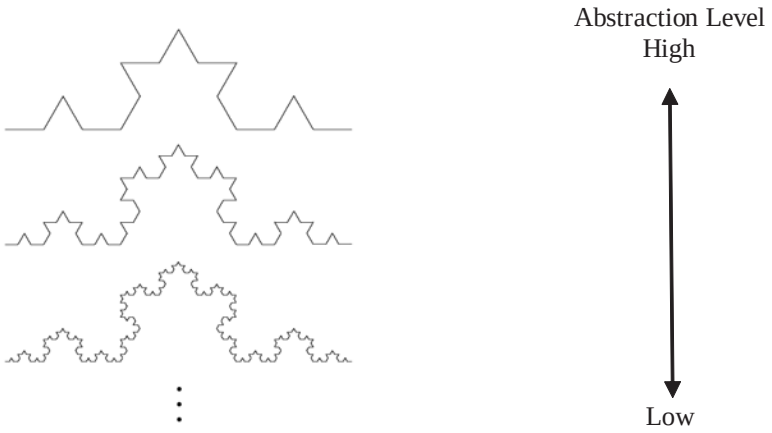


Figure 5. Fractal view for logical tree at different abstraction levels

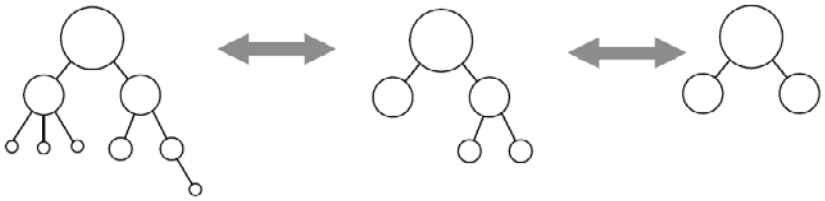
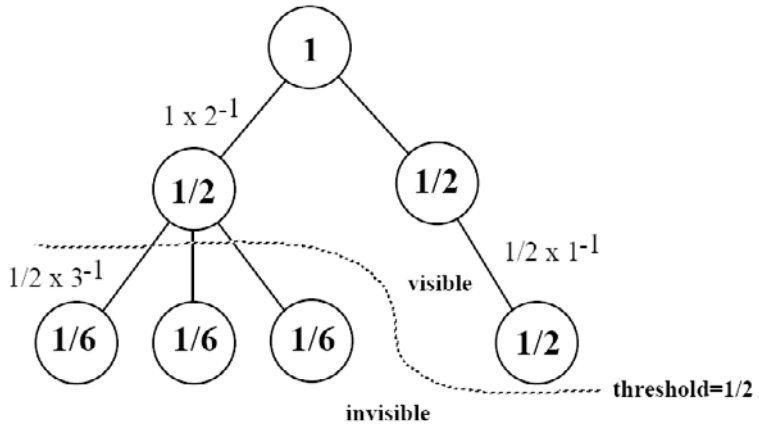


Figure 6. An example of the propagation of fractal values



Fractal view is a fractal-based method for controlling information displayed (Koike, 1995). Fractal view provides an approximation mechanism for the observer to adjust the abstraction level and therefore control the amount of information displayed. At a lower abstraction level, more details of the fractal object can be viewed.

A physical tree is one classical example of fractal objects. A tree is made up of many subtrees; each of them is also a tree. By changing the scale, the different levels of abstraction views are obtained (see Figure 5). The idea of fractal tree can be extended to any logical tree. The degree of importance of each node is represented by its fractal value. The fractal value of focus is set to 1. Regarding the focus as a new root, we propagate the fractal value to other nodes with the following expression:

$$\begin{cases} Fv_{root} = 1 \\ Fv_{child\ node\ of\ x} = C\,\dfrac{Fv_x}{N_x^{\frac{1}{D}}} \end{cases}$$

where Fv_x is the fractal value of node x; C is a constant between 0 and 1 to control rate of decade; N_x is the number of child nodes of node x; and D is the fractal dimension.

A threshold value is chosen to control the amount of information displayed; the nodes with a fractal value less than the threshold value will be hidden (see Figure 6). By change the threshold value, the user can adjust the amount of information displayed.

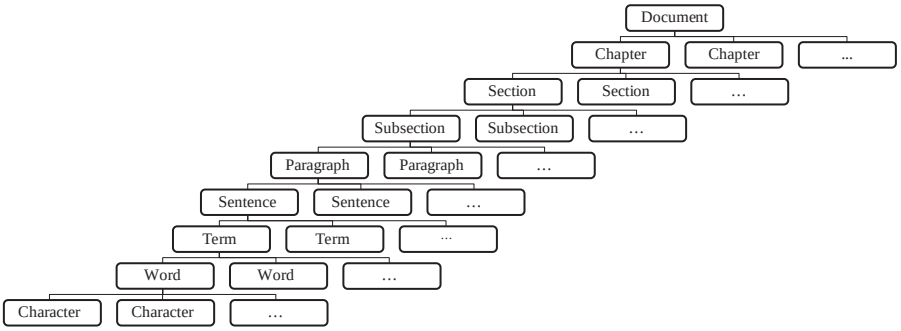
Fractal Summarization

Advance summarization techniques take the document structure into consideration to compute the probability of a sentence to be included in the summary. Many studies (Endres-Niggemeyer, Maier, & Sigel, 1995; Glaser & Strauss, 1967) of human abstraction process have shown that the human abstractors extract the topic sentences according to the document structure from the top level to the low level until they have extracted sufficient information. However, most traditional automatic summarization models consider the source document as a sequence of sentences, but ignore the structure of document. Some summarization systems may calculate sentence weight partially based on the document structure, but they still extract sentences in a linear space. In conclusion, none of the Traditional Summarization Models is entirely based on document structure. *Fractal summarization model* is proposed here to generate summary based on document structure. Fractal summarization generates a brief skeleton of summary at the first stage, and the details of the summary on different levels of the document are generated on demand of the users. Such interactive summarization reduces the computation load in comparing with the generation of the entire summary in one batch by the traditional automatic summarization, which is ideal for m-commerce.

Fractal summarization is developed based on the fractal theory. In our fractal summarization, the important information is captured from the source text by exploring the hierarchical structure and salient features of the document. A condensed version of the document that is informatively close to the original is produced iteratively using the contractive transformation in the fractal theory. Similar to the fractal geometry applying on the British coastline where the coastline includes bays, peninsulas, sub-bays, and sub-peninsulas, a large document has a hierarchical structure with several levels, chapters, sections, subsections, paragraphs, sentences, and terms. A document considered as *prefractal* that is a fractal structure in the early stage with finite recursion only (Feder, 1988).

A document can be represented by a hierarchical structure as shown in Figure 7. A document consists of chapters. A chapter consists of sections. A section may consist of subsections. A section or subsection consists of paragraphs. A paragraph consists of sentences. A sentence

Figure 7. Prefractal structure of document



consists of terms. A term consists of words. A word consists of characters. A document structure can be considered as a fractal structure.

At the lower abstraction level of a document, more specific information can be obtained. Although a document is not a true mathematical fractal object since a document cannot be viewed in an infinite abstraction level, we may consider a document as prefractal. The smallest unit in a document is character; however, neither a character nor a word will convey any meaningful information concerning the overall content of a document. The lowest abstraction level in our consideration is a term.

The fractal summarization model applies a similar technique as fractal view and fractal image compression (Barnslet, 1988; Jacquin, 1993). An image is regularly segmented into sets of non-overlapping square blocks, called range blocks, and then each range block is subdivided into sub-range blocks, until a contractive mapping can be found to represent this sub-range block. The fractal summarization model generates the summary by a simple recursive deterministic algorithm based on the iterated representation of a document. The original document is represented as a fractal tree structure according to its document structure. The system first calculates the sentence significance score based on the summarization features for each sentence. After that, it computes the sum of the normalized sentence significance score of all the sentences under each range block as its Range-Block Significance Score (*RBSS*). The fractal value (*Fv*) of range-block *r* is computed based on the *RBSS* as follows:

$$Fv(r) = \begin{cases} 1 & \text{if } r \text{ is root} \\ C\, Fv(\text{parent of } r) \times \left(\dfrac{RBSS(r)}{\sum_{x \in \text{sibling of } r} RBSS(x)} \right)^{\frac{1}{D}} & \text{otherwise} \end{cases}$$

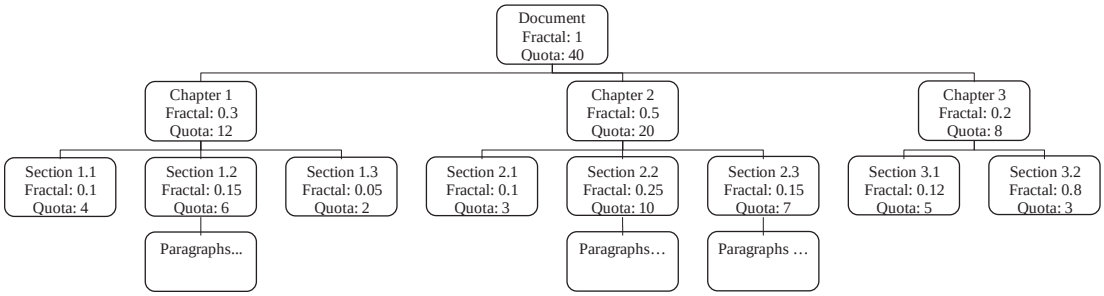
Given a document, a user will specify compression ratio to specify the amount of information displayed. The summarization system calculates the number of sentences to be extracted as summary accordingly, and the system assigns the number of sentences to the root as the quota of sentences. The quota of sentences is allocated to child nodes by propagation—that is, the quota of parent node is shared by its child nodes directly proportional to the fractal value of the child nodes.

The quota is then iteratively allocated to child nodes of child nodes until the quota allocated is less than a threshold value and the range block can be transformed to some key sentences by traditional summarization methods.

Figure 8 demonstrates an example of a fractal summarization model. The detail of the algorithm is shown as Algorithm 1.

The *compression ratio* of summarization is defined as the ratio of number of sentences in the summary to the number of sentences in the source document. It was chosen as 25% in most literature because it has been proven that extraction of 20% of sentences can be as informative as the full text of the source document (Morris, Kasper, & Adams, 1992); those summarization systems

Figure 8. An example of a fractal summarization model



Algorithm 1. Fractal summarization algorithm

Fractal Summarization Algorithm

1. Choose a Compression Ratio.
2. Choose a Threshold Value.
3. Calculate the Sentence Number Quota of the summary.
4. Divide the document into range blocks.
5. Transform the document into fractal tree.
6. Set the current node to the root of the fractal tree.
7. **Repeat**
   - 7.1 **For** each child node under current node,
     Calculate the fractal value of child node.
   - 7.2 Allocate Quota to child nodes in proportion to fractal values.
   - 7.3 **For** each child nodes,
     If the quota is less than threshold value
     Select the sentences in the range block by extraction
     Else
     Set the current node to the child node
     Repeat Step 7.1, 7.2, 7.3
8. **Until** all the child nodes under current node are processed

can achieve up to a 96% precision (Edmundson, 1969; Kepiec et al., 1995; Teufel & Moens, 1997). However, Teufel and Moens (1998) pointed out that high-compression ratio abstracting is more useful, and 49.6% of precision is reported at 4% compression ratio. In order to minimize the bandwidth requirement and reduce the pressure on computing power of mobile devices, the default value of compression ratio is chosen as 4%. By the definition of compression ratio, the sentence quota of the summary can be calculated by the number of sentences in the source document times the compression ratio.

A threshold value is the maximum number of sentences that can be extracted from a range block, if the quota is larger than the threshold value, and the range block must be divided into sub-range block. Document summarization is different from image compression: more than one attractor can be chosen in one range block. It is proven that in the summarization by extraction of a fixed number of sentences, the optimal length of summary is three to five sentences (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999). The default value of threshold is chosen as 5 in our system.

Summarization Features in Fractal Summarization

The weights of sentences under a range block are calculated by the traditional summarization methods described in the former section. However, the traditional summarization features cannot fully utilize the fractal model of a document. In traditional summarization mode, the sentence weight is static through the whole summarization process, but the sentence weight should depend on the abstract level at which the document is currently viewing at, and we will show how the summarization features can integrate with the fractal structure of a document.

Thematic Feature in Fractal Summarization

Among the thematic features proposed previously, the *tfidf* score of a keyword is the most widely used approach; however, in the traditional summarization, it does not take into account the document structure, therefore modification of the *tfidf* formulation is derived to capture the document structure and reflect the significance of a term within a range block.

The *tfidf* score of term t_i is calculated as followed:

$$w_{ij} = tf_{ij} \log_2\left(\frac{N}{n}|t_i|\right)$$

where w_{ij} is the weights of term t_i in document d_j, tf_{ij} is the frequency of term t_i in document d_j, N is the number of documents in the corpus, n is the number of documents in the corpus in which term t_i occurs, and $|t_i|$ is the length of the term *ti*.

Many researchers assume that the weight of a term remains the same over the entire document. However, Hearst (1993) thinks that a term should carry a different weight in a different location of a full-length document. For example, a term appears in chapter A once and appears in chapter B many times; the term is obviously more important in chapter B than in chapter A. This idea can be extended to other document levels: if you look at the document level, a specific term inside a document should carry the same weight; but if you look at a chapter level, a specific term inside a chapter should carry the same weight, but the a specific term inside two chapters may carry different weights.

As a result, the *tfidf* score should be modified to different document levels instead of the whole document. In the fractal summarization model, the *tfidf* should be defined as term frequency

Table 1. tfidf score of the term "Hong Kong" at different document levels

	Term frequency	Text block frequency	No of Text Block	*tfidf* Score
Document-Level	1113	1	1	1217.00
Chapter-Level	70	23	23	70.00
Section-Level	69	247	358	105.95
Subsection-Level	16	405	804	31.83
Paragraph-Level	2	787	2626	5.48
Sentence-Level	1	1113	9098	4.03

within a range block inversely proportional to frequency of range blocks containing the term; for example:

$$w_{ir} = tf_{ir} \log_2\left(\frac{N'}{n'}|t_i|\right)$$

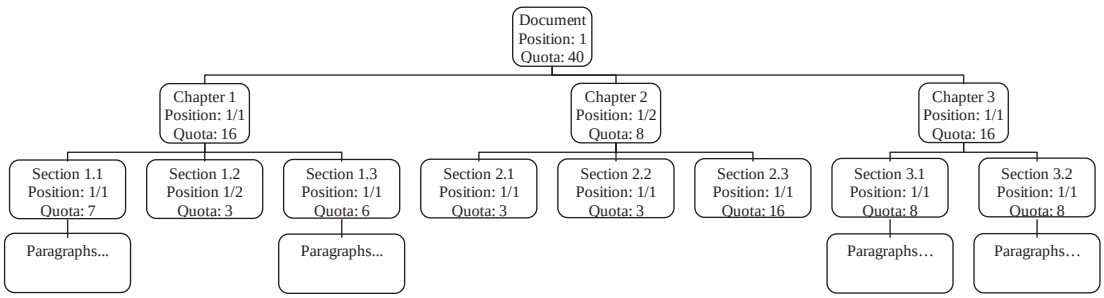
Here, w_{ir} is the weights of term t_i in range block r, tf_{ir} is the frequency of term t_i in range block r, N' is the number of range blocks in the corpus, n' is the number of range blocks in the corpus in which term t_i occurs, and $|t_i|$ is the length of the term *ti*.

Taking the Hong Kong in the first chapter, first section, first subsection, first paragraph, first sentence of the Hong Kong Annual Report 2000 as an example (see Table 1), the *tfidf* score at different document levels differ significantly; the maximum value is 1217.00 at the document level and the minimum is 4.03 at the sentence level.

Location Feature in Fractal Summarization

Traditional summarization systems assume that the location weight of a sentence is static, where the location weight of a sentence is fixed. How-

Figure 9. Fractal summarization with location feature only



ever, the fractal summarization model adopts a dynamic approach; the location weight of a sentence depends on which document level one is viewing.

It is known that the significance of a sentence is affected by the location of the sentence inside a document. For example, the sentences at the beginning and the end of a document are usually more important than the others. If we consider the first and second sentences in the same paragraph at the paragraph level, the first sentence has much more impact on the paragraph than the second sentence. However, the difference of importance of two consecutive sentences is insignificant at the document level. Therefore, the importance of the sentence due to its location should depend on the level we are considering.

In the fractal summarization model, we calculate the location weight for a range block instead of individual sentence; all the sentences within a range block will receive the same location weight. The location weight of a range block is $1/p$, where p is the shortest distance of the range block to the first or last range block under the same parent range block. Consider the previous example of generic fractal summarization model (Figure 8), where the new quota system is changed to Figure 9 if only the location feature is considered.

Heading Feature in Fractal Summarization

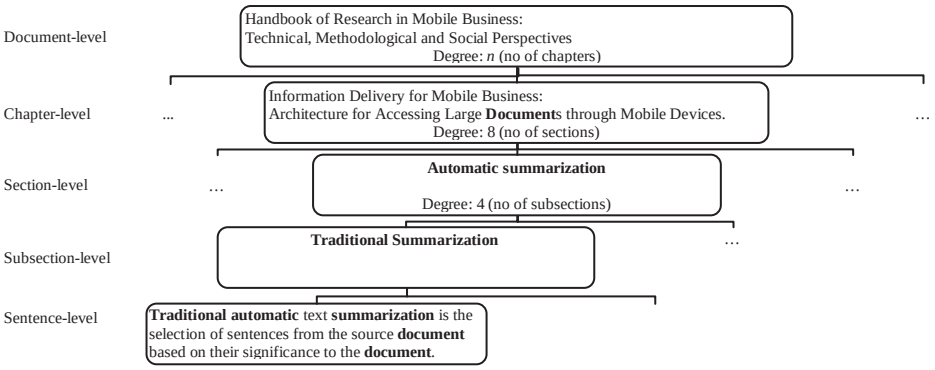
During summarization, a sentence containing a term in its headings is considered as more important. The heading weight of a sentence is dynamic and depends on which level we are currently looking at in the document. At a different abstraction level, some headings should be hidden and some headings must be emphasized.

Taking the first sentence from the first chapter, first section, first subsection, and first paragraph as an example, if we consider the document level, only the document heading should be considered. However, if we consider the chapter level, then we should consider the document heading as well as the chapter heading. Since the main topic of this chapter is represented by the chapter heading, the terms appearing in the chapter heading should have a greater impact on the sentence. Most of the internal nodes above the paragraph level in the document tree usually associate with a heading, and there are two types of headings—structural and informative. The structural headings indicate the structure of the document only, but not any information about the content of the document; for example, "Introduction," "Overview," and "Conclusion" are structural headings. The informative headings can give us an abstract of the content of the branch, and they help us to understand the content of the document and are used for calculation of heading weight. On the other hand, the structural headings can be easily isolated by string matching with a dictionary of those structural headings, and they will be used for cue feature.

The terms in the informative headings are very important in extracting the sentences for summarization. Given a sentence in a paragraph, the headings of its corresponding subsection, section, chapter, and document should be considered. The significance of a term in the heading is also affected by the distance between the sentence and the heading in terms of depth in the hierarchical structure of the document. Propagation of fractal value (Koike, 1995) is a promising approach to calculate the heading weight for a sentence.

The first sentence of the subsection "Traditional Summarization" in this chapter is taken as an example to illustrate the propagation of the heading weight (see Figure 10). The heading of this book is "Handbook of Research in Mobile Business: Technical, Methodological, and Social Perspectives." Assume that there are n chapters, and the heading of this chapter is "Information Delivery for Mobile Business: Architecture for Accessing Large Documents through Mobile Devices." The heading of the fourth section in the chapter is "Automatic Summarization." The

Figure 10. Example of heading feature in fractal summarization



heading of the first subsection is "Traditional Summarization." The first sentence in the subsection is: "Traditional automatic text summarization is the selection of sentences from the source document based on their significance to the document." To compute the heading weight of the sentence, we shall propagate the term weight of the terms that appear in both the sentence and the headings based on the distance between the headings and the sentences, and the number of text units of the intermediate nodes.

$$w_{heading} = w_{heading\ in\ document} + w_{heading\ in\ chapter} + w_{heading\ in\ section} + w_{heading\ in\ subsection}$$

where

$w_{heading\ in\ document} = 0$

$w_{heading\ in\ chapter} = (w_{\text{“}document\text{”}} \times 2)$ *in chapter heading* / (8×4)

$w_{heading\ in\ section} = (w_{\text{“}automatic\text{”}} + w_{\text{“}summarization\text{”}})$ *in section heading* / 4

$w_{heading\ in\ subsection} = (w_{\text{“}traditional\text{”}} + w_{\text{“}summarization\text{”}})$ *in subsection heading*

Cue Feature in Fractal Summarization

The abstracting process of human abstractors can help us understand the cue feature at different document levels. When human abstractors extract the sentences from a document, they will follow the document structure to search the topic sentences. During the searching of information, they will pay more attention to the range block with headings containing some bonus words such as "Conclusion," since they consider it as a more important part in the document and they extract more information for those important parts. The cue feature of the heading sentence is usually classified as the rhetorical feature (Teufel & Moens, 1998).

As a result, we propose to consider the cue feature not only at the sentence level, but also at other document levels. Given a document tree, we will examine the heading of each range block by the method of cue feature and adjust their quota of entire range block accordingly. This procedure can be repeated to sub-range blocks until the sentence level.

Experimental Result

It is believed that a full-length text document contains a set of subtopics (Hearst, 1993), and a good quality summary should cover as many subtopics as possible; the fractal summarization model will produce a summary with a wider coverage of information subtopics than the traditional summarization model.

The traditional summarization model extracts most sentences from a few chapters. Using the Hong Kong Annual Report 2000 as an example (see Table 2), the traditional summarization model extracts 29 sentences from one chapter when the sentence quota is 80 sentences, and a total of 53 sentences are extracted from the top three chapters, out of total 23 chapters; not one sentence is extracted from eight of the chapters. However, the fractal summarization model extracts the sentences distributively from each chapter. In our example, it extracts a maximum of eight sentences from one single chapter, and at

Table 2. Number of sentences extracted by two summarization models from Hong Kong Annual Report 2000

Chapter ID	Chapter Title	Fractal Summarization	Traditional Summarization
1	Hong Kong: Asia's World City	6	3
2	Constitution and Administration	4	1
3	The Legal System	2	0
4	The Economy	5	14
5	Financial; and Monetary affairs	8	29
6	Commerce and Industry	6	10
7	Employment	2	2
8	Primary Production	1	0
9	Education	2	1
10	Health	1	0
11	Social Welfare	1	0
12	Housing	1	0
13	Land, Public Works and Utilities	4	0
14	Transport	5	3
15	Infrastructure	1	0
16	The Environment	4	1
17	Travel and Tourism	1	1
18	Public Order	5	2
19	Communications, the Media and Information Technology	6	6
20	Religion and Custom	2	0
21	Recreation, Sport and the Arts	5	3
22	Population and Immigration	3	1
23	History	5	3

least one sentence is extracted from each chapter. The standard deviation of sentence number extracted from chapters is 2.11 sentences in fractal summarization vs. 6.55 sentences in traditional summarization. Researchers believe that a good summary should find diverse topic areas in the text and reduce the redundancy of information contents in the summary (Nomoto & Matsumoto, 2001). Fractal summarization extracts the sentences distributively, therefore it finds diverse topic areas and reduces the redundancy of information at the same time.

A user evaluation is conducted. Ten subjects were asked to evaluate the quality of summaries of 23 documents generated by fractal summarization and traditional summarization. Both summaries of all documents are assigned to each subject in random order without telling the generation methods of the summaries. The results show that all subjects consider the summary generated by fractal summarization method as a better summary. In order to compare the result in greater detail, we calculate the precision as the number of relevant sentences in the summary accepted by the user, divided by the number of sentences in the summary (see Table 3). The fractal summarization can achieve up to 91.25% precision and 87.16% on average, while the traditional summarization can achieve up to a maximum 77.50% precision and 67.00% on average. The one-tailed T-test has shown that the precision of the fractal summarization model outperforms traditional summarization significantly, at a 99% confidence level.

VISUALIZATION OF FRACTAL SUMMARIZATION

The summary generated by the fractal summarization model is represented in a hierarchical tree structure. The hierarchical structure of summary is suitable for visualization on mobile devices, and it can be further enhanced by displaying the sentences in different font sizes. A summary displayed in a small area without visualization effect is difficult to read. Displaying the sentences in different font sizes according to their importance can help users to focus on important information.

WML is the markup language supported by wireless mobile devices. The basic unit of a WML file is a deck; each deck must contain one or more cards. The card element defines the content dis-

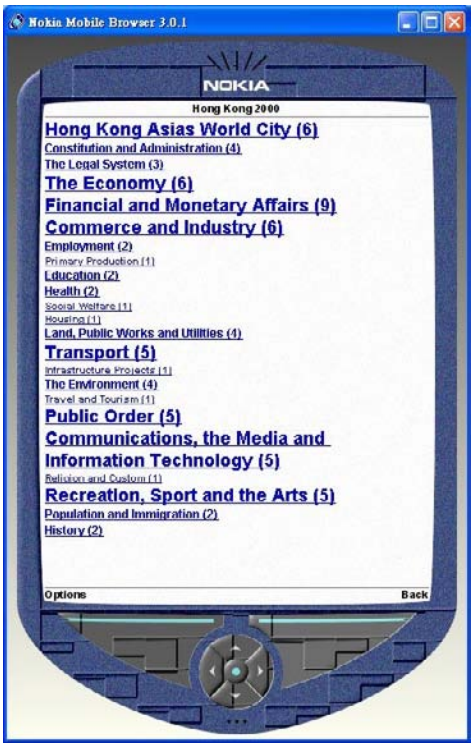
Table 3. Precision of two summarization models

User ID	Fractal Summarization Model	Traditional Summarization Model
User 1	81.25%	71.25%
User 2	85.00%	67.50%
User 3	80.00%	56.25%
User 4	85.00%	63.75%
User 5	88.75%	77.50%
User 6	81.25%	61.25%
User 7	91.25%	76.25%
User 8	86.25%	58.75%
User 9	85.00%	65.00%
User 10	87.50%	72.50%

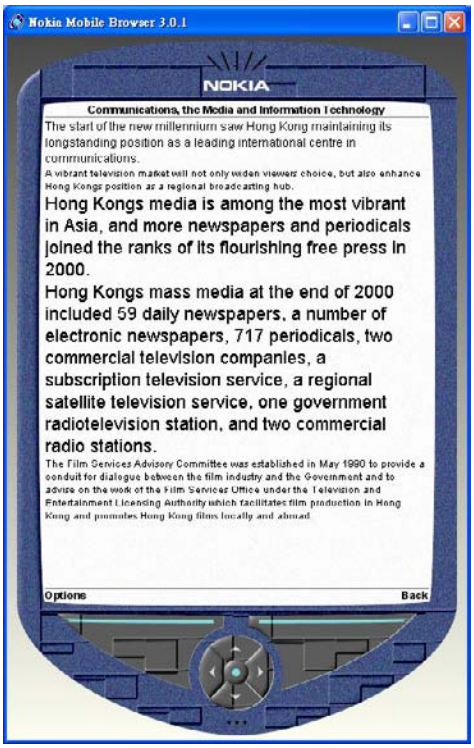
played to users, and the card cannot be nested. Each card links to another card within or across decks. Nodes on the fractal tree of the fractal summarization model are converted into cards, and anchor links are utilized to implement the tree structure. Given a card of a summary node, there may be a lot of sentences or child nodes. A large number of sentences in a small display area make them difficult to read. In our system, the sentences are displayed in different font sizes according to their significance. We have implemented the system with a three-scale font mode available for WML. The sentences or child nodes are sorted by their sentence weights or fractal value and separated evenly into three groups. The group with highest value is displayed in "Large" font size, and the group with middle value and the group with lowest value are displayed in "Normal" and "Small" font sizes respectively.

The prototype system using Nokia Handset Simulator is presented in Figure 11. The document we are using is the Hong Kong Annual Report 2000. There are a total of 23 chapters in the annual report, eight of them are in large font, which means that they are more important; the rest are in normal font or small font according to their

Figure 11. Screen capture of WAP summarization system



a. Hong Kong Anual Report 2000



b. Chapter 19 of the Hong Kong Annual Report 2000, "Communications, Media and Information Technology"

importance to the report (Figure 11a). The number inside the parentheses indicates the number of sentences under the node that are extracted as part of the summary.

The main screen of the Hong Kong Annual Report 2000 gives the user a general idea of overall information content and the importance of each chapter. If the user wants to explore a particular node, he or she can click the anchor link and the mobile device sends the request to the WAP gateway; the gateway then decides whether to deliver another menu or the summary of the node to the user depending on its fractal value and quota allocated. Figure 11b shows the summary of Chapter 19 of the Hong Kong Annual Report 2000, "Communication, the Media, and Information Technology."

A handheld PDA is usually equipped with more memory, and the complete summary can be downloaded as a single WML file to the PDA through local synchronization. To read the summary, the PDA is required to install a standard WML file reader.

FINANCIAL NEWS DELIVERY ON MOBILE DEVICES

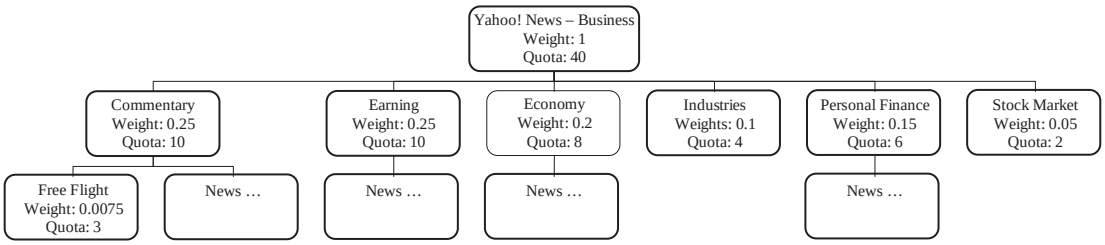
The fractal summarization model summarizes the documents based on hierarchical document structure. In addition to a large text document, many other information sources also exhibit hierarchical document structure. Due to the large amount of information available, it is difficult to browse these information sources on mobile devices. Automatic summarization is a possible solution. Theoretically, fractal summarization is capable of summarizing all of these information sources, as long as the calculation of fractal value is well formulated. As financial news is critical in decision making, we shall modify the fractal value formula of generic fractal summarization in order to summarize the financial news, and we shall demonstrate financial news delivery with fractal summarization on mobile devices.

Fractal Summarization of Financial News

The fractal summarization model performs summarization based on hierarchical document structure. In addition to large text documents, a lot of other documents also exhibit hierarchical tree document structure, such as Web sites and newspapers. The fractal summarization model is capable of summarizing these documents based on their structure and their relationships in categorization; therefore it is a powerful tool in providing m-services of real-time information delivery. At present, many electronic news delivery services have been provided. An example of the fractal summarization model being used to summarize the financial news available from the Internet is presented in this section.

A newspaper is one of the documents that exhibits the well-defined hierarchical document structure. At present, there are many electronic news delivery services provided for PC, and most of them provide summarization tools to help the user search information, such as the Lycos Financial Feed System with a summarization system from Diyatech, and YellowBrix with Inxight's Summarizer. However, summarizers for the PC platform are not adaptable to mobile devices directly. Moreover, the existing commercial summarizers are indeed extracting the first few sentences from the document or using the primitive summarization model without considering the hierarchical structure of documents or the organization of information. Yahoo!News is one of the most popular online content providers. There are 21 categories in Yahoo!News. Moreover, each of the categories will be subdivided into subcategories. Take the Business category as an example (see Figure 12). This category contains financial news and is subdivided into six

Figure 12. Fractal summarization of Yahoo! News–Business category



subcategories, namely, Economy, Stock Markets, Earnings, Personal Finance, Industries, and Commentary. Each subcategory contains about 10 news articles. Each news article is a tree structure by itself. For some longer news articles, there may exist more than one section, each section contains a few paragraphs, and each paragraph contains sentences.

The fractal summarization of Yahoo!News is very similar to the fractal summarization of a large text document; only some minor modifications are required to demonstrate the characteristics of Yahoo!News.

- First, the headings of categories and subcategories do not have a direct impact on the content of news under the branch; it serves for classification purpose only. As a result, the heading method will consider the headings of news articles only. In addition, the headings of categories can be used for personalization of news delivery, the user can set his preference of each category in advance, and the system will adjust the weights accordingly. Alternatively, the preference can be constructed by auto-learning of a machine in the middle tier. The WAP gateway can analyze the reading behavior of the user and predict the user's preference.
- The location feature in traditional summarization assumes that the text unit in the beginning or ending is more important. The news articles inside a subcategory are sorted in chronological order. The most recent news is usually considered as more important. Therefore, we propose calculating the location weight of a news article by its chronological position in the subcategory or the time lag between the news event and browsing time. However, when the system traces the summarization tree down to a node inside a news article, the generic location method in fractal summarization will be adopted.
- In order to provide a glimpse of every article, each news article will receive a sentence quota with at least one sentence.

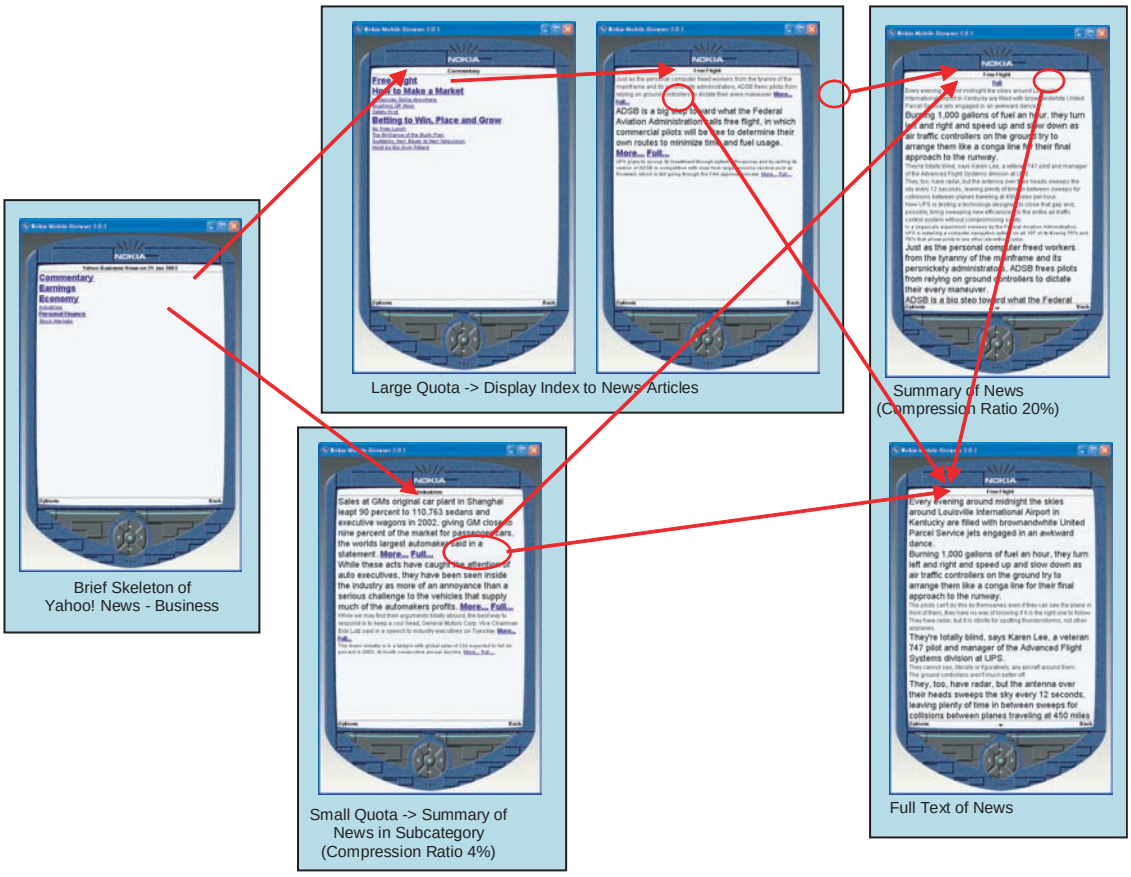
Financial News Delivery to Mobile Devices

In order to minimize the bandwidth requirement and reduce the pressure on computing power of mobile devices, the summarization of Yahoo!News will be conducted in two levels. As high-compression ratio abstracting is more useful (Teufel & Moens, 1998) and can save network bandwidth, the fractal summarization system generates a

brief skeleton of summary, with compression ratio equal to 4% at the first stage. The details of the summary at different levels of the news tree are generated on demand by users.

When the mobile device retrieves the financial news from Yahoo!News–Business, the system will first show a card containing six subcategories of the Business category (see Figure 13). In the Figure, three subcategories of the Business category are displayed in large font, which means that they are more important; the rest are in a normal or small font according to their importance. The skeleton of news gives the user a general idea about how the news articles are organized, and the user can decide in which subcategory to go into detail. When the user clicks the anchor link of subcategory, the WAP gateway will deliver a card depending on the quota allocated. If a large quota is allocated to the subcategory, the system will show another card containing an index of news article. However, if the quota is less than a threshold value of five sentences, the system will show a card with the summary of all news articles in the subcategory. In the summary page, when the user clicks the

Figure 13. Financial news delivery system on mobile devices



anchor link 'More' at the end of sentences, the system will generate the summary for the corresponding news articles with a compression ratio of 20%, because it has been proven that the extraction of 20% sentences can be as informative as the full text of the source document (Morris et al., 1992). On the other hand, the user can click the anchor link 'Full' to view the full text of the news articles. Such interactive summarization reduces the computation load, when comparing it with the generation of the entire summary in one batch by the traditional automatic summarization, which is ideal for m-services.

CONCLUSION AND FUTURE DIRECTION

Mobile business is a promising addition to the electronic commerce by the adoption of portable mobile devices. However, mobile computing should not be limited to user-centered m-service applications only; it should be extended to decision making in an organization. With a fast-paced economy, organizations need to make decisions as fast as possible, and access to large text documents or other information sources is important during decision making. Unfortunately, there are many shortcomings of the mobile devices, such as limited resolution and narrow bandwidth. In order to overcome the shortcomings, fractal summarization and information visualization are proposed in this chapter; these are critical in decision support in an m-organization. Fractal summarization creates a summary in the hierarchical tree structure and presents the summary to the mobile devices through cards in WML. The adoption of keyword feature, location feature, heading feature, and cue feature are discussed. Users may browse the selected summary by clicking the anchor links from the highest abstraction level to the lowest abstraction level. Based on the sentence weight computed by the summarization technique, the sentences are displayed in different font size to enlarge the focus of interest and diminish the less significant sentences. Such visualization effect draws users' attention to the important content. The three-tier architecture is presented to reduce the computing load of the mobile devices. The proposed system creates an information visualization environment to avoid the existing shortcomings of mobile devices for mobile business.

In its current stage, fractal summarization is capable of processing textual information only. However, there is a lot of information available in multimedia formats on the Web. Information delivery of multimedia documents will be one of key research topics in the near future. As multimedia documents require a much higher bandwidth than textual documents, this problem cannot be resolved solely by the current steaming technology. Summarization of multimedia documents is required for information delivery to mobile devices. The research of spoken document summarization and video summarization has been started (Vasconcelos & Lippman, 1998; Zechner & Waibel, 2000). It would be a great challenge to move the proposed model to multimedia documents. The summarization of multimedia documents is complementary to the proposed model. Nowadays, most of the mobile devices are speech based. With the summarization of spoken documents, the information can be easily delivered to speech-based mobile devices. This will certainly increase the popularity of the proposed model.

REFERENCES

Barnsley, M. F., & Jacquin, A. E. (1988, November). Application of recurrent iterated function systems to images. *Proceedings of the Conference on SPIE Visual Communications and Image Processing (VCIP'88),* Cambridge, MA (vol. 1001, pp. 122-131).

Baxendale, P. B. (1958). Machine-made index for technical literature—An experiment. *IBM Journal of Research and Development, 2*(2), 354-361.

Brown, M. H., & Weihl, W. E. (1996, October). Zippers: A focus + context display of Web pages. *Proceedings of the World Conference of the Web Society (WebNet'96),* San Francisco, CA.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001a, May). Seeing the whole in parts: Text summarization for Web browsing on handheld devices. *Proceedings of the 10th International Conference on the World Wide Web (WWW10),* Hong Kong, China (pp. 652-662).

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001b, March). Accordion summarization for end-game browsing on PDAs and cellular phones. *Proceedings of the SIGCHI Conference on Human Factors in Computing System (CHI 2001),* Seattle, WA (pp. 213-220).

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001c, June). Text summarization of Web pages on handheld devices. *Proceedings of the Workshop on Automatic Summarization 2001 in conjunction with the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001),* Pittsburgh, PA.

Buyukkokten, O., Garcia-Molina, H., Paepcke, A., & Winograd, T. (2000, April). Power browser: Efficient Web browsing for PDAs. *Proceedings of the SIGCHI Conference on Human Factors in Computing System (CHI 2000),* Hague, The Netherlands (pp. 430-437).

Edmundson, H. P. (1969). New method in automatic extraction. *Journal of the ACM, 16*(2), 264-285.

Endres-Niggemeyer, B., Maier, E., & Sigel, A. (1995). How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing and Management, 31*(5), 631-674.

Feder, J. (1988). *Fractals.* New York: Plenum.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory; strategies for qualitative research.* New York: Aldine de Gruyter.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999, August). Summarizing text documents: Sentence selection and evaluation metrics. *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99),* Berkeley, CA (pp. 121-128).

Harman, D. K. (1992). Ranking algorithms. In W.B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 363-392). Englewood Cliffs, NJ: Prentice-Hall.

Hearst, M. A. (1993, June). Subtopic structuring for full-length document access. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93),* Pittsburgh, PA (pp. 56-68).

Jacquin, A. E. (1993). Fractal image coding: A review. *IEEE, 81*(10), 1451-1465.

Kepiec, J., Pedersen, J., & Chen, F. (1995, July). A trainable document summarizer. *Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'95),* Seattle, WA (pp. 68-73).

Koike, H. (1995). Fractal views: A fractal-based method for controlling information display. *ACM Transactions on Information Systems, 13*(3), 305-323.

Lam-Adesina, M., & Jones, G. J. F. (2001, September). Applying summarization techniques for term selection in relevance feedback. *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'01),* New Orleans, LA (pp. 1-9).

Lin, Y., & Hovy, E. H. (1997, March). Identifying topics by position. *Proceedings of the Workshop*

on Intelligent Scalable Text Summarization in conjunction with the 5th Conference on Applied Natural Language Processing Conference (ANLP'97), Washington, DC (pp. 283-290).

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.

Mandelbrot, B. (1983). *The fractal geometry of nature*. New York: W. H. Freeman.

Mani, I. (2001, November). Recent development in text summarization. *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01),* Atlanta, GA (pp. 529-531).

Morris, A. H., Kasper, G. M., & Adams, D. A. (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Information System Research, 3*(1), 17-35.

Nomoto, T., & Matsumoto, Y. (2001, September). A new approach to unsupervised text summarization. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01),* New Orleans, LA (pp. 26-34).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513-523.

Teufel, S., & Moens, M. (1997, July). Sentence extraction as a classification task. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent and Scalable Text Summarization,* Madrid, Spain (pp. 58-68).

Teufel, S., & Moens, M. (1998, March). Sentence extraction and rhetorical classification for flexible abstracts. *Proceedings of the 1998 AAAI Spring Symposium on Intelligent Text Summarization,* Palo Alto, CA (pp. 16-25).

Vasconcelos, N., & Lippman, A. (1998, June). A spatiotemporal motion model for video summarization. *Proceedings of the IEEE Computer Society Conference on computer Vision and Pattern Recognition (CVPR'98),* Santa Barbara, CA (pp. 361-366).

Yang, C. C., & Wang, F. L. (2002, December). Document summarization on handheld device: An information visualization tool for mobile commerce. *Proceedings of the First Workshop on E-Business (WEB2002) of the International Conference on Information Systems (ICIS 2002),* Barcelona, Spain.

Yang, C. C., & Wang, F. L. (2003, July). Fractal summarization: Summarization based on fractal theory. *Proceedings of the 26th Annual International ACM SIGIR Conference: Research and Development in Information Retrieval (SIGIR 2003),* Toronto, Canada (pp. 391-392).

Yang, C. C., & Wang, F. L. (2003, May). Automatic summarization for financial news delivery on mobile devices. *Proceedings of the 12th International Conference on the World Wide Web (WWW2003),* Budapest, Hungary (pp. 391-392).

Yang, C. C., & Wang, F. L. (2003, May). Fractal summarization for mobile devices to access large documents on the Web. *Proceedings of the 12th International Conference on the World Wide Web (WWW2003),* Budapest, Hungary (pp. 215-224).

Zechner, K., & Waibel, A. (2000, July). DiaSumm: Flexible summarization of spontaneous dialogues in unrestricted domains. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000),* Saarbruecken, Germany (pp. 968-974).

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 247-266, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.18
Resource-Based Interdependencies in Value Networks for Mobile E-Services

Uta Wehn Montalvo
TNO Strategy, Technology and Policy, The Netherlands

Els van de Kar
Delft University of Technology, The Netherlands

Carleen Maitland
Pennsylvania State University, USA

ABSTRACT

The advent of new electronic platforms, such as fixed and mobile Internet, is forcing firms from a range of industries to come together in so-called value networks for the provision of innovative e-services. Firms from different industries have widely varying resources. Our analysis is aimed at specific types of interdependencies, relating the actors' own and others' resource contributions to the value creation involved in bringing the service about. To better understand these interdependencies, we draw on theories about firm resources and interorganizational relations. We analyze the importance and relevance of different resources in a number of case studies of mobile information and entertainment services in terms of the actors' resources and contributions to value in the provision of such mobile services. In the cross-case comparison, we contrast the power structures in the different value networks and identify similarities and differences in terms of the types of industrial players that assume positions of greater or lesser importance. We conclude with a discussion of the implications for value network research.

INTRODUCTION

The advent and adoption of new electronic platforms, such as fixed and mobile Internet, provides a basis for so-called e-services. E-services have been defined as "any asset that is made available via the Internet to drive new revenue streams or create new efficiencies" (Piccinelli et al., 2001, p. 100). The rapid evolution of these services has left many issues unresolved. The problems of interest to us include uncertainty with respect to the complex networks that are involved in delivering these services. In environments of increasing electronic interaction, the value chain concept, where materials are moved sequentially down a supply chain, has been replaced by the value network, a dynamic network of partnerships and information flows (Bovel & Martha, 2000), changing as customer preferences change. This phenomenon also is taking place in the telecommunications industry (Fi & Whalley, 2002; Maitland et al., 2002; Sabat, 2002). We aim to understand the interdependencies among actors involved in delivering mobile services in general, and mobile information and entertainment services and location based services in particular, in terms of their contribution to value creation. To this end, we adopt a resource-based perspective.

We consider a number of innovative cases of mobile information and entertainment services. Such services involve the delivery of information and entertainment content to a mobile user. Since these services typically require collaboration of a range of actors across different sectors, our analysis encompasses the entire value network of firms involved in making the service available. Whilst research on value networks for mobile services could be approached from several angles, including network formation, strategic management, and so forth, here, we focus on resources and interdependencies. We investigate the actual constellation of actors; what are their resources, how are they interdependent, and what do they contribute to the value network?

The chapter is structured as follows. We begin with a brief review of relevant literature to provide a basis for our analysis of several mobile information and entertainment services. In particular, we examine the interdependencies among actors in the value networks and how their contribution to value creation determines their strategic position within the network. These tools are then used to analyze each of the five case studies of specific services. In the cross-case analysis, we collate and discuss the findings from the cases. We conclude with implications of our research for the literature on value networks and point to further areas of research.

THEORETICAL CONTEXT

A fundamental aspect of a value network is that it accomplishes the directed utilization of resources in the provision of a product or service. In the following subsections, we derive a basis for our analysis of interdependencies in mobile information and entertainment services. The aim is to arrive at an analytical tool that can be used to understand the interdependencies among actors involved in delivering such services in terms of their contribution to value creation. This will provide important insights into the configuration and dynamics of actors in value networks.

We begin with a definition of resources, and, given the context of value networks, we include a discussion of the resource-based view and its links to strategic alliances. Next, we look at interactions among organizations in interorganizational relations and, more specifically, value networks for the provision of mobile services. Finally, we consider different classes of interdependencies, focusing on the strategic position of firms within the value network and not within the market. We will argue that the configuration of actors is based on their resource-based contribution to value. We conclude this section with a summary of the

analytical tools to be employed in the analysis of the case studies.

Resources

In this section, we first consider the concept of resources in detail in order to arrive at a definition/classification of resources for the analysis of interdependencies.

Definition of Resources

Resources have been studied from many perspectives, and the concept can be conceived very broadly to include almost everything in an organizational (capital, labor, infrastructure, technology, knowledge, processes, routines, capabilities) and interorganizational setting (relationships, etc.). Hoskinson et al. (1999) review a range of studies by researchers from different disciplines that analyzes resources giving rise to competitive advantage.

With our focus on the role of resources as they relate to interdependencies in a network of actors, we find the distinction between *tangible* and *intangible* resources (Itami & Roehl, 1987; Wernerfelt, 1984) most useful. Haanes and Fjeldstad (2000) identify tangible resources as concrete and tradeable, factories, technology, capital, raw material, and land and intangible resources as difficult to transfer, skills, knowledge, relationships, culture, reputation, competencies. Essentially, this distinction parallels the two types of resources—*property-based* and *knowledge-based*—identified by Miller and Shamsie (1996). Building on this distinction, Das and Teng (2000) have identified three salient characteristics of resources in the resource-based literature and arrive at a matrix of resources that illustrates specific kinds of resources in each category. The basis for this classification is the reasoning that alliances need to be formed in order to obtain resources featuring imperfect mobility, imitability, and substitutability. Imperfect mobility refers to the difficulty and cost of moving certain resources from one firm to another and obtaining them from the owner. Imperfect imitability and imperfect substitutability imply the difficulty of obtaining similar resources elsewhere. Complementing this with the external assets identified by Porter (1991) (reputation and relationships), we arrive at the following illustration of resources (see Table 1).

Resources, Strategic Alliances and Value Networks

Resources have been considered in a range of different literatures, which play a particularly central

Table 1. Typical resources (based on Das and Teng [2000, p. 42] and Porter [1991])

Resource Characteristics	**Resource Type**	
	Property-Based Resources	Knowledge-Based Resources
Imperfect Mobility	Human resources	Organisational resources (e.g. culture, reputation, relationships)
Imperfect Imitability	Patents, contracts, copyrights, trademarks, and registered designs	Technological and managerial resources, skills
Imperfect Substitutability	Physical resources	Technological and managerial resources, skills

role in the resource-based view (RBV) of the firm and in the resource dependence literature. More recently, links have been established between the RBV literature and the role of resources in strategic alliances (Appelman, 2004; Das & Teng, 2000). We consult these to arrive at a definition (classification scheme) of resources for our analysis of interdependencies in value networks of mobile information and entertainment services.

The focus of the resource-based view (RBV) is the resources possessed by the firm. The RBV stresses value maximization through the integration of resources. Successful firms are those firms that are able to acquire and maintain valuable idiosyncratic resources for competitive advantages (Oliver, 1997).

The resource-based view has been applied mainly to the individual firm to analyze various resources possessed by the firm, but increasingly also in strategy research. Recently, the resource-based view also has been linked to a network perspective, specifically by considering the resource-based view in the context of strategic alliances: "the resource-based view suggests that the rationale for alliances is the value-creation potential of firm resources that are pooled together" (Das & Teng, 2000, p. 56).

The application of the resource-based view to research on strategic alliances provides the link with value network research in focus here. Strategic alliances can be regarded as a category of interorganizational relations and networks. The common premise is that it is precisely the complementarity of resources that necessitates the formation and evolution of both strategic alliances and value networks, and that none of the actors can make all the necessary components available for product development or service provision. "The resource-based logic suggests that the competitive advantage of alliances is based on the effective integration of partner firms' valuable resources" (Das & Teng, 2000, p. 48). A resource-based perspective of the actors, therefore, provides a relevant basis to examine interdependence in the value network. From a resource-based perspective, paraphrasing Das and Teng (1998) on strategic alliances, value networks are about combining resources that an individual firm cannot provide all on its own, yet are critical for the provision of a mobile service.

Interdependencies

It has long been argued that all firms are embedded in one or more networks in which they collaborate with others to create value and in order to service the markets (Granovetter, 1985). Network boundaries are not easily defined, because mostly there is no overarching purpose for the interactions. As noted in Maitland et al. (2003a, 2003b), this is different for so-called value networks, where the boundaries of the network can be more clearly distinguished by identifying the actors involved in the provision of a specific service. In a value network, the interaction among actors is goal-directed (i.e., the provision of a service) and cannot be assumed to be influenced merely by the individual actors' intention to influence each other. Value networks imply interdependencies (which may differ in their form and extent) among the organizations involved in it. Our analysis is aimed at specific relationships and interdependencies within the value network (i.e., the actors' own and others' resources) rather than in terms of products, markets, and competitors. Gadde et al. (2003) have argued that each actor has a unique position in the network that is perceived differently by the different actors in the network, because all have different relationships. We are interested in a more objective assessment of the different actors' positions within the value network on the basis of the resources and their relevance or importance to value creation in a given network.

In social systems and social interactions, interdependence exists when one actors does not entirely control all of the conditions necessary for the achievement of an action or for obtaining

the outcome desired from the action. … Interdependence characterizes the relationship between the agents creating an outcome, not the outcome itself. (Pfeffer & Salancik, 1978, p. 40)

Theories on strategic management and resource dependence often have regarded interdependencies among organizations as inherently negative. Emphasis, therefore, is placed on how to manage interdependencies, on the implication of different coordination mechanisms (Ebers, 1999), and on strategies to restructure the conditions of interdependence (Mintzberg, 1979, 1983; Nassimbeni, 1998). In order to analyze dependencies in industrial networks, Håkansson (1987) and Håkansson and Waluszewski (2002) present a network model, inspired by strategic management theory, with three dimensions: (1) actors, (2) activities, and (3) resources, whereby actors perform activities and control resources. Activities are used to change other resources in different ways. These three elements are assumed to be related to each other as networks (i.e., actors related to other actors, activities related to other activities, and resources related to other resources). In addition, these networks are closely connected in an overall network. The interdependence between various relationships in the network implies that a certain actor's change in behavior also influences the position of other actors (Axelsson, 1987).

The distinctions and relations of the dimensions in this network can provide a basis for our analysis of interdependencies in value networks and the process of value creation. Activities within the value network bring together different types of actors and resources and create (different) relationships of (inter)dependency.

Several forces are identified, binding the three networks together (actor network, activities network, and resource network) (Håkansson & Johanson, 1984): (1) functional interdependence (actors, activities, and resources as a system that is functionally related), (2) power structure (actor power based on activities and resources), (3) knowledge structure (activities' design and resource use bound together by actors' knowledge), and (4) time-related structure (network as a product of its history). For our analysis of the strategic position of actors within the value network in terms of their contribution to value creation, the second type (power structure) is of greatest interest for our analysis. Actor power is assumed to be based on the activities and resources of a particular actor. In particular, we argue, that actor power stems from the characteristics (i.e., degree of mobility, imitability, and substitutability) of the resources. To typify the power structure among the actors in a value network, we propose a distinction between *essential*, *network-specific*, and *generic* resource contributions to value creation, ranging from greater to lesser relevance to value creation in the network, based on resource characteristics. We define *essential* resources as resources that are indispensable to the value network and the service it provides. These resources cannot be replaced without affecting the existence of the service, and they are highly immobile and difficult to imitate or substitute. *Network-specific* resources are crucial for the service that the value network provides, yet their replacement would be possible without affecting the service directly. They are fairly mobile and possible to be imitated or substituted. *Generic* resources are required for the provision of the service, but they are so general that they could be replaced fairly easily without impacting the service. They are reasonably mobile and imitable or substitutable.

This distinction provides a basis for defining different partner types in the value network: *structural*, *contributing*, and *supporting* partners (ranging from greater to lesser actor power, depending on the kind of resources they contribute), thus identifying the nature of interdependencies in a given network and the strategic position of actors within the network (Ballon & Hawkins, 2003).

Summary and Conclusions on Theoretical Framework

Summarizing the above discussion, this section provides a brief overview of the key concepts and their definitions that will be used in the subsequent analysis of mobile information and entertainment services. In order to unpack interdependencies in value networks for the provision of mobile information and entertainment services, we adopt Hakansson's distinction of networks of actors, activities, and resources as a functionally related system.

To capture the importance or relevance of different resources to value creation in a given network, we have proposed a distinction between *essential*, *network-specific*, and *generic* resource contributions to the value network. Each of these contributions may be in the form of tangible (property-based) or intangible (knowledge-based) resources. For our analysis, the following matrix (see Table 2) will be used to map out the different resources in a given value network and their relevance or importance to it. In each case study, we consider the actors and their resource contribution to the network.

Given our interest in resources from an interdependency perspective, our focus is not on all possible resources that a partner may possess. Rather, we consider resources of partners in terms of their contribution to the value network (i.e., to the provision of the specific service). With this approach, we also are able to counter criticisms of the resource-based view (Foss, 1998) by looking beyond the individual resource and considering how resources are clustered and how they relate; in this case, in the provision of a mobile service provided by a value network.

The distinction of different resource contributions provides a basis to label different partner types in each value network: *structural partners* provide essential resources; *contributing partners* add network-specific resources; and *supporting partners* contribute generic resources to the process of value creation. This allows us to identify the nature of resource-based interdependencies in a given network and the strategic position of actors within the network. At this level of analysis, it will be possible to carry out a cross-case comparison of power structures in different value networks and identify similarities and differences in terms of the types of industrial players that assume positions of greater or lesser importance.

Table 2. Partner types and resource contributions in a value network

Partner type	Resource contribution to value network		Actor
structural	**essential**	*tangible*	
		intangible	
contributing	**network-specific**	*tangible*	
		intangible	
supporting	**generic**	*tangible*	
		intangible	

CASE STUDIES: INTERDEPENDENCIES IN VALUE NETWORKS FOR MOBILE INFORMATION AND ENTERTAINMENT SERVICES

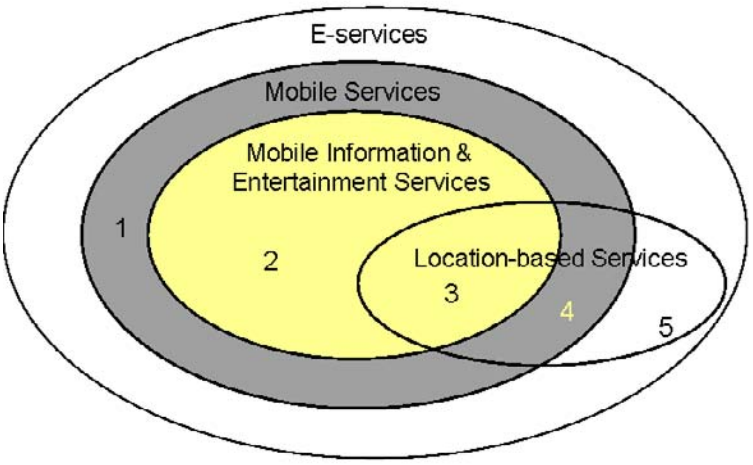
This section presents the analysis of the case studies of five mobile information and entertainment services. First, we set out the scope of the empirical research with a brief introduction and definition of mobile information and entertainment services, followed by an outline of the methods used and an overview of the five services that were selected as case studies. Then, each service is introduced and analyzed in turn. The findings of the cases are collated and discussed in the cross-case analysis.

Mobile Information and Entertainment Services

The e-services discussed in this research are limited to mobile information and entertainment services. As depicted in Figure 1, *mobile information* and *entertainment services* (category 2) are a subset of the broader category of *mobile services* (category 1), which are simply services made available to mobile users independent of the type of network (i.e., GPS, public switched mobile network, etc.). As defined here, mobile information and entertainment services require a connection to a network, which, in turn, is connected to the Internet. Currently, the dominant mode of access is through the mobile telecommunications network infrastructure connected to the fixed public switched network.

We define *mobile information and entertainment services* as the delivery of information and entertainment from specially formatted content sources (e.g., Internet sites, SMS, MMS) via the mobile telecommunication network to a mobile user. The terms *value added services in mobile commerce* and *mobile information and entertainment services* often are used synonymously. What is important is that parties other than the network

Figure 1. Mobile information and entertainment services domain



operator are involved to make the service available to customers.

In this research, we also consider information and entertainment services that are based on location information. The use of location information has the potential to enable a whole range of new services and requires the involvement of a new kind of actors such as geographic information system (GIS) suppliers. In general, location based services can be offered through the mobile telecommunications network (category 3), independent of this network (4), and also in a fixed environment (5). Of interest to this research are services offered in the domain of category 2 (mobile information and entertainment services) and category 3 (location-based mobile services) offered over the mobile telecommunications network.

Method

Within the context of mobile information and entertainment services, five services were selected as case studies. The services were offered to end-users in three different European countries: the Netherlands, Germany, and Sweden. To understand the service network composition and the dynamic among actors in terms of interdependencies and resources, interviews were held during the summer and fall of 2002. For each service, depending on the network size, interviews were held at two to five firms. At each firm, the interviewees were typically managers in charge of the relationship with the external partners associated with the particular service. Data from interviews also were supplemented with information gathered from company Web sites, through industry reports, and, in some cases, through other academic literature. The scope of the five service networks was defined by their relationship to end-customers. Services for which end-users were charged were chosen. In the appendix, we first present a table with an overview of the cases. Then, a completed matrix (based on Table 2) is shown for each of the five services in Tables 5 through 9.

Case Studies

In this section, each service is introduced and analyzed in turn. The findings of the cases are collated and discussed in the cross-case analysis. First, the two non-location-based services are discussed.

My Babes

My Babes is a Dutch i-mode service that allows a customer unlimited access to a variety of genres of erotic content for a monthly subscription. Customers can view photos in different categories (topless, bikini, etc.), access games (Stripjack and HotOrNot), and store their favorite photos in a photo album for easy reference. The actors are KPN Mobile, iMedia, and Internet-based raw content suppliers. iMedia is a media firm that purchases content through market-based transactions with Internet firms and then modifies the content to meet the standards for the i-mode service.

Analysis of the My Babes Case:

The i-mode cases have many of the resource contributions in common. The network operator contributes a large number of the essential resources (see Table 5). The operator takes care of the network, platform, billing, marketing, and partner network concept, and is involved with customer support. This is all part of the i-mode concept. Another element of the i-mode concept is that the operator controls a procedure to enforce content quality management. Thus, the resource contribution of the content suppliers is influenced through this relationship. In the My Babes case, iMedia acts as an intermediary that developed the specific service concept. In the i-mode model, intermediaries propose services to the i-mode staff and are approved or rejected. One of the items

they are judged on is feasibility, and, hence, they must have their downstream partners identified. Once the service is accepted, the intermediary is responsible for developing the content to the specifications of the operator, which may change according to demand and feedback from consumers. This ongoing content development, which is a network specific resource, places iMedia in the contributing partner role. In this role, the firm must contend with an abundant supply of pornographic material (raw content) and provide the value-added service of matching the raw content to the tastes of the i-mode target market as well as editing the content so that it can be considered erotic rather than pornographic. The intermediary also handles customer support problems. While the operator deals with problems related to the data access service, iMedia is responsible for any problems specifically related to the service. In this capacity, iMedia functions as a supporting partner. Also appearing as a supporting partner are the raw content suppliers, which are left to supplying the generic resources. In the case of My Babes, the generic content is a class of pornographic images that can be easily transformed into erotic content that does not contain depictions of sexual acts. This content tends to be of higher quality (settings, models, etc.), and, therefore, the intermediary requires a raw content supplier that can discriminate pornographic content quality. Despite this caveat, this function is in demand across a number of industries and, hence, is considered a generic resource. Thus, this case presents a clear picture of a dominant network operator and an intermediary that appears as a structural partner but also plays contributing and supporting roles. The raw content suppliers are supporting and can be easily replaced, as occurred after only four months of operation.

Radio538[1]

The Radio 538 ringtune case is also a Dutch i-mode service. A monthly subscription to the Radio 538 ringtunes service allows customers to download five ringtones from a variety of categories: music, voices, and sounds. Radio 538 branded their service as *ringtunes* to distinguish the service from other ringtone services on the KPN Mobile i-mode portal. The i-mode handset, manufactured by NEC, allows customers to store a total of 13 polyphonic (16-chord) ringtones. Radio538 is a Dutch media firm that owns and operates a popular radio station. The ringtones are developed by several means that include the participation of the Radio538 DJs, Tutch, and Jingle Hell, which turns popular music into ringtones. Permission to use the popular songs for ringtones is obtained through a copyright clearinghouse, BumaStemra, and the software that makes the ringtones available via the i-mode service is provided by Faith.

Analysis of the Radio538 Case:

Like the My Babes case, the network operator, KPN Mobile, contributes a large number of resources. However, there are quite a number of differences with the former case with respect to the other actors in this value network. In the Radio538 case, the partner type identification is not so clear (see Table 6). Many actors are involved in content development with each actor offering its very own specific contribution. There is a strong intermediary, Tutch, responsible for the service conception and design and the ringtune application provision. Tutch, however, is invisible in the market, since the well-known media firm radio station Radio 538 provides the branding. KPN Mobile wanted to pursue a relationship with Radio538 because Radio538 has access to the targeted customers. Initially, the ringtones created by Radio538 were expected to match the taste of the target market; subsequently, the diversity provided by the DJs' ringtones added to their popularity. However, there were other motives for pursuing a relationship with Radio538, given Radio538's national radio coverage and, hence, national brand recognition in the i-mode

target market. The media company Radio 538 also develops content, since it decides which songs are considered to be hits and the source for ringtunes, facilitates DJs to record remarkable quotes for ringtunes, and produces sounds for ringtunes. The DJs and the music makers are the providers of generic content. The firm Jingle Hell turns the music, voices, and sounds into ringtunes and, therefore, develop the content further. Customers are supported by the network operator as well as by the intermediary Tutch, depending on the kind of question they have. Often, customers first address the media firm Radio 538, and subsequently, they are seamlessly transferred to Tutch. Because of the nature of the service, other specialized resources also are involved, such as software for ringtones and a copyright clearinghouse. In this case, the handset provider is explicitly mentioned as a contributing partner, since not all handsets have the capability of storing the 13 polyphonic ringtones. The handset provider was involved in the process of service development.

This case presents a picture of a core triangle existing of a dominant network operator, a strong intermediary that appears as a structural partner but also plays contributing and supporting roles, and a media firm classified as a structural partner for providing the brand. The raw content suppliers are supporting partners and are only indirectly involved by producing hit songs, voices, and sounds, which all can be easily replaced. The structural partners, the network operator, and the intermediary base their positions on both tangible and intangible resources. The contributions of the media company and the content developer consist mainly of intangible resources. All but one supporting partners contribute tangible resources such as raw content, software, and chips.

Case 3: Finder

Finder is a location-based i-mode service offered by E-Plus in Germany. The service enables the consumer to find the nearest hotel, restaurant, taxi, or ATM. The content and geographical information are updated on a regular basis and stored in databases on the application platform. When a customer sends a request for information to the application platform, its position information is combined with the content and geographical information, and the customer receives the desired information. Actors include the operator E-Plus and Webraska, a worldwide provider of location-based services and telematics software solutions. Webraska also serves as the intermediary between E-Plus and the content providers, together with the geographical information provider. Cell Point provides the positioning equipment.

Analysis of the Finder Case:

In this third i-mode case, the analysis of the resource contributions shows again the same list of resources for the operator (see Table 7). Since this is a location-based service, the operator, E-Plus, also contributes the user positioning. Once more, the operator is the structural partner in the network that cannot be replaced. The intermediary, Webraska, develops the location-based application and controls essential and network-specific resources as well as generic resources, thus appearing as a structural, contributing, and supporting partner. Webraska is a specialist in location-based applications and is involved in the primary process of real-time geocoding the requested information. Webraska aggregates the content provided by five other content developers and the GIS provider. The five content developers have specific knowledge in a content domain such as business information, restaurants, fastfood, taxi, and financial information. They receive their content from raw content providers. Webraska needs to update geographical information to create a LBS, and the GIS provider is a contributing partner that provides this tangible asset to Webraska.. Webraska chose Navtech because they are a major player in digital maps in the US and Europe.

The network operator E-Plus maintains contact with the customer. The customer is probably not even aware of the intermediary Webraska. If customers find an error in the information, they send this information to E-Plus, and E-Plus forwards it to Webraska.

The handset providers are labeled as contributing partners, since the design of the handset is part of the i-mode concept. Especially for a location-based service such as Finder, it is important that the screen is suitable to display maps. Another hardware provider is Cell Point, which provides positioning equipment, a generic resource required for the service but is easily replaceable.

To conclude, the i-mode cases show a similar pattern of a dominant network operator, an intermediary that appears as a structural partner, but which also plays contributing and supporting roles, and one or two other structural or contributing partners. The raw content suppliers are only supporting, and all can be easily replaced. The structural partners provide both tangible and intangible resources. With a few exceptions, the supporting partners provide mainly tangible resources.

Case 4: LBS Directory[2]

The LBS directory service is a location-based service offered via WAP and SMS. It offers directory-type location information for ATMs, taxis, cinema, hotels, restaurants, events, emergency pharmacies, and fast food. The service is produced in two steps. First, the content is aggregated, aligned technically (in terms of file formats), geocoded, checked for quality assurance, and then transferred at regular intervals to the operator in an ongoing process. The second step consists of the actual provision of the service (i.e., receiving a service request from the user, positioning the user, matching the request with appropriate content, and passing the response with routing information back to the user). These two levels of implementation are a result of the service design and implementation that was driven by the operator. Actors consist of the operator, an intermediary, and a group of content providers specifically chosen to provide predetermined content categories.

Analysis of the LBS Directory Case:

The analysis of the resource contributions in this case indicates a noticeably large number of essential resources all being contributed by one actor, the network operator (see Table 8). This is due to its intention to learn as much as possible about the different aspects of providing a location-based service. The operator is the structural partner in the network that cannot be replaced without the service ceasing to exist. The operator conceived the service, designed the network in terms of partners and roles, and, as service provider of the LBS directory, provides the branding of the service. The LBS directory is marketed as a service of this operator so that the identity of the other actors in the network is almost entirely hidden from the customers (information about their involvement in the service is available on the Internet). In essence, the operator carries out all the activities that imply some form of customer contact (i.e., billing, marketing, customer support, and service provision). Other essential contributions of the operator are the provision of the network on which the service runs and the user positioning.

The financial resources contributed by the operator to run the service stem largely from the man months invested in developing the service in-house in terms of product management and application development. Nevertheless, the extent to which these investments would be lost if the operator were to withdraw from the value network is very limited, because it engaged in these in-house activities specifically to learn how to run a location-based service, a goal already achieved. The knowledge gained will not be lost, unless most of the staff in whom it is embedded was

to leave the company immediately. Moreover, the investments in these in-house development activities also were considered to be necessary and most cost effective because, at that time, external developers were not conceived to have the required know-how about developing a location-based service. However, the operator also was aware that the knowledge contributions and gains made in this project would not have been in vain *only* if the know-how is applied again (i.e., to offer other location based services).

The intermediary controls several network-specific and generic resources, appearing as both a contributing and a supporting partner. This position is due to the intermediary's explicit strategy to provide generic resources (i.e., finalized content and content development) that are typically supplied to the value network by the supporting partners, but for which no adequate content provider could be identified in the market. Thus, in the long run, the intermediary could be able to replace at least some of the content providers and add to its own importance in the network.

As a contributing partner, it has established the service-specific portfolio of content partners for the network, and it constitutes the single point of contact for both the network operator and the content providers. Content quality management implies a range of checks and procedures to align and standardize the content from different sources such as completeness of required fields, spell checking, and address correction. Regardless of existing geocodes in the content databases, all content is geocoded according to one standard (i.e., the points/events of interests are enhanced with the X/Y coordinates of their actual geographical location). Finally, the intermediary also provides technical customer support for queries about the LBS directory. These queries are passed on to the intermediary by the operator and the solution, or response from the intermediary is passed to the customer via the operator. Other resources contributed to the LBS directory are related to its technical competence; for example, working with the database formats preferred by the operator and aligning content from a diverse range of content providers that are able to submit their input in whatever format suits them.

Essentially, the content providers are supporting partners in this value network, left to supply generic resources (i.e., content in specific categories). Each of them develops and then provides one type of content (in which they are typically market leader) to the content aggregator (the intermediary). The content they produce can be used and sold in a range of projects, so no technical adjustments are required to their content that may imply extra costs.

In summary, this presents a clear picture of the relative position of the different actors in this network, showing a dominant network operator as the sole structural partner, the intermediary as both, a contributing and a supporting partner, and the content providers as supporting partners. All partner types in this network base their position on both tangible and intangible resources.

Case 5: Botfighter

Botfighter is the world's first location-based mobile game that uses mobile positioning information from an operator's network and is played using a standard GSM phone with SMS capabilities. On a Web site, the player designs a robot, which will be used to carry out a mission. The mission, which is obtained through the phone or Web site, involves another player, either a friend or one that is randomly assigned. Information concerning the location of the opponent is provided through the robot's radar system (the mobile handset). Botfighter's service network includes both companies and end-users, who provide content via the game's Web site. The service was conceived by It's Alive!, which maintains the game and organizes the Web site and the geographical information. The game, along with other Telia content, is hosted on a platform by Mobilaris. Ericsson provides the positioning equipment.

Analysis of the Botfighter Case:

Overall, a mixed picture of the relative position of different actors in this value network arises (see Table 9). Essential resource contributions are made by a number of different actors, who seem to form a core of structural partners to produce the service—Telia, It's Alive!, Ericsson, and Mobilaris. Telia, the mobile operator, provides the infrastructure, marketing, and branding, as well as customer support. It integrates the various technologies that are necessary to offer the service to the end-customer[3] and also contributes the billing relation. While the service idea was proposed by It's Alive!, Telia conceived the value network design and allocates revenue streams to the other actors within the network. At the start of the project, the cooperation between the two companies was intensive, addressing technical issues, graphical interface for the Web presence, and integration aspects.

It's Alive! appears at all levels, contributing essential, network-specific, and generic resources. Its essential contributions consist of the game service conception and design and the application provision. It also maintains the Web site and the application, which are network-specific resource contributions. Furthermore, the finalized content (e.g., missions) provided by It's Alive! is a supporting resource. It's Alive! has its own GI provider (Cartesia) and integrates the GI data into the Botfighter Web site. While It's Alive! did not use the GIS server from Telia at the time of the empirical research, discussions were planned whether It's Alive! may use Telia's GIS server. For Telia, this would mean consistency in terms of recognizable maps across its service offerings. The function of Its'Alive!' as an intermediary between network operator and other partners (typically content providers) is less apparent than in the other cases. There is only one formal content provider (Cartesia) aside from the end-users who can act as informal content providers via the Web site.

While the botfighter application could be integrated fully into the network of Telia, Telia decided to run the application on a platform that can be used as middleware. This service management platform, bridging end-user services, and the complexity of the mobile network infrastructure, are provided by Mobilaris. Mobilaris only has a relationship with Telia to provide the platform. The lack of a formal relationship between Mobilaris and It's Alive!, despite the fact that the botfighter application needs to be programmed according to the API (Application Protocol Interface) of Mobilaris, is striking and suggests that Telia wants to exert control over the interaction of its value network partners.

Ericsson provides the positioning technology to offer a location-based service. This Mobile Positioning System (MPS) enables the whereabouts of mobile phones to be made known to providers of location-based services. It is the outcome of a joint venture between Telia and Ericsson called Team Positioning that was established with the aim of providing Telia with the best possible system for services based on GSM-based positioning. As such, it constitutes an essential resource contribution to the botfighter value network. By participating in this value network, Ericsson is able to learn from a network operator and its end-customer requirements in order to enhance the MPS. The obtained know-how provides insight and arguments for them when selling their product to other network operators. The generic resource contributions by Cartesia (i.e., the geographical data that is built into the botfighter Web site) and Genuity (i.e., hosting the botfighter Web site) are easily substituted.

An unusual actor in this network is the user of the game, whose involvement via the Web site (i.e., to design the robot) means that he supplies raw content. Hence, the end-users are included in the network as content providers.

Summary of Findings: Cross-Case Analysis

To summarize the findings from the individual cases and to assist with the cross-case analysis, Table 3 provides an overview of the partners in the different cases. In all cases, a network operator and an intermediary can be distinguished, who contribute a variety of resources, although the intermediary function is less apparent in the botfighter case. In addition, a content supplier, as a raw content supplier or as a supplier of adapted content, is present in all cases. Besides those three kinds of partners, different case-specific partners appear. The end-user is explicitly referred to only in the botfighter case, where the end-user takes on a more active role as an informal content provider by contributing his or her own resource (input to design the robot) to the value network than end-users in the other four cases presented in this chapter.

As far as network operators are concerned, they appear as structural partners in all five cases. Natsuno (2003) argued that the decisive difference between Japan, the US, and Europe is that neither of the latter two had a telecommunications provider like DoCoMo with the will to grow a new business and service based on a comprehensive view of the ecosystem. It seems that Europe started to follow the example of Japan.

In four out of our five cases, the intermediaries appear as structural partners. The crucial resource that allows them to claim such a strong position in their respective network is service conception and design. The only case where the intermediary does not control this resource is the LBS directory, where the network operator initiated the service and kept hold of the service design.

The position of the content providers in our value networks for mobile information and entertainment services is, perhaps surprisingly, of lesser importance. These services are designed to deliver content, and, therefore, content could have been expected to show up as an essential resource. However, content providers appear never as structural partners that can easily assert their place in the network. In our cases, content providers are either contributing or supporting partners that can be replaced fairly easily in their respective networks.

The same resource (e.g., customer support) can vary in importance in the different networks. The implication is that the possession of a particular resource, with the exception of the possession of the network, does not necessarily propel the actor into a specific position within the network. It is the composition of resources that is important. Moreover, with respect to the distinction between tangible and intangible resources, we observe that in several cases (i.e., Botfighter, Finder, Radio538), the supporting partners are limited to providing tangible resources, whereas the contributing and structural partners in all cases are, with few exceptions, providing tangible and intangible resources. While intangible resources are particularly immobile and difficult to imitate or substitute, these partners base their position in the value network on the combination of both tangible and intangible resource contributions.

Finally, customer support is split into two. In the i-mode cases, this depends on the kind of customer query, whether it is network-/i-mode-related or service-specific. In the case of the LBS directory, it depends on the facility that is required, either the customer support facility or technical support. In this case, the close contact with the customer (the interface facility) constitutes an essential resource, whereas the skill to carry out the technical support is a contributing resource.

CONCLUSION

Content is King was the adagio when convergence among telecommunication and media industries seemed inevitable not so long ago. However, analysis of the case studies in this chapter shows that this period is gone. Mobile e-services are

Table 3. Cross case findings

	My Babes i-mode case	**Radio538** i-mode case	**Finder lbs** i-mode case	**Botfigther** lbs game	**LBS directory**
Structural partner(s) [essential resources]	**Network Operator** *Platform Provisionr Billing Marketing Content Quality Partner network concept Customer Support* **Intermediary** *Service Conception & Design*	**Network Operator** *Platform Provision Billing Marketing Content Quality Partner network concept Customer Support* **Intermediary** *Service Conception & Design Application Provision* **Media** *Branding*	**Network Operator** *Platform Provider Billing Marketing Content Quality Partner network concept Customer Support* **Intermediary** *Service Conception & Design Application Provision Real time geocoding*	**Network Operator** *Billing Partner network concept* **Intermediary** *Service Conception & Design Application Provision* **Platform Provider** **Positioning & equipment Vendor**	**Network Operator** *Platform Provision Billing Marketing Customer Support GIS application User positioning Partner network concept Service Conception & Design Branding*
Contributing partner (s) [network-specific resources]	**Intermediary** *Finalised content Content development*	**Intermediary** *Finalised content* **Media** *Content development* **Content developer** **Handset provider**	**Intermediary** *Content aggregation* **Content developer** *Finalised content Content development* **GIS Provider** **Handset provider**	**Intermediary** *Application maintenance Website maintenance*	**Intermediary** *Batch geo-coding Content Quality Portfolio management Technical support Single point of contact Content aggregation*
Supporting partner(s) [generic resources]	**Intermediary** *Customer Support* **Raw content suppliers**	**Intermediary** *Customer Support* **Raw content suppliers** **Hardware and Software suppliers** **Legal right clearing**	**Intermediary** *Customer Support* **Raw content suppliers** **Positioning Equipment Vendor**	**Intermediary** *Finalised content* ***Content Quality*** **GI Provider** **Web hoster** **End-user** ***Informal content***	**Intermediary** *Finalised content Content Quality* **Content suppliers** *Finalised content Content Quality*

offered jointly by a network operator, an intermediary, and content providers. The network operator is the structural partner that controls the value network. The intermediary plays an important role in coordinating the production of specific services on the network, with resources like application provision, geocoding, and content aggregation. The service concept further enhances their position in their network. In everyday life, the intermediary is referred to as service or application provider. And the content provider? The activities related to content show a dispersion of effort in which content supply alone can be divided among many actors. But more essential ones, such as content development, are contributed by other actors, typically the intermediaries. We conclude that the content providers are not even princes or princesses; content provision alone does not propel them into a strong position in the value network.

While it may appear obvious that the network operators have a strong position in the value network, this is due largely to their external facing orientation. Even in the case where the intermediary would act as the dominant structural partner, it may still appear that the operator is in this position. Thus, the theoretical framework constructed in this chapter provides a basis for a detailed analysis, and, as such, it presents a valuable tool that extends the view beyond mere network operator dominance to confirm and categorize the resource-based status of all actors involved in such value networks, based on their contributions to the service. Furthermore, this framework, by focusing on resource contributions, provides a mechanism for understanding the role of end-users in such networks. Moreover, this framework serves as a useful tool for comparing and contrasting different value networks for mobile e-services in terms of their resource-based configuration and dynamics among actors.

We also note that the findings of the analysis presented in this chapter are limited to the current market situation, and that, in the future, the positions of the various partners may change. Therefore, analyses such as the one presented here should be performed from time to time in order to assess the extent to which any changes have occurred.

ACKNOWLEDGMENTS

The research reported in this chapter is part of the BITA (Business Models for Innovative Telematics Applications) project. In the BITA project, the Telematica Instituut cooperated with Delft University of Technology and TNO-STB. We acknowledge the efforts of our colleagues who worked on this part of the project, including Sander Hille and Edward Faber of the Telematica Instituut; Richard Hawkins, Pieter Ballon, and Wouter Hoff of TNO-STB; and Harry Bouwman of Delft University of Technology.

REFERENCES

Appelman, J.H. (2004). Governance of global interorganizational tourism networks. Erasmus Research Institute.

Axelsson, B. (1987). Supplier management and technological development. In H. Håkansson (Ed.), *Industrial and technological development—A network approach* (pp. 128-176). London: Croom Helm.

Ballon, P., & Hawkins, R. (2003). From business models to value networks [*TNO/STB working paper].* Delft, NL: Schoemakerstraat 97, 2600JA.

Bovel, D., & Martha, J. (2000). From supply chain to value net. *Journal of Business Strategy, 20*(4), 24-29.

Das, T.K., & Teng, B.S. (1998). Resource and risk management in the strategic alliance making process. *Journal of Management, 24*(1), 21-42.

Das, T.K., & Teng, B.S. (2000). A resource-based theory of strategic alliances. *Journal of Management, 26*(1), 31-61.

Ebers, M. (1999). The dynamics of inter-organizational relationships. In S.B. Andrews, & D. Knoke (Eds.), *Network in and around organization* (pp. 31-56). Stamford, CT: Jai Press Inc.

Foss, N. (1998). The resource-based perspective: An assessment and diagnosis of problems. *Scandinavian Journal of Management, 14*(3) 133-149.

Gadde, L.E., Huemer, L., & Håkansson, H. (2003). Strategizing in industrial networks. *Industrial Marketing Management, 32*, 357-364.

Granovetter, M. (1985). Economic action and social structure: A theory of embeddedness. *American Journal of Sociology, 9*(3), 481-510.

Haanæs, K., & Fjeldstad, O. (2000). Linking intangible resources and competition. *European Management Journal, 18*(1), 52-62.

Håkansson, H. (Ed.). (1987). *Industrial and technological development—A network approach*. London: Croom Helm.

Håkansson, H., & Waluszewski, A. (2002). *Managing technological development: IKEA, the environment and technology*. London: Routledge.

Hoskisson, R., Hitt, M., Wan, W., & Yiu D. (1999). Theory and research in strategic management: Swings of a pendulum. *Journal of Management, 25*(3), 417-456.

Itami, H., & Roehl, T. (1987). *Mobilizing invisible assets*. Cambridge: Harvard University Press.

Li, F., & Whalley, J. (2002). Deconstruction of the telecommunications industry: From value chains to value networks. *Telecommunications Policy, 26*, 451-472.

Maitland, C., Bauer, J., & Westerveld, R. (2002). The European market for mobile data: Evolving value chains and industry structure. *Telecommunications Policy, 26*, 485-504.

Maitland, C., van de Kar, E.A.M., Wehn de Montalvo, U., & Bouwman, H. (2003a). Mobile information and entertainment services: Business models and service network. *Proceedings of the Second International Conference on Mobile Business*, Vienna, Austria.

Maitland, C., van de Kar, E.A.M., & Wehn de Montalvo, U. (2003b). Network formation for mobile information and entertainment services. *Proceedings of the 16th Bled Electronic Commerce Conference*, Bled, Slovenia.

Miller, D., & Shamsie, J. (1996). The resource-based view of the firm in two environments: The Hollywood film studios from 1936 to 1965. *Academy of Management Journal, 39*, 519-543.

Mintzberg, H. (1979). *The structuring of organizations: A synthesis of research*. London: Prentice-Hall.

Mintzberg, H. (1983). *Power in and around organizations*. London: Prentice Hall.

Nassimbeni, G. (1998). Network structures and co-ordination mechanisms—A taxonomy. *International Journal of Operations & Production Management, 18*(6), 538-554.

Natsuno, T. (2003). *The i-mode wireless ecosystem*. West Sussex, UK: John Wiley & Sons.

Oliver, C. (1997). Sustainable competitive advantage: Combining institutional and resource-based views. *Strategic Management Journal, 18*(9), 697-713.

Pfeffer, J., & Salancik, G.R. (1978). The external control of organizations—A resource dependence perspective. New York: Harper & Row.

Piccinelli, G., Di Vitantonio, G., & Mokrushion, L. (2001). Dynamic service aggregation in electronic marketplaces. *Computer Networks, 37*(2), 95-109.

Porter, M. (1991). Towards a dynamic theory of strategy. *Strategic Management Journal, 12*, 95-117.

Sabat, H. (2002). The evolving mobile wireless value chain and market structure. *Telecommunications Policy, 26*, 505-535.

van de Kar, E.A.M., Maitland, C., Wehn de Montalvo, U., & Bouwman, H. (2003). Design guidelines for mobile information and entertainment services, based on the Radio538 ringtunes i-mode service case study. *Proceedings of the Fifth International Conference on Electronic Commerce*, Pittsburgh, Pennsylvania.

Wernefelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal, 5*, 171-180.

ENDNOTES

1 This case is more extensively described in Kar van de et al. (2003).

2 Fictitious service name; case is anonymous.

3 The botfighter application as well as 10 other services that Telia offers are connected to the middleware platform for location-based services of Mobilaris called Pacific Ocean. The Pacific Ocean platform is connected to all the supporting systems of Telia, such as the billing system, customer support system, SMS and WAP gateways, and the GIS server.

APPENDIX

Table 4. Overview of MIES case studies

	General MIES		***Location-based MIES***		
Service	**i-mode MyBabes**	**i-mode Radio 538**	**i-mode Finder**	**LBS directory**	**Botfighter**
Content	Erotic pictures and games	Ringtones	Find-the-nearest	Find-the-nearest	Multi-actor Game
# interviews (# firms)	2 (2)	3 (3)	6 (4)	6 (4)	5 (5)
Country	The Netherlands	The Netherlands	Germany	confidential	Sweden
Network	GPRS	GPRS	GPRS	GSM and GPRS	GSM and GPRS
Device	i-Mode handset (NEC)	i-Mode handset (NEC)	i-Mode handset (NEC)	Any mobile Phone	Any mobile Phone Website
Interface	cHTML	cHTML	cHTML	WAP and SMS	SMS

Table 5. MyBabes overview of partner types and contributions to the network

Partner type	**Resource contribution to value network**		**KPN Mobile**	**iMedia**	**Internet sites**
Structural	essential	tangible	Network provision, Platform provision, Billing provision,		
		intangible	Marketing, i-mode concept, Quality management, Customer support	Service conception & design	
Contributing	network-specific	tangible		Finalised content	
		intangible		Content development	
Supporting	generic	tangible			Raw content
		intangible		Customer support	

Table 6. Radio538 overview of partner types and contributions to the network

Partner type	Resource contribution to value network		KPN Mobile	Tutch	Radio538	Jingle Hell	DJ's/ music-makers	Faith	NEC (Toshiba)	Yamaha	BUMA Stemra
Structural	essential	*tangible*	Network provision, Platform provision, Billing provision	Application provision							
		intangible	Marketing, i-mode concept, Quality manage-ment, Customer support	Service conception & design	Brand-ing						
Contributing	network-specific	*tangible*		Finalised content					Hand-sets		
		intangible			Content develop-ment	Content develop-ment					
Supporting	generic	*tangible*					Raw content	Ring-tune software		Chips for handset for ring-tunes	
		intangible		Customer support							Legal right issues

Table 7. Finder overview of partner types and contributions to the network

Partner type	Resource contribution to value network		E-Plus	Webraska	Schober, varta, foot-food, taxi, fovium	Navtech	NEC, Toshiba	Cell Point	Schober intern. and other raw content suppliers
Structural	essential	*tangible*	Network provision, Platform provision, Billing provision, User positioning	Application provision, Content geocoding real time					
		intangible	Marketing, i-mode concept, Quality management Customer support	Service conception & design					
Contri-buting	network-specific	*tangible*		Content provider/ aggregator	Finalised content	Geographic information providing	Handsets		
		intangible			Content development				
Supporting	generic	*tangible*						Positioning equipment	Raw content
		intangible		Customer support					

Table 8. LBS directory resource contributions to the network per actor

Partner type	Resource contribution to value network	Company X	Company Y	Co. A	Co. B	Co. C	Co. D	Co. E	Co. F
Structural	essential	*tangible* Network, User position information, Billing provision, Marketing, Customer support facility, GIS application, LBS directory application Middleware platform Financial resources for manpower & knowledge development							
		intangible Service conception & design, Value network design, Branding							
Contributing	network-specific	*tangible*	Portfolio of content partners						
		intangible	Geo-coding of content in batch process, Quality management, Technical support Technical competence, Single point of contact for NO & CPs						
Supporting	generic	*tangible*	Finalised content	Fina-lised content	Fina-lised content	Fina-lised content	Fina-lised content	Fina-lised content	Fina-lised content
		intangible	Content develop-ment	Content develop-ment	Content develop-ment	Content develop-ment	Content develop-ment	Content develop-ment	Content develop-ment

Table 9. Botfighter partner types and contributions to the network

Partner type	Resource contribution to value network	Telia	It's Alive!	Mobilaris	Ericsson	Cartesia	Genuity	End-user
Structural	essential / *tangible*	Network provision, User positioning, Billing provision	Application provision	Platform provider	Positioning vendor, Equipment provider			
	essential / *intangible*	Value network design, Branding, Marketing, Customer support	Service conception & design					
Contributing	network-specific / *tangible*		Application maintenance Website maintenance					
	network-specific / *intangible*							
Supporting	generic / *tangible*		Finalised Content			GI provider	Web hosting	Raw content
	generic / *intangible*							

This work was previously published in the International Journal of E-Business Research, Volume 1, Issue 3, pp. 1-20, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 6.19
Channel Choices and Revenue Logics of Software Companies Developing Mobile Games

Risto Rajala
Helsinki School of Economics, Finland

Matti Rossi
Helsinki School of Economics, Finland

Virpi Kristiina Tuunainen
Helsinki School of Economics, Finland

Janne Vihinen
Helsinki School of Economics, Finland

ABSTRACT

In this chapter, we explore the revenue logics and related product distribution models of mobile game developer companies. Mobile gaming is facing a transformation in both technical infrastructures and business models as it grows at a very fast pace. The former change originates from the technological shift of the environment of use; for example, from specific game consoles toward mobile phone platforms. The latter change relates to the possibility of delivering and playing games online, which affects both the distribution partnerships and the revenue stream options of mobile game vendors. We present a set of possible business models for game developers and concentrate on the possible combinations of revenue logics and distribution models for different games.

INTRODUCTION

The worldwide number of digital phone (GSM and PCS) subscribers has increased from 140 million in 1996 to approximately 900 million at the end of 2002 (GSMdata, 2002). Concurrently, the number

of PC users is reaching a saturation point at around 400 million. From 2002 to 2004, the difference between phones and computers has continued to diminish with the arrival of Java-enabled phones and with a larger number of phones that support Web browsing and e-mail applications.

As mobile phones are rapidly turning into software platforms capable of supporting gaming, many handset manufacturers, operators, and game developers see the opportunity for mobile games. However, the recent downturn of investments into the enhanced cellular networks makes it challenging for companies to develop and deploy new advanced games. Furthermore, many aspects of the new business models, including revenue logics and distribution models for these new entertainment services, are still unproven. The mobile game market is expected to grow from $124 million in 2001 to exceed $4 billion in 2006 (Ovum, 2002). Today, most of the mobile gaming activity is in Asia-Pacific, particularly in Japan and South Korea, where there are tens of millions of subscribers of mobile entertainment services. However, we can expect that Europe and the US will soon see growth in these areas, as well.

Mobile games can be played with mobile phones; PDAs (Personal Digital Assistants), such as Palm or iPaq; Web-enabled phones; or other handheld game devices. In Europe, the development of mobile services has been characterized largely by technology push (Nurmi et al., 2001), but the future success of mobile services will strongly be affected by the ability of businesses to offer, already at an early stage, the right products and services to consumers (Anckar & D'Oncau, 2002). Experiences with PC-based Internet and Japanese mobile iMode services emphasize the role of entertainment services as a significant factor in the growth of mobile network usage. Games and entertainment services are important application areas for information industry as a whole (Shapiro & Varian, 1999), and, as the third generation mobile phone networks proliferate, demand for these services will increase rapidly.

In this chapter, we look at the mobile game scene and introduce a framework for analyzing software business models within it. We then develop the model further for mobile games and use it to discuss the revenue logics of mobile game developers. In the last section of this chapter, we summarize and draw conclusions on the discussed aspects of mobile games.

TYPES OF MOBILE GAMES

Generally, the existing games for mobile handsets are either server-based or stand-alone games. Server-based mobile games can be divided further into WAP, SMS, and Java games. Java games also can be used as stand-alone games. All of these games can be either single- or multi-player games. Multi-platform games, in turn, are a subset of games that can be played in conjunction with online, PC, and console versions.

A report of Durlacher Research (2001) suggests that mobile games can be classified by their operating and distribution platform into three types: stand-alone, server-based, and streamed. These games can be either downloadable from a server or preinstalled by a vendor or distribution partner.

- **Stand-alone games** do not require a network connection in order to play the game. As they run on mobile terminal, the user does not have to pay for data transmission after downloading the game. The games are restricted by the storage and operating capacity of mobile devices.

  An Example of Stand-Alone Games: Nokia Snake

  Snake was the first stand-alone game that was preinstalled in Nokia's mobile handsets in 1998. Nokia owns the intellectual property rights for the application and has developed

it in-house. Therefore, Nokia can install the game for free in any Nokia handset.

The idea of the original Snake was to catch more and more points with the snake steered by the player, making the snake longer and longer. At first, the player chooses the game level, which defines the speed of the snake. Finally, when the snake hits the wall or its own body, the game ends. The second version of Snak*e* was similar to the first one, but the game field had more complex shape instead of a simple box, also containing extra figures that may give extra points to the user. Both versions can be played by two simultaneous users through an infrared connection. A number of active players of the game have formed a group competing with each other in the game. The Snake community is not as well known as the ones that the most famous console

- **Server-based games** usually require connection to the service provider's server while the game is played. The server contains the information of the game's current status.

  An Example of Server-based Games: Who Wants to Be a Millionaire?™ by Codetoys

  Codetoys develops mobile entertainment services for mobile operators, mobile portals, and other service providers. Codetoys' games are based on internationally recognized brands. The supported technical platforms are SMS, WAP, and iMode. One of its main products is an interactive mobile game based on the popular television show, Who Wants to Be a Millionaire?™. In addition to delivering the actual mobile game services, Codetoys provides the game platform, content, user statistics, and advice for marketing the game for mobile markets. The game can be played on all types of digital mobile phones.

  Who Wants To Be A Millionaire?™ is a server-based game that is connected to the service provider's server while the game is played. The server contains information on the game's current status. The game is designed to follow the original television show concept as closely as possible. The mobile version is a multiple-choice game with a minimum of 2,000 questions. The game has 15 questions with four answers each, three lifelines, and the chance to walk away exactly as in the TV show. The object of the game is to answer 15 subsequent questions right and earn a million points. All players have a chance to make it to the hall of fame.

- **Streamed games** use advanced video decoding systems for delivering audio and visual data from servers to terminals. Streamed games require a certain minimum bandwidth for data transfer, but they will provide more advanced graphics and audio for the games on terminals that do not have the processing power for rendering demanding visual data.

  An Example of Streamed Games—MatchEm TV Chat

  Finland- and Hong Kong-based MatchEm Ltd. develops software systems called MiTV Tools that enable real-time interactivity between television and wireless handsets. The company's customers are TV and production companies that are willing to create interactive content into the programs delivered via analog TV, teletext, Digital TV, or the Web. MatchEm's software products are modular and can be adapted and customized for different purposes. The illustrated TV chat is one of the services based on MatchEm's iMatch product.

  The vision of MatchEm is that mobile messaging should never be a stand-alone channel

Figure 1. MatchEm TV chat registration and use

of communication, because the medium is too limited in its ability to deliver a robust, complete message. Rather, the medium should be used to extend the presence of a company or an event into an additional channel. Companies with a physical presence, a television, or even a Web site will be able to leverage mobile media to extend their presence to be anywhere the user is at any time.

One of the MatchEm's products is a mobile TV chat service (see figure 1), in which the end user can register a nickname and send anonymous text messages to a television show. After registration, the service confirms the availability of the name and allows sending messages to the live TV chat or to the other users. The registered users can define their profiles with user-specific information such as age and sex, which encourages the other users to contact each other. The chat service is branded as 4Date by Finnish local TV channel 4.

A CONSTRUCTIONAL VIEW OF BUSINESS MODELS

Rajala et al. (2001) have developed a conceptual software business model framework and practical tools for analyzing and comparing different business models in the software industry. We will use a subset of that framework to identify and describe the revenue logics of mobile games and related product characteristics and distribution models applied in the selected mobile game businesses. Our purpose is to identify alternative revenue logics that are technically possible, economically sustainable for various players, and that could be acceptable for the customers.

According to Rajala et al, (2001), the business model of a software vendor can be viewed as an action plan derived from strategic objectives of a company with a given product and service offering in a given market. Accordingly, a single business model deals with a single product/market situation. Consistent with the recent literature on business models (Amit & Zott 2001; Hedman & Kalling, 2003; McHugh, 1999; Morris, et al., 2004), Rajala et al. (2001, 2003) describe a business model as a

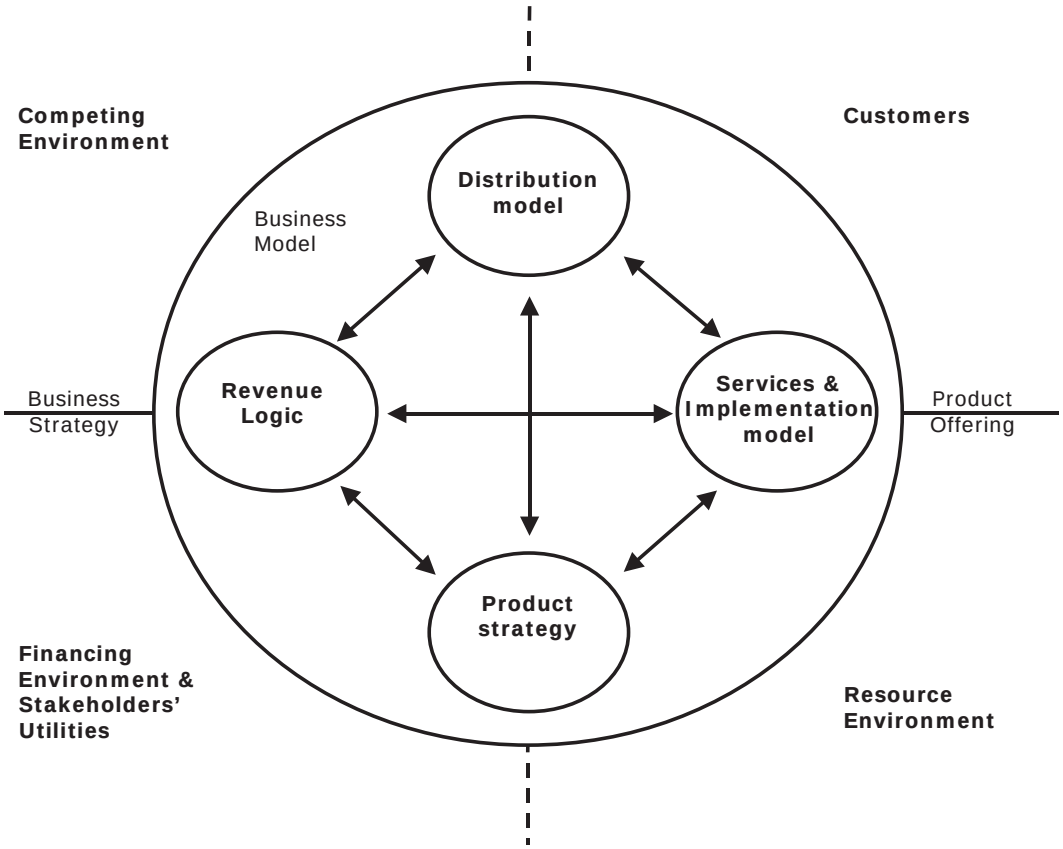
combination of different functional elements of product development, revenue, sales, marketing, services, and implementation. This model includes four basic elements that can further contain several options inside them. A construction of these elements is presented in Figure 2.

In Figure 2, the business model of a software vendor is depicted by four key elements:

- The *product strategy* describes what the core product offering is and how the development of the core product of a company is organized.
- The *revenue logic* describes how the company finances its operations; in other words, how and from what sources the revenue is generated.
- The *distribution model* describes how the marketing and sales have been organized, what are the characteristics of the channels of distribution, and who are the sellers and marketers of the product.
- The *services and implementation model* explains how the product offering is made available to the end users as a working solution.

A software company has multiple options to structure each of the elements in its business model. All of these elements are tightly intercon-

Figure 2. Elements of a business model (Rajala, Rossi et al., 2001)



nected with each other and cannot be analyzed in isolation. Therefore, even though our main focus here is on the revenue logic of a software company developing mobile games, we will look first at the product proposition and distribution model aspects of the business model.

Product Strategy for Mobile Games

The concepts of product strategies and product offerings are discussed widely in the literature of marketing (Cravens, 1987; Kotler et al., 1996). According to Cravens (1987), a product strategy consists of deciding how to position a product offering (e.g., specific product, line, or mix) to serve its target market, setting strategic objectives for the product offering, selecting a branding strategy, and developing and implementing strategies for new and existing products. Kotler, et al. (1996) describe a product offering on three levels. The *core product* is the essential benefit that the customer is really buying; the *actual product* includes the features, styling, quality, brand name, and packaging of the product offered for sale. Furthermore, they point out that the *augmented product* is the actual product plus the various services offered with it, such as warranty, installation, maintenance, and delivery. In the mobile game industry, these dimensions of product offerings are to be considered in both the business-to-business and business-to-consumer settings.

From the business model perspective, a defining characteristic of mobile game software as a product is that it is not a physical but an information product. Information, or digital, products have unique cost characteristics, differing largely from those of a physical product. A digital product is typically expensive to produce but very cheap to reproduce (Shapiro & Varian, 1999). In the mobile game industry, we can see that variable costs of single pieces of mobile game software are typically small, as there are no capacity constraints, and marginal costs are less than average cost. Thus, declining average costs create significant economies of scale for the producer. As the infrastructure and development tools of mobile game software evolve, the development costs of these products decrease. Simultaneously, the expected product life cycles of mobile games are sped up, and the barriers to market entry of new actors ease. In addition to having a direct effect on the game product strategies, these factors also affect the revenue logic of game software producers.

There are several dimensions in game offerings that can be used to analyze and compare different types of mobile games. First, analysis of product offerings can be made according to the intended usage scenario. This kind of analysis emphasizes the position and role of the offering in the value-creating network that produces and delivers the game offering for the end customers. According to this view, game software platforms and tools are outlined basically in different positions and roles in the industry-level value system, through which game components and final games are made available for end users. Secondly, the type and structure of the game product offering can be considered, for example, with the level of similarity of the product offering across multiple customers or customer groups and its potential distributed through different channels of distribution. This view emphasizes the potential to gain scale economies through serving a wide customer base with the same products. Third, the product development method, including various alternatives of in-house development vs. subcontracting, networking, and other forms of external development activities, can be used as a basis of industry-level classification. For example, the structure of the total offering may consist of one or more modules, including both product and service components. The structural aspect of a software product component includes the product architecture (i.e., component-based, single-core application, etc.) and the modularity in the sense of design and development. The modularity of

a product potentially affects the chances for its collaborative development, including different approaches to in-house and external development.

According to the business model framework of Rajala, et al. (2003), the generic product strategy options of software vendors can be divided into five main classes, as presented in Figure 3. This classification is based on the architecture of the product offering on the level that is thought to be useful in studying different revenue logics related to specific types of product offerings.

As seen in the Figure 3, the generic options for software product offerings range from customer-specific models, where customers' needs are met with tailor-made solutions, to standardized product-oriented models, including approaches for creating universal software products and standardized online services. Between these extreme alternatives, there may be, for instance, development of parameterized system products, uniform core products, or modular product families consisting of universal software components.

Server-based and streamed mobile games typically are affiliated with online services, while stand-alone games may range from uniform core products to modular product families. In the emerging market of mobile game software, we also can identify single component-based game products provided in collaboration with different partners. These components include, for example, graphics libraries or toolkits for game environments or other game components. Along with the development of the market for mobile game products, we can see an increasing diversity of game product offerings ranging from tool and platform offerings to final game solutions and product line offerings that consist of complementary games.

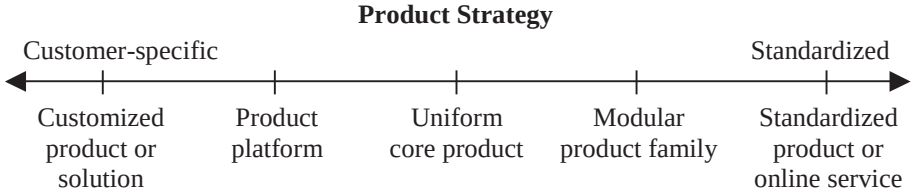
Distribution of Mobile Games

The development of the mobile games market has created new models for conducting business that affect all aspects of product distribution, marketing, and billing. A particularly important aspect of this new business paradigm is its impact on mobile distribution channels. Channel efficiency, channel marketing, and channel conflict are serious concerns for game developers and publishers alike.

We can identify two aspects that can serve as classification schemes in distinguishing among different distribution models of mobile game: the positioning and role of the game vendor in the mobile game value network and the complexity of the distribution system as defined by the length of the distribution channel. According to our view, the type of the distribution model and the length of the channel of distribution as one of its defining aspects strongly affect the available revenue logic options of mobile game producers.

Channels of distribution are divided in the literature, for example, into short and long channels (Lewis & Trevitt, 1996). In the mobile

Figure 3. Product strategy options (Rajala et al., 2003)



game business, short channels of distribution that typically consist of three or fewer than three stages of supply chain may include, for example, game developers, game publishers, and mobile operators. On the other hand, long channels of distribution may consist of more than three stages of the supply chain and typically include game developers, game publishers, aggregators, mobile portals, and mobile operators.

Short channels can offer companies possibilities to:

- Better control the sales of the product
- Monitor product sales relatively quickly and easily
- Assist dealers' operations with advertising and promotion material
- Offer discounts and other incentives to dealers and retailers

There also are several potential disadvantages with short channels. First, the large retailers and operators will be in a better position when bargaining with the developer. Moreover, the distribution and marketing costs might increase, because the developer may have to supply to several distribution channels (Bask, 1999).

Long channels of distribution are called traditional methods of distribution channels, which are common to a wide range of products. Lewis and Trevitt (1996) identify both advantages and disadvantages with long channels related to the following themes:

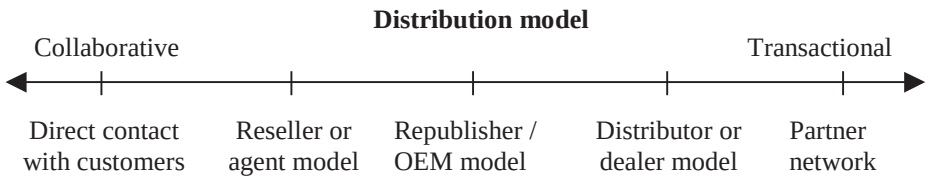
- Retailers can realize all the benefits of dealing with the wholesaler
- Consumer can buy goods individually or in small quantities
- Customers usually have to pay higher prices because small retailers buy goods in bulk

There are various possibilities to reach different types of customers. For example, a firm can sell the same products through different channels, different products and services through the same channel, or different products and services through different channels.

Firms can benefit from multi-channel distribution in a variety of ways. First, it can allow them to better adapt to changing customer needs and shopping patterns. Such adaptive capability has proven useful, for example, when attempting to respond to novel distribution channels such as the Internet. Second, companies with broad product lines can benefit from a multi-channel strategy, because it is unlikely that any single channel will be optimal for all products. Third, firms with excess manufacturing capacity can benefit from additional outlets when existing channels are saturated with supply (Webb & Didow, 1997).

Use of several distribution channels and complex distribution systems offer many poten-

Figure 4. Distribution model options (Rajala et al., 2003)



tial benefits to game suppliers but present some managerial challenges, as well. Multiple channels place competing demands on internal company resources such as capital, personnel, products, and technology. Moreover, the various distribution channels may compete with each other for the same customers in the marketplace, increasing the likelihood of intermediary dissatisfaction and customer confusion (Webb, 2001).

The framework of Rajala, et al. (2003) includes an aspect of the business model that deals with distributing and providing the offerings to customers. Here, the distribution model describes how the marketing and sales of the product and service offering has been organized and identifies the sellers and marketers of the product and service offering. The elementary ways of marketing a software product and service offering can been organized as illustrated in Figure 4.

In the mobile games industry, the distribution models typically include either pre-installation of the game into the mobile handset or downloadable games provided by a mobile network operator. These distribution models also are strongly tied to the pricing models. With the current GSM networks, only the smallest applications could be downloaded over the network. On the other hand, these are the only games where the network operator can act as the software distributor. Pre-installed games come as a free supplemental product. Markets are emerging for software sold separately in memory cards as commercial off-the-shelf (COTS) software.

Revenue Logic of Mobile Game Businesses

The revenue logic within a business model describes the way the software business generates revenue and profit. The different approaches to capture revenue range from different methods of pricing to different sources of revenue and different things sold. The revenue logic can include both sales revenues and other sources of financing. Here, we will focus on just the revenue element, assuming it includes the cost structure of both the offering and operation.

High initial cost and nearly zero marginal cost characterize the production and dissemination of information-intensive products (Mahadevan, 2000; Shapiro & Varian, 1999). In addition to the various revenue stream alternatives described previously, a software vendor, as in any other organization that sells electronically delivered products, has unique characteristics of the information economy to exploit. For instance, in the case of digital products, it is possible to use a range of pricing alternatives based on user segments and user-selectable options. Varian (1995) has argued that if the willingness to pay is correlated with some observable characteristics of the consumers, such as demographic profile, then it could be linked to the pricing strategy. One strategy is to bundle goods to sell to a market with heterogeneous willingness to pay (Mahadevan, 2000).

In the retail business of mobile games, a revenue model in which games are available on a subscription basis with monthly fees has met with success in Japan and for some online games services. Payment schemes allowing hardcore users to pay a flat fee for unlimited use and pay-per-use options for casual gamers make sense for maximizing volume. For the most part, only operators can employ this model at present, since they control billing for all end-user wireless services. In addition, strategies based on service subscription payment in addition to network connection charges may largely be problematic in increasing customer base.

Generic approaches to revenue logic in the software business are identified by Rajala et al. (2003) as follows:

- *Licensing*, which means license sales and royalties are the main source of revenue.
- *Revenue sharing* with distribution partners or profit sharing with users.

- *Loss-leader pricing*, which means giving something for less than its value. This is done, for example, in order to increase customer base for later revenue or to support sales of some other part of the product/service offering.
- *Media model*, where the revenue is based on advertisement sales either through advertisement in the user interfaces of software or by selling user information for advertisers.
- *Effort-, cost-, or value-based pricing* is a common approach in customized or tailor-made software solutions and made-to-order software projects.
- *Hybrid models* as various combinations of the previous points.

In the following paragraphs, we discuss selected approaches to potential revenue logics in the mobile game business.

Licensing

Licensing is the most common revenue model in the mobile industry (Durlacher Research Ltd., 2001). This revenue model is identified by Hecker (1999) as being a part of the standard software business model. It involves selling the customer the right to use the software. In licensing, there are many alternatives, including per-user, per-machine, per-concurrent user, or site licensing. Revenue structure may include some amount of upfront payment for the integration of the wireless solution itself and revolving license payments over the life cycle of the contract. This may depend on the number of the users or number of applications. Unfortunately for application providers, network operators often retain a major share of revenues.

Revenue Sharing

Revenue sharing is a common practice between partners in the channels of distribution (e.g., between game developer and mobile operator). Instead, profit sharing is usually limited to B-to-B settings only between the user and producer of a piece of software. Profit sharing is essentially a form of licensing in the sense that it also involves selling the right to use the software. However, in this model, the software provider's revenue is tied to its customer's performance when using the software.

A logical choice of a model for wireless games companies is based on revenue sharing with network operators, who provide the backbone for transmission game data. Empirical observations indicate that network operators often retain a major share of the revenue. However, in some cases, the service provider may charge the end user directly without an operator taking their share of the data transfer revenue.

Loss Leader Model

The loss leader pricing model here means giving something for less than its value. This is done in order to increase the customer base for later revenue or to support sales of some other part of the product/service offering. An example of loss leader revenue logic is a model in which the software is provided for free, and revenue is collected through selling related products or services to the users, or from the sales of complementary offerings to other customers. Hecker (1999) introduces the term *support selling* to illustrate cases in which revenue is collected through media distribution, branding, training, consulting, custom development, or after- sales support. Glynn (1999) notes that offerings provided for free are not merely an incentive, but, ideally, they also stimulate the usage of fee services. Examples of this approach can be seen in games that are preinstalled in mobile devices. These games can be developed in-house by the handset manufacturer or outsourced to a third-party software company.

Media Model

Hagel and Armstrong (1997) point out that the media revenue model, for instance, is an essential part of the virtual community business model. In a media model, the software is used to collect a group of users. For example, access to this group of users may be sold to third parties for advertising purposes. The media model involves a multitude of arrangements, in which third parties can be provided with information about the users, users are provided with information about the services of third parties, and the software acts as the mediator. This approach will produce interesting revenue opportunities in the future, as the user segments of mobile games make interesting target groups for many advertisers, and they may share some preferences or demographical and cultural characteristics.

SUMMARY AND CONCLUSION

In this chapter, we have focused on channel choices and revenue logics of mobile game producers. Using parts of the business model framework developed by Rajala et al. (2003), we have discussed potential revenue models of mobile game developers and identified examples of these in the mobile game industry.

In the existing and potential delivery models of mobile games, revenue can be collected either from user licenses, from royalties based on sold copies, from transactions concluded while the game is played, or from both. Furthermore, the possibility to deliver mobile games online offers a way to reach a large number of users. However, this possibility strengthens the role of mobile operators as distributors in the mobile gaming businesses, because the delivery and billing processes play a key role in the revenue logic. It seems, therefore, that the current the winners in the industry are the telecom operators, who dominate the end-user interface. They have close relationships with the users and viable delivery and billing mechanisms. However, they have not yet been able to create a critical mass of users or the volume needed to convert mobile entertainment into a profitable business. These circumstances create opportunities for other players with new business models.

The expected growth of the entertainment business combined with increased mobility offer a number of opportunities even for small software companies that develop mobile games. Research of viable business models for these companies is highly necessary. There is an emergent need to identify and analyze the success factors and key characteristics of business models of companies developing mobile games. This will improve the understanding of business models as well as provide valuable information for the companies involved in the mobile entertainment services industry.

REFERENCES

Amit, R., & Zott, C. (2001). Value creation in e-business. *Strategic Management Journal, 22*(6/7), 493-520.

Bask, A.H. (1999). *Third party relationship in logistics services.* Helsinki: Helsinki School of Economics.

Cravens, D.W. (1987). *Strategic marketing.* Homewood, IL: Richard D. Irwin, Inc.

Durlacher Research Ltd, E.P.O. (2001). *UMTS report: An investment perspective*. Durlacher Research Ltd.

Glynn, S. (1999). Making money from free services. Mercer Management Consulting.

Hagel, J.I., & Armstrong, A.G. (1997). *Net gain—Expanding markets through virtual communities.* Boston, MA: Harvard Business School Press.

Hecker, F. (1999). *Setting up shop: The business of open-source software*. Open Resources.

Hedman, J., & Kalling, T. (2003). The business model concept: Theoretical underpinnings and empirical illustrations. *European Journal of Information Systems, 12*, 49-59.

Kotler, P., Armstrong G., Saunders, J., & Wong, V. (1996). *Principles of marketing.* Hertfordshire, UK: Prentice Hall.

Lewis, R., & Trevitt, R. (1996). *Intermediate retail & distribution*. London: Hodder and Stoughton.

Mahadevan, B. (2000). Business models for Internet-based e-commerce: An anatomy. *California Management Review, 42*(4), 55-69.

McHugh, P. (1999). Making it big in software—A guide to success for software vendors with growth ambitions. Rubic Publishing.

Morris, M.,. Schindehutte, M., & Allen, J. (2004). The entrepreneur's business model: Toward a unified perspective. *Journal of Business Research* (forthcoming).

Rajala, R., Rossi, M., & Tuunainen, V.K. (2003). A framework for analyzing software business models. *Proceedings of the 11th European Conference on Information Systems*, Naples, Italy.

Rajala, R., Rossi, M., Tuunainen, V.K., & Korri, S. (2001). *Software business models—A framework for analyzing software industry.* Helsinki: The National Technology Agency of Finland.

Shapiro, C., & Varian, H.R. (1999). *Information rules, a strategic guide to the network economy.* Boston: Harvard Business School Press.

Varian, H.R. (1995). *Pricing information goods.* MI: University of Michigan.

Webb, K.L. (2001). Managing channels of distribution in the age of electronic commerce. *Industrial Marketing Management, 31*, 95-102.

Webb, K.L., &. Didow, N.M. (1997). Understanding hybrid channel conflict: A conceptual model and propositions for research. *Journal of Business-to-Business Marketing, 4*, 39-78.

This work was previously published in Managing Business in a Multi-Channel World: Success Factors for E-Business, edited by T. Saarinen, M. Tinnila, and A. Tseng, pp. 220-234, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 6.20
3G Mobile Virtual Network Operators (MVNOs): Business Strategies, Regulation, and Policy Issues

Dimitris Katsianis
National and Kapodistrian University of Athens, Greece

Theodoros Rokkas
National and Kapodistrian University of Athens, Greece

Dimitris Varoutas
National and Kapodistrian University of Athens, Greece

Thomas Sphicopoulos
National and Kapodistrian University of Athens, Greece

Jarmo Harno
Nokia Research Center, Finland

Ilary Welling
Nokia Research Center, Finland

ABSTRACT

Digital convergence brings new players in the telecom market and the Mobile Virtual Network Operators (MVNO) are an alternative way for companies to enter the 3G telecom market and start offering services. This chapter aims to contribute to the assessment of the market conditions, architectures and potential for profitable business cases of MVNOs aiming to operate in the mature and competitive markets. The results and conclusions provide guidelines for the wide

audience of mobile market players and media companies, spanning telecom operators to regulators and academia. In the following, the necessary background information is presented, quantitative figures such as Net Present Value, pay-back period, investment cost, revenues and running cost for different MVNO business cases are estimated and compared. The MVNO's impacts on a MNO operator and the effects of MVNO collaboration with a WLAN operator are analyzed with the same method and figures.

INTRODUCTION

The traditional barriers between separate sectors (so far) like telecom and broadcast companies, as well as fixed and mobile operators, are no longer so distinct.

Digital convergence will appear at different levels, such as user terminal, backbone network technology, tariffs and even at business or commercial levels. It seems that in few years the separation between mobile and fixed markets and between telecom and broadcast companies will disappear, allowing many agents to compete in a single telecom market.

As the licensing phase of 3G networks reaches a more mature level and the telecommunications operators are investigating the business perspectives of 4G networks, there is an increased interest worldwide from enterprises, active or not in the telecommunications sector without a 3G license, to become part of the 3G value chain, as it is considered a business opportunity with exceptional or acceptable profit margins. However, the economic and technical requirements imposed upon 3G licensees act as an economic burden to 3G developments and therefore the questions of better and more rapid market exploitation of licenses have already arisen and business collaborations are sought after. This situation encourages solutions without a radio access network via the network operations or service provision market channel. Especially for those without a 3G license, a new channel of entering and participating into the mobile business is the Mobile Virtual Network Operator (MVNO) channel. MVNOs initially appeared in the 2G market reflecting the self-evident interest of companies to enter the telecom market and start offering services.

Companies from different sectors, working or not in the mobile sector, as a first step to enter the market and start offering services can use the channel of MVNO, which is complementary either to service provision channel or to operator channel.

According to their origination, companies can be classified into three categories (Lillehagen, et al., 2001). First, those who already have business in the communication sector, second, those with business outside the communication sector and last, companies with business inside the Information and Communication Technology (ICT) sector but not as telecommunication operators (media and broadcast companies).

The interest from the companies that are already activated in the telecom sector is originated from their need to enter new markets and to increase their total market share. Operators with only fixed networks want to expand into the mobile sector because they experienced a substitution from fixed to mobile telephony and a reduction of their traffic while the total mobile traffic increased. Mobile operators already want to expand in order to increase their geographical coverage (domestic or international) in areas where they don't own a license. In this case main business sectors such as marketing, billing and customer care are shared by both networks in order to reduce the operational cost of the overall network. Furthermore, some network elements that the company already owns reduce the cost of the initial investment.

Companies inside the ICT but not in the communications sector, e.g., Internet Service Provid-

ers (ISPs), content providers and media companies, seek to increase their sales by introducing new services to the customers. ISPs foresee that users want to have access everywhere, and that means that beside the fixed broadband Internet, they must develop wireless broadband Internet solutions as well. Content providers want to be able to offer richer content through the broadband 3G mobile networks. Companies from the broadcast sector discern that the transmission of media content is no longer their exclusive right. So in order to gain back their market power, these companies must find a way to enter the new convergence scenery.

There are also companies outside the ICT sector that want to become MVNOs because they want to be able to provide mobile services to their customers (e.g., financial institutions, automotive industry, etc.) or want an extra sales channel to promote their brand names or products (e.g., consumer electronics companies) (OMSYC, 2004).

Between these three groups, large differences concerning their drivers to become an MVNO exist, since they have different business models, different characteristics and positions in the market. Early studies regarding both the Mobile Network Operators (MNOs) (Katsianis, et al., 2001) and the MVNOs (Varoutas, et al., 2006), have shown that market factors such as population density, customers type, timing of entry and penetration levels by new entrants will determine which strategy can be used in different areas and at different stages of market development.

But, in spite of MVNOs abilities and strategies, their competitiveness in the 3G mobile business will be severely limited if MNOs, which effectively control the available frequencies, the network infrastructure and the operation facilities, charge monopoly prices for their services. Due to the fact that in many cases MNOs are vertically integrated in the 3G market, they may also have incentives to restrict access to the facilities required by MVNOs through the imposition of prices, which will make the MVNO business case totally unprofitable for enterprises wishing to enter the market and effectively compete for 3G customers.

Of course, as the mobile market becomes more competitive and the regulatory framework more mature for such cases, the cost-based approach to charge MVNOs for their access to a 3G network would become less necessary, but it could circumscribe MNOs' incentives to invest in infrastructure. These arguments should be assessed within the context of the overall objective of promoting and strengthening the competitive framework for mobile services, which is the prime rationale for allowing MVNOs to operate in the market in the first place (ITU, 2001).

This chapter aims to contribute to the assessment of the market conditions, the architecture and the potential for profitable business cases of MVNOs aiming to operate in the mature and competitive European markets focusing on either wide market or lucrative market segments. Starting in Section 2 with the necessary background information regarding the existing and foreseeable business models for MVNOs, as well as the regulatory and access issues worldwide, the chapter in Section 3 addresses interesting questions regarding the market potential, the critical factors affecting the profitability of MVNOs, such as access prices, but also the impact of MVNOs on the associated MNOs trying to identify the win-win situations. The business opportunities of WLAN as an access technology for MVNOs' users are also presented and discussed. Section 4 outlines future trends and research questions aiming to contribute to further development of issues addressed in this Chapter. Finally, Section 5 summarizes the results and conclusions to provide guidelines for the wide audience of mobile market players, spanning telecom operators to regulators and academia.

BACKGROUND

MVNO Business Models

In many countries, the reserved 2G frequency band is nearly saturated, at least in the urban areas, so not much room is left for new players to enter the telecom market. The existence of a market for mobile services is more than confirmed today and there are always more people who see in 3G the opportunity to take part in a big game.

In the cellular world there are two main routes in order to provide cellular services: network operation & service provision channels. In order to supply services through an owned network, radio spectrum is required; a resource that is limited in supply.

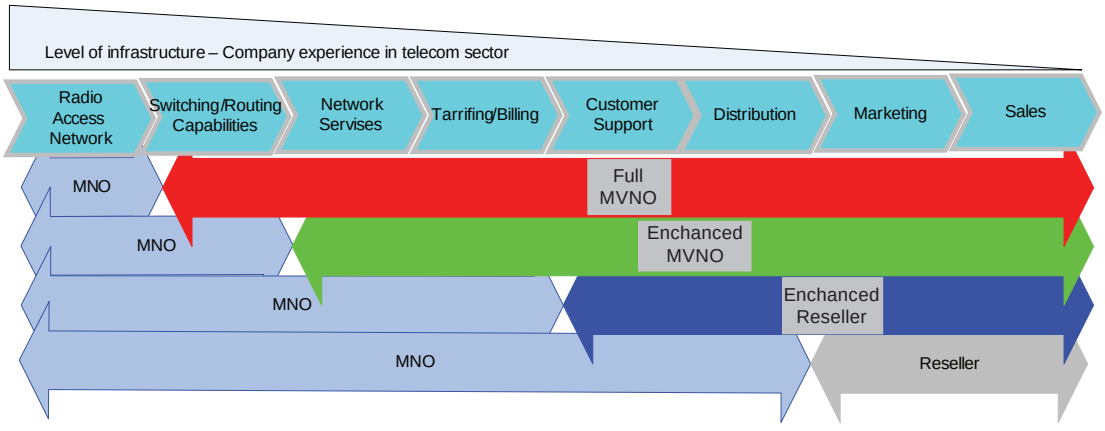
So, one of the possible ways to enter the world of 3G services, which have drawn considerable attention, is that of the MVNO model, meaning the reach of commercial agreements with already existing MNOs.

There is a lot of discussion and an obvious doubt over the degree to which an operator is considered to be virtual in comparison with actual network operators. The definition of the term "virtual operator" varies from each point of view. For example, certain analysts affirm that an MVNO must necessarily have a network code and SIM cards of its own (OVUM). On the other hand, the UK Regulator (OFCOM) considers as an MVNO those without a SIM card, which can be also considered as Enhanced Service Providers (ESPs) (OFCOM, 1999). A general definition may be that an MVNO is an operator that provides cellular services (data or voice) without owning spectrum access rights. From the customers' point of view, there is no distinction between the two operators, since an MVNO looks like any MNO, but a MVNO does not imply ownership or operation of base station infrastructure. Figure 1 illustrates the MVNO idea compared to other mobile business schemes.

There are different scenarios for an MVNO approach and consequently different architectures for the MVNO such as (Table 1):

- A full MVNO, with its own SIM card, network selection code and switching capabilities as well as service center but without

Figure 1. MVNO types and value chain



ownership of any radio spectrum (OVUM). The main difference from the other business models is the ability to operate independently from the MNOs network facilities and the full control over the design of services and tariff structures. Furthermore, the possibility to select the host network (MNO) could be an option in the foreseeable future.

- Enhanced Service Provider ESP or Indirect Access MVNO (IA-MVNO) with its own SIM card, core network (circuit switched and/or packet) and service facilities, e.g., with its own IN or IP application servers (OFCOM, 1999). Some independence from the host MNO exists and under some constraints the provider can design services and differentiate its products.
- Wireless Service Provider (SP) without own core network and SIM card, basically an Internet portal providing wireless IP services based on the MNOs access and core network.
- Resellers just offering pre-packed cellular services to end users.

It is foreseen that MVNOs can act as an important driver for the emerging 3G market since potentially they would offer customers additional service baskets. In addition, since users are rather indifferent about the network infrastructure but put their emphasis on user interfaces and services, an operator with carefully selected content and good marketing strategy could drive forward the market if specific target groups, such as family users, are addressed.

The mobile market is, without doubt, considered important, especially the 3G, and very promising by a lot of firms; but in view of the surrounding uncertainty, "big" new entrants will be content with taking the opportunity of the virtual model, at least for the time being.

In many countries, the 3G licenses have been assigned to existing 2G operators, so there is enough place for new players, although in 3G the emergence of just one or two new operators can be witnessed. The new entrants, who in many cases will be quite well known brand names with considerable experience in the marketing, distribution and management of customer relations,

Table 1. Types of MVNOs

	Reseller	Service Provider	Enhanced Service Provider	Full MVNO
SIM Card	No	No/Yes	Yes	Yes
Interconnection/ Roaming	No	No	No	Yes
Value Added Services	No	No/Yes	Yes	Yes
independence from Host MNO	High	High	Medium	Low
Control over tariffs	No	No	Medium	High
Size of own network infrastructure	None	None	Medium	High

will focus prevailingly on services of great added value and, above all, on m-commerce.

The new virtual operators can be a threat for new entrants in the 3G markets. In a mass market environment, where social, emotional and cultural criteria often prevail, each new entrant will attract a lot of attention. It is nonetheless reminded that the network operator, following an agreement such as that of the Virgin model (which according to OFCOM is an ESP (OFCOM, 1999)), obtains new clients and new earnings without investing actually anything additional in infrastructure and in the management of new services.

In the 3G world, it could become clearer that the above described prospect can bring to 3G MNOs immediate revenues so as to recover at least a part of the investment made as soon as possible. Furthermore, the MVNO offers prospects of expansion at an international level in markets that, up to now, have been closed to the participation of operators from different areas. For example, the European 2G operators could become virtual operators on the CDMA networks in U.S. and vice versa. In reality, both in the long and short run, the MVNO model represents a profitable option for all parties involved. The fear of losing customers to a virtual operator could in fact make mobile network operators lose an opportunity for development.

Regulatory Framework

Whether, and to what extent, regulatory intervention is necessary is still under discussion as national regulators in different countries have initiated discussions and consultation about the MVNO regulation. In the European Union (EU), until now there is no directive that obliges MNOs to grand access to MVNOs to use their 3G networks. Currently, while there is a tendency in favor of MVNOs, no major regulation actions have been undertaken. The ones in favour of the regulation believe that no MNO will provide access to MVNOs unless there is a regulators intervention. On the other hand, MNOs have very high profit margins and in some cases significantly over costs. The current regulation, as already interpreted by some national regulatory authorities, gives them the power to enforce an access obligation on existing operators following the paradigm of "local loop unbundling."

Regulatory intervention, especially in terms of pricing and access rights, is an important factor for MVNO success. Without it, the MVNO model depends only on the commercial negotiations and agreements between the MVNO and the network operators. Since MNOs are the owners of the desired radio spectrum, they will refuse to negotiate and so it may be difficult for companies, especially those originating outside the telecom sector, to enter the market, because traditional MNOs see the MVNO idea as a possible threat capable to shrink their market share and therefore their revenues. Regulators in general show little sign of intervention and hope that the market itself will achieve the desired agreements between the two parts.

In the EU, the National Regulatory Authority (NRA) in each member country is responsible for the determination of the framework that allows (or doesn't) an MVNO to enter the market. The EU defines an operator with market share over 40 percent as a Significant Market Player (SMP) and forces them to provide access to other minor network providers. Each NRA decides if an SMP exists but the EU can always intervene and take the final decision. If the local NRA determines that a MVNO is truly a network operator and a SMP status exists, then it is possible for the MVNO to enter the market although the appropriate regulatory framework does not exist.

Issues surrounding the MVNO concept have not been discussed in great detail, and hence most regulators are not yet in a position to provide statements of policy.

OFCOM recently assessed the state of policy development on MVNOs in other European countries and found that, with a few exceptions, it is premature for European regulators. In 2004, OFCOM decided that all the licensed operators in UK are SMP but the market is competitive, so there are no requirements for operators to grant access to their networks to MVNOs. The German NRA (Bundesnetzagentur) also determined than no operator is a SMP. In Italy, although the NRA (AGCOM) has adopted a decision at 2000 that permitted MVNOs to enter the mobile market, it afterwards decided that the conditions to amend the regulatory framework will not occur until the end of 2009. In Sweden, PTS forced 3G operators to allow MVNOs to access their networks (Analysys, 2002), while in Norway, the regulator (NPT) decided that MNOs are not obliged to give access to virtual network operators (MVNO pricing in Finland, 2005). On the hand, the French Telecommunications Regulator (ART) in 2002, decided that it had no power to force a licensed mobile operator to sign MVNO agreements. In 2005, the Spanish regulator (CMT) issued a license to launch services as long as the MVNO reaches an agreement with Spain's mobile network operators and in Ireland, the NRA (ODTR) has issued a 3G license to an operator who will host MVNOs.

The regulator of Hong Kong, the Office of the Telecommunications Authority (OFTA), has indicated that 3G networks should be opened up to MVNOs. In an analysis paper based on an industry-wide consultation (OFTA, 2001), OFTA proposed a 3G licensing framework based on an "open network" requirement. Under this requirement, 3G service provision would be separated from network operation in order to enhance competition in services and provide customers with more choices and price packages. Successful bidders of 3G licenses have been required to make at least 30 percent of their network capacity available to unaffiliated MVNOs and content and service providers but the term capacity is not defined. Furthermore, any successful bidder that currently operates a 2G network must agree to offer domestic roaming service to all new entrants. OFTA requires each MVNO to have its own Mobile Switching Centers (MSCs) and gateways, billing and customer care sections, to provide SIM cards and to be able to offer interconnection with other networks as well as roaming services.

Access Charges

The basic key cost element affecting the profitability of this telecommunication business model is the structure of the interconnection cost models.

At immature markets, as 3G, or when the competition is ineffective, cost-based prices are desirable. Yet, the determination of costs is debatable (Leive, 1995; Melody, 1997) and it is not clear which is the best methodology or even if the resulting prices are consistent with what happens in the competitive mobile market.

The additional costs associated with mobile network elements that do not exist in fixed-line networks are the main reason why termination costs are higher on mobile networks than on fixed-line (OECD, 2000). However the cost of fixed-mobile interconnection is similar to those of interconnection in general (fixed to fixed, termination fees included) with the only exception being the different investments needed and the rapid changes in the technology (possible interconnection with additional schemes, WiMAX or similar technology).

According to the ITU (ITU, 2001) the methodologies that may be applied to the determination of interconnection charges rates include:

- Different forms of long-run incremental cost methodologies, such as Long-Run Average Incremental Costs (LRAIC), Total Element

Long-Run Incremental Costs; (TELRIC) and Total Service Long-Run Incremental Costs (TSLRIC);
- Different forms of Fully Distributed Costs (FDC);
- Efficient Component Pricing Rule (ECPR); and
- Hybrid forms, such as LRIC, subject to FDC-based caps.

A study of Europe-wide mobile costs for the European Competitive Telecommunications Association (ECTA) (Analysys, 2000), revealed the controversy and the sensitivity of costing methodologies in the rapid changing mobile market. The Long Run Incremental Costs (LRIC) methodology used in this analysis indicated that Mobile operators charge 40% to 70% above their LRIC costs. However, operators argue that LRIC methodology is not appropriate for dynamic and rapidly growing markets (Clark, 2002).

MVNOs, in order to enter the 3G market, have a choice of different strategies as already described in the previous sections. The nature of MVNO and the extent to which it is engaging in interconnection or pure resale of network capacity should be reflected in the pricing principles that apply to the provision of services. So, a full MVNO with an extensive network of its own, will only make minimum use of the MNO's infrastructure and should be granted to interconnection on the same basis as the MNO.

The ability and the attractiveness of MVNOs to offer competition will be severely limited if network providers, who effectively control facilities, are in a position to charge monopoly prices for their services. Because network providers are in many cases vertically integrated into the competitive 3G market, they may also have incentives to restrict access to the facilities required by competitors through the imposition of prices which make it unprofitable for MVNOs to enter the market and effectively compete for 3G customers.

It is widely agreed that cost-based charging for access to a 3G operator's network by MVNOs would become less necessary as the market becomes more competitive and mature. It has also been claimed that cost-based access charges for MVNOs could damage incentives to invest in infrastructure, particularly in the early stages of investment in 3G systems. These arguments should be assessed within the context of the overall objective of promoting and strengthening the competitive framework for mobile services, which is the prime rationale for allowing MVNOs to operate in the market in the first place.

Market factors such as population density, customer type, timing of entry and penetration levels by new entrants will determine which strategy is used in different areas and at different stages of market development. Relying solely on full facilities-based competition to deliver competing 3G services may not provide 3G service competition to all end users, given the costs involved in duplicating a full network deployment throughout all areas of a country. As such, service-based competition through the resale of network capacity will be an important element of the overall state of competition in the 3G market.

Currently the EU obliges companies with a market share of over 50 percent to open their networks to other users at a cost-plus-margin-based price and for the moment, only KPN Mobile is in this position. Other licensed operators with market shares of more than 35 percent do not have to charge on a cost-plus-margin basis, so leasing from them could be more expensive.

OFCOM takes the view that the logical principle for MVNO charging would be retail-minus which sets an interconnection price by looking at foregone costs and deducting these from the retail price. The costs foregone would be those associated with customer care, billing, provision of value-added services, etc. OFCOM concludes that simple resale of 3G capacity can encourage entry of efficient service providers of retail 3G services.

EVALUATION OF 3G MVNO BUSINESS STRATEGIES

Business Cases, Technoeconomic Methodology and Assumptions

Based on market studies and associated reports about companies which have expressed their interest to enter the market, several MVNO business profiles can be foreseen. The existing similarities lead to the grouping of these profiles into two main business profiles: those focusing on network operations and those focusing on service provisioning. Different demand models and service penetration rates must be defined in order to take into account these two different cases for an MVNO. This business classification will lead to specific service packages offered by these potential MVNOs and will be attributed to MVNO business profiles.

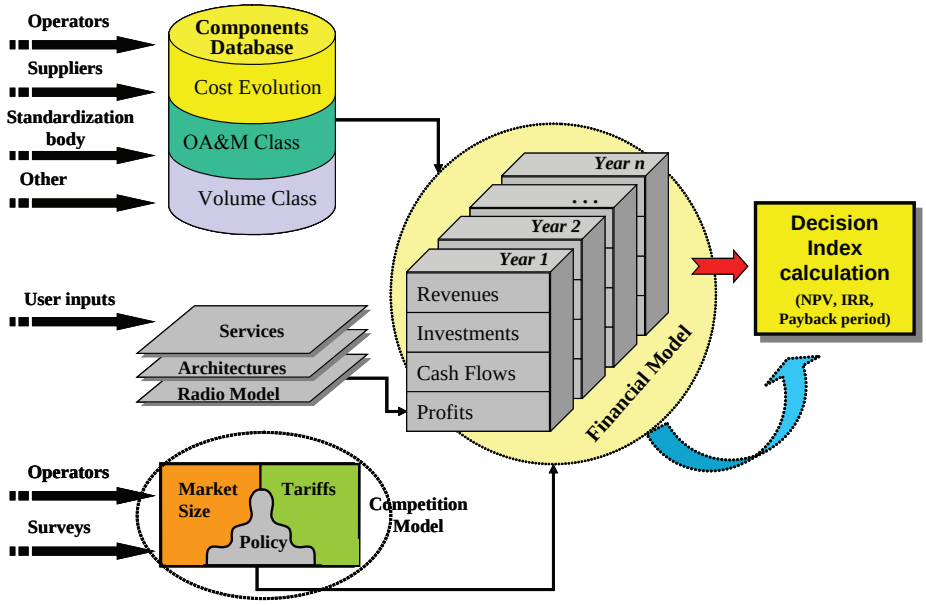
The techno-economic modeling was carried out using the TONIC tool, which has been developed by the IST-TONIC project (IST-TONIC, 2000). This tool is an implementation of the techno-economic modeling methodology developed by a series of EU co-operation projects in this field. The tool has been extensively used in several techno-economic studies among major European telecom organizations and academic institutes (Katsianis, et al., 2001; Monath, et al., 2003; Varoutas, et al., 2003; Varoutas, et al., 2006).

The tool calculates revenues, investments, cash flows and other financial results for the network architectures for each year of the study period. It consists of a dimensioning model for different architectures that is linked to a database containing the cost figures of the various network elements and the cost evolution of them over time. An analytical description of the methodology and a similar tool can be found in Ims (1998), while a more detailed description of the tool used in this analysis is presented in Katsianis, et al. (2001).

Structure of the Tool for the Techno-Economic Evaluations

The main principles of the methodology used in the techno-economic tool are analyzed in Figure 2. The cost figures for the network components have been collected in an integrated cost database, which is the "heart" of the model. This database is frequently updated with data obtained from the major telecommunication operators, suppliers, standardisation bodies and other available sources. These data concern the initial prices for the future commercial networks components as well as a projection for the future production volume of them. The cost evolution of the different components derives from the cost in a given reference year and a set of parameters which characterises the basic principles of each component. The cost evolution of each component in the database, the estimations for the OA&M cost and the production volume of the component are incorporated in the database according to the model described in Appendix I. As a next step in the network evaluation, a specification of the services that will be provided to the consumers is needed. The network architectures for the selected set of services are defined, and a geometrical model (Olsen 1999) is used in order to calculate the length of the cables as well as the civil works for their installation (database data) for wireline access networks. In wireless networks, a radio model is incorporated in order to calculate the coverage used. The future market penetration of these services and the tariffs associated with them, according to each operator's policy, are used for the construction of the market evolution model as also defined in Appendix I. The operator tariff policy could be taken into account by modifying the tariff level in conjunction with the expected penetration of the offered services. Data from statistics or surveys can be easily integrated into the tool when formulas measuring the impact of tariff level to the saturation of the services are available.

Figure 2. Techno-economic methodology (Katsianis, et al., 2001)



The data are inserted into the financial model and the revenues, investments, cash flows and profits (or other financial results) for each year, during the project's study period, are calculated.

In this study, the full MVNO scenario has been analyzed, assuming that the MVNO owns exactly the same elements as the MNO, except the radio access part where the infrastructure of MNO is used. The dimensioning of the MNO UMTS network is performed starting from the coverage requirements set by the subscriber distribution information and WCDMA radio link characteristics. The obtained capacity for components and leased line is compared to the WCDMA radio interface capacity calculated from average busy hour need per subscriber. If the usage based capacity is higher than the coverage based, the additional components stations are added to the network build-out. The core network elements and their needed capacity are calculated from the base station distribution and traffic amount or number of served subscribers, depending on the limiting factor in the element capacity.

MVNO Business Cases

The MVNO business cases have exploited useful insights from previous 2G and 3G business cases (Varoutas, et al., 2003; Katsianis, et al., 2001). In order to compare the different scenarios and models, economic indicators such as NPV, IRR and payback period are presented (Appendix I).

A discount rate of 10% is selected in order to calculate discounted cash flows, which take into account the cost of capital and the expected risk-free return from investments in the telecom

business. The value used reflects a mean value among the major European Telecommunication Operators.

The study period is ten years and the modelling focuses on two area scenarios: a large European country (population of 70 million) characterised, for example, by Germany and France and a small European country (population of 5 million) exemplified by Scandinavian countries like Norway or Finland. The models are not exactly representative of any defined country, but rather share typical demographic characteristics among these countries. The countries differ on several points in addition to their geographical and demographic features. The geographical approach been examined for full coverage in all areas icluding the rural ones. In Table 2 the charactersistics of the areas covered are illustrated. Note that the overall size of the surface area isn't the sum of all the sub-areas because certain areas (mountain tops, etc.) do not need to be covered.

The subscriber saturation level is estimated to be higher in the Nordic country type —95% versus 90% in the large country type. Second, usage differs in that the Scandinavian users are assumed to have 20% greater usage than their counterparts in the large country. Last, terminal subsidies are four times larger per new subscriber in the large country type than in the small country type. The country types differ in several points, in addition to their geographical and demographic features. The operator in the large country is assumed to have significant license costs and the two profiles for each type of country are differentiated in terms of greater usage and ARPU.

In the first case, it is considered the business profile of a telecom operator or a power company without a spectrum access license aiming to be a Full MVNO using the existing infrastructure in order to complement or expand its business to other market areas and services like B2 in Sweden, Kingston in UK, One.Tel in The Netherlands, etc. This will be the *Operator-like* MVNO business profile. This kind of MVNO takes advantage of issues such as initial market share, lower training costs, etc.

Table 2. Large and small countries demographics

CountryType	*Large*	*Small*	*Description*
Area size	370,000	330,000	Size of surface area of the country (km^2)
Area dense	185	17	Size of dense urban area (km^2)
Area urban	2,960	264	Size of urban area (km^2)
Area suburban	37,000	3,300	Size of suburban area (km^2)
Area rural	303,400	264,000	Size of rural area (km^2)
Population dense	50,000	50,000	Number of inhabitants in dense urban area per km^2
Population urban	4,000	4,000	Number of inhabitants in urban area per km^2
Population suburban	1,000	1,000	Number of inhabitants in suburban area per km^2
Population rural	40	3	Number of inhabitants in rural area per square km (during busy hour)
Total Population	65,000,000	5,500,000	Total population

In the other profile, the MVNO has high brand-value with an existing large customer base aiming to expand its business in the mobile area and, therefore, aims to attract market share from the other MNOs. Consequently, the churn effects must be taken into account. In this case, several advantages (e.g., marketing costs) exist and disadvantages (e.g., leased lines costs and personnel costs) are the key elements. This is actually a *Service-oriented* MVNO business profile.

Evaluation of Business Cases

3G MVNO

Based on the previously described assumptions, an analysis for the profitability of the two full MVNO profiles, both in large and small countries, has been conducted and presented in Varoutas, et al. (2006). In Table 3 the main economic results for the different scenarios are presented.

In Varoutas, et al.(2006), it has been revealed that companies planning to provide 3G services can benefit from acceptable NPV and IRR figures. In more detail, operators investing in MVNO rollout benefit from more or less the same payback period and rather attractive economic figures. It has also been denoted that the investments are more or less proportional to the population for the large country but almost double for the small one. This difference is based on the necessity to offer coverage and, therefore, in the small country equipment that is not fully utilized is purchased. The figures are for rather pessimistic market shares (all are considered more or less new entrants) and surely MVNO can expect more optimistic results.

For the case of a small country, the initial position of the MVNO in the 2G world is mandatory for a successful business in the emerging 3G market. On the other hand, stronger service differentiation is followed by larger investments while the payback period remains the same.

The breakdown of total investments in the large country case confirms that the bulk of the OPEX is accounted for the interconnection costs. The running costs include leased lines, interconnection costs, terminal subsidies, employee and training cost, marketing and maintenance cost.

Table 3. Summary of the basic results (Varoutas, et al., 2006)

Country type	*Large*		*Small*	
MVNO type	**Operator – like**	**Service Oriented**	**Opera-tor - like**	**Service Oriented**
NPV (M€)	111	332	259	28
IRR	12%	15%	40%	14%
Rest Value (M€)	48	39	5	2
Pay-back period (years)	8.2	7.7	5.0	7.6
Number of customers	4,800,000	3,600,000	640,000	210,000
Total mobile penetration - end	90%	90%	95%	95%
Total UMTS penetration - end	76%	76%	80%	80%
Investments (M€)	144	121	55	49

MVNO Impact to a MNO (UMTS) Operator

In this scenario, the impact of an Operator-like MVNO to its MNO is analysed and discussed. In this case, the MNO has increased costs since there are additional customers in the network but the benefit comes from the interconnection cost that the MVNO operator pays in order to use the UMTS network.

The selection of the appropriate value for the interconnection price between MVNO and MNO has been based on data from operators and reports. The situation where the interconnection cost is 50% increased yield to negative NPV and non-acceptable IRR and payback period for the MVNO case. This could be the turning point for this business case and the MVNO must have hard negotiation with the MNO in order to keep the interconnection costs as low as possible. On the other hand, the regulators should protect the new entrants as MVNOs and ensure that the interconnection price level will boost the overall competition although it remains a good profit for the MNO.

The main economic results for the two basic scenarios are illustrated in Table 4 and Figure 3. These results show that companies that intend to provide UMTS services can have acceptable NPV and IRR figures when they support an additional MVNO in their network as well.

The economic figures (Figure 4) reveal that the revenues stream (for the MNO) from the MVNO operation exceeds the required investment and operation cost. The logical explanation for that lies in the fact that the operators are going to build UMTS networks that are capable of serving more than the expected customers due to regulation implications. This obligation is based on the necessity to offer coverage, and therefore they purchase equipment that is not fully utilized.

In this business case it has been assumed that the MNO's market share remains 30%. It is logical to assume that the MVNO will gain some

Table 4. Summary of the basic results (MVNO impact to MNO). (LC= Large country, High=High licenses fees, Wno=without WLAN, Impact=with a MVNO)

Country type	*Large*		*Small*	
	Higher license fees and MVNO	***High license fees without WLAN***	***Low license fees and MVNO***	***Low license fees without WLAN***
NPV (MEuros)	9,825	5,639	1,278	635
IRR	23.4%	18.8%	53.2%	38.6%
Rest Value (MEuros)	3,606	3,479	255	239
Payback period	6.8	7.1	6.0	6.3
Investments (MEuros)	7,432	7,308	381	363
Running costs (MEuros)	23,007	22,561	2,396	2,329
Revenues (MEuros)	55,682	46,475	5,602	4,182
Revenues-Running	32,675	23,914	3,206	1,853

Figure 3. Financial indexes for different cases

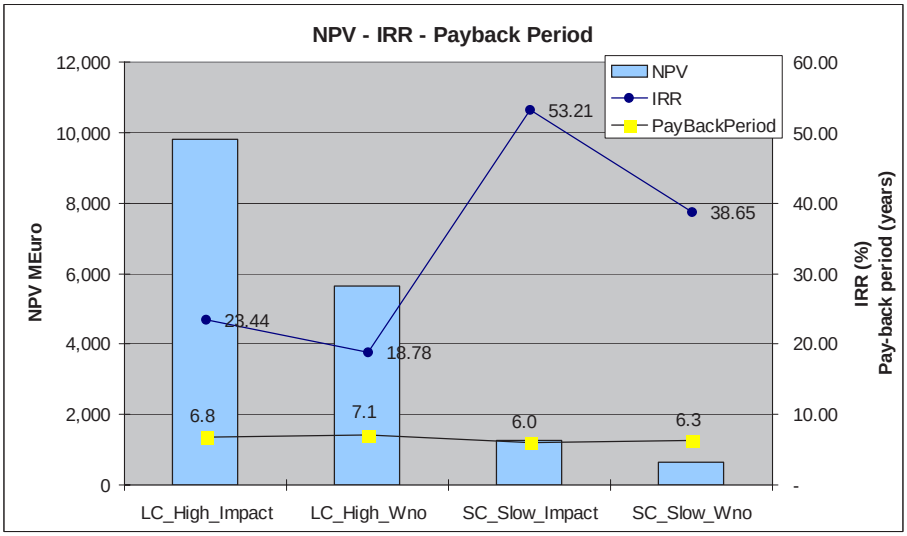
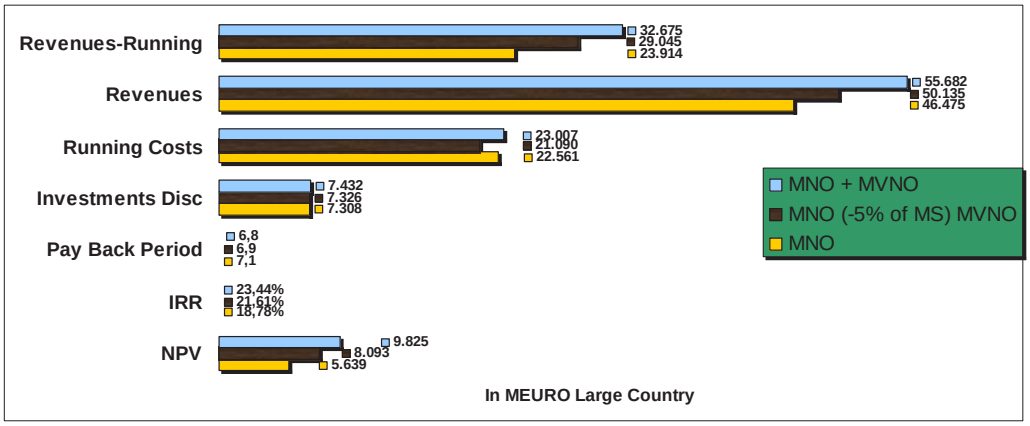


Figure 4. Basic financial indexes for a large country (including impact on MNO market share)



customers from its MNO. It has been calculated that 5% of MNO customers in the large country case will become MVNO customers in the future. So a part of the MVNO's customers came from the potential MNO's share. Of course the MNO makes special agreements with its MVNO, so that as many as possible of the MVNO customers are out of the competitors share and not from its own potential customers. This market share's losses influence the running cost positively, since

fewer customers must be served via the network. Furthermore, the revenues and NPV values are greater than in the basic case (without the MVNO) due to the interconnection cost.

Concluding, MVNO can have a positive impact to MNO even if it reduces its market share. MNOs have many benefits from their "marriage" with MVNOs and can overcome any strict coverage obligations or even pessimistic market forecasts.

MVNO as a WLAN Operator

In this case the MVNO deploys his own broadband wireless network (WiFi or WiMAX) in order to cut off the high connection costs that limit its ability to offer customers additional broadband services. An Operator-like MVNO and WLAN operator have been studied both in a large and a small European country. The case of a licensed UMTS+WLAN operator has been studied in Varoutas, et al. (2003).

The main economic results for the basic scenarios are illustrated in Table 5 and Figure 5. The WLAN MVNOs have a larger revenues stream since the WLAN operation will act as an additional service for its existing customers. This occurs due to better usage patterns of its customers and associated service consumption with only small additional investments needed. The WLAN operation could be the logical step for a MVNO since the investments are minimal and the additional potential revenues are in the scale of MEuros. In the large country, the NPV is almost three times more than in the basic case (without WLAN) whereas in the small country 30% greater. This occurs due to the larger number of potential customers that an operator can serve in a large country.

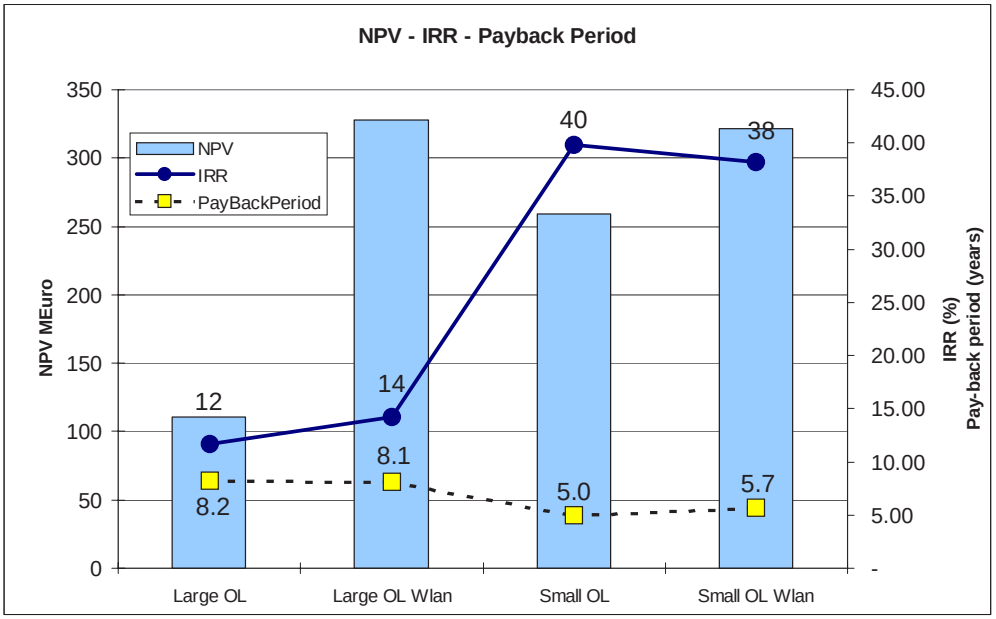
The main difference between running costs in both large and small country types are the marketing costs, because these are associated with the population. Furthermore, the additional running cost for the WLAN operation is negligible, especially in the small country.

The economic results reveal that investments are roughly proportional to population in the two country types. The population ratio is almost 14:1. It can be observed that the enhancement of the MVNOs' service basket due to the provision of broadband services via WLANs can act as a

Table 5. Summary of the basic results for an operator-like MVNO with and without WLAN operations

Country type	*Large*		*Small*	
	Without WLAN	***With WLAN***	***Without WLAN***	***With WLAN***
NPV (MEuros)	111	328	259	322
IRR	11.68%	14.18%	39.77%	38.22%
Rest Value (MEuros)	48	116	5	139
Pay-back period	8.2	8.1	5.0	5.7
Investments	144	194	55	141
Running Cost	23,070	24,042	3,080	3,092
Revenues	25,192	27,035	3,950	4,241
Revenues-Running	2,122	2,993	870	1,148

Figure 5. Basic financial indexes for all cases (OL=Operator-like).



significant leverage to its business. The MVNO can almost double its economic figures with almost negligible investments. Taking into account the positive impact of MVNOs to a MNO, the additional revenues of MVNO due to broadband services can reverse the negative attitude of a MNO to MVNOs.

Regarding the other cost structures, analysis shows that usage and tariff levels have great impact. The tariff and usage levels are the most critical parameters for the economic criteria of NPV and IRR outside the interconnection related costs. The model links revenues to usage levels, which means that a 50% increase in revenue corresponds to a 50% increase in usage. Under these circumstances, it would be expected that network costs would increase accordingly. However, since network costs are essentially dictated by coverage constraints and not by capacity ones, an increase in usage leads only to greater revenues, while the corresponding increase in costs is minimal, and relates to core network elements.

FUTURE TRENDS

The future trends for MVNOs depend on several factors. First, the regulation continues to have an important role, as the market entry is in many cases dependent on protection against overcharged radio network capacity. Once a virtual operator has reached a considerable share of subscribers, it has the bargaining power to approach several MNOs and does not need such protection.

Another aspect is the business model applied by the MVNO. The target may be the corporate market, where high IT competence and resources are needed, but, on the other hand, with deep co-operation, a large customer base can be won in one deal. In the future, MVNO business in the

consumer market may be based more and more on bringing in large loyalty groups, whether relating to brand, lifestyle or other social of ethnic grouping.

The trend for many of the most important mobile service providers has led toward increasing infrastructure, and, in this way, strengthening their market power. Full MVNOs, with their own MSCs, can easily connect to several radio networks, and negotiate over termination fees. As more and more services are eventually being realized, through 2.5G, 3G and WLAN type of networks, MVNOs can start to compete not only with price but also with intelligent IP solutions. Integration of the traditional voice with the new services will definitely be crucial, whether VoIP or other solutions are used.

The MVNO business channel will be a viable financial path of digital convergence schemes. Several alternatives examined within this chapter, like pure 3G MVNO and MVNO with WLAN, will support digital convergence with the same technical approach as the traditional operators will implement.

This leads to the convergent service realm, where those owning certain type of access networks, whether fixed or mobile, or not owning any, can complement their access methods with the virtual mode to serve their customer base to the fullest. As the access means should be possible to select optimally for the particular situation and application, no player can any longer provide all possible networks physically everywhere; as a result, liaison is going to have increasing importance. Business opportunities are eventually going to migrate more and more from the network provisioning to the service provisioning and integration.

A trend toward market differentiation and service customization is taking place, and this gives opportunities for different kind of MVNOs. On the other hand, some smaller virtual operators can serve the niche groups, the virtual enablers (MVNEs) providing the technical platforms for them. Regional players would have more potential in the future world, as well as the pan-European and even international MVNOs. With international presence, the virtual operators would solve even the roaming conditions for their benefit, in the way towards a converging world.

After 3G licencing, especially in Europe, the case of a new entrant having its own 3G license in a small and large country market can be considered, in light of the new developments in regulation, technology, convergence and better understanding of the market demands. This business model is very important and follows the same rules as the model described in this chapter. Major effort should be expended in order to identify the hurdles that have prevented entities from becoming new entrants. Understanding the cost structure of a new entrant and identifying parameters related to longer than seven years pay-back periods are issues for further study.

The possibility of building a CDMA450 network as a comparison scheme to the 3G MVNO could be also analyzed. This scenario could be pointed out since CDMA450 has been one of the skyrocketing sectors in the global wireless communication industry in the past few years. It has attracted keen interest in the industry because the initial driving force for CDMA450 was the urgent need to find a digital replacement for the ageing NMT450 analogue cellular systems that had been widely deployed, not just in the Nordic countries but also in many countries in Central and Eastern Europe. In addition, CDMA450 inherits all the technical and service advantages of the CDMA2000® system. This means that technology can become a 3G solution for some operators without UMTS licenses before becoming an MVNO player. A Greenfield CDMA450 operator could be entering the 2G and 3G mobile networks market. In this case study the economics for two build out strategies--a full country coverage case and a rural rollout case could also be studied.

Fixed line operators, in order to provide fixed-mobile convergence services, could also become MVNOs, using either their own infrastructure (and become a full MVNO) or just reach an agreement with a MNO to use the wireless access network (only service providers).

CONCLUSION

The interest of companies, either working in the mobile sector or not, entering this market is self-evident and many of them are looking for specific channels to start offering services under a Digital Convergence scheme. The channel of MVNO is either complementary to a service provision channel or operator channel but is still a ways off from taking part in this big game.

Acceptable business opportunities can be observed through calculations in terms of forecasted and actual mobile penetration across Europe. Agreements with MNOs for spectrum usage and interconnection give MVNOs enough space for business opportunities and acceptable profit margins. As the Digital Convergence path evolves, the MVNO channel will be more attractive for companies left out from the licencing.

Both infrastructure costs (which are high due to difficulties in obtaining volume discounts) and interconnection costs are too critical for the success of MVNOs. Interconnection costs, which could be the turning point for the business cases and the MVNO must have hard negotiation with the MNO in order to keep them as low as possible. On the other hand, the regulators should protect the new companies and ensure that the interconnection price level will boost the overall competition.

Marketing and entry costs in general can be a burden for a potential MVNO, but this can be overcome by means of a high brand firm or a company already operating. Although revenues from the provision of broadband services are missing from current MVNO business plans, this could be another opportunity for the MVNO to expand its business in the future. In reality, the MVNO way to 3G represents a profitable option for all parties involved and a key enabler for technology, network and services providers.

Future MVNOs should also benefit from WLAN if they complement their offer with WLAN services and in addition MNO profitability could be increased within the "financial" support of the concentrated to retail business MVNO.

Furthermore, different technology schemes and different business profiles could be studied in order to clearly achieve the optimal strategy and policies in the mobile era always including solutions for non regulated licensed operators offering even convergence applied services.

ACKNOWLEDGMENTS

This work has been partially supported from the European CELTIC/ECOSYS project, a PYTHAGORAS Grant from the Greek Ministry of Education and a PENED grant from the Greek Ministry of Development (General Secretariat for Research and Technology). Authors would like to acknowledge the fruitful comments and contributions from their colleagues from NOKIA Coorporation, Telenor AS R&D, France Télécom R&D, Helsinki University of Technology and University of Athens.

The authors would also like to thank the reviewers for their fruitful comments and suggestions.

REFERENCES

AGCOM. (2001). *Italian regulator of the telecommunication market.* Retrieved from http://www.agcom.it

Analysys. (2000). *Economic studies on the regulation of the mobile industry.* Final report for ECTA: Analysys.

Analysys. (2002). *The future of MVNOs.*

ART. *The French telecommunications regulator.* Retrieved from http://www.art-telecom.fr/eng/index.htm

Bundesnetzagentur. *German Federal network agency.* Retrieved from http://www.bundesnetzagentur.de/

Clark, V. (2002). Business & regulatory: Ecta pressures regulators on fixed-mobile access charges: Total Telecom.

CMT. *Spanish telecommunications market commission.* Retrieved from http://www.cmt.es

Ims, L. (1998). *Broadband access networks introduction strategies and techno-economic evaluation.* Chapman & Hall.

IST-TONIC. (2000). *Techno-economics of ip optimised networks and services.* IST: EU.

ITU. (2001). *Mobile virtual network operators.* ITU.

Katsianis, D., Welling, I., Ylonen, M., Varoutas, D., Sphicopoulos, T., & Elnegaard, N. K., et al. (2001). The financial perspective of the mobile networks in europe. *IEEE Personal Communications, 8*(6), 58-64.

Leive, D. M. (1995). *Interconnection: Regulatory issues.* Geneva: ITU.

Lillehagen, A., Armyr, L., Hauger, T., Masdal, V., & Skow, K.-A. (2001). An analysis of the MVNO business model. *Telektronikk,* (4), 7-14.

Melody, W. H. (1997). *Telecom reform: Principles, policies and regulatory practices,*Technical University of Denmark.

Monath, T., Elnegaard, N. K., Cadro, P., Katsianis, D., & Varoutas, D. (2003). Economics of fixed broadband access network strategies. *IEEE Communications Magazine. 41*(9), 132-139.

MVNO pricing in Finland. (2005). The Ministry of Transport and Communications of Finland.

NPT. *Norwegian NRA.* Retrieved from http://www.npt.no

ODTR. *Irish commission for communications regulation.* Retrieved from http://www.odtr.ie/

OECD. (2000). *Cellular mobile pricing structures and trends (No. DSTI/ICCP/TISP(99)11/FINAL).*

OFCOM. *Ex OFTEL, the UK NRA.* Retrieved from http://www.ofcom.org.uk/

OFCOM. (1999). *Statement on mobile virtual network operators.*

OFTA. (2001). *Open network: Regulatory framework for third generation public mobile radio services in Hong kong* (Discussion Paper).

Olsen, B. T., Zaganiaris, A., Stordahl, K., Ims, L.A., Myhre, D., Overli, T., et al. (1996). Technoeconomic evaluation of narrowband and broadband access network alternatives and evolution scenario assessment. *IEEE Journal Selected Areas in Communications, 14*(8), 1203-1210.

Olsen, B. T. (1999). OPTIMUM – a techno-economic tool, *Telektronikk, 95*(2/3).

OMSYC. (2004). *MVNO in europe benefits and risks of co-opetition.*

OVUM. *Virtual mobile services: Strategies for fixed and mobile operators.*

PTS. Swedish national post and telecom agency. Retrieved from http://www.pts.se

Varoutas, D., Katsianis, D., Sphicopoulos, T., Loizillon, F., Kalhagen, K. O., Stordahl, K., et al. (2003). Business opportunities through umts-wlan networks. *Annales Des Telecommunications-Annals of Telecommunications, 58*(3-4), 553-575.

Varoutas, D., Katsianis, D., Sphicopoulos,T., Stordahl, K., & Welling, I. (2006). On the eco-

nomics of 3G mobile virtual network operators (MVNOs). *Wireless Personal Communications, 36*(2), 129-142.

APPENDIX I

Cost Evolution of the Network Components

The cost prediction curve depends on a set of parameters such as reference cost at a given time, the learning curve coefficient that reflects the type of component, penetration at the starting time and penetration growth in the component's market. The cost database contains estimation on these parameters for all components and generates cost predictions based on the extended learning curve. The forecast function for the evolution of the relative accumulated volume $n_r(t)$ is illustrated in Equation (1) (Olsen 1996).

$$n_r(t) = \left(1 + e^{\left\{ h\left[n_r(0)^{-1} - 1 \right] - \left[\frac{2 \cdot h\ 9}{\Delta T} \right] \cdot t \right\}} \right)^{-1} \quad (1)$$

The expression for $n_r(t)$ can be substituted into a learning curve formula Equation (2) yielding the final expression for price versus time in the cost database.

$$P(t) = P(0) \cdot \left[n_r(0)^{-1} \cdot n_r(t) \right]^{\log_2 \cdot K} \quad (2)$$

where $n_r(0)$ is the relative accumulated volume in year 0. The value of $n_r(0)$ should be equal to 0.5 for components that exist in the market and their price is expected to be further reduced due to aging rather than due to the production volume (i.e., very old products--many years in the market). From estimations in industrial telecommunication network components, $n_r(0)$ could be 0.1 for mature products and 0.01 for new components in the market.

P(0) is the price in the reference year 0, ΔT is the time for the accumulated volume to grow from 10% to 90%, and K is the learning curve coefficient. K is the factor that causes reduction in price when the production volume is doubled. The K factor can be obtained from the production industry, mainly the suppliers. For a component (with constant nr(0)=0.1) when the ΔT is equal to 10 years and K is equal to 0.98, Equation (2) gives almost 2% of reduction in the price of the component per year for the first 10 years. If ΔT is five years, this reduction is almost 4% per year for the first five years. All the above described values have been extensively used (Ims, 1998) for the evaluation of telecommunications investment projects.

OA&M Approach

The OA&M approach is divided into three separate components. Conceptually, the three components are defined as follows:

1. The cost of repair parts.
2. The cost of repair work.
3. The Operation and Administration cost for each service cross-related to the number of customers or to the number of critical network components.

The formula for calculating OA&M cost is given by Equation (3) (Olsen 1999).

$$(OA\ \&\ M)_i = \frac{V_{i-1} + V_i}{2} \cdot \left(P_i \cdot R_{class} + P_l \cdot \frac{MTTR}{MTBR} \right) + OA \quad (3)$$

The first term in the parenthesis represents the cost of repair parts, the second term is the cost of repair work while OA represents the Operation and Administration cost. V_i is the equipment

volume in year i, P_i is the price of cost item in year i, Rclass is the maintenance cost percentage for every cost component, P_l is the cost of one working hour, MTTR is the mean time to repair for the cost item in question and MTBR is the mean time between failures for the cost item in question. In order to implement the calculation of the OA&M cost, classes for MTTR and MTBR are defined in the database of the Tool as well as cost for P_I and P_i.

DEMAND FORECASTS

A logistic model is used to perform demand forecasts. This model is recommended for long-term forecasts and for new services. To achieve a good fit, a four-parameter model, including the saturation level, is used.

The model is defined by the following expression:

$$Y_t = M / (1 + \exp (\alpha + \beta t))^{\gamma}$$

where the variables are as follows:

Y_t :Demand forecast at time t
M :Saturation level
t : Time
α, β, γ : Parameters

The parameters α, β, and γ cannot be estimated simultaneously by ordinary least-squares regression since the model is non-linear in the parameters. Instead, a stepwise procedure is used to find the optimal parameter estimates. The saturation level M is estimated, and is a fixed input to the forecasting model.

TECHNO-ECONOMIC TERMS

The objective of a business case for this network is to estimate investments, revenue, operating cost, general administration cost and taxes. The network is expected to generate revenue throughout the lifetime of the product. Depreciation, operating cost, general administration cost and taxes are deducted from the revenue stream. To assess the model, the cash flow is calculated by adding back the depreciation to the income (net). The business case is evaluated according to four conventional criteria: Net Present Value (NPV), Internal Rate of Return (IRR), cash balance and payback period.

The Net Present Value (NPV) is today's value of the sum of resultant discounted cash flows (annual investments and running costs), or the volume of money, which can be expected to receive over a given period of time. If the NPV is positive, the project earns money for the investor. It is a good indicator for the profitability of investment projects, taking into account the time value of money or opportunity cost, which is expressed in the discount rate (10 percent in most cases).

The Internal Rate of Return (IRR) is the interest rate calculated on an investment and income (resultant net cash flow) that occur over a period of time. If the IRR is greater than the discount rate used for the project, then the investment is judged to be profitable. This criterion is especially useful in comparing projects of different type and size. The Internal Rate of Return gives a good indication of "the value achieved" with respect to the money invested.

The Cash Balance curve (accumulated discounted Cash Flow) generally goes deeply negative because of high initial investments. Once revenues are generated, the cash flow turns positive and the Cash Balance curve starts to rise. The lowest point in the Cash Balance curve gives the maximum amount of funding required for the project. The point in time when the Cash Balance turns positive represents the Payback Period for the project.

This work was previously published in Strategies and Policies in Digital Convergence, edited by S. Park, pp. 50-69, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.21
A Mobile Portal Solution for Knowledge Management

Stefan Berger
Universität Passau, Germany

Ulrich Remus
University of Erlangen-Nuremberg, Germany

ABSTRACT

This chapter discusses the use of mobile applications in knowledge management (mobile KM). Today more and more people leave (or have to leave) their fixed working environment in order to conduct their work at changing locations or while they are on the move. At the same time, mobile work is getting more and more knowledge intensive. However, the issue of mobile work and KM is an aspect that has largely been overlooked so far. Based on requirements for mobile applications in KM an example for the implementation of a mobile KM portal at a German university is described. The presented solution offers various services for university staff (information access, colleague finder, campus navigator, collaboration support). The chapter is concluded by outlining an important future issue in mobile KM: the consideration of location-based information in mobile KM portals.

INTRODUCTION

Today many working environments and industries are considered as knowledge intensive, that is, consulting, software, pharmaceutical, financial services, and so forth. Knowledge management (KM) has been introduced to overcome some of the problems knowledge workers are faced by handling knowledge, that is, the problems of storing, organizing, and distributing large amounts of knowledge and its corresponding problem of information overload, and so forth. Hence, KM and its strategies aim at improving an organization's way of handling internal and external knowledge in order to improve organizational performance (Maier, 2004).

At the same time more and more people leave (or have to leave) their fixed working environment in order to conduct their work at changing locations or while they are on the move. Mobile business tries to address these issues by providing (mobile) in-

formation and communication technologies (ICT) to support mobile business processes. However, compared to desktop PCs, typical mobile ICT, like mobile devices such as PDAs and mobile phones, have some disadvantages (Hansmann, Merk, Niklous, & Stober, 2001):

- Limited memory and CPU – Mobile devices are usually not equipped with the amount of memory and computational power in the CPU found in desktop computers.
- Small displays and limited input capabilities – for example, entering a URL on a Web-enabled mobile phone is cumbersome and slower than typing with a keyboard.
- Low bandwidth – in comparison to wired networks, wireless networks have a lower bandwidth. This restricts the transfer of large data volumes.
- Connection stability – due to fading, lost radio coverage, or deficient capacity, wireless networks are often inaccessible for periods of time.

Taking into account the aforementioned situation one must question whether current IT support is already sufficient in order to meet the requirement of current knowledge-intensive mobile work environments. So far, most of the off-the-box knowledge management systems are intended for use on stationary desktop PCs and provide just simple access from mobile devices. As KMS are generally handling a huge amount of information (e.g., documents in various formats, multimedia content, etc.) the management of the restrictions described above become even more crucial. In addition, neither an adaptation of existing knowledge services of stationary KMS nor the development of new knowledge services according to the needs of mobile knowledge workers is taking place.

The goals of this chapter are to identify the main issues when mobile work is meeting knowledge management. In particular the focus lies on mobile knowledge portals, which are considered to be the main ICT to support mobile KM. Further on the applicability of these suggestions is shown with the help of a mobile knowledge portal that was implemented at a German university.

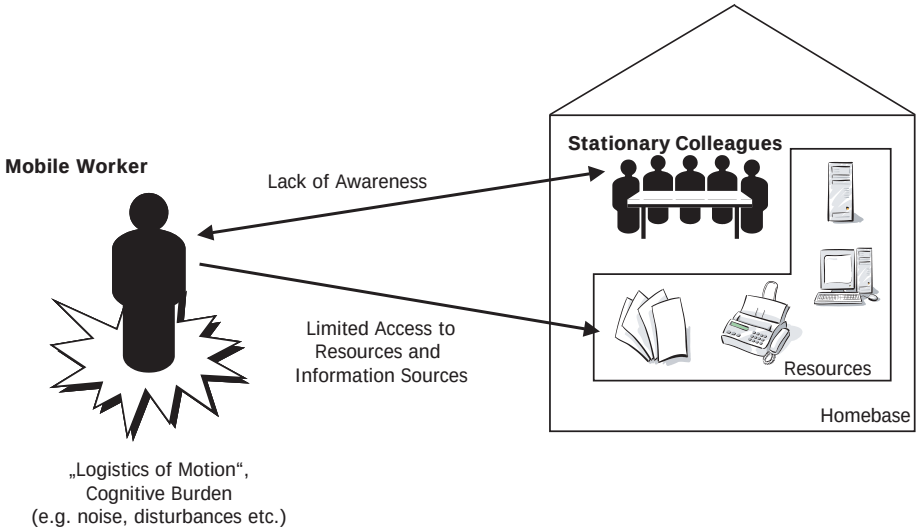
The chapter is structured as follows: Section two will detail the understanding about mobile KM and derive important requirements to be fulfilled. In section three mobile knowledge portals are then described as main ICT to support tasks in mobile KM. As an example the mobile KM portal of the University of Regensburg is presented (section four) whereas section five shows location orientation as the next step in mobile KM. Finally, section six concludes this chapter and gives an outlook on future research issues within the field of mobile KM.

KNOWLEDGE MANAGEMENT MEETS MOBILE WORK

A mobile working environment differs in many ways from desk work and presents the business traveler with a unique set of difficulties (Perry, O'Hara, Sellen, Brown, & Harper, 2001). In the last years several studies have shown that mobile knowledge workers are confronted with problems that complicate the fulfillment of their job (Figure 1).

Mobile workers working separated from their colleagues often have no access to the resources they would have in their offices. Instead, business travelers, for example, have to rely on faxes and messenger services to receive materials from their offices (Schulte, 1999). In case of time-critical data, this way of communication with the home base is insufficient. Bellotti and Bly (1996) show in their survey about knowledge exchange in a design consulting team that it is difficult for a mobile team to generally stay in touch. This is described as "Lack of Awareness." It means that a common background of common knowledge and shared understanding of current and past

Figure 1. Problems related to mobile work



activities is missing. This constrains the exchange of knowledge in teams with mobile workers. In addition, mobile workers have to deal with different work settings, noise levels, and they have to coordinate their traveling. This "Logistics of Motion" lowers their ability to deal with knowledge-intensive tasks (Sherry & Salvador, 2001) while on the move. The danger of an information overflow increases.

Mobile KM is an approach to overcome these problems. Rather than adding to the discussion of what actually is managed by KM—knowledge workers, knowledge, or just information embedded into context—in this chapter, mobile KM is seen as KM focusing on the usage of mobile ICT in order to:

- provide **mobile access** to KMS and other information resources;
- generate **awareness** between mobile and stationary workers by linking them to each other; and
- realize **mobile KM services** that support knowledge workers in dealing with their tasks (Berger, 2004, p. 64).

The next section reviews the state of the art of KMS and reviews if it meets these requirements.

MOBILE KM PORTALS

Currently, many KMS are implemented as centralistic client/server solutions (Maier, 2004) using the portal metaphor. Such knowledge portals provide a single point of access to many different information and knowledge sources on

the desktop together with a bundle of KM services. Typically, the architecture of knowledge portals can be described with the help of layers (Maier, 2004). The first layer includes data and knowledge sources of organizational internal and external sources. Examples are database systems, data warehouses, enterprise resource planning systems, and content and document management systems. The next layer provides intranet infrastructure and groupware services together with services to extract, transform, and load content from different sources. On the next layer, integration services are necessary to organize and structure knowledge elements according to a taxonomy or ontology.

The core of the KMS architecture consists of a set of knowledge services in order to support discovery, publication, collaboration, and learning. Personalization services are important to provide a more effective access to the large amounts of

Figure 2. Tasklist, Calendar, and Discussion Board of Open Text's Livelink Wireless (Open Text, 2003, p. 12)

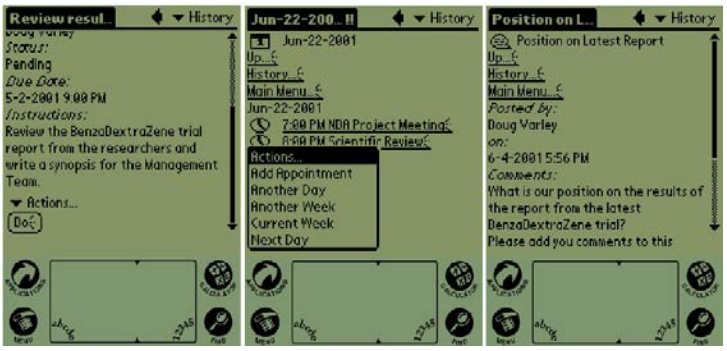


Figure 3. Automatic text summarization (Open Text, 2003, p. 11)

content, that is, to filter knowledge according to the knowledge needs in a specific situation and offer this content by a single point of entry (portal). In particular, personalization services together with mobile access services become crucial for the use of KMS in mobile environments.

Portals can be either developed individually or by using off-the-shelf portal packages, for example, Bea WebLogic, IBM Portal Server, Plumtree Corporate Portal, Hyperwave Information Portal, or SAP Enterprise Portal. These commercial packages can be flexibly customized in order to build up more domain-specific portals by integrating specific portal components (so called portlets) into a portal platform. Portlets are more or less standardized software components that provide access to a various amount of applications and (KM) services, for example, portlets to access ERP-systems, document management systems, personal information management.

In order to realize mobile access to knowledge portals, portlets have to be implemented as mobile portlets. That means that they have to be adapted according to technical restrictions of mobile devices and the user's context. At the moment, commercial portal packages cannot fulfill sufficiently the needs of mobile KM. Most of the systems are enhanced by mobile components, which are rather providing mobile access to stationary KM services instead of implementing specific mobile KM services.

Hyperwave's WAP (Wireless Application Protocol) Framework, for example, enables mobile users to browse the Hyperwave Information Portal with WAP-enabled devices. The Wireless Suite of Autonomy is a WAP-based solution with the focus on awareness-generating features such as peoplefinder and community support.

At present, the most comprehensive support for mobile KM is provided by the Livelink portal from Opentext Corporation. With the help of the Wireless Server users can access discussion boards, task lists, user directories (MS Exchange, LDAP, Livelink User Directory), e-mail, calendar, and documents (Figure 2). In addition, it provides some KM services specially developed for mobile devices, for example, automatic summarization of text. Hence even longer texts can be displayed on smaller screens (Figure 3).

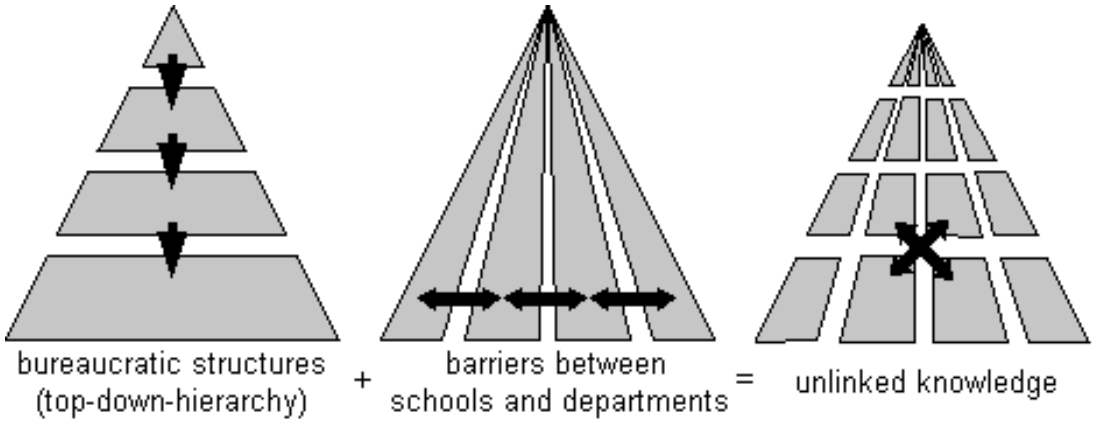
EXAMPLE: A MOBILE KM PORTAL FOR A GERMAN UNIVERSITY

In recent years German universities, which are financed to a large extent by public authorities (federal states and federal government), have been severely affected by public saving measures. As a result lean, efficient administrative procedures are more important than ever. KM can help to achieve these objectives. One example is to provide easy accessible expert directories, where staff members with certain skills, expertise, and responsibilities can be located ("Person XY is responsible for third-party funding") in order to support communication and collaboration.

However, there are several reasons why the access to information of this type is limited at the University of Regensburg. First, there is the decentralized organizational structure. All together about 1,000 staff members are working in 12 different schools and about 15 research institutes at the university, serving about 16,000 students. Because most of the organization units are highly independent, they have their own administrations and the exchange of knowledge with the central administration is reduced to a minimum. Likewise there is hardly an exchange of knowledge between different schools and departments. As a result, knowledge that would be useful throughout the whole university is limited to some staff members ("unlinked knowledge," Figure 4).

A second problem is that many scientific staff members work on the basis of (short-term) time contracts. This leads to an increasing annual labor turnover, comparable to the situation that consulting companies are facing. Important knowledge about past projects, courses, and

Figure 4. Unlinked knowledge because of independent organization structures



scientific results is lost very easily. Due to this fact, a high proportion of (new) staff members are relatively inexperienced to cope with administration processes, which can be described as highly bureaucratic and cumbersome.

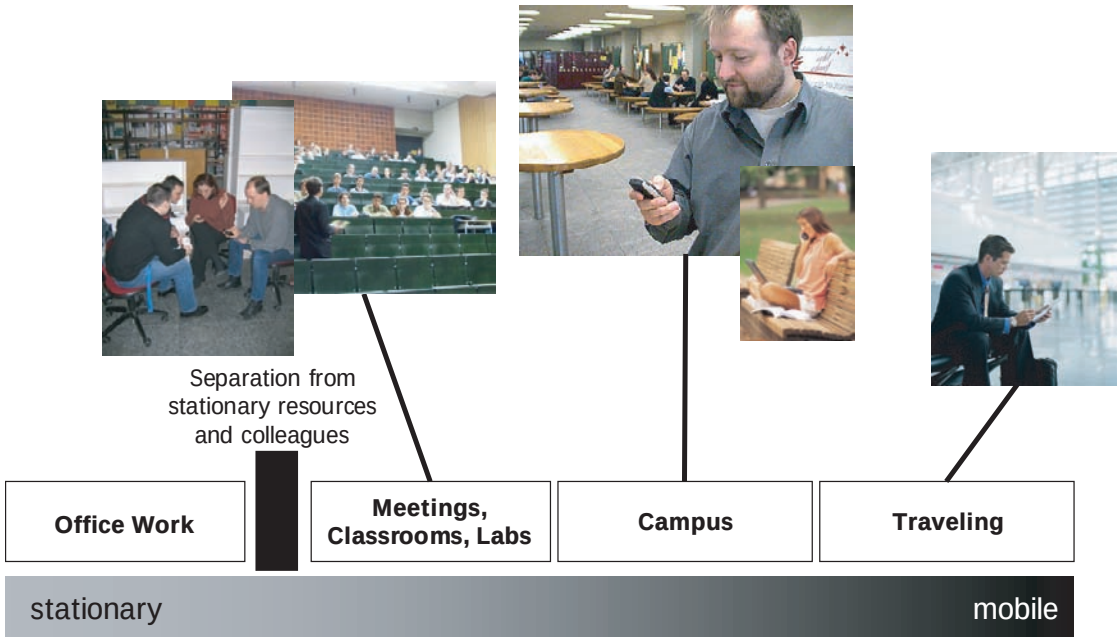
To overcome these problems—the lack of communication between departments and the need to provide specific knowledge (i.e., administrative knowledge) for staff members—the University of Regensburg decided to build up a knowledge portal called U-Know (Ubiquitous Knowledge). U-Know is meant to be a single point of access for all relevant information according to the knowledge needs described above. When conducting a knowledge audit it became obvious that a large amount of knowledge is needed when knowledge workers are on the move, that is, working in a mobile work environment. Staff is frequently commuting between offices, meeting rooms, laboratories, home offices; they attend conferences; and sometimes they are doing field studies (e.g., biologists or geographers).

Hence the picture of one single resource-rich office has to be extended towards different working locations, where a large number of knowledge-intensive tasks are carried out as well (Figure 5). Consequently the considered solution should meet these "ubiquitous" knowledge needs of current work practices at a university.

The portal should support staff members by managing the following:

1. **Documented knowledge**: A knowledge audit was conducted in order to obtain a better picture of knowledge demand and supply. This was mainly done with the help of questionnaires and workshops where staff members were asked to assess what kind of (out-of-office) information is considered as useful.
2. **Tacit knowledge**: In order to support the exchange of tacit knowledge (which is difficult to codify due to the fact that this knowledge lies solely in the employees' heads, often embedded in work practices and processes), the considered KM solution should enable communication and cooperation between staff members.

Figure 5. Knowledge demand in "mobile" situations



In order to meet these requirements U-Know should offer the KM services in Figure 6.

The services can be categorized into information, communication, collaboration, and search. The first category comprises all services that are responsible to manage simple information in the knowledge base. By invoking these services staff members obtain the information they need to perform their daily tasks, for example, news, notifications about changes in rooms, or phone numbers. A very important part of this section is the yellow pages (Figure 7) where all staff members are listed. This list can be browsed by names, departments, fields of research, and responsibilities.

Frequently asked questions (FAQ) answer questions that are typically asked by new staff members. The Campus Navigator helps locate places and finding one's way around the campus. Each room at the university carries a doorplate with a unique identifier. After entering a starting point in form of the identifier and a destination in form of the name of a person, of an office (e.g., "Office for Third-Party Fundings," "Academic Exchange Service"), or just another room number, the shortest way to the destination is calculated and shown on maps of different sizes (Figure 8).

Communication-oriented features like e-mail, short message service (SMS), and discussion boards are intended to support the exchange of tacit knowledge between staff members.

To foster collaboration, for example, in temporary project groups, staff members can initiate workgroups by inviting colleagues via SMS or e-mail to join a virtual teamspace. After forming a workgroup the participants can use their

Figure 6. Features of U-Know

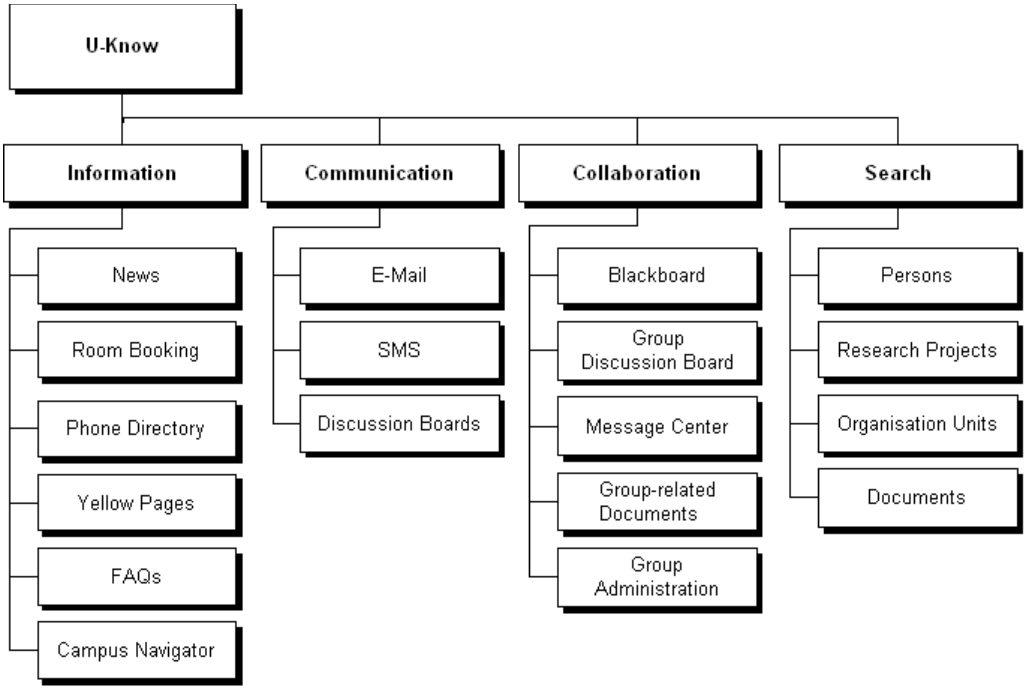


Figure 7. U-Know yellow pages

Figure 8. U-Know Campus Navigator

teamspace for (electronic) group discussions and sharing documents. The blackboard displays all recent events, including new group members, new files, discussion entries, and administrative actions that are taken. In the search section queries can be limited to persons, research projects, organization units, or documents.

To support different networks there are several ways to access the portal. University staff can use the campuswide Wi-Fi network with Wi-Fi–capable devices. Users can also deploy a mobile phone and access the portal via a GSM-network and the Wireless Application Protocol (WAP). Hence it is possible to use the portal even when users are outside the university, for instance, at a conference. The phone directory or the yellow pages can be accessed via voice as the entry of longer words may be cumbersome in many situations. An integrated speech-recognition system "translates" the user's spoken words into database requests and the results back into speech.

LOCATION ORIENTATION AS NEXT STEP IN MOBILE KM

Generally, there is agreement about the distinction between human- and technology-oriented KM approaches which basically reflects the origin of the approaches. KM research should try to bridge the gap between human- and technology-oriented KM. Many authors have propagated a so-called "holistic" approach to KM. However, so far these authors leave it to the interpretation of the reader what such an approach might look like. The examples in the last column of Table 1 should be seen as a step towards detailing this approach which is called "bridging the gap" KM. In Table 1 this classification (Maier, 2004; Maier & Remus, 2003) is enhanced towards the consideration of mobile KM. As mobile KM is mainly focusing on instruments and systems, other dimensions like strategy, organization, and economics are not considered in this table.

Table 1. Mobile KM approaches (gray highlighted cell is covered by U-Know)

	Technology-oriented instruments and systems	**Human-oriented instruments and systems**	**Bridging the Gap instruments and systems**
Mobile Access	Mobile access to content, e.g. knowledge about organization (e. g. Campus Navigator), processes, products, internal studies, patents, online journals by using mobile devices focusing on services for presentation (e.g. summarization functions, navigation models) and visualization	Mobile access to employee yellow pages, skill directories, directories of communities, knowledge about business partners using mobile devices focusing on asynchronous E-Mail, Short Message Service (SMS) and synchronous communication (Chat), collaboration and cooperation, community support	Mobile access to ideas, proposals, lessons learned, best practices, community home spaces (mobile virtual team spaces), evaluations, comments, feedback to knowledge elements using mobile devices focusing on profiling, personalization, contextualization, recommendation, navigation from knowledge elements to people
Location-orientation	Adaptation of documented knowledge according to the user's current location	Locating people according to the user's location, e.g. locating colleagues, knowledge experts	Personalization, profiling according to the user's location and situation, providing proactive mobile KM services

In order to structure mobile KM, one can distinguish two dimensions: mobile access and location orientation. Mobile access is about accessing stationary KMS whereas location orientation explicitly considers the location of the mobile worker. The field of location-oriented KM draws attention from research in mobile KM, ubiquitous computing, location-based computing, and context-aware computing (Lueg & Lichtenstein, 2003).

So far, the implemented solution provides mobile access to a broad range of different knowledge sources in a mobile work environment. University staff can use the KM services provided by U-Know in order to access information, to find colleagues, to navigate the campus, to collaborate, and so forth. These KM services mainly support the human-oriented KM approach. In fact, typical knowledge services were adapted with regard to the characteristics of mobile devices, that is, small display, bandwidth, and so forth.

However, an adaptation of these services according to the user's location has not taken place yet, whereas a customization of services according to the location of the user would enable a mobile knowledge portal to supply mobile knowledge workers with appropriate knowledge in a much more targeted way. At the same time,

information overload can be avoided, since only information relevant to the actual context and location is filtered and made available. Think of a researcher who is guided to books in a library according to his/her own references but also according to his/her actual location.

Currently, common "stationary" knowledge portals are ill-suited to support these new aspects of KM derived from a location-oriented perspective (Berger, 2004). One reason is that the context, which is defined by the corresponding situation (tasks, goals, time, identity of the user) is still not extended by location-oriented context information (Abecker, van Elst, & Maus, 2001).

Location-oriented knowledge services could contribute to

- **More efficient business processes**: Shortcomings arising from mobility can be compensated by considering location-oriented information. Times for searching can be reduced due to the fact that information about the location might restrict the space of searching (e.g., an engineer might get information about a system that he/she is currently operating). Possibly, redundant ways between mobile and stationary work place are omitted when the information is already provided on the move.
- **Personalization**: When considering the user's location information can be delivered to the user in a much more customized and targeted way (Rao & Minakakis, 2003). For example, an engineer in a production hall is seeking information about outstanding orders, whereas close to machines he might need information about technical issues or repair services. In addition, location-oriented information might be helpful to locate other "mobile" colleagues who are nearby.
- **New application areas**: The integration of common knowledge services together with location-oriented mobile services may also extend the scope for new applications in KM, for example, the use of contextual information for the continuous evolution of mobile services for mobile service providers (Amberg, Reus, & Wehrmann, 2003). One can also think of providing a more "intelligent" environment where information about the user's location combined with sophisticated knowledge services adds value to general information services (e.g., in museums, where customized information to exhibits can be provided according to the user's location).

To build up mobile knowledge portals that can support the scenario described above, mobile portlets are needed that can realize location-oriented KM services. In case of being implemented as proactive services (in the way that a system is going to be active by itself), these portlets might be implemented as push services. In addition, portlets have to be responsible for the import of location-oriented information, the integration with other contextual information (contextualization), and the management and exploitation of the location-oriented information. Of course, the underlying knowledge base should be refined in order to manage location-oriented information.

With respect to mobile devices, one has to deal with the problem of locating the user and sending this information back to the knowledge portal. Mobile devices might be enhanced with systems that can automatically identify the user's location. Depending on the current net infrastructure (personal, local, or wide area networks), there are many possibilities to locate the user, for example, Wi-Fi, GPS, or radio frequency tags (Rao & Minakakis, 2003).

CONCLUSIONS AND OUTLOOK

The example of U-Know shows some important steps towards a comprehensive mobile KM solution. With the help of this system it is pos-

sible to provide users with KM services while being on the move. With its services like yellow pages, messaging features, and so forth, it creates awareness among remote working colleagues and thus improves knowledge sharing within an organization.

With respect to the acceptance of U-Know, two user groups can be distinguished. The first group is characterized by users who already own a mobile device, especially a PDA, in order to organize their appointments and contacts (personal information management). They are the main users of the system because they perceive the additional KM-related services as an extension of the capabilities of their devices. In contrast, staff members who did not use mobile devices for their personal information management are more reluctant to adopt the new system.

The Wi-Fi access soon became the most popular way of accessing the system. This is because of several reasons. Most of the staff members are actually working on the campus and the Wi-Fi access is free of charge for university members. Another reason is probably the higher bandwidth (and therefore faster connections) of Wi-Fi in comparison to a GSM-based access via WAP. Nevertheless, it can be assumed that decreasing connection fees and higher bandwidths of 3G-Networks (UMTS) would encourage staff to use the system from outside the university.

However, in order to fully meet the requirements of mobile KM in the near future, mobile KM portals have to be enhanced with mobile knowledge services that consider location-oriented information. Current work needs once more to address the adaptation of mobile services, the consideration of the user and work context for KM, and the design of highly context-aware knowledge portals.

REFERENCES

Abecker, A., van Elst, L., & Maus, H. (2001, July 13–16). *Exploiting user and process context for knowledge management systems.* Workshop on User Modeling for Context-Aware Applications at the 8th International Conference on User Modeling, Sonthofen, Germany.

Amberg, M., Remus, U., & Wehrmann, J. (2003, September 29–October 2). Nutzung von Kontextinformationen zur evolutionären Weiterentwicklung mobiler Dienste. *Proceedings of the 33rd Annual Conference "Informatics 2003," Workshop "Mobile User - Mobile Knowledge - Mobile Internet,"* Frankfurt, Germany.

Belotti, V., & Bly, S. (1996). Walking away from the desktop computer: Distributed collaboration and mobility in a product design team. *Proceedings of CSCW '96* (pp. 209–218). Boston: ACM Press.

Berger, S. (2004). *Mobiles Wissensmanagement. Wissensmanagement unter Berücksichtigung des Aspekts Mobilität.* Berlin: dissertation.de.

Grimm, M., Tazari, M.-R., & Balfanz, D. (2002). Towards a framework for mobile knowledge management. *Proceedings of the Fourth International Conference on Practical Aspects of Knowledge Management 2002* (PAKM 2002), Vienna, Austria.

Hansmann, U., Merk, L., Niklous, M.S., & Stober, T. (2001). *Pervasive computing handbook*. Berlin: Springer.

Lueg, C., & Lichtenstein, S. (2003, November 26–28). *Location-oriented knowledge management: A workshop at the Fourteenth Australasian Conference on Information Systems (ACIS 2003),* Perth, Australia.

Maier, R. (2004). *Knowledge management systems, information and communication technologies for knowledge management.* Berlin: Springer.

Maier, R., & Remus, U. (2003). Implementing process-oriented knowledge management strategies. *Journal of Knowledge Management, 7*(4), 62–74.

Open Text Corporation. (2003). *Livelink Wireless: Ubiquitous access to Livelink Information and Services* (White paper). Waterloo, Canada: Author.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: understanding access anytime, anywhere. *ACM Transactions on Human-Computer Interaction, 8*(4), 323–347.

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61–65.

Schulte, B.A. (1999). *Organisation mobiler Arbeit. Der Einfluss von IuK-Technologien.* Wiesbaden, Germany: DUV.

Sherry, J., & Salvador, T. (2001). Running and grimacing: The struggle for balance in mobile work. *Wireless world: Social and interactional aspects of the mobile age* (pp. 108–120). New York: Springer.

This work was previously published in Unwired Business: Cases in Mobile Business, edited by S. Barnes and E. Scornavacca, pp. 173-186, copyright 2006 by IRM Press (an imprint of IGI Global).

Chapter 6.22
Strategies of Mobile Value-Added Services in Korea

Jin Ki Kim
Korea Areospace University, Korea

Heasun Chun
The State University of New York at Buffalo, USA

ABSTRACT

As the growth of the mobile market decreases and the market competition intensifies, mobile carriers have been trying to find new business models to retain their profits and expand their business boundaries. Development of value-added services increases the chances of keeping the growth with mobile carriers. This chapter discusses the motivation of mobile value-added service in terms of value chain and mobile adoption. Six mobile value-added services presented in Korea are introduced: (1) short messaging service (SMS), (2) personalized call-ring service, (3) mobile music service, (4) mobile video service, (5) mobile payment (m-payment), and (6) mobile games. The major characteristics of those value-added services are discussed with "4Cs": (1) customization, (2) content-focused, (3) connectedness, and (4) contemporary. This chapter also discusses digital multimedia broadcasting (DMB) as a new value-added service and the impacts of value-added services on the mobile market. This chapter is concluded with three plausible strategies of mobile carriers: (1) real-time, market-responding strategy, (2) content-focused market strategy, and (3) various bundling service.

INTRODUCTION

Worldwide, the number of mobile subscribers reached 1.7 billion in 2004 (International Telecommunications Union [ITU], 2006). The compound annual growth rate (CAGR) from 1980 to 2004 is 59.54%. The number of subscribers keeps increasing due to the increase of subscription in the under-developed and developing countries. How-

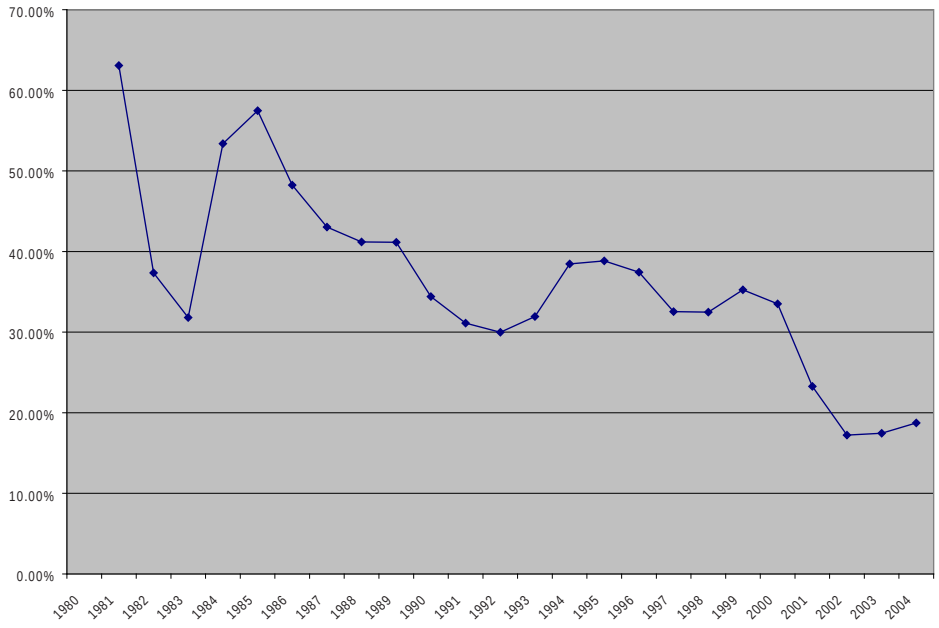
ever, recently the growth rate of subscription has decreased. Since 2002, the growth rates dropped to under 20% (See Figure 1). It means that the mobile service market is approaching the mature stage. In several European and Asian countries, penetration ratios are around 80-100%. According to ITU World Telecommunications Indicator 2004, 45 out of 170 countries which reported the penetration ratio of mobile service shows more than 70% (ITU, 2006).

From the perspective of a mobile carrier, the reduced growth rate means a decrease in potential profits. In terms of market competition type, the decrease of growth rate leads to retention-based competition, not to subscription-based competition. From the financial view, the margin would be reduced. Mobile carriers would concentrate on retention of subscribers and on an increase of billing payments per subscriber. For this reason, the focus is now on average revenue per user (ARPU) and attempts are being made to increase ARPU by introducing premium services. Premium services are defined as services that provide added value for which the service provider can charge a premium (Brenner, Grech, Torabi, & Unmehopa, 2005).

What kinds of value-added services can be technically provided? Can they contribute to the profit of mobile carriers? And which kinds of comparative strategies can make sense in the market? Those questions become major issues which should be answered regarding those value-added services.

In this chapter, current trends and strategies of value-added services to keep or increase ARPU of customers for mobile carriers are discussed. This chapter is structured into seven sections as follows: in the second section the motivations of

Figure 1. Growth rate of number of mobile subscribers worldwide



mobile value-added services have been discussed, in terms of value chain and mobile adoption. The case studies on the current value-added services form the content of the third section, including SMS, personalized call-ring service, mobile music service, mobile video service, m-payment, and mobile games. The fourth section highlights the characteristics of current trends of value-added services. DMB for a new value-added service is introduced in the fifth section. In the sixth section, the impacts of value-added services on the mobile market are discussed. Concluding remarks with plausible strategies are presented in the final section.

MOTIVATIONS OF MOBILE VALUE-ADDED SERVICES

The telecommunications industry is structured by the economic, regulatory, and technical aspects. The shift from second generation (2G) to third generation (3G) mobile induced several changes in those aspects.

From the economic perspective, the costs of standardization, R&D, the significant costs of the licenses for spectrum, the possibility for network sharing, and the uncertainty surrounding the potential revenue streams for 3G mobile are major concerns. Licensing of spectrum, competition policy, and network sharing agreements are influential factors in the view of regulation. From the view of the technology, the evolution of mobile services has two components. First, voice-only has changed into multimedia-capable communications since the 3G mobile network has more capacity which is devoted to data communications. Data communications on the mobile network have larger portions than before. Second, the closed and dedicated network moved to the open network which is based on the Internet. The 3G mobile network is based on an all IP network.

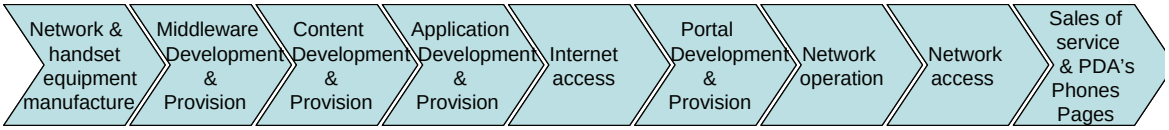
Those kinds of changes impact the value chain of the mobile service industry which is characterized by a more complex and multi-faceted production platform (or industry value chain). It is necessary to understand the value chain of the mobile industry and its trends of changing in the future in order to check the potential growth of mobile communication services (Maitland, Bauer, & Westerveld, 2002; Sabat, 2002; Steinbock, 2003). Figure 2 shows an example of the value chain of 3G mobile services.

In the market which is diverse and multi-faceted, mobile carriers can take two different approaches: product differentiation and market integration.

Regarding on product differentiations, there are three kinds of options: horizontal, vertical, and cross-market differentiation (Geng & Whinston, 2001).

- **Horizontal differentiation:** Sellers can avoid a price war by targeting different consumer groups with various preferences, tastes, or occupations

Figure 2. Third generation value chain (Maitland et al., 2002)



- **Vertical differentiation:** Sellers can differentiate their products in a way that consistently affects all consumer valuations
- **Cross-market differentiation by tying:** A seller can use a bundling strategy and its monopoly in one market to attack competitors in another

There are two kinds of integration: vertical and horizontal integration. Mobile carriers can take the benefit from the economy of scale from the horizontal integration. From the view of vertical integration, the development of various value-added services contributes to mobile carriers' leading roles in the new value chain. Multi-faceted and specialized value chain means a possibility of the loss of competitive advantage which mobile carriers had held for a long time. The possibility could lead to losing the leadership in the value chain, and, as the result, losing the market and the profitability. Thus, in order to keep their market leadership, mobile carriers should integrate adjacent components in the value chain.

Provision of value-added services is the first step to forward and backward expansion of business domains in the value chain. Through the process, mobile carriers can learn how to survive in the market in which creativity is a major competitive advantage. Lessons from the learning process can grant mobile carriers a useful chance of expanding their business scope into a broad media industry.

Another clue that shows the importance of value-added services is found in the literature of mobile adoption. Major influential factors on mobile adoption include call quality, tariff level, handsets, brand image, income, and subscription duration (H.-S. Kim & Yoon, 2004; J. K. Kim, 2005; M.-K. Kim, Park, & Jeong, 2004; Middleton, 2002). Switching cost and switching promotion are also found to be critical factors (M.-K. Kim et al., 2004). Under mobile number portability (MNP), the quality of service and price had more affect on customers' intention to retain or churn than before launching the MNP (J. K. Kim, 2005).

Mobile value-added services is one of the components to measure the customer satisfaction for mobile services (J. Lee, Lee, & Feick, 2001). In addition, perceived usefulness, ease of use, price, and speed of use are known as the most important determinants of adoption of multimedia mobile services (Pagani, 2004). In recent studies value-added service is included as one of influential factors (H.-S. Kim & Yoon, 2004; J. K. Kim, 2005; M.-K. Kim et al., 2004). The market needs have shifted from fulfilling their basic needs, such as call quality, tariff, and handset, to satisfying their upgraded needs, such as various service features, personalized services, and qualified customer services.

Therefore, how to develop value-added service, how to implement them, and how to react to changes in customer preferences will be critical factors that influence the performance of mobile carriers and their strategic positions for future business. To capture strategic implications for developing value-added services, in the following section, we discuss the experiences in the Korean mobile market in which various value-added services were launched and are being developed.

Table 1. SMS revenues in Korea mobile market (billion dollars) (Source: K.-M. Lee, 2005)

Corporate	2003	2004
SK Telecom	1.87	2.47
KTF	1.03	1.17
LG Telecom	0.43	0.61
Total	3.33	4.25

MOBILE VALUE-ADDED SERVICES IN KOREA

Short Messaging Service (SMS)

SMS is a text communication available on mobile phones that permits the sending of short messages. Once a message is sent, it passes through a Short Message Service Center (SMSC) to reach a roaming customer. Multimedia Messaging Service (MMS) is an advanced messaging service of SMS. It extends text messaging to include various multimedia data, such as longer texts, image clips, audio, or video clips. Currently, MMS is popularly used to transmit multimedia data from camera phones to other mobile phones or Internet accounts.

Due to its capacity of transmitting multimedia data through a mobile network, MMS can be applied to various business items. For example, a mobile printing service of camera phone photos is gaining popularity with the development of camera phones. The users who take a picture by camera phones can send their photos by following directions on the browser. Another trend in MMS is convergence with messenger services via fixed communication networks. *Cool Shot*, a joint PC-mobile SMS service of KTF, allows the customer to simultaneously check messages and reply through both SMS and PC pop-up windows (KTF, 2005e). Even when customers do not have SMS-enabled phones, they send their messages by typing text that will be converted as a voice message in the Internet messenger programs. Recently, SMS and MMS offer online billing and payment services in association with Internet banking systems. A customer who registers his/her accounts on Internet banking systems or the bill requester's server can receive electricity or gas bills and confirm the payment through SMS. It has a strong potential in customers' convenience because it does not need to have m-payment chips in their phones. In the near future, MMS is expected to replace SMS, which provides new opportunities to maximize revenues in the value chain of the mobile industry.

SMS and MMS are very rapidly developing from 2002, the first year of MMS services. MMS is expected to continue its sharp increase at an average growth rate of 108.4% from 2002 to 2007. In 2004, the revenue of SMS was 4.2 billion dollars and the number of SMS messages was 332 billion, which was increased by 27.5% and 31.5%, respectively, from a year ago (K.-M. Lee, 2005).

The rapid growth of SMS in Korea is related to various payment plans for heavy users of SMS. KTF and SK Telecom launched *Bigi Egg Unlimited Text Price Plan* and *Ting Text Price Plan*, respectively, to cater to the trend of teenagers who prefer text-messaging to voice communication, which allows a customer to adjust the rate of phone calls and SMS at $.02 per SMS and $.03 per 10 second voice call in his/her price plans (monthly price ranging from 14 to 26 dollars), according to the users needs. SMS are particularly popular among teenagers and young adults. According to Consumer Protection Board (2004), 23% of teenagers are heavy users of SMS, sending over 50 messages per one month, and 87% of teenagers are sending over 10 text messages to their friends and families. The average number of SMS per user is 29.11 per month. The ARPU of SMS was monthly $2.6 per customer.

Personalized Call-Ring Service

Ring-back tone is typically used to refer to the audible ringing that is heard on the telephone line by the calling party after dialing and prior to the call being answered at the distant end (Wikipedia, 2006b). Recently this form of ring-back tone has transformed as "personalized call-ring service." With personalized call-ring service, callers will hear an audio selection applied to the telephone line that has been previously determined by the called party. Personalized call-ring service is a kind of value-added service which customers can

choose their call rings, such as music, voice, and sound instead of providing a simple mechanic ring-back tone in general.

Personalized call-ring service is operated by servers of mobile carriers. Equipment is installed in the telephone network to enable replacement of the standard ring-back tone with a personalized audio selection. Mobile carriers keep their music source codes which come from content providers. When a user selects a certain music source code, a database of the mobile carrier keeps the sound source code. When a request has been made, the database queries servers by the code and then the sound source is provided to the caller.

The personalized call-ring service is called *Coloring* in Korea because the personalized sound makes personality colored. There are several brand names for that service, such as *Coloring, Tooling, Ring to you*, *Feeling, Ringo*, and so on. Among them, *Coloring* is the popular name due to that is the first provided brand name. *Coloring* has the largest service which has 8.2 million paying subscribers as of 2006. Users can choose their own sounds by their preferences. They also select sounds by time and numbers of the person called. It is very interesting that *Coloring* which sends sounds to the called party, is more popular than ring-back to which the caller listens. In 2005 the numbers of downloading *Coloring* was 6.8 million which is much more than ring-back which has 4.7 million.

Experience of implementing personalized call-ring services grants Korean mobile carriers a chance of exploiting the international market. Table 2 shows some cases of exporting personalized call-ring services by Korean mobile carriers.

KTF creates a new concept in karaoke with the release of *Magic, Chilo, Joy*, which customize the phone with a song sung by the user through on-line/off-line and fixed/wireless networks. It allows high-quality MP3 musical accompaniment and the option to send karaoke ring tones as a present to another user. Service grows into customer-participation content services in wallpapers, ring-back tones, and so forth (KTF, 2005f).

Personalized call-ring service is contributing to mobile carriers' financial performance. Three mobile carriers have revenues of $8-20 million in 2005. According to a study, the World Cup 2006 is seen as an opportunity to promote 3G which will generate $6.35 billion in revenue, with text-based services and downloads, such as ring tones and logos (3GNewsroom.com, 2006).

As the functionality of mobile handsets has been improved, higher quality of services can be provided. Mobile carriers are trying to develop higher technology. For example, SK Telecom

Table 2. Some cases of exporting personalized call ring services by Korean mobile carriers (Source: KTF, 2005c; SKTelecom, 2004a)

Date	Mobile carrier	Imported mobile carrier (country)	Deal size (million dollars)
Apr-03	SK Telecom	S-Telecom (Vietnam)	1.7
Jun-03	SK Telecom	Mobile-1, SingTel (Singapore)	3
Dec-03	SK Telecom	Smart (Philippines)	1.5 (additional 3.0)
Jul-04	SK Telecom	Telkomsel (Indonesia)	1.5
2005	KTF	PT Mobile-8 Telecom (Indonesia)	2

Table 3. Digital music market in Korea (Source: Music Industry Association of Korea [MIAK], 2005)

	2000	2001	2002	2003
Ring back and Call ring services	30.6	62.7	129	176.8
Streaming (WEB, MP3)	9.4	18.8	3.6	4.4
Others (VoD, Mobile)	9.5	8.6	1.5	3.8
Sum (million dollars)	49.5	90.1	134.1	185

reached an agreement for jointly developing an audio CODEC technology with Coding Technologies (CT) of Germany to increase the service quality of its *Coloring* service (SKTelecom, 2006).

Ring-back tone and call-ring services have a major portion of the digital music market in Korea. The market increased by approximately 80% annually and has about 95% share in this market (see Table 3).

Mobile Music Service

Mobile music service refers to a value-added service of mobile telephone service, which users can download music files into their mobile music service-enabled devices. PC, MP3 phone, and MP3 player are popular mobile music service-enabled devices. PC supports download and streaming services through the mobile music Web site and mobile music players. Through an MP3 phone, users can enjoy music by transmitting music files downloaded through its Web site to their mobile phone. When users connect to the wireless Internet service on their mobile phones, users download, stream, and search for the music they want to enjoy. Users can also listen to music by receiving the music files they want in the mobile music service-enabled MP3 players.

Three Korean mobile carriers started providing their own mobile music services, such as *MelOn* (SK Telecom), *Dosirak* (KTF), and *musicOn* (LG Telecom) from November 2004, May 2005, and July 2005 respectively. *MelOn* utilizes Digital Right Management (DRM) technology which prevents illegal distribution and use of wired and wireless integrated networks, platforms, and digital content (SKTelecom, 2004c). *MelOn* service is provided by pay-per-downloading and by monthly flat rate. The number of paying subscribers reached more than 600,000 as of December 2005. *Melon*

Table 4. Mobile music market in Korea (Source: Daishin Security, 2005)

	2004*	2005**	2006**	2007**	2008**
Revenue of mobile carriers by mobile digital music (million dollars)	80	88	100	120	140

* *Estimated;* ** *Forecasted*

Shop, a one-stop shopping mall in which customers can purchase items related to music, opened on December 2005 (SKTelecom, 2005a).

KTF launched its music portal service, which offers a unified service allowing users not only to listen to both Korean and foreign music, but to also spice up their phones with ring tones and callback tones. It has a 900,000-tune database and digital rights for 480,000 tunes. KTF has contracts with 90% of Korea's music property rights owners (KTF, 2005d). KTF has attracted 350,000 members to *Dosirak* ("lunchbox" in English) just 2 months after its release; 120,000 are paying subscribers among them (KTF, 2005b).

Most young singers first release their music on the mobile music market. Music producers can gauge the success of a new single through the mobile music market. In addition, technology is advancing to provide various high quality services. A Portable Multimedia Player (PMP) phone that lets users enjoy audio and streaming video through a mobile phone was first made available in November 2005 (SKTelecom, 2005d).

Mobile music market of mobile carriers in Korea is in the growing stage. Annually the increase by $20 million will be forecasted.

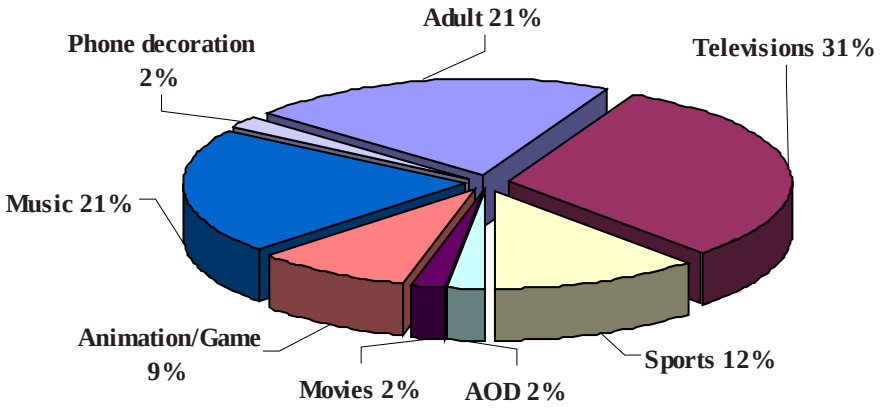
Mobile Video Service

3G networks offer the capacity and capability to transmit richer mobile multimedia to users, such as video phone functions, multimedia messages, music on demand (MoD), video on demand (VoD), TV broadcasting, and the Internet. Korea launched the first commercial CDMA 2000 1x EV-DO service in the world that ensures 144Kbps data speed for LAN-quality video streaming service. Korea currently experiences the transfer toward W-CDMA that ensures DVD-quality video streaming services through mobiles.

Fimm and *June*, as KTF's and SK Telecom's brand names of premium multimedia services, provide varied multimedia content menus of movie channels with downloadable mobile movies; music channels with music videos and the latest music; and broadcasting channels to receive TV programs.

Sports, adult videos, and short soap operas are most popular contents on mobiles. Retransmission of existing broadcasting channels is also available on mobiles, including terrestrial, cable, and digital satellite channels as real-time services. These services account for 66% of the total

Figure 3. Fimm's mobile video revenue (Source: Song, 2005)



revenue of value-added services of KTF. Figure 3 shows *Fimm*'s revenues of mobile multimedia service sector in the fiscal year of 2004. The most dominant revenue resource, mobile television channels, provided 31% of *Fimm*'s mobile multimedia revenue. The second largest sector was adult services and music services, each of which reached 21% of total revenue. This financial result may prove that streaming services are more profitable revenue resources than simple downloadable services, because the revenue of audio-on-demand was only 2%, whereas music streaming service was 21% of total revenue.

In order to retain subscribers, each network operator should secure various contents to cater to the customer's taste. Therefore, each network operator forms a strategic alliance with mobile video contents providers. KTF formed a strategic alliance with Sports Online to supply broadcasting services of American Major League from 2003. This service replayed the main matches from the 2003 major league baseball season through *Fimm*. To produce high-quality mobile movies only for *June*, SK Telecom also formed a strategic alliance with iHQ, a multi-entertainment company, in February 2005 by acquiring 21.66% stake.

Not including wireless data transfer charges, the rate of each downloadable or streaming video is ranging from $0.5 to $1.2. In addition, it is possible to use wireless data service including video streaming and downloading services without limits at a monthly fixed rate of $30. In order to receive television channels of existing broadcasters, the users should pay only additional $6-10. *Fimm Free* service offers 36 channels of Skylife satellite broadcasting and 4 channels of terrestrial broadcastings at $10.

The revenues from mobile video in Korea grew to $59 million in 2003 from $23 million in 2001 (KETI, 2004). In August 2004, over 1.5 million users of mobile-video-enabling devices registered on KTF and the number of users of *Fimm* was over 603,000.

Mobile Payment

M-payment is a payment method for goods or services with a mobile device such as a phone, PDA, or other such device. These devices can be used in a variety of payment scenarios. Typical usage entails the user electing to make a m-payment, being connected to a server via the mobile device to perform authentication and authorization, and subsequently being presented with confirmation of the completed transaction (Wikipedia, 2006a).

The m-payment business has several service paradigms: the payments can be included in the user's mobile phone bill or a separate "mobile wallet" can be used, where the user makes deposits and withdrawals on a mobile money account governed by the mobile operator. Another solution makes use of the mobile phone only as a digital identifier, which is then used to access a digital bank account probably governed by a financial institution rather than the mobile operator.

KTF has launched the world's first exclusive mobile commerce mobile phone, which can easily settle credit card accounts anywhere. The new phone sports an "IC chip" that stores all kinds of credit information. The K-merce phone is used to exclusively settle payments. Users can conveniently settle accounts via IrFM or RF by using the K-merce phone (SPH-X8500). The IC chip card will be issued from the credit card company and inserted into the socket at the back of the phone (KTF, 2002). SK Telecom launched a mobile transaction payment service called Moneta. This system uses an installed IC chip (Smart Chip) in a cellular phone that can be used online as well as off-line (SKTelecom, 2003d). SK Telecom issues an IC chip that has functions such as a membership card, e-money, and ID card, among others. This service offers a prepaid transportation fee payment card function (SKTelecom, 2003b). SK Telecom launched a chip-based mobile banking service in March 2004 with major Korean banks. It would increase synergy effect by establishing

a win-win business model between a telecom company and banks. It adopted the SEED for standard security module of a banking IC chip (SKTelecom, 2003a).

In December 2003, SK Telecom started its Liquid Screen Small Payment Service that allows settlement of account charges through ray signals captured on a liquid screen. Any customer who uses a color screening handset can use this service as a method of payment transaction. Users download the exclusive service program which will then generate the rays that flash with special patterns on the cellular liquid screen. This ray acknowledges on a special receiver that it is connected to a PC through the USB port, and is then automatically linked to a server to create a legitimate and secure approval procedure for making payment transactions (SKTelecom, 2003c).

SK Telecom introduced Korea's first mobile bank (m-bank) international roaming service. Customers can use this m-banking service while in Beijing and Shanghai of China, by using their m-bank handsets (SKTelecom, 2004e). SK Telecom issued Moneta IC chip card with all credit card packages issued by Samsung Card (SKTelecom, 2004b). SK Telecom offered an instant mobile lottery purchase service, apartment subscription service, stock trading service, and so forth. (SKTelecom, 2004d). A joint effort by Tong Yang Investment Bank and SK Securities implements an IC chip-based stock trading service. This makes a total of three chip-based mobile financial services. These services offer increased transaction speed, as well as security, compared to a traditional wireless application protocol (WAP)-based mobile stock trading service (SKTelecom, 2004f). Customers are allowed to conduct banking transactions and stock trading with a single chip installed in their mobile phone (SKTelecom, 2005c).

Three Korean mobile carriers make t-money services available for mobile users with Korea Smart Card Co. (KSCC). The service launched in June 2005. The mobile t-money service would allow payment for public transport using t-money on all three Korean mobile carriers. T-Money is a payment system built by KSCC for public transport in metropolitan areas. Users of this service will not have to go to kiosks to "top off" their transport cards, but rather just use wireless Internet to transfer money from a registered bank account. Users can also check the amount remaining on the transport smart cards and use a refund service (KTF, 2005a). The m-bank service establishes a cooperative business model between a mobile communications operator and a financial firm by the sharing of their roles.

Mobile Games

Mobile games are a gaming service available on mobile devices such as mobile phones, PDAs, and other devices. Through the mobile devices, users can download game programs or use real-time role playing games (RPGs) similar to online games on fixed broadband networks.

In Korea, mobile games are very popular entertainment. Marketing Insight, a consumer research institute, reported that 14 million Koreans play mobile games, accounting for 40% of the total mobile phone users. In addition, around 2.3 million users play mobile games everyday, accounting for 6.2 % of the total mobile phone users (Moon, 2005). Korea's mobile game industry has been sharply increased with annual growth of more than 45 %. The revenues from mobile games in Korea grew to $2.2 billion in 2004 from $1.0 billion in 2002 (Atlas Research, 2004), accounting for the growth rate of 88.4%.

Table 5 shows that 62% of mobile game users are affected by peer-group influence on making a purchase decision of game contents and the average usage time is 1 hour 38 minutes per day, which indicates that mobile games are low-involvement products (Ahn, 2004). It is mainly because most of current mobile phones do not support the real-time interaction, so the users are usually not absorbed in game-playing through mobiles. For

Table 5. Motivations and information sources about mobile games (Source: Ahn, 2004)

Motivations	**(%)**
Experience on PC game	14
New games	45
Curiosity	36
Advertising and promotions	2
Boredom	63
Friend's recommendation	62
Direct usage experience	**35**
Information Sources	(%)
Friends	72
Game magazines	20
Television programs about mobile games	4
Advertising	3
Game Web site	0

these reasons, mobile game users still use simple board games or arcade games more than RPGs or strategic games.

However, according to the development of mobile devices and game contents, the mobile game industry is now evolving into RPGs and 3D games. SK Telecom has launched a mobile game portal site called *GXG* in April 2005 to offer various 3D converting games. 3D games are three-dimensional games that users can enjoy virtual reality as if they are exploring the virtual game place. The representative games of SK Telecom are *Mavinogi*, *Mu*, and *Ragnarok*, of which price ranges from $3 to 3.7 per each game downloaded. If consumers contract for the *Nate Free Flat Rate Plan* of $14 per month, they can download all *GXG* games without additional call charges. SK Telecom is developing mobile 3D game phones with Qualcomm and Samsung at the average of $400, which is $100-200 lower than current popular game phones such as IM-8300 and IM-8100. KTF also invested $8 million on their mobile gaming portal *GPANG*, in a strategic alliance with NHN that is Korea's second largest Internet portal and that operates *Hangame*, a successful game portal on fixed networks. In addition, KTF has launched additional 100-300 new games every year and has invested in the development of interactive games, since customers are easily bored with simple games. Over 50% of revenue in KTF's mobile game part comes from new games.

The introduction of multimedia online role playing games (MMORPGs) will contribute to expanding the existing mobile game industry. Since heavy users of online games are accustomed to the PCs and consoles which permit real-time interaction with other users, game operators should design their graphics and interaction technologies to attract the online gamers. Once new MMORPGS games through mobile phones attain the awareness from the heavy users, they will enjoy the strong royalty and stable revenues from the heavy online game users. As mobile games evolve into MMORPGs, the ARPU of mobile games will be expected to increase because of the propensity of high royalty and longer usage time in MMORPGs.

CHARACTERISTICS OF MOBILE VALUE-ADDED SERVICES

In the previous section, we discussed six major value-added services in Korea which impact the ARPU of mobile customers. In this section, characteristics of mobile value-added services are discussed.

M-commerce shows the similar aspects to mobile value-added services. Siau, Lim, and Shen (2001) shows that m-commerce has four features, such as ubiquity, personalization, flexibility, and dissemination. Customers can get

any information they are interested in, whenever they want regardless of where they are, through Internet-based mobile devices (*ubiquity*). M-commerce applications can be personalized to represent information or provide services in ways appropriate to the specific user (*personalization*). Mobile users may be engaged in activities, such as meeting people or traveling, while conducting transactions or receiving information through their Internet-enabled mobile devices (*flexibility*). Some wireless infrastructures offer an efficient means to disseminate information to a large consumer population (*dissemination*).

From the market trend in Korea and the discussion about m-commerce, we derived four characteristics of the trends of value-added services in mobile communication market: (1) customization, (2) content-focused, (3) connectedness, and (4) contemporary. We call them the "4Cs" for mobile value-added services.

Figure 4. Data ARPU of Korean mobile carriers (Source: Huh, 2006)

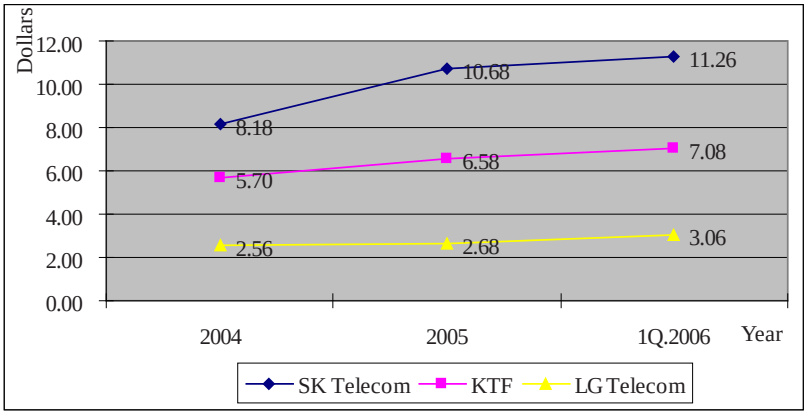
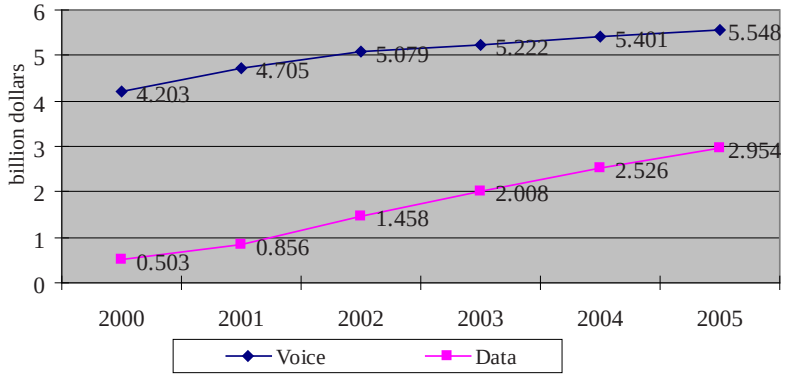


Figure 5. Voice vs. data revenues of Korean mobile carriers (Source: ETRI, 2004)



Customization

Most value-added services have various customized plans to meet different customer needs. This is an effort to satisfy the demand of personalization for customers by their mobile service. This effort helps customers who have concerns about their mobile services and handsets. It increases the involvement of customers with their services. For the call-ring service customers can use a song sung by themselves through online/off-line and fixed/wireless networks.

There are some differences in service preferences among mobile consumer segments. Major segmentation factors are age, gender, and culture. Whereas there is no significant difference between male and female in terms of mobile Internet subscription, gender does affect usage time, ARPU, and their service preferences. Women who have used mobile Internet in 2005 have been overtaking men in usage time. Women use an average of 58.2 mobile Internet minutes a week, compared to an average of 47.4 minutes a week for men (National Internet Development Agency [NIDA], 2005). Female teenagers tend to more frequently use SMS and phone decorations than male teenagers, while male teenagers tend to more frequently use mobile games than vice versa (Consumer Protection Board [CPB], 2004). These results reflect gender differences in mobile usage patterns, which indicate that two groups require distinct segmentation strategies.

Age also influences the amount of use and service preferences. Young adults and teenagers are a distinct segment for mobile service providers. They are more likely to be technology friendly and to try new technology and services when they become available. They usually have a high willingness to pay for entertainment services. However, in terms of mobile banking and some information services, older age groups are more likely to use the services (NIDA, 2005).

To design customized services, mobile carriers in Korea released various service packages and price plans appropriate for each segment. For instance, KTF and SK Telecom provide *Bigi Egg Unlimited Text Price Plan* and *Ting Text Price Plan* as exclusive rating systems for female teenager's communication trends. It allows subscribers aged from 13 to 18 to adjust the rate of phone calls and SMS according to their needs. LG Telecom also released five types of *Human Special Packages* which distinguish each contemporary life style according to various usage patterns.

Content-Focused

Mobile service is no more the commonly standardized telephone service. It is more than a voice-only service. It is similar to an Internet portal site. As the success of an Internet portal site depends on whether the site has sufficient information and is updated frequently, the success of mobile value-added service can be made by the capacity which can meet their customer needs. Having key content is more important than a better network. As Figures 4 and 5 show, Korea's mobile industry is moving toward a value-added services market in terms of ARPU. During the last 2 years, voice revenue has been stagnated while data revenue has rapidly increased.

Connectedness

Connectedness can be defined as the psychological need for social relationships. The need motivates people to maintain connection with others so as to keep utilizing reliable peer-to-peer communication technologies. Nardi et al. (2000) reported that people monitor the presence of colleagues through online messenger services even when they have no special intention to communicate with others. In mobile services, the need for social relationship is a still important motivation for people to use the services. From the perspective of consumer's psychology, the desire for connectedness is directly linked to the amount and frequency of mobile service usages, which

is capable to raise the revenues in the mobile industry. For example, young adults and teenagers, who have the highest willingness to pay for the services, recognize their mobile phones as a personal device to communicate with friends and peer groups. They are susceptible to peer-group influence and highly consider peer evaluation and social presence when utilizing mobile services. Therefore, they show high demand for peer-to-peer connectedness in their usage pattern of mobile services, such as SMS, messengers, and real-time mobile games.

In this sense, mobile carriers should consider their desire for connectedness when they design a new value-added service. Only the new value-added services enabling "connectedness orientated communication" will be a key business in the mobile market, and the marketing strategies that gratify the needs will contribute to boost the revenues (Rettie, 2003, p. 3). In Korea, three mobile carriers developed membership marketing and promotions exclusive for each segment that allow discounts available for restaurant, amusement parks, movie theaters, shopping malls and more. For example, KTF's successful *Bigi* membership promotion provides the subscribers of *Bigi* price plans with various participation programs such as educational classes and workshops, which generates virtual communities among young subscribers. What is more, the subscribers who contract *Bigi Egg Present call plan* can receive extra phone usage time as gifts from other subscribers. This strategy has been effective in preventing the subscriber loss and in increasing usage frequency.

Contemporariness

The mobile market no longer resides in the infrastructure industry. If mobile carriers can not meet their customers' likeliness in real time, they can hardly survive in this market. It is because the key success factor in this market is how quickly they satisfy the customers' needs, regardless of the technology they provide. As the customers' preferences change very often, the key concern of mobile carriers is how fast they can develop their new value-added services to meet their customer needs.

FUTURE TREND: DIGITAL MOBILE BROADCASTING

One of future trends in mobile service will be DMB. DMB systems are designed to provide television and radio programs to mobile phones with high-resolution and secure connections. In 2005, Korea launched terrestrial DMB (T-DMB) as a free mobile broadcasting and satellite DMB (S-DMB) as a subscriber-based mobile broadcast. S-DMB started its commercial services in May 2005 with 11 video channels and 21 audio channels and T-DMB started in January 2006 with 7 video channels and 12 audio channels.

The difference from the existing EV-DO services is that it is utilizing a new broadband network to offer mobile real-time broadcasting services, which enables high-speed movements without disconnections. There are two kinds of DMB according to its technology type and network configuration; S-DMB and T-DMB. S-DMB is based on code division multiplexing (CDM) similar to CDMA mobile communication technology and S-Band (2.630-2.655 GHz), whereas T-DMB is based on orthogonal frequency division multiplexing (OFDM) for digital TV standard in Europe and VHF Ch 8 (180-186MHz) and Ch 12 (204-210MHz). Due to stable reception and mobility, DMB is becoming a new value-added service to satisfy consumer demands for mobile broadcasting and to provide the mobile industry, recently experiencing the slow growth rate, with new revenue sources. TU Media, the world's first DMB operator, launched its nationwide S-DMB service in May 2005 and acquired 100,000 customers in 2 months and 22 days and additional 100,000 customers in 4 months. TU Media is collecting

more than 2,400 users a day (SKTelecom, 2005b). TU Media now provides 37 channels including 11 video and 26 audio channels.

Since S-DMB and T-DMB services share characteristics of both broadcasting and mobile telecommunication, they were expected to lead to fierce competition in several existing markets, such as broadcasting market, mobile market, and high-speed Internet market. Therefore, the impact of DMB on existing mobile value-added services was controversial and was considered as a challenge for existing network operators to come up with successful revenue models. Current statistics of Korea's DMB reports that new DMB services have helped the continuing growth of mobile industry. The basic effect of introducing DMB turns out to open a new media market, not to transmit existing broadcasting services through mobile phones. In addition, three mobile network operators and existing broadcasters, who composed consortiums for DMB services, benefit from the new services more profit than losses by compensating the decreasing growth rate of voice ARPU. Despite of maturation in mobile phone services, KTF reported that they have attracted 117,000 new DMB subscribers with 200% increased T-DMB's ARPU in the last year, when over all data ARPU increased $.60. SK Telecom also reported that last year's data ARPU was $17, increased by 31 % from 2004 and DMB ARPU was $3.3, which indicated that S-DMB has no negative effect on data ARPU. These results indicate that DMB contributes to the growth of mobile market without cannibalization. New DMB services are allowing three mobile carriers to find the next revenue sources by awakening a dormant market, making the stagnating mobile market a "blue ocean" (W. C. Kim & Mauborgne, 2005).

To promote DMB adoption and usages of mobile services, operators and vendors should secure enough content supply to attract new subscribers. According to KBI's survey (2006), people subscribe for entertainment and killing time, and sports, entertainments, and drama are still popular in DMB services, like traditional broadcasting media. In particular, sports are the top programming choice across all mobile video services including S-DMB, T-DMB, and video downloading through MMS. For all DMB subscribers, the average preference score of sports was 3.35 on a 5-point scale. Also, the gratification score of sports was also 3.35 in S-DMB and 3.5 in T-DMB. During the World Baseball Classic in 2006, sales of DMB mobile phones have surged up to 3,000 per a day, a 200% increase in daily sales (KBS, 2006). According to Visiongain, an industry research company, a 1 month football tournament generates $6.35 billion in revenue only with text-based services and downloads in a 3G network, which implies that sports game relay is a key generator to boost DMB adoption and usages. That is, it will be critical factors of the mobile carriers' performance to secure the supply of sports and entertainment programs.

From the perspective of usage pattern, DMB users are more similar to mobile phone users than traditional broadcasting audiences. Although their program preferences are as same as that of traditional audience, 82.3% of the total respondents have propensity to watch the programs alone, which indicates that people consider DMB media as personal devices. After adopting mobile devices enabling DMB reception, the users utilize their mobile phones longer than before, but reduce the usages of traditional televisions and communication with their families. This means that DMB functions as a revenue generator or a new market exploiter of mobile carriers.

IMPLICATIONS

As the mobile market has shifted from 2G to 3G, a lot of changes took place in the mobile market. The value chain has been divided into several sub-components, which results in various market opportunities. For mobile carriers, the new mo-

bile market has more potentialities through new revenue models. The new potentialities come into view as several value-added services such as mobile video, game, music, and other new services. The advent and growth of value-added services in mobile market has changed the structure of mobile market.

First, the subject of market appeal has changed from new customers to existing customers. At the time when the market has growing enormously, players in this market have more concerns on inducing new customers rather than on the retention of existing customers.

Generally, promoting new subscribers leads to higher financial benefit than keeping existing customers. However, as the current mobile market growth closes to saturation point, the market players concern more about how to retain existing customers and how to increase the net revenue rate from the customers than how to attract new customers. In a saturated market, competition among players has become fiercer and the differences of service quality among competitors have been diminished. Therefore, it is hard for a player to get a significantly competitive position in relation to others. Mobile players in a saturated market should focus on value-added services for existing customers, in order to increase net revenue rate.

Second, as a result of the first implication, market players have shifted from a general strategy to differentiated strategy. Many textbooks on industrial organization (Carlton & Perloff, 2000; Tirole, 2002) explain this kind of change as a dynamics of market structure. When an innovative product or service is introduced in a market, innovators take and use the product or service and offer their services to the public. If the product or service has popularity, the market grows rapidly and encourages potential market players to enter the market. The new entry of players may boost market growth and stimulate the diffusion of product or service in a certain time period. When the market approaches a matured stage, the intensity of competition becomes high and the margins of market players have diminished. At that point, market players need to make a decision for their future business. Typically economists explain there are two kinds of strategies. The first one is a trial to get competitive advantage through cost-saving which is called cost-leadership. The other is exploiting new markets through product differentiations. Mobile carriers are at that place that they need to make a decision. Various value-added services are the examples of product differentiation.

The third one is the change of regulation. Telecommunications service had been classified as a utility until mid-1980s. Even though competition has been introduced partially, the telecommunications market has been under rigorous regulation. It is because telecommunication services have an aspect of natural monopoly in which a bigger company with the economy of scale has the competitive advantage. However, as the value chain of the mobile industry has been divided into several components, the market power of network carriers has been reduced and, as the result, the necessity of market regulation has been diminished. Thus, competition type also has changed from network-based competition to market-based competition.

Fourth, mobile market structure has changed from supply-based to demand-based in the perspective of economics. That means the era in which customers have to wait to enjoy telecom services has passed. There exists excessive supply in this market. The network sunk costs can not be rewarded. The market concern has moved to customers' demand.

Fifth, the telecommunications market is traditionally regarded as a network business that is driven by technology. However, now the market has more concerns about services and marketing. Creativity to meet consumers' needs is much more spotlighted than higher technology.

Sixth, as a result of the change of value chain, the mobile service market has multi-aspect competition rather than the one-dimensional competition

in the "old days." Companies from adjacent industries such as ISPs, satellite platform operators, and other broadcasters are entering this converged market and competing with mobile carriers. For instance, iPod of Apple computers providing music services in a mobile environment can be a potential competitor to the mobile carriers, who are trying to provide entertainment with mobility. Therefore, the boundary of the market is going to be vague. In the near future, interconnection between adjacent businesses which were far from the telecom sector becomes a critical competitive factor in the market.

Therefore, it is the time of mobile carriers to come up with new strategies satisfying various demands of their customers.

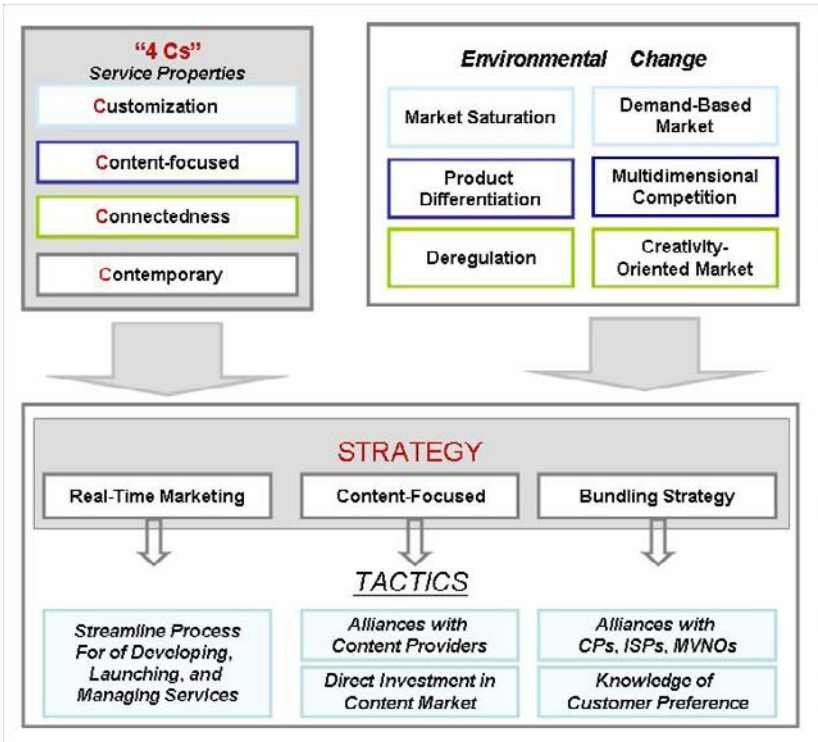
NEW STRATEGIES FOR MOBILE VALUE-ADDED SERVICES

As we discussed in the previous section, the characteristic of the mobile communications market has changed from common needs to a variety of market needs. Considering service properties and environmental changes, we recommend three new strategies for mobile carriers as the conclusions of our discussion (see Figure 6).

Strategy 1: Real-Time Market-Responding Strategy

The market environment has changed so rapidly. The entry barrier to this market has been lowered.

Figure 6. Mobile value-added service development strategy



If mobile carriers can not respond the needs of their customers' quickly, the business opportunity no more waits for the mobile carrier. Mobile carriers can not rely on their networks anymore. There are huge numbers of alternate networks that are waiting for the market opportunity.

Mobile carriers should adapt to the market dynamics in order to survive in this market. The development of digital technology accelerates launching new converged services, enables mobile carriers to meet various customer segments, and encourages regulators to change their regulatory frameworks into competition-oriented ones. Thus, mobile carriers should be prepared to respond to dynamic environmental changes such as competitors' new products, the change of customer preferences, and regulation change.

One of the tactics for this strategy is to streamline the process of developing, launching, and managing the new value-added services. The timing of launching services is a critical factor to gain the market. In particular, certain services have a short life cycle and the market responses are also spontaneous. A systematic and efficient process to develop and launch value-added services is essential.

Strategy 2: Content-Focused Market Strategy

Mobile networks are converging into the IP-based broadband network. The capacity and speed are close to each other. Even though the competition by the network is still going on, the pattern of competition has changed into what mobile carriers can provide on their networks from what kinds of networks they have. If mobile carriers can not have competitive advantage from their networks, they will be trying to keep their markets by product differentiation. That means which content they can provide is more critical. Having "killer" content will be a key success factor.

The transformation from a network company to a network-based content company is a challenge for the traditional mobile carriers. Content, not network will be in the center of strategic decision making. If a company has outstanding content, the company can use any network with network contract and transmit their content to their potential customers. As a result, a company that has excellent content can have bargaining power against a traditional network operator.

Two different tactics come up for the strategy. The first is the alliance with various content providers. Even though it is an indirect method, mobile carriers can respond to market needs more quickly without much risk. However, through this process, mobile carriers can lose their leading role in the value chain of the market. The second is the direct entry to the content market. It can be risky because they do not have sufficient experiences in that market, it can hardly respond to market needs quickly, and they can be exposed to the whole risk of failure of developing killer content. However, mobile carriers can keep their leading role in the value chain and get whole rewards from the success. "Higher risk and higher return," as many mobile carriers already tried to, the combination of two different tactics will be suitable under the recent market environment.

Strategy 3: Various Bundling Strategy

As we discussed before, a characteristic of mobile value-added service is customization. The underlying assumption for that threat is that customers have personalized preference on mobile value-added service. In order to meet their personalized demands, a variety of bundling services should be ready to provide. Without a scheme to mix various service features, it is hard to satisfy customers' personalized needs.

Because "variety" is a key factor in the market of mobile value-added services, mobile carriers should have a strategy to provide a variety of bundled services which meet the specific demands of customers in an efficient way. How many vari-

ous bundled services mobile carriers have could be a critical competitive advantage in the future mobile market.

In order to provide various bundling services, there are several kinds of tactics.

The first one is the preparedness of a various combination of service features with a strategic network with related companies. Strategic alliances with contents providers, ISPs, and mobile virtual network operators (MVNOs) can be a major strategic decision.

The second tactic is knowledge on customer preferences.

Most content providers or ISPs are less regulated than traditional mobile network carriers. They are small sized and their process of decision making is efficient. Thus they can respond the market more rapidly than network carriers.

Through the state-of-the-art techniques to capture the demands, a quick response mechanism for the changes of customer preferences is necessary.

REFERENCES

3GNewsroom.com. (2006, January 22). *World Cup to promote 3G*. Retrieved from http://www.3gnewsroom.com/3g_news/jan_06/news_6615.shtml

Ahn, S. (2004). Mobile business strategy of SKT. SKTelecom.

Atlas Research. (2004). [Statistics] mobile contents market trend in Korea. *Mobile Contents & Application* (p. 2).

Brenner, M. R., Grech, M. L. F., Torabi, M., & Unmehopa, M. R. (2005). The open mobile alliance and trends in supporting the mobile services industry. *Bell Labs Technical Journal, 10*(1), 59-75.

Carlton, D. W., & Perloff, J. M. (2000). *Modern industrial organization* (3rd ed.). Reading, MA: Addison-Wesley.

Consumer Protection Board (CPB). (2004). *Summary: A survey of teenagers' mobile phone and wireless Internet usage pattern.* Author.

Daishin Security. (2005). *Daishin equity report (SK Telecom).* Author.

ETRI. (2004). *Korea's mobile services gross revenue*. Author.

Geng, X., & Whinston, A. B. (2001). Profiting from value-added wireless services. *Computer, 34*(8), 87-89.

Huh, W. (2006, May 10). *The gap of data ARPU among mobile carriers.* Retrieved from http://www.fnnews.com/view?ra=Sent0901m_01A&corp=fnnews&arcid=0920722524&cDateYear=2006&cDateMonth=05&cDateDay=10

International Telecommunications Union (ITU). (2006). *World telecommunications indicators.* Author.

KBS. (2006). *Baseball boom in sports and marketing.* Retrieved May 11, 2006, from http://english.kbs.co.kr/life/trend/1388891_11857.html

Korea Electronics Technology Institute (KETI). (2004). *Mobile contents market of 2004 in Korea*. Seoul: Author.

Kim, H.-S., & Yoon, C.-H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy, 28*(9-10), 751-765.

Kim, J. K. (2005). *Mobile subscribers' willingness to churn under the mobile number portability (MNP).* Paper presented at the the Eleventh Americas Conference on Information Systems, Omaha, NE.

Kim, M.-K., Park, M.-C., & Jeong, D.-H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunications Policy, 28*(2), 145-159.

Kim, W. C., & Mauborgne, R. (2005). *Blue ocean strategy: How to create uncontested market space and make the competition irrelevant*. Boston: Harvard Business School Press.

Korea Broadcasting Institute (KBI). (2006). Trends and strategy of digital multimedia broadcasting in Korea. *KBI Focus, 6*(6).

KTF. (2002, November 4). *KTF launches the world's first mobile payment handset based on IC chip*. Retrieved from http://www.ktf.com/front/IR/eng/

KTF. (2005a, April 27). *3 mobile carriers partner for release of mobile t-money service*. Retrieved from http://www.ktf.com/front/IR/eng/

KTF. (2005b, July 25). *KTF's music portal "Dosirak" finds strong foundation in the market.* Retrieved from http://www.ktf.com/front/IR/eng/

KTF. (2005c, May 25). *KTF expands service in Indonesia*. Retrieved from http://www.ktf.com/front/IR/eng/

KTF. (2005d, May 24). *KTF opens music portal service Dosirak*.Retrieved from http://www.ktf.com/front/IR/eng/

KTF. (2005e, August 17). *Mobile SMS now on PC messengers!* Retrieved from http://www.ktf.com/front/IR/eng/

KTF. (2005f, June 14). *Singing a song, making it my ringtone!* Retrieved from http://www.ktf.com/front/IR/eng/

Lee, J., Lee, J., & Feick, L. (2001). The impact of switching costs on the customer satisfaction-loyalty link: Mobile phone service in France. *Journal of Services Marketing, 15*(1), 35-48.

Lee, K.-M. (2005, September 22). Mobile network operators, 2,354billion KWON revenue of CID and SMS. Retrieved from http://issuei.com/sub_read.html?uid=1608&section=section1&section2=

Maitland, C. F., Bauer, J. M., & Westerveld, R. (2002). The European market for mobile data: Evolving value chains and industry structures. *Telecommunications Policy, 26*(9/10), 485-504.

Middleton, C. A. (2002). *Exploring consumer demand for networked services: The importance of content, connectivity, and killer apps in the diffusion of broadband and mobile services.* Paper presented at the Twenty-Third International Conference on Information Systems (ICIS), Barcelona, Spain.

Moon, B. (2005, November 8). *Mobile game addiction spreading.* Retrieved from http://www.donga.com/fbin/output?sfrm=1&u=200511080069

Music Industry Association of Korea (MIAK). (2005). *Statistics of digital music market in Korea*. Author.

National Internet Development Agency of Korea (NIDA). (2005). *Summary: Mobile Internet consumer research 2005.* Author.

Nardi, B., Whittaker, S., and Bradner, E. (2000). Interaction and outeraction: Instant messaging in action. In *Proceedings of Conference on Computer-supported Cooperative Work*, (pp. 79-88). New York: ACM Press.

Pagani, M. (2004). Determinants of adoption of third generation mobile multimedia services. *Journal of Interactive Marketing, 18*(3), 46-59.

Rettie, R. (2003). *Connectedness, awareness and social presence*. Paper presented at the 6th Annual International Workshop on Presence.

Sabat, H. K. (2002). The evolving mobile wireless value chain and market structure. *Telecommunications Policy, 26*(9/10), 505-535.

Siau, K., Lim, E.-P., & Shen, Z. (2001). Mobile commerce: Promises, challenges, and research agenda. *Journal of Database Management, 12*(3), 4-13.

SKTelecom. (2003a, December 9). *SK Telecom establishes the national standard for mobile banking services* [Press release]. Seoul, Korea: Author.

SKTelecom. (2003b, October 27). SK Telecom launches "MONETA Membership Pack Service" for teenagers [Press release]. Seoul, Korea: Author.

SKTelecom. (2003c, December 22). SK Telecom launches "Liquid Screen Small Payment" service [Press release]. Seoul, Korea: Author.

SKTelecom. (2003d, August 12). SK Telecom starts "Moneta Online Payment Service" [Press release]. Seoul, Korea: Author.

SKTelecom. (2004a, July 27). SK Telecom exceeds ten millions US dollars in export sales of call ring (so called Coloring) service solution [Press release]. Seoul, Korea: Author.

SKTelecom. (2004b, August 18). SK Telecom forms strategic alliance with Samsung Card for mobile payment service [Press release]. Seoul, Korea: Author.

SKTelecom. (2004c, November 15). SK Telecom presents a new paradigm for promoting the digital music market [Press release]. Seoul, Korea: Author.

SKTelecom. (2004d, September 30). SK Telecom starts "M-Bank Service" in a joint effort with KB [Press release]. Seoul, Korea: Author.

SKTelecom. (2004e, March 16). SK Telecom starts M-Bank international roaming service [Press release]. Seoul, Korea: Author.

SKTelecom. (2004f, October 18). Stock trading chip service launched as part of a succession of mobile banking service advances [Press release]. Seoul, Korea: Author.

SKTelecom. (2005a, December 15). Melon subscribers hit four million mark [Press release]. Seoul, Korea: Author.

SKTelecom. (2005b, October 14). Satellite DMB customers reach over 200,000 [Press release]. Seoul, Korea: Author.

SKTelecom. (2005c, April 19). SK Telecom introduces mobile multi-use financial chip [Press release]. Seoul, Korea: Author.

SKTelecom. (2005d, October 31). SK Telecom releases the world's first PMP phone [Press release]. Seoul, Korea: Author.

SKTelecom. (2006, February 16). SK Telecom executes an agreement for jointly developing a music file CODEC technology with Germany's CT company [Press release]. Seoul, Korea: Author.

Song, M. (2005). *KT's media business and content strategy*. Retrieved May 3, 2006, from http://ct.kaist.ac.kr/file/seminar/20051025_Song.pdf

Steinbock, D. (2003). Globalization of wireless value system: From geographic to strategic advantages. *Telecommunications Policy, 27*(3/4), 207-235.

Tirole, J. (2002). *The theory of industrial organization*. Cambridge, MA: The MIT Press.

Wikipedia. (2006a). *Mobile payment*. Retrieved from http://en.wikipedia.org/wiki/Mobile_payment

Wikipedia. (2006b). *Ringback*. Retrieved from http://en.wikipedia.org/wiki/Ringback

This work was previously published in Global Mobile Commerce: Strategies, Implementation and Case Studies, edited by W. Huang, Y. Wang, and J. Day, pp. 265-285, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 6.23
Semantic Location Modeling for Mobile Enterprises

Soe-Tsyr Yuan
National Chengchi University, Taiwan

Pei-Hung Hsieh
STPRIC, National Science Council, Taiwan

ABSTRACT

A location model represents the inclusive objects and their relationships in a space and helps engender the values of location based services (LBS). Nevertheless, LBS for enterprise decision support are rare due to the common use of static location models. This chapter presents for enterprises a framework of dynamic semantic location modeling that is novel in three ways: (1) It profoundly brings location models into enterprise business models; (2) with a novel method of dynamic semantic location modeling, enterprises effectively recognize the needs of the clients and the partners scattered in different locations, advancing existing business relationships by exerting appropriate service strategies through their mobile workforces; (3) through the location model platform of information sharing, enterprises are empowered to discover potential business partners and predict the values of their cooperation, gaining competitive advantages when appropriate partnership deals are made by enterprise mobile workforces. This proposed framework has been implemented with the J2EE technology and attained the positive evidences of its claimed values.

INTRODUCTION

With the advent of wireless communication technologies, the era of mobile enterprises unfolds. Many international enterprises like IBM, Sun, HP, and Microsoft are vying to develop mobile enterprise servers and solution architectures. According to a Cutter report, 57% of the employees in the enterprises worldwide were regarded as the "mobile workforce" in 2005 (Ericsson Enterprise, 2002). Accordingly, following the e-business

trend, competitive advantages built on wireless technologies in dynamic mobile environments are now widely recognized by enterprises.

The conventional perception of mobile enterprises is that enterprise users are able to have personalized, seamless access to enterprise applications and services from anyplace and at anytime, regardless of the devices employed, in order to facilitate the tasks at hand (Bouwman et al., 2005; Ericsson Enterprise, 2002).

Subsequently, location is an inherent feature of many mobile services. Location-based services (LBS) are information services that exploit knowledge about where an information device user is located. According to Ovum, an analyst and consulting company, the market for LBS will grow to $12 billion by 2006. Existing LBS primarily rest on four categories of services (Varshney, 2000): (1) safety (e.g., emergency services, roadside assistance); (2) navigation and tracking (e.g., vehicle navigation, asset tracking, people tracking); (3) transactions (e.g., location-sensitive billing, zone-based traffic calming); and (4) information (e.g., yellow pages, location-based advertising). The main idea behind the former three categories is locating targeted objects for provision/consumption of certain external resources. The last category then focuses on targeted advertising, linking nearby consumers/buyers and providers/sellers to facilitate additional revenue generation (Polyzos, 2002; Ververidis & Yuan & Peng, 2004; Yuan & Tsao, 2003). LBS has been a hot area of research because mobility of information device users leads to the generation of user location information that subsequently drives a slew of new services.

Moreover, enterprise decision support (Bolloju, 2003) is often regarded as: (1) use of corporate data to derive and create higher level information and knowledge, (2) integration of organizational information to support all departments and end users, and (3) provision of tools to transform scattered data into meaningful business information. Enterprises utilizing geometrical data are often the likes of logistic companies of which LBS mainly rests on the provision of support on navigation and the tracking of their employees (shipping vehicles) or clients. For instance, logistic delivery planning locates shipping vehicles based on geometric models: static location models (MapInfo, http://www.mapinfo.com/products/Features.cfm; RITI Technology Inc., http://www.elocation.com.tw) to know all inventories in transit and enable efficient logistic deliveries (Varshney, 2000). Nevertheless, it is rare to perceive LBS as enterprise decision support in attaining higher level information and knowledge. *It naturally comes to a question of how to marry enterprise decision support with LBS so as to deeply utilize the business data together with the geometric data.* In searching for the answer to the question, there is a need to identify the reasons behind the limited extent of this marriage. (Afterwards, this sort of marriage is named *enterprise-based LBS*.)

In this research, we believe the possibilities behind this limitation are (1) the integration of enterprise business models and existing location models is difficult; and (2) the limitation of existing location models hinders additional development on enterprise-based LBS.

With the aforementioned suppositions, *this chapter aims to present a framework of dynamic semantic location modeling (DSLM) that shows certain integration of enterprise business models and the proposed location model (that surmounts the problems encountered in static location models), realizing enterprise-based LBS* (e.g., the location-sensitive decisions of potential strategic partners required in the expansion of enterprise alliance networks). The DSLM framework is believed to encourage the development of myriad research on enterprise-based LBS in the future. This chapter will first discuss the limitations of existing location models and then present the DSLM framework, followed by some evaluation results and conclusion.

LOCATION MODELING

Existing methods for location modeling are twofold (Domnitcheva, 2001): the first one is geometric modeling that is built upon the geometric coordinate system. The other is symbolic modeling that represents locations with symbols and symbol sets.

Each location modeling method has its pros and cons. Geometric modeling (static location models) has the advantages of high accuracy and easy communication between different kinds of platforms. However, geometric modeling requires reference points and mappings between information objects and geometric coordinate objects. On the other hand, symbolic modeling represents locations with location object names (e.g., 11th Park in Taipei), each of which unfolds as a set containing the objects residing in the designated location. Symbolic modeling accordingly is easy to comprehend, but requires effort in managing the naming of the location objects and the handling of the ranges and the overlaps of the location objects[1] (vLeonhardt, 1998).

While exerting geometric models (static location models) for enterprise-based LBS, there are two primary problems encountered:

- **Meaningless syntactic information:** A mobile enterprise application system can attain only *syntactic* information objects regarding a given location. For instance, when a salesperson queries the system for product sales information of a designated branch office, he may get numerous sales figures for the product at the designated location, but do not know whether these figures imply good sales or bad sales. For situations that salespeople are capable of judging the performance of these figures, the judgments cannot be wisely retained for facilitating subsequent relevant decision making (that however appreciates these *semantic* judgments).
- **No seamless information exchange/integration:** When the exchange or integration of location-sensitive information is intended by enterprises, this might give rise to the need of a middleware for the information translation when enterprises employ different static local models. The rationale is twofold: (1) the mapping between information objects and coordinate objects in a static location model is fixed (*static*), and thus it is hard to interoperate the information objects exchanged; (2) this fixed mapping also creates difficulties in the merging of the two static location models when tight enterprise relationships are attempted (i.e., the *dynamic* expansion of existing location models).

From the above discussion, there are two vital desired features for enterprise-based LBS: "semantic" and "dynamic." *"Semantic" indicates that an enterprise can define its own objects, object values, object relationships in a location model* (Pradhan, 2002). *"Dynamic" then denotes that a location model can grow and adapt with the enterprise interactions, building "dynamic links" between locations [6].* These two features drive the necessity of the development in a new method of location modeling[2] in order to shed light on advanced enterprise-based LBS.

This chapter presents DSLM that unfolds itself as a new location modeling method and is the first attempt integrating enterprise business models and the proposed location model so as to realize an advanced enterprise-based LBS. The contributions of DSLM are threefold (denoted by BOLM, PNLM, and LMP) and outlined in Table 1. BOLM, PNLM, and LMP differ with each other mainly in the scopes of their functions. BOLM and PNLM can endow enterprise mobile workforce with location-sensitive decision information about their clients or potential partner enterprises. On the other hand, LMP furnishes enterprises of an industry with a platform in which location-sensitive new potential partners

Table 1. The DSLM solutions in mobile enterprise decision support

Solution	Within Enterprise ***Business-Oriented Location Model (BOLM)***	Between Enterprises ***Partner-Network Location Model (PNLM)***	Within Industry ***Location Model Platform (LMP)***
Main Function	Assist an enterprise mobile workforce to understand the business relationships with clients in certain locations.	Endow an enterprise mobile manager with the knowledge of the benefits of the cooperation with potential partner enterprises in a certain location area so as to attain satisfactory cooperation contracts or deals.	Assist an enterprise to search for potential partner enterprises to cooperate in certain location areas (of different location regions).
Benefit	Employ proper location-sensitive strategies to better utilize the enterprise's resources.	1. Location-sensitively attain the cooperation relationships between enterprises. 2. Expand the service scope and range through cooperation between enterprises.	Realize an information-sharing platform between enterprises in different locations.

can be identified. Their details will be described in Section 3.2, 3.3, and 3.4, respectively.

As for other existing enterprise location services, they mainly focus on the potentials of RFID by managing readers, filtering and aggregating raw RFID data, and facilitating data exchange among the supply chain partners (Bouwman, Haaker, & Faber, 2005). There is a need to search for advanced work for location intelligence (that combines spatial-data collection with advanced analysis and visualization methods to transform sitting to knowledge) (Grimes, 2005).

THE FRAMEWORK OF DSLM

DSLM aims to fulfill a certain integration of enterprise business models and location models in terms of the three DSLM solutions. These solutions involve enterprise clients, enterprise partners, and a platform enabling the search of new partners. Accordingly, interoperability and decision-support aid are the key characteristics of DSLM. This section starts with the description of the ontology employed in DSLM (that defines the semantics required to represent our location models and to enable location intelligence in enterprise interoperability) followed by the three DSLM solutions addressed in Table 1.

DSLM Ontology

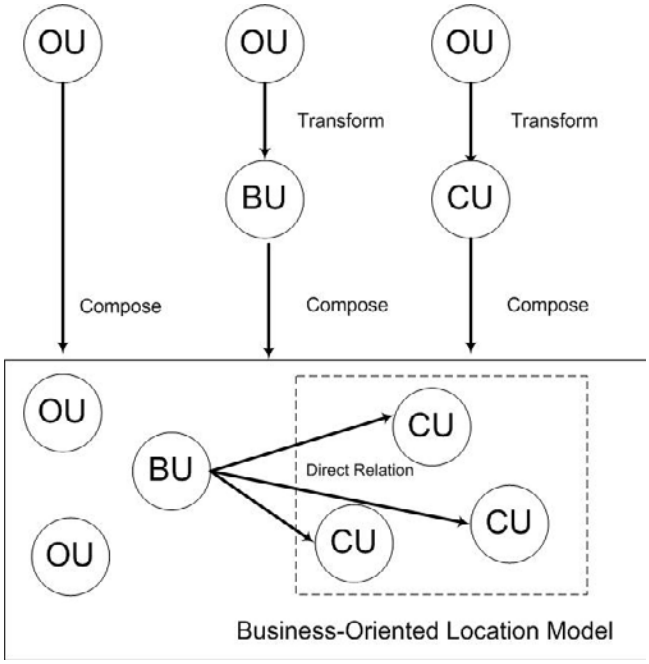
DSLM fits in the category of symbolic modeling, but the relationships between symbols and symbol sets can be changed dynamically. The DSLM ontology is a shared ontology that is regarded as

the interchange format, enabling common access to enterprise operational data (Jasper & Uschold, 1999). DSLM ontology defines objects, object relationships, and relationship measurements. The following subsections will detail these terms.

Objects

DSLM ontology defines four types of objects (original unit, business unit, client unit, and business-oriented location model) as shown in Figure 1 and defined in Definition 1:

Figure 1. Objects in DSLM



Definition 1.

OU(Y)：Y is the Original Unit in the DSLM

BU(C)：C is the Business Unit of the DSLM

CU(D,C)：D is the client of Business Unit C in the DSLM

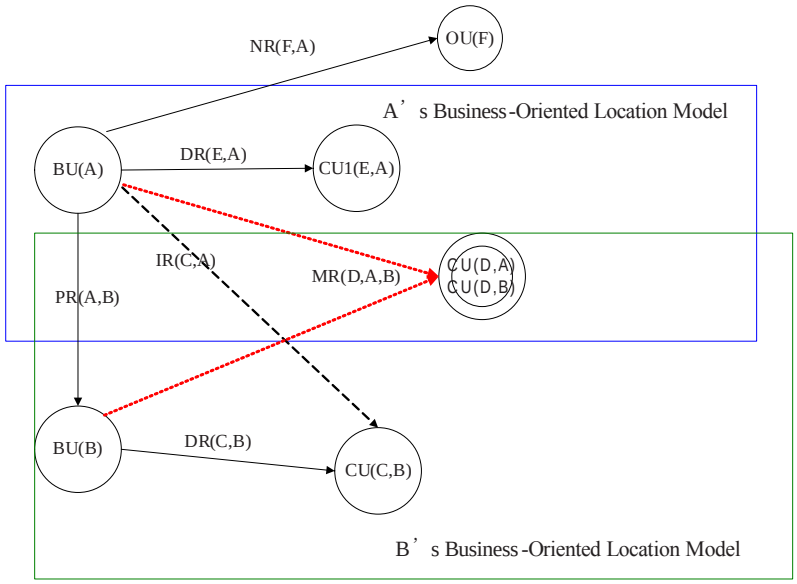
$$BOLM(C) = def \exists_{X_1, X_2, \ldots, X_n} \text{ is } CU(X_1, C), CU(X_2, C), \ldots, CU(X_n, C)$$
$$\exists_{Y_1, Y_2, \ldots, Y_m} \text{ is } OU(Y_1), OU(Y_2), \ldots OU(Y_m) \in BOLM(C)$$
$$\bigwedge_{i=1}^{n} \bigwedge_{j=1}^{m} ((BU(C) \wedge CU(X_i, C) \wedge OU(Y_j)) \Big|_{DR(x_i, C)}$$

- **Original unit (OU):** An entity in a map that is not at all referenced in the location model of an enterprise because of no business relationship between the enterprise and the entity. For instance, if there is no business relationship between a freight company A and a bookstore B, then B will be regarded as an OU in A's Business-Oriented Location Model.
- **Business unit (BU):** Upon the construction of a business-oriented location model for an enterprise, the OU representing the enterprise transforms into a BU.
- **Client unit (CU):** An OU representing a client of the enterprise (constructing its business-oriented location model) transforms into a CU.
- **Business-oriented location model (BOLM):** The BOLM of an enterprise is comprised of the BU (representing the enterprise), the CUs (denoting all of the clients of the enterprise, and OUs (symbolizing the entities without business relationships with the enterprise). For instance, if a logistic company C has three clients (D, E, F), then C's BOLM is composed of 1 BU representing C and 3 CUs denoting D, E, F, and a couple of OUs. The relationship (DR relationship) between the enterprise and its clients will be detailed in the next subsection.

Object Relationship

Object relationships stand for the relationships between the BU, CUs, and the BOLM. There are a variety of relationships being modeled: direct relationship (DR), indirect relationship (IR), no relationship (NR), multiple relationship (MR), partner relationship (PR), located in (LI), and not-located in (NI) as shown in Figure 2 and defined in Definition 2:

Figure 2. Object relationships in DSLM



Definition 2.

A is a Business Unit in BOLM(A)
B is a Business Unit in BOLM(B)
F represents an Original Unit
E represents a Client Unit

NL(F, A) : F is not located in BOLM(A)

NR(F, A) : F has no relation with BU(A)

LI (E, A) : E is located in BOLM(A)

PR(A, B) : BU(A) and BU(B) has partner relation

$DR(x, A) = (LI(x, BOLM(A)) \wedge CU(x, A))$

$IR(y, A) = \exists_z (BU(z) \wedge (NL(y, BOLM(A)) \wedge PR(A, z) \wedge CU(y, z))$

$MR(z, A, B) = (LI(z, BOLM(A)) \wedge LI(z, BOLM(B) \wedge PR(A, B) \wedge CU(z, A) \wedge CU(z, B))$

- **Direct relationship (DR):** A relationship denoting the direct business relationship between a client and the enterprise such as DR(C, B) as shown in Figure 2 in which C is a client of the enterprise B.
- **Partner relationship (PR):** A business relationship between two enterprises such as PR(A, B) as shown in Figure 2.
- **Indirect relationship (IR):** A relationship between a client C (of the enterprise B) and the enterprise A that is formed because of a Partner Relationship between A and B such as IR(C, A) as shown in Figure 2.
- **Multiple relationship (MR):** A relationship between a client and multiple enterprises that have the partner relationship such as MR(D, A, B) as shown in Figure 2 in which the client D is a client of both A and B (that further have the partner relationship with each other).
- **No relationship (NR):** A relationship other than any of the aforementioned relationships.
- **Located in (LI):** An inclusive relationship between a BU (CU, or OU) and a BOLM.
- **Not-Located in (NI):** A non-inclusive relationship between a BU (CU, or OU) and a BOLM.

Relationship Measurement

In order to differentiate the relationships for the purpose of decision support, the relationship DR, IR, and MR are associated with measurements. These measuring are based on the values of certain object attributes that an enterprise concerns such as distance[3] from the enterprise, average revenue and average order. Algorithm 1, Algorithm 2, and Algorithm 3 exemplify certain algorithms for calculating the relationship measurements:

- **DR measurement:** Between the direct clients (CU) of an enterprise (BU), DR measurements aim to differentiate the clients. Algorithm 1 exemplifies one possible way of such differentiation that is accomplished

Algorithm 1. An exemplar of DR measurement

Function Direct_Relation_Measurement (BU, CU)

1. Select significant attributes A_i that characterizes the relationship between BU and CU, and transform their values to Semantic Levels SL_i according to BU's subjective judgment.
2. Assign weight W_i to all chosen attributes according to the levels of their significance to BU.
3. DR = $\sum_{i=1}^{n} SL_i * W_i$

Note：This algorithm only exemplifies a linear measurement. Non-linear measurements can be employed in Step 3 as well.

Algorithm 2. An exemplar of IR measurement

Function Indirect_Relation_Measurement (Source Enterprise BU, Target Enterprise BU, CU)

1. Select significant attributes A_i from the client CU's Source Enterprise BU, and transform their values to Semantic Levels SL_i according to Target Enterprise BU's subjective judgment.
2. Assign weight W_i to all chosen attributes according to the levels of their significance to Target Enterprise BU.
3. IR = $\sum_{i=1}^{n} SL_i * W_i$

Note：This algorithm only exemplifies a linear measurement. Non-linear measurements can be employed in Step 3 as well.

through the calculation of a weighted sum of the values of the client's attributes chosen by the enterprise.

- **IR measurement:** Between the indirect clients (CU) of an enterprise (Target Enterprise BU) because of its partnership with another enterprise (Source Enterprise BU), IR measurements intend to distinguish the indirect clients by calculating a weighted sum of the CU's attribute values gathered

Algorithm 3. Example of MR measurement

Function Multiple_Relation_Measurement (Source Enterprise BU, Target Enterprise BU, CU)

1. Calculate DR(Source Enterprise BU, CU).
2. Source Enterprise BU calculates DR'(Target Enterprise BU, CU) by using the CU's attributes and data retained in Target Enterprise BU.
3. MR = DR' - DR

from Source Enterprise BU.[4] However, the weights are assigned from the point view of Target Enterprise BU (instead of from Source Enterprise BU's as shown in Algorithm 2).

- **MR measurement:** Given a MR (in which a client CU is associated with Source Enterprise BU and Target Enterprise BU by the MR bindings), MR measurements aims to further discriminate these bindings in terms of different originating perspectives (i.e., from the perspective of Source Enterprise BU). Algorithm 3 shows the method for a MR measurement from the perspective of Source Enterprise BU. This MR measurement represents a strength difference between the DR measurement (of Source Enterprise BU and CU) and the DR' measurement (of Target Enterprise BU and CU) for which the retrieval of CU's data retained in Target Enterprise BU is made). In other words, from the perspective of Source Enterprise BU, a MR measurement reveals an important message about the subjective relative strength (with respect to Target Enterprise BU) in regard to the relationship with the client CU. For instance, it manifests a stronger relationship that Source Enterprise BU has with CU than that of Target Enterprise BU when the MR measurement is less than zero.[5]

Mobile Enterprise Decision Support Using DSLM

This section describes the three DSLM solutions (BOLM, PNLM, LPM) mentioned in Table 1. Each of the solutions supplies relevant decision-support aids and leads to certain integration of enterprise business models and enterprise location models as described in Section 3.1.

BOLM

A business-oriented location model (BOLM) (as defined in Definition 1) represents a location model that is composed of the objects and the relationships that are embodied with semantics and are able to be dynamically expanded and updated as the myriad enterprise relationships develop with the clients. The construction of a BOLM for an enterprise involves the calculation of the DR measurements (i.e., the semantics perspective) with respect to the enterprise clients and evolves these DR measurements with con-

tinuous interactions between the enterprise and the current clients engaged (i.e., the dynamic perspective).

The application of a BOLM (e.g., a mobile workforce deciding the service priorities for clients in a certain location area) accordingly involves consulting these relationship measurements together with additional myriad considerations (attributes) of service requests (e.g., request distance, emergency, profit, etc.). For simplicity, a liner weighted scheme is exerted on these service attributes to attain the proportion of the significance share for a given service request (invoked by a given client) besides the other proportion of the significance share coming from relationship measurements, followed by another liner weighted scheme combining both significance shares.

The following exemplifies the BOLM application (that subsequently will be evaluated in Section 4.1):

- A logistic enterprise A has seven clients (spread over different regions of a given area): B, C, D, E, F, G, H (that simultaneously make requests to Enterprise A for its services by the temporal order of {C, F, D, E, H, B, G}. Assume their distances to a mobile workforce (shipping vehicle) of Enterprise A (arriving in this designated area) are increasingly ordered as follows: {C, E, F, H, D, B, G}.
- Based on the client attributes and their weights shown in Table 2, the DR measurements of BOLM for these clients are calculated as shown in Table 3 (with Algorithm 1).

Table 2. The attributes (and their corresponding weights) considered in DR measurements

DR Attributes	**Average Shipment**	**Average Revenue**	**Average Positive Feedback**
Weights	0.5	0.2	0.3

Note: The weights (0.5, 0.2, 0.3) represent the relative degrees of importance considered by Enterprise A when differentiating its clients in terms of the three chosen attributes.

Table 3. Results of DR measurements

	B	**C**	**D**	**E**	**F**	**G**	**H**
Average Shipment	7	5	4	7	2	1	3
Average Revenue	3	5	3	5	4	6	7
Average Positive Feedback	4	5	6	3	3	6	7
DR Measurement	5.3	5	4.4	5.4	2.7	3.5	5

Note: A SL value (e.g., a value ranging from 1 to 10) for a designated attribute denotes a subjective performances score with respect to the attribute from the viewpoints of Enterprise A.

- While the mobile workforce wirelessly accesses the enterprise's BOLM for the decision of an appropriate arrangement to serve the clients requests, additional myriad attributes of service requests (request distance, emergency, and profit) are taken into account as shown in Table 4. Subsequently, the DR measurements and the service request considerations are combined as shown in Table 5, and the decision of the priorities of the clients to serve is then determined.
- Table 6 contrasts the BOLM service-request arrangement with the others' (First-In-

Table 4. Service request considerations

		B	C	D	E	F	G	H
Request distance	0.5	7	1	5	1	2	7	2
Emergency	0.1	3	3	6	1	5	3	4
Profit	0.4	4	4	7	2	1	4	3
Weighted SUM		5.4	2.4	5.9	1.4	1.9	5.4	2.6

Table 5. The resulting service request arrangement by the BOLM method

		B	C	D	E	F	G	H
DR Measurement	0.8	5.3	5	4.4	5.4	2.7	3.5	5
Service Request Considerations	0.2	5.4	2.4	5.9	1.4	1.9	5.4	2.6
Weighted SUM		5.62	4.48	5.19	4.6	2.54	3.88	4.52
Order		1	5	2	3	7	6	4

Table 6. A contrast between the different service-request arrangements

	B	C	D	E	F	G	H
First-In-First-Out	6	1	3	4	2	7	5
Shortest-Distance-First	6	1	5	2	3	7	4
BOLM	1	5	2	3	7	6	4

First-Out and Shortest-Distance-First), manifesting that the BOLM method takes on a different service-request arrangement (that will be shown to outperform First-In-First-Out and Shortest-Distance-First in Section 4.1).

PNLM

Partner-network location model (PNLM) enables the realization of the benefits of cooperation between enterprises residing in different location regions, in terms of the expanded market share of clients or the increased relationships with clients. This realization is able to assist an enterprise mobile manager to negotiate with potential partner enterprises of a certain location area regarding their cooperation contracts or deals.

Suppose a PNLM is formed because of the cooperation between enterprises A and B. The PNLM from A's perspective is then defined as in Definition 3. A picturesque view of this PNLM is shown in Figure 3. (Figure 4 then shows that of the PNLM from B's perspective.) In other words, PNLM is constructed out of a PR relationship between A and B, and subsequently IR and MR are generated. Since IR and MR are directional relationships, a PNLM accordingly is formulated as a directional model (i.e., from the perspective of A (or B)). The PNLMs from different perspectives differ with each other in terms of the different measurements calculated.

The benefits of exerting PNLM in an enterprise are exemplified by two scenarios as shown below:

- **Competitors cooperation scenario:** Two competitive enterprises (such as the former Compaq and HP) cooperate through PNLM that facilitates things such as expanding market share in myriad location regions, perceiving client relationships development, recognizing their services overlap, discerning the increase in their service scope, and so forth. This scenario emphasizes the importance of the number of clients increasing because of cooperation.
- **Vertical supply chain scenario:** A manufacture enterprise residing in southern Taiwan seeks a northern logistic enterprise to cooperate through PNLM for providing better services to the manufacture enterprise's

Definition 3.

$$\begin{aligned}
&\text{PNLM(A,AB)} = \text{def } \exists x_1, x_2, x_3, ..., x_m \text{ is CU}(x_1,\text{A}), \text{CU}(x_2,\text{A}),\ldots, \text{CU}(x_m,\text{A}) \\
&\qquad \exists y_1, y_2, y_3, ..., y_n \text{ is CU}(y_1,\text{B}), \text{CU}(y_2,\text{B}),\ldots, \text{CU}(y_n,\text{B}) \\
&\qquad \exists z_1, z_2, z_3, ..., z_p \text{ is CU}(z_1,\text{A}), \text{CU}(z_2,\text{A}),\ldots, \text{CU}(z_p,\text{A}) \\
&\qquad\qquad \text{also is CU}(z_1,\text{B}), \text{CU}(z_2,\text{B}),\ldots, \text{CU}(z_p,\text{B}) \\
&\text{A and B have PR(A, B)} \\
&\bigwedge_{i=1}^{m}\bigwedge_{j=1}^{n}\bigwedge_{k=1}^{p} (\text{BU(A)} \wedge \text{BU(B)} \wedge \text{CU}(x_i,\text{A})\Big|_{DR(x_i,A)} \wedge \text{CU}(y_j,\text{B})\Big|_{IR(y_j,A)} \wedge \\
&\qquad\qquad \text{CU}(z_k,\text{A})\Big|_{MR(z_k,A,B)})
\end{aligned}$$

Figure 3. PNLM of enterprise A and B in A's point of view

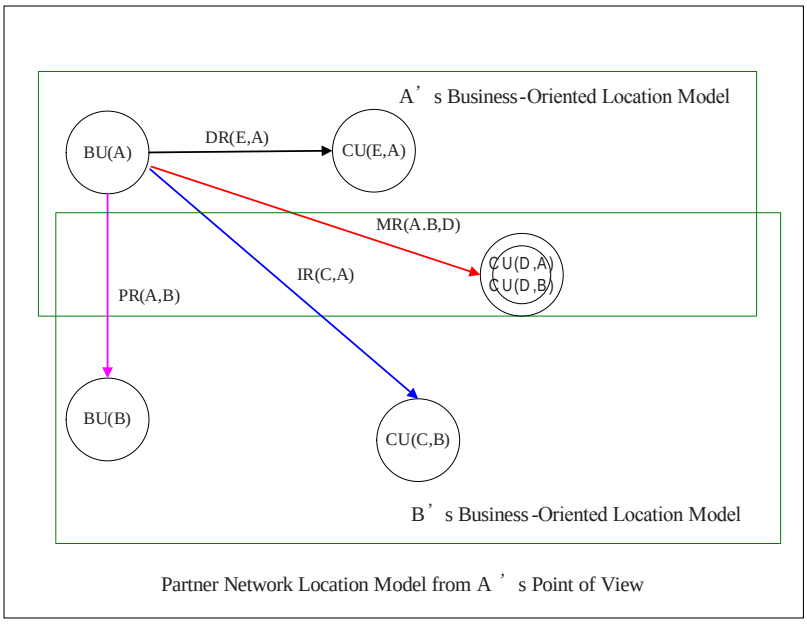
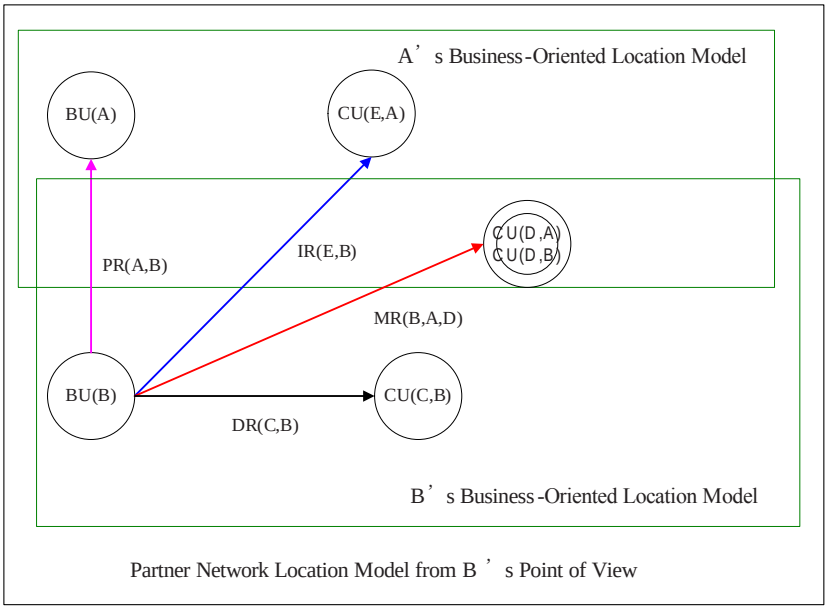


Figure 4. PNLM of enterprise A and B in B's point of view



clients in northern Taiwan. This scenario then emphasizes the increased relationship measurements because of the cooperation.

With the two aforementioned scenarios, a PNLM performance evaluation algorithm is provided (as shown in Algorithm 4) for evaluating the performance of the cooperation with a Target Enterprise BU from the perspective of a Source Enterprise BU. The performance evaluation is represented as a vector comprising a CAN value and a SIM value. *The CAN value denotes the increase in the number of the clients because of the cooperation between Source Enterprise BU and Target Enterprise BU, and the SIM value then stands for the increased relationship measurements because of the cooperation (i.e., the sum of the IR and MR measurements).*

The rationale behind this performance vector is twofold: (1) Different enterprises might have different objectives in the cooperation (as exemplified in the above scenarios) and thus the performance vector is unfolded as a vector of a CAN value and a SIM value (instead of a single scalar); (2) rendering different [CAN, SIM] vectors (corresponding to different Target Enterprises) on a 2-dimension space, it is easy to snatch the various strengths between different cases of enterprise cooperation (i.e., an exemplar of advanced location intelligence derived from dynamic semantic reasoning and computation).

As follows is an exemplar regarding the computation of the CAN and SIM values:

- In Figure 5, (a) shows the picturesque view of Enterprise αof clients {B, C, D, E, F}; (b) shows that of Enterprise βof clients {A, B, C, F, H}; (c) then denotes the picturesque view of PNLM(α,β) that is a directional view from Enterpriseα's perspective.
- Table 7 shows the DR measurements in Enterprise α's BOLM and in Enterprise β's BOLM; Table 8 then exhibits those IR measurement and MR measurement that are generated from the creation of PNLM(α,β).
- Since CAN denotes the increase in the number of the clients because of the cooperation between Source Enterprise α and Target Enterprise β, the CAN value of PNLM(α,β)

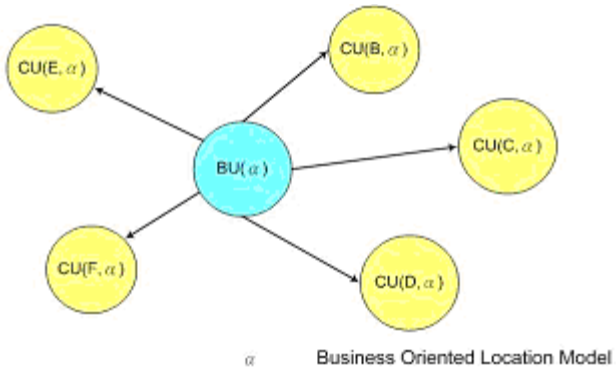
Algorithm 4. PNLM performance evaluation

Function PNLM_Performance (PNLM, Source Enterprise BU, Target Enterprise BU)

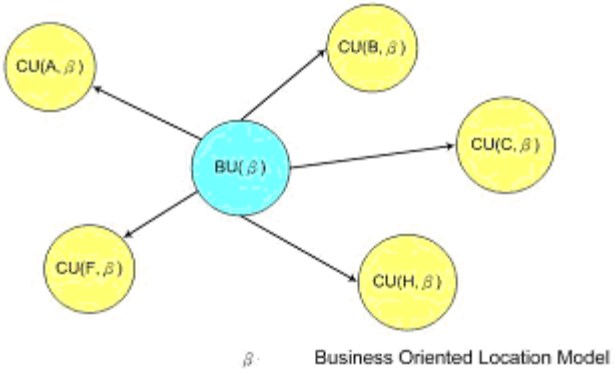
1. From the Source Enterprise's perspective, calculate the increase in CU because of the given PNLM and give rise to a statistics named a CAN value.
2. From the Source Enterprise's perspective, calculate the increased amount of measurements in relationships because of IR and MR encountered. This amount is named a SIM value.
3. Set the PNLM performance vector with respect to the Target Enterprise BU as a vector of [CAN, SIM].

Note: If Source Enterprise BU cannot attain relevant client's attribute values (CU) from Target Enterprise BU during the calculation of the relationship measurements, then this CU would be considered as an OU. Source Enterprise BU subsequently calculates the relationship measurements in terms of the OU's attribute values.

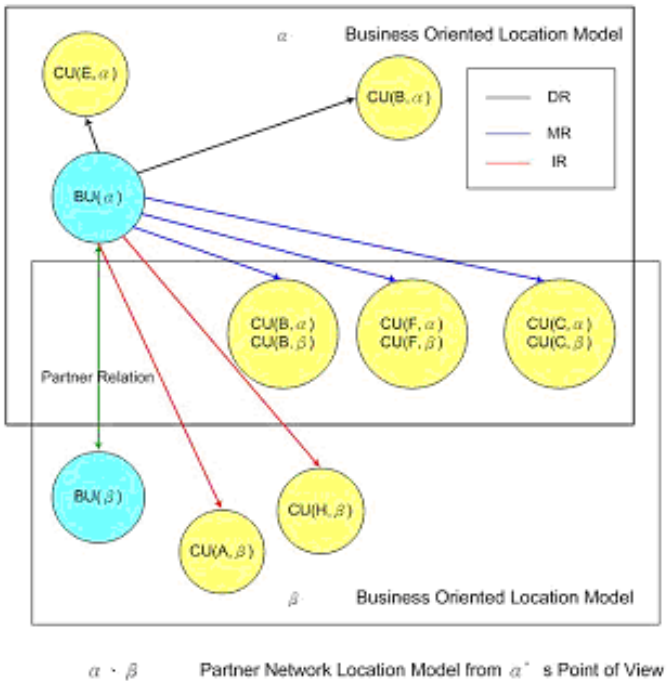
Figure 5. (a) α's BOLM; (b) β's BOLM; (c) PNLM(α,β)



(a)



(b)



(c)

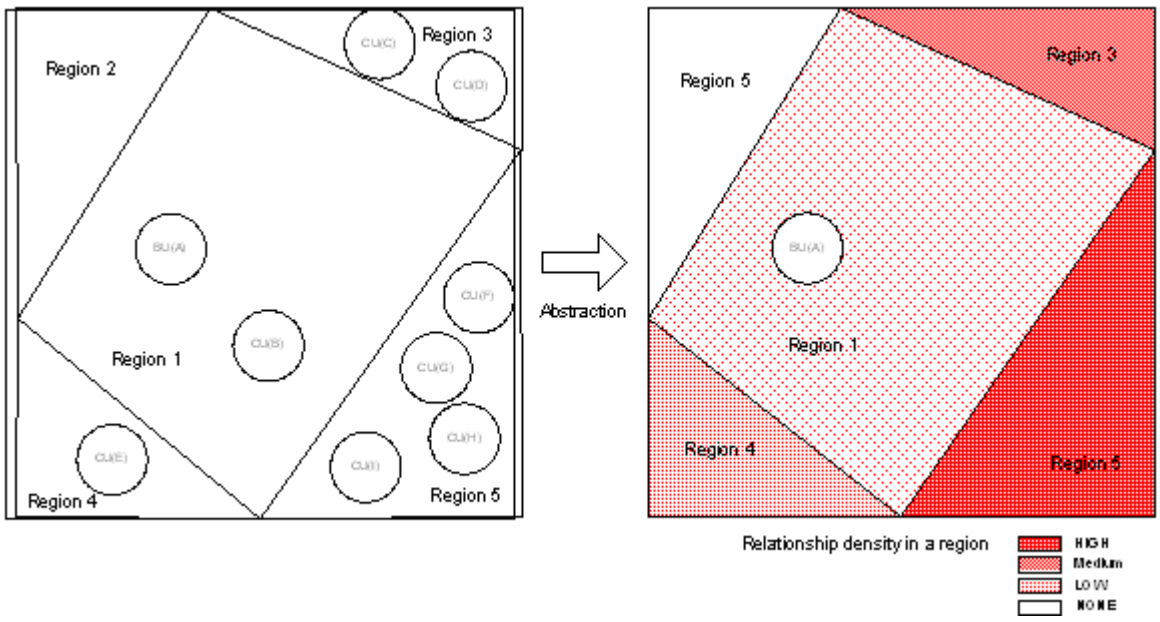
Table 7. The DR measurements presumed for Enterprise αand Enterprise β

	A	B	C	D	F	H	I	J
α	N/A	2	7	4	3	N/A	N/A	N/A
β	6	4	3	N/A	5	1	N/A	N/A

Table 8. The IR and MR measurements presumed from the perspective of Enterprise α

	A		B		C		D		F		H		I		J	
	IR	MR	IR	MR	IR	MR	IR	MR	IR	MR	IR	MR	IR	MR	IR	MR
α	4	N/A	N/A	2	N/A	-4	N/A	N/A	N/A	2	1	N/A	N/A	N/A	N/A	N/A

Figure 6. Example of BOLM abstract



accordingly equals to 2 (that arises because of the clients in{A, H}becoming the indirect clients of Enterprise α due to the partnership with Enterprise β).

- Since SIM represents the increased relationship measurements (i.e., the sum of the IR measurements and the MR measurements) because of the cooperation, the SIM value of PNLM(α,β) accordingly equals to 5 (that is calculated by summing the values {4, 2, -4, 2, 1}[6] that denote those IR and MR measurements attained in enterprises {A, B, C, F, H} respectively).

LMP

Location model platform (LMP) is a platform for the exchange of BOLM abstracts. In other words, the shared BOLM abstracts empower the search of potential enterprises to cooper-

Algorithm 5. BOLM abstract construction algorithm

Function BOLM_Abstract_Construction (BOLM)

1. From BOLM, identify all CUs that have direct relationship (DR) to the business.
2. In each geographical region, sum up all DR measurements and multiply this sum with the number of CUs in the region, obtaining a scalar representing a Region Relationship (RR).

   $\forall$ Region R_j € BOLM(C), *j=1,…,n; CU(*x_p*,C)*$\subset R_j$, *p=1,…,m;*

   $RR_j = (\sum_{p=1}^{m} DR(x_p, C)) * m$
3. Total Region Relationship (TR) is the sum of all the RRs.

   $TR = \sum_{j=1}^{n} RR_j$
4. Region Relationship Percentage is defined as the percentage of a designated RR to TR.

   $RRP_j = \frac{RR_j}{TR} \times 100\%$, *j = 1,…..,n*;
5. Assign Semantic Levels (region density) to RRPs :
   - High Density 100% > RRP > = $\frac{100\%}{\text{total region}}$
   - Medium Density $\frac{100\%}{\text{total region}}$ > RRP > = $\frac{100\%}{m \times \text{total region}}$
   - Low Density $\frac{100\%}{m \times \text{total region}}$ > RRP > 0
   - None RRP=0

   Note : m is a tuning parameter determined by the platform designer.
6. Label the region density (the semantic level of RRP) in every region.
7. Return the labeled abstract.

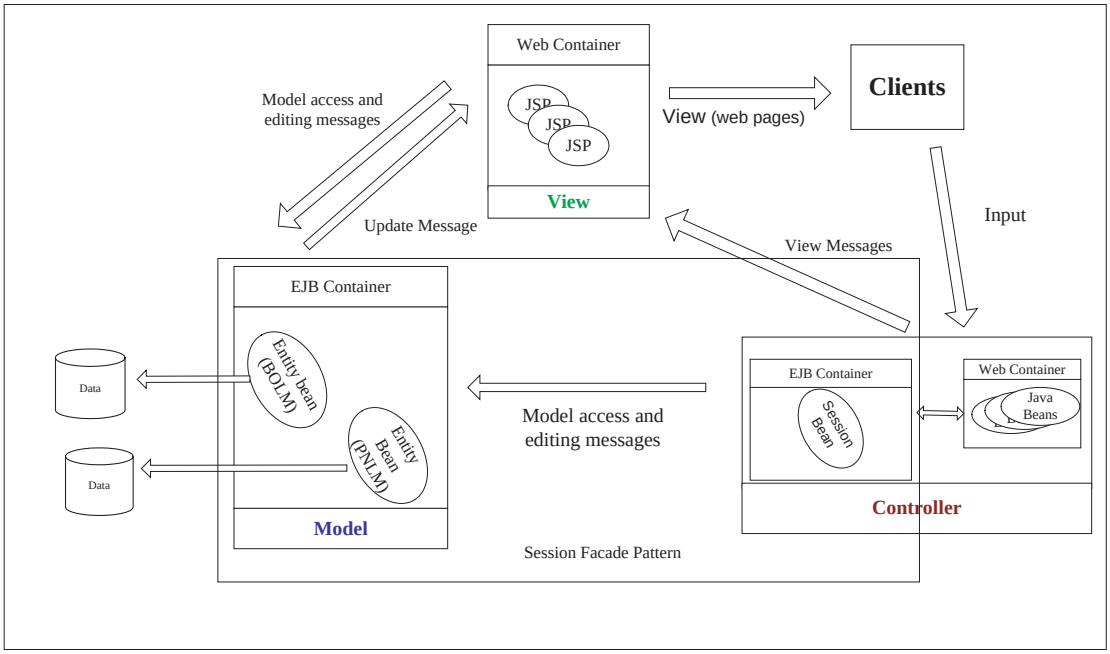
ate without exposing enterprises' confidential and private information (i.e., another exemplar of advanced location intelligence derived from dynamic semantic reasoning and computation). Figure 6 shows an example of BOLM abstracts, and Algorithm 5 lists the algorithm for producing BOLM abstracts.

A BOLM abstract unfolds as a distribution of various business sizes on a designated geographical coverage that is composed of a certain number of geographical regions (as shown in Figure 6). The business size in a region is represented with a certain semantic label (High Density, Medium Density, or Low Density) that is allocated according to the relative strength of ongoing business unfolding in the designated region with respect to that developing in all of the regions. Ongoing business is measured by the size of the clients and the size of the relationship measurements.

EVALUATION

Our DSLM is implemented using the service-oriented architecture (Machiraju, 2001). J2EE and Enterprise JavaBeans technology are used to develop the DSLM system (as shown in Figure 7). The Model View Controller (MVC) pattern is exerted to hinge on a clean separation of objects into one of three categories: **models** for maintaining data (BOLM EJB Entity Bean and PNLM EJB Entity Bean), **views** for displaying all or a portion of the data (JSP Javabeans), and **controllers (EJB Session Beans)** for handling events that affect the model or view(s). Because of this separation, multiple views and controllers can interface with the same model. Even new types of views and controllers that never existed before can interface with a model without forcing a change in the model design.

Figure 7. DSLM J2EE implementation architecture



In this section, different sets of experiments are employed to realize the claimed contributions of the three solutions (BOLM, PNLM and LMP) in Section 4.1, 4.2 respectively. Section 4.3 then provides a short discussion of the evaluation results. Although the full-scope justifications won't be available until fielded experiments (i.e., attaining long-term observation of the DSLM performance in relevant enterprises/industries) are constructed, these results anew shed light on future enterprise LBS.

BOLM Evaluation

This evaluation unfolds itself by exerting a logistic enterprise BOLM example on the task of service request arrangement (as shown in Figure 8) in order to show the increased values brought by the decision support of BOLM.

In this logistic enterprise BOLM example, six types of clients (that are commonly perceived as shown in Table 9) dynamically generate requests to the enterprise (of which its mobile workforce with their shipping vehicles are responsible for fulfilling the requests of the clients). Each request is composed of a variety of attribute values (such as request distance, request unit price, and request quantity) that are also dynamically generated (given the assumption that there are limits set for request distance and request quantity in this example).

This example compares three different methods for the task of service request arrangement in terms of the average resulting value to the enterprise:

- **First-In-First-Out:** Serving requests by the order of the request sequence.
- **Far-Distance-Based:** Serving requests by the decreasing order of the request distances.[7]

Figure 8. BOLM experiment system architecture

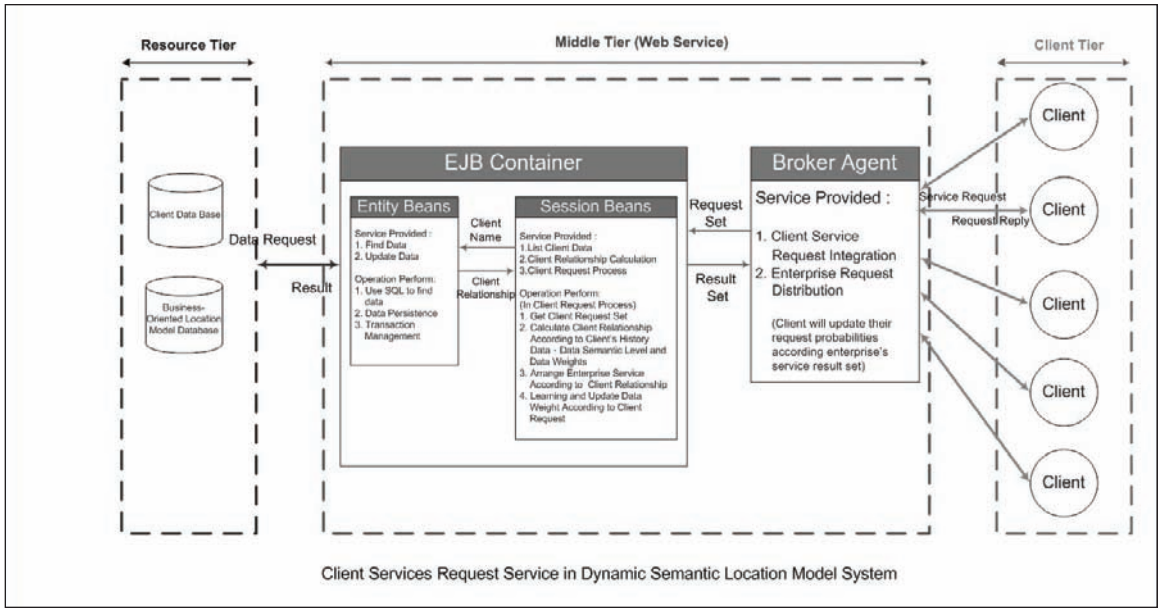


Table 9. Client request type in experiment environment

Attribute Client Type	Request Distance	Request Quantity	Request Unit Price
1	Long	High	High
2	Short	High	High
3	Long	Low	High
4	Short	Low	High
5	Long	High	Low
6	Short	High	Low

**(Note: Request Distance: "Long" represents the range from 500 to 1300 and "short" from 100 to 499; Request Quantity: "High" represents the range from 50 to 130 and "Low" from 10 to 49; Request Revenue = Request Quantity * 1.5 (with High Request Unit Price); Request Revenue = Request Quantity (with Low Request Unit Price); Request Revenue is also multiplied by 3.5[8] while Request Distance is Long)*

- **BOLM:** Serving request by the order of client relationship measurements.

In Figure 8, Broker Agent pools 10 clients[9] requests (forming a request set) and sends them to the enterprise service arrangement method periodically. The request-sending magnitude of a client controls how often this client will post requests to the enterprise. Clients will tune their request-sending magnitude in the following ways: *if the enterprise rejects a client's request, the client will tune down the request-sending magnitude, but will raise this magnitude vise versa.*[10] The value of a request set (i.e., the 10 pooled client request per period) will be calculated with equation (1) (in which the value of a request set for the logistic enterprise is proportional to the revenue received but reverse proportional to the distance transported and the quantity carried).

*Request Set Value = Total Request Revenue / (Total Request Distance * Total Request Quantity)* (1)

A reply set is the arrangement results (with respect to a given request set) returned by the enterprise service arrangement method.[11] Equation (2) computes the value of the reply set.

*Reply Set Value = Total Reply Revenue / (Total Reply Distance * Total Reply Quantity)* (2)

Request Set Value *and* Reply Set Value *are two metrics employed to evaluate the performance of the service arrangement methods. High* Request Set Value *indicates the continuity of intensive business opportunities, and high* Reply Set Value *then denotes quality arrangement between service requests.*

Distinguished from First-In-First-Out *and* Far-Distance-Based, *the BOLM method employs LMS weight update rule [4,17] to* evolve the weights of the service-request attributes (as shown in Table 4) for the purpose of adaptively serving clients in light of the dynamic magnitudes of their service requests. *This adaptation aims at adjusting the*

weights toward the direction of high Request Set Value *and high* Reply Set Value. *The weight learning equation is shown in equation (3).*

*Weight = Weight + learning rate * (Request set Value / Reply set Value)*Xi* (3)

*Note: learning rate = 0.1; If the weight of the distance attribute is under tuning, then Xi represents the sum of distance in reply set.

With the evaluation experiment setup addressed, evaluation results show that the choice of the request arrangement method will affect the magnitudes of client requests, request set values, and reply set values. On the other hand, *a good request arrangement method should be able to stably generate competitively high client request set values and enterprise reply set values by continuously arranging the requests of the clients to serve prosperously.*

We gradually experiment up to 1,000 client request sets (an exemplar of a dynamically generated request set is shown in Table 10). That is, there are 10,000 client requests in total sent to each request arrangement method. An investigation of which method can stably generate higher client request set values and enterprise reply values is explored. Figure 9 then shows the evaluation results in which the BOLM method outperforms the other two methods (FIFO and Far-Distance-Based) throughout the whole experiment process (i.e., from a small number of request sets to a great number of request sets). Furthermore, the two values stay quite stable throughout the experiment process with the BOLM method.

Figures 10, 11, and 12 show the client request magnitude changes (i.e., the dynamics of the distribution of the requests) for the three request arrangement methods. In the FIFO method, the client types 2, 4, and 6 that make Low distance requests are the higher request magnitude ones (i.e., with wider ranges along the dimension of Sum of Request Magnitude). In contrast, the client types 1, 3, and 5 are the higher magnitude ones in the Far-Distance-Based method. However, in the BOLM method, the client types 1, 2, and 5 are the higher magnitude ones. This is due to request revenue being computed by request distance and request quantity. The client types 1 and 2 have High request quantity and the client type 5 has Higher distance compensation than the client type 6. In other words, the BOLM method is able to come up with valuable clients to serve.

Figure 9. The evaluation results for the three request arrangement methods

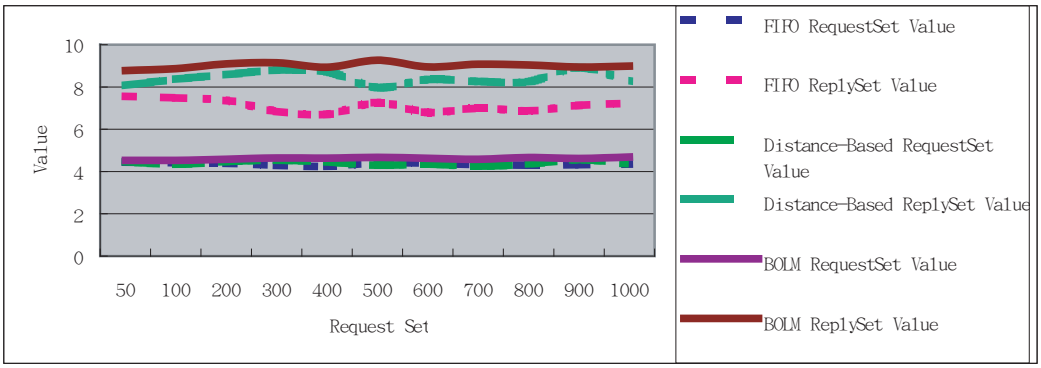


Table 10. Exemplars of a request set and the reply set obtained by the BOLM method

Client Request Set are :
The No.0 request set is :client = 2; distance = 404; revenue = 189; quantity = 126
The No.1 request set is :client = 3; distance = 1156; revenue = 182; quantity = 52
The No.2 request set is :client = 1; distance = 608; revenue = 309; quantity = 59
The No.3 request set is :client = 6; distance = 143; revenue = 61; quantity = 41
The No.4 request set is :client = 5; distance = 817; revenue = 73; quantity = 21
The No.5 request set is :client = 1; distance = 1069; revenue = 346; quantity = 66
The No.6 request set is :client = 6; distance = 139; revenue = 58; quantity = 39
The No.7 request set is :client = 4; distance = 181; revenue = 66; quantity = 66
The No.8 request set is :client = 2; distance = 319; revenue = 169; quantity = 113
The No.9 request set is :client = 6; distance = 142; revenue = 42; quantity = 28
request distance = 4978.0
request revenue = 1495.0
request quantity = 611.0
request set value = 4.915244095295898
++
The Reply Set after Enterprise process are :
The request set Enterprise accepted : client = 5; distance = 817; revenue = 73; quantity = 21
The request set Enterprise accepted : client = 1; distance = 608; revenue = 309; quantity = 59
The request set Enterprise accepted : client = 1; distance = 1069; revenue = 346; quantity = 66
The request set Enterprise accepted : client = 2; distance = 404; revenue = 189; quantity = 126
The request set Enterprise accepted : client = 2; distance = 319; revenue = 169; quantity = 113
The request set Enterprise rejected : client = 3; distance = 1156; revenue = 182; quantity = 52
The request set Enterprise accepted : client = 6; distance = 143; revenue = 61; quantity = 41
The request set Enterprise accepted : client = 6; distance = 139; revenue = 58; quantity = 39
The request set Enterprise accepted : client = 6; distance = 142; revenue = 42; quantity = 28
The request set Enterprise rejected : client = 4; distance = 181; revenue = 66; quantity = 66
reply distance = 3641.0
reply revenue = 1247.0
reply quantity = 493.0
reply value = 6.947024896198523

Figure 10. Client request magnitude changes in the FIFO method

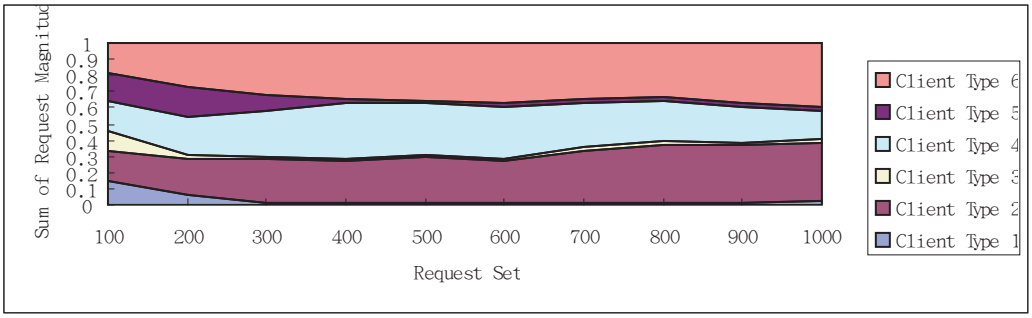


Figure 11. Client request magnitude changes in the far-distance-based method

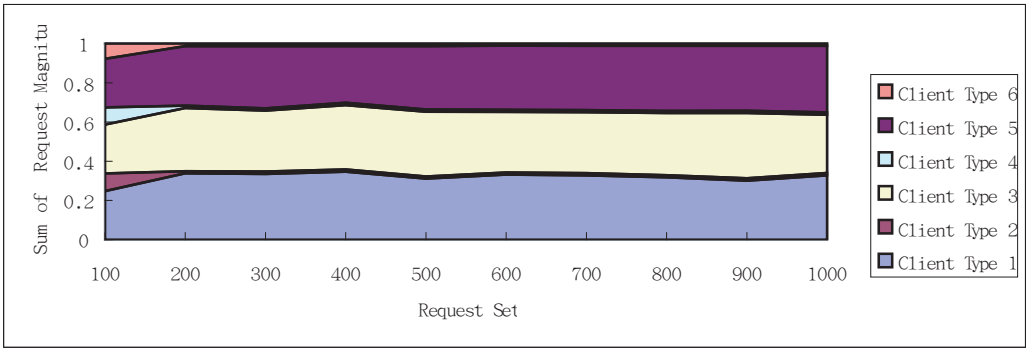
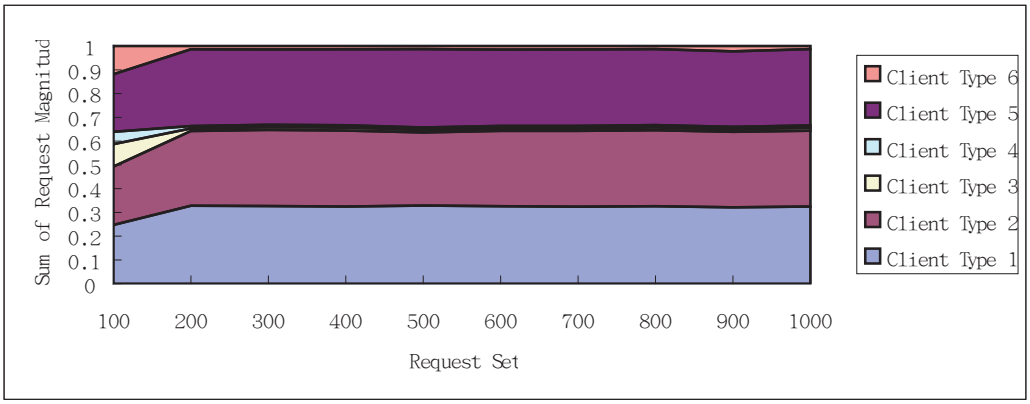


Figure 12. Client request magnitude changes in the BOLM method



Evaluation of PNLM and LMP

The application of PNLM and LMP involves factors considered in the decisions of seeking and evaluating enterprises for intercooperation or interoperability (that usually induces monumental policies, practices, contracts on enterprise alliance). However, this section aims at delivering certain prospects realized by the existence of PNLM and LMP in terms of two scenarios and their positive evaluation results (that will be detailed in Section 4.2.1 and 4.2.2, respectively).

These scenarios unfold themselves around a common geographical setting and circumstance, detailed as follows:

- As shown in Figure 13, in the geographical setting of Taipei city there are five geographical regions (identified by Region1 to Region5) covering the 36 smaller geographical areas (those in a 6*6 coordinated plane).
- Four enterprises (Enterprise1 to Enterprise4) are exerted for manifesting the two scenarios. These enterprises are presumed to be cooperative when conditions are met. Figure 14 exemplifies a fragmented portion of the client data of an enterprise. In this figure, each client record is composed of its location values (Location, LX, LY)[12] and its attribute values (PROP_A, PROP_B, PROP_C, PROP_D). Without loss of generality, these attributes are merely represented by symbols (A, B, C, D) (that can be tailored to a customized set of attributes in an industry and an enterprise of this industry can exert only a subset of the attributes for measuring its client relationship as shown in Figure 15(a) to (b), and the attribute values are randomly generated.
- A set of general principles underlying the selection of enterprises to cooperate [1,7,10] is employed to evaluate the performance of PNLM and LMP. The following are the principles: compatible (conforming to the needs), complementing (complementing with each other either in functions

Figure 13. An exemplar of the geographical setting

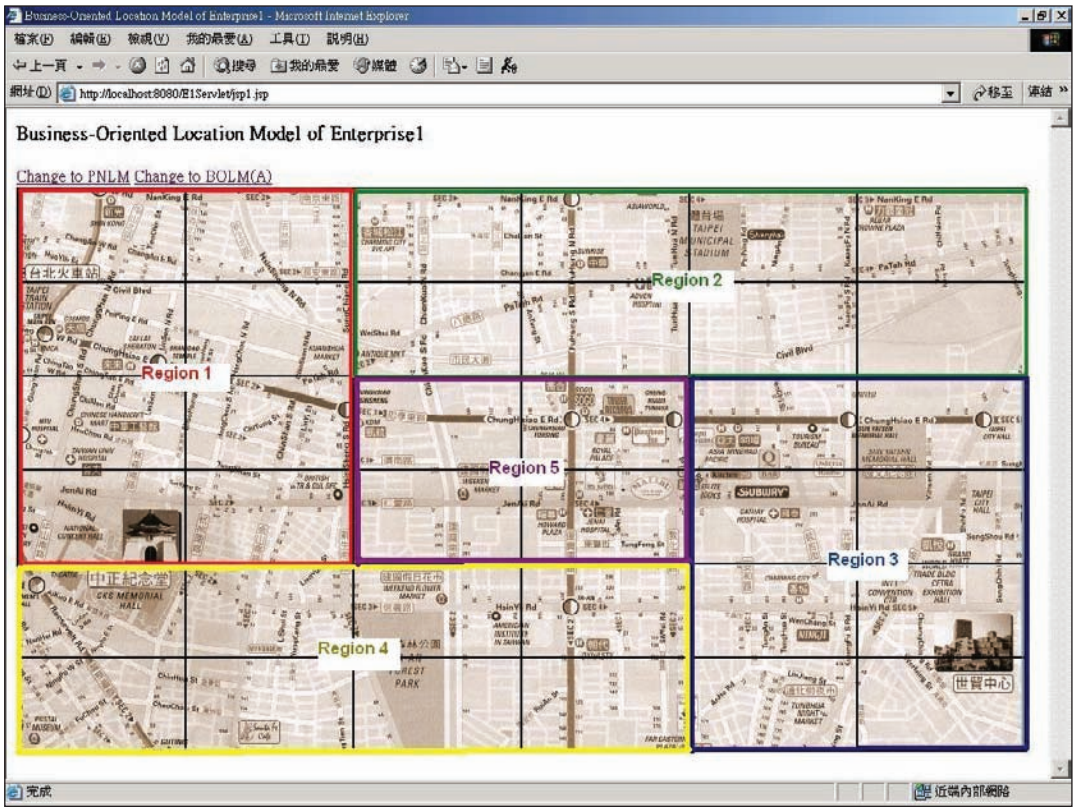


Figure 14. A fragmented portion of the client data for an enterprise

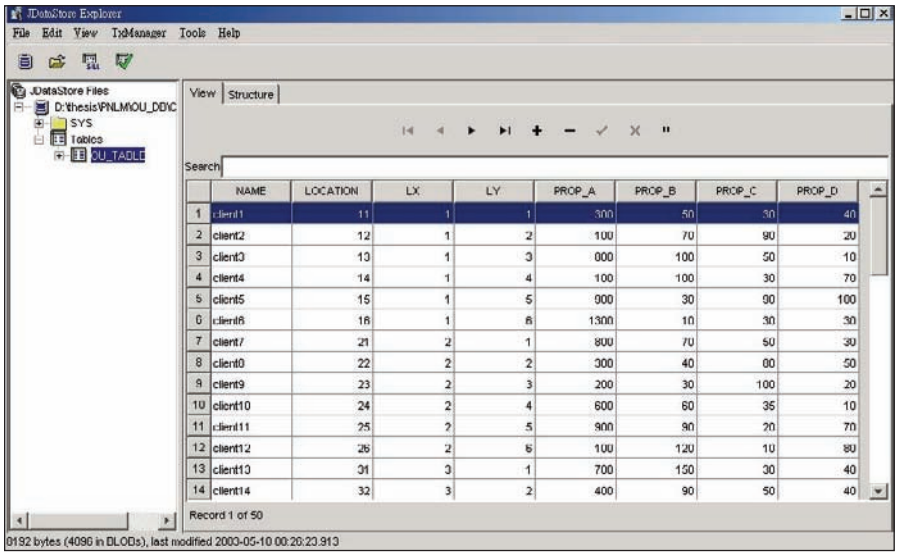
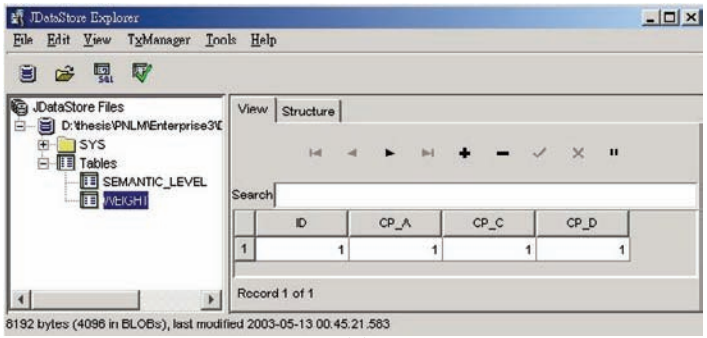
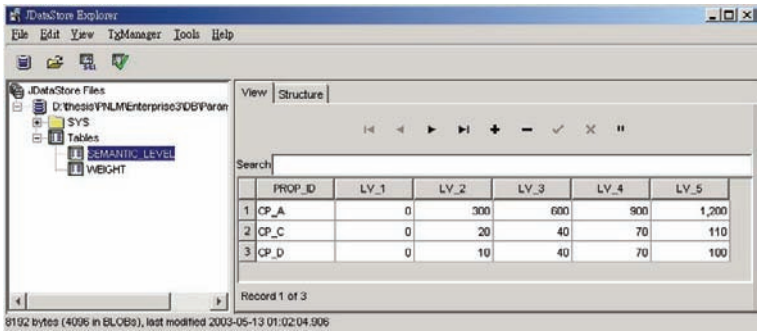


Figure 15. (a) shows the selection of the attributes (together their weights) in measuring the client relationship for Enterprise3; (b) then shows the mappings to semantic labels from the attribute values (e.g., the semantic label of 2 is assigned when the values of the attribute A are less than 600 but greater than 300).



(a)



(b)

or in operational locations), analytic (measuring the merits of the cooperation), and feasible (examining the feasibility of the cooperation), and homogeneous (sharing the same competitive goal).

Scenario 1

This scenario goes as follows: Enterprise1 is a logistic delivery company not residing in Taipei city and is searching for good regional operation representatives in Taipei.

This search and evaluation can be facilitated with LMP and PNLM and unfold as follows:

- Attain a set of candidates of regional representatives from LMP based on the needs and the conditions. For instance, Enterprise1 requires a representative that fairly engages in commercial activities (exhibiting its adequate business connection) in Region3 and Region5. This need can be fulfilled by looking up LMP (as shown in Figure 16) for the enterprises that are of mediocre densities of clients (e.g., densities roughly larger than 0.1 but less than 0.2) in Region3 and Region5, and the set of candidates accordingly comprises Enterprise3 and Enterprise4.
- Analyze the candidates with the measurements of CAN and SIM attained from PNLM. In other words, assess the values (the quantity of the indirect client size and the quality of their clients in terms of the relationship sum) brought to Enterprise1 through the partnership with Enterprise3 (Enterprise4). Figure 17 shows PNLMs associated with Enterprise1&3 and Enterprise1&4, and exhibits the (CAN, SIM) value vector of (20, 188^{13}) for the partnership with Enterprise3 and (20, 182) for Enterprise4. Accordingly, Enterprise3 outperforms Enterprise4 from the perspective SIM. This information can furnish Enterprise1's mobile managers with a valuable starting point to continue alliance

Figure 16. An exemplar of LMP

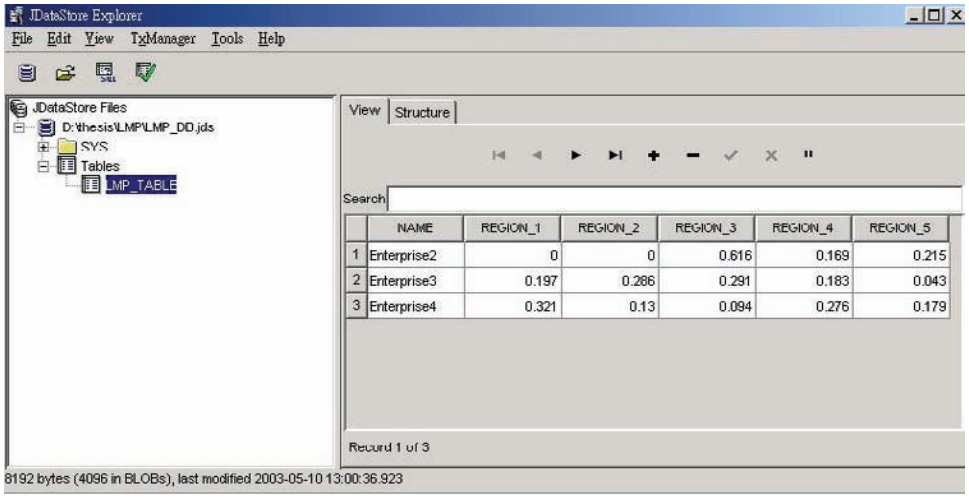
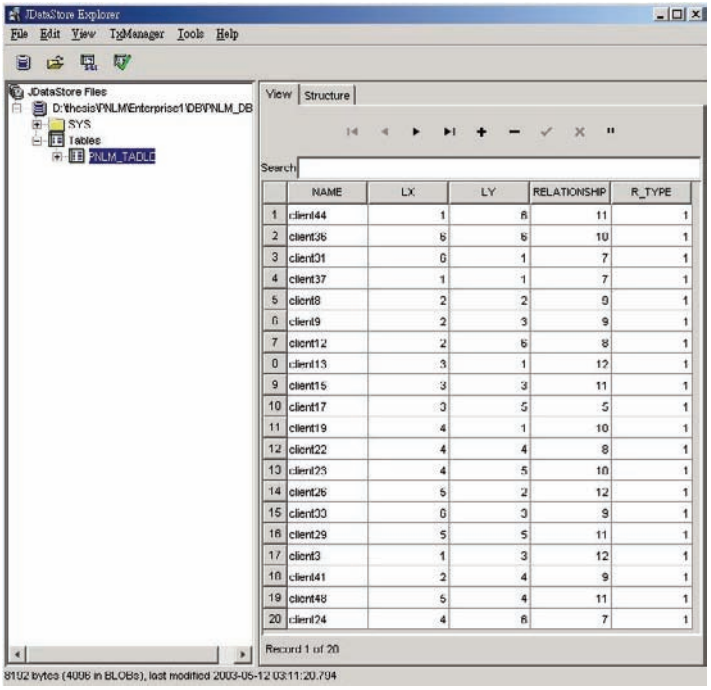
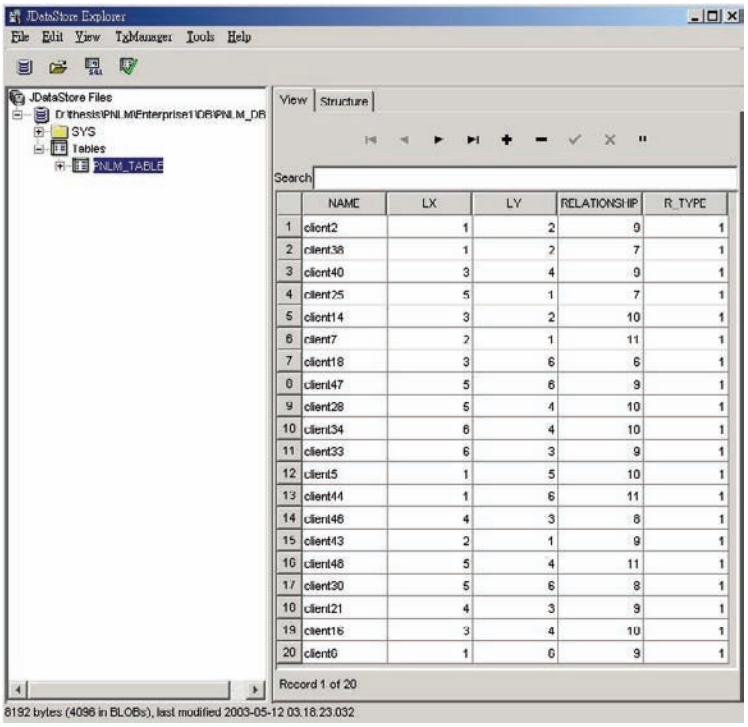


Figure 17. (a) Exemplifies the PNLM data associated with Enterprise1&3 and (b) then exemplifies that of Enterprise1&4



(a)



(b)

contract negotiation with Enterprise3 or Enterprise4 together with the consideration of additional alliance criteria.

- In this scenario, PNLM and LMP fully (partly) satisfy the following cooperative principles: compatible (LMP assists in generating the cooperation candidates in designated regions), analytic (PNLM measures the values CAN and SIM), and feasible (PNLM's ClickPoint function equips the evaluating task a close look of the to-be-clients).

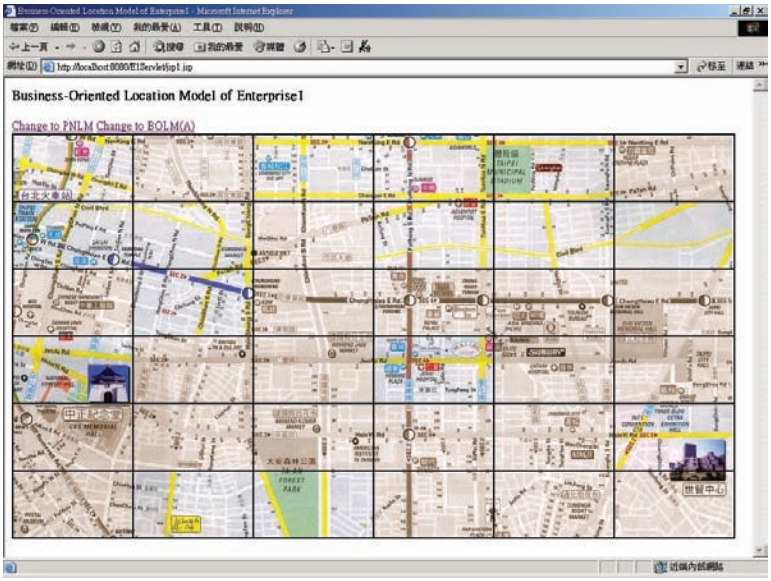
Scenario 2

This scenario proceeds as follows: Enterprise1 is a logistic delivery company and has clients spread over the Taipei geographical regions. Enterprise2, Enterprise3, and Enterprise4 simultaneously are soliciting the partnership with Enterprise1. Due to the limited resource, Enterprise1 has to select the one partner among them. Furthermore, Enterprise1 prefers a complement partner in Region3. This scenario differs from the previous one in Enterprise2-4 initiating the requests and thus PNLM sufficing to assist the decision making. Moreover, this scenario investigates the application of the cooperation criteria of the compliment principle.

This investigation with PNLM unfolds as follows:

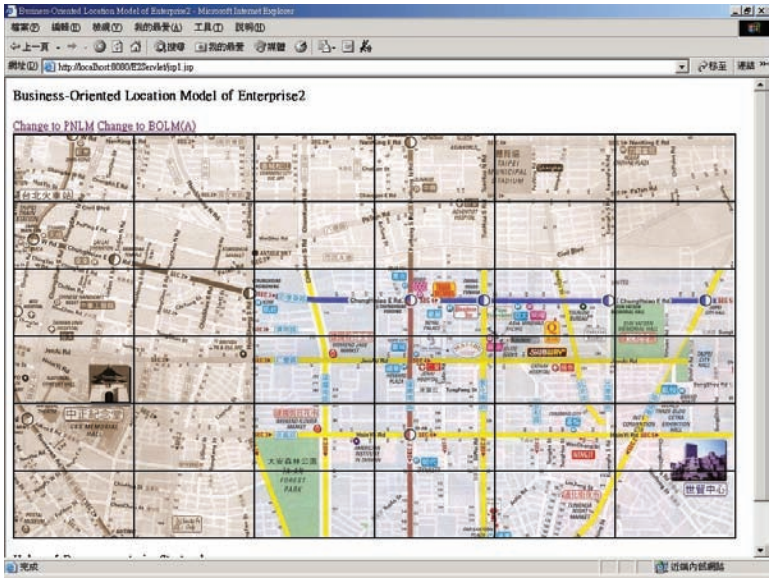
- In each BOLM of Figure 18, yellow-colored lines indicate the existence of clients in designated geographical areas colored nonblack/white (in contrast to the black/white geographical areas representing none-client areas).

Figure 18. The picturesque views of (a) Enterprise1's BOLM; (b) Enterprise2 's BOLM; (c) PNLM of Enterprise1&2

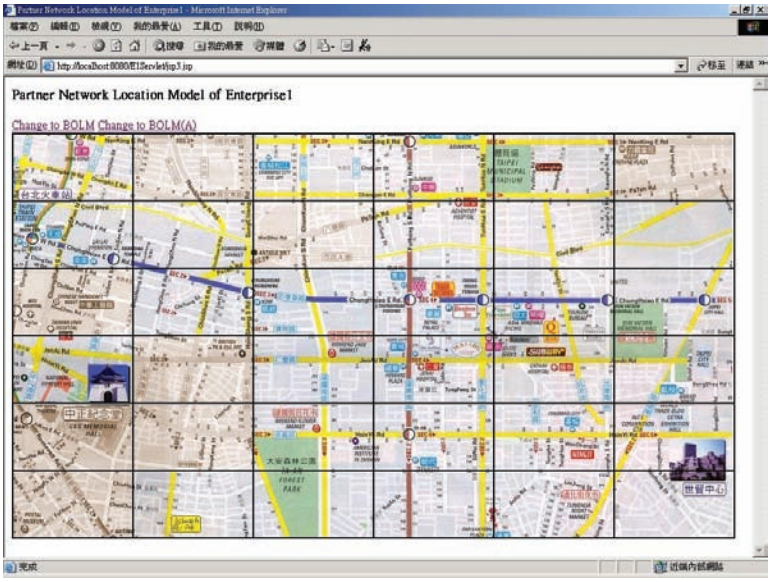


(a)

Figure 18. continued



(b)



(c)

- Figure 18(a) shows the clients of Enterprise1 spread over Region1&2, while Figure 18(b) exhibits those of Enterprise2 unfolding in Region3&5. The PNLM of Enterprise1&2 manifests clients spreading over Region1&2&3&5 (Region3&5 circled to indicate the addition of regions imposed by the partnership with Enterprise2 as shown in Figure 18(c), and enables the Click Point enquiries rendered on the colored areas for the details of selected clients (as shown in Figure 19).
- Figure 20(a) shows the fragmented data associated with the PNLM of Enterprise1&2 that manifests the (CAN, SIM) value vector of (18, 163) based on the BOLMs of Enterprise1 and Enterprise2 as shown in Figure 20(b) and Figure 20(c), respectively.
- The aforementioned steps are repeated for attaining the PNLMs of Enterprise1&3 and Enterprise1&4 (as shown in Figure 21) and the resulting values of CAN and SIM are presumed in Table 11.
- Table 11 summarizes the (CAN, SIM) value vectors for Enterprise2-4, enlightening certain clues to the decision making. That is, Enterprise2 prosperously serves the needs better than Enterprise3-4 because of its higher SIM in Region3 (for complementing Enterprise1) besides its overall high values of CAN and SIM. This analysis endows Enterprise1's mobile managers with valuable decision knowledge while negotiating contract deals with Enterprise2 (or Enterprise3/Enterprise4).
- In this scenario, PNLM fully (partly) satisfy the following cooperative principles: compatible (PNLM assists in the selection of a partner), complementing (PNLM locates a complement partner in Region3), analytic (PNLM measures the values of CAN and SIM), and feasible (PNLM's ClickPoint function equips the evaluating task a close look of the to-be-clients).

Figure 19. Click point enquiry (of the details of a client) and its reply window

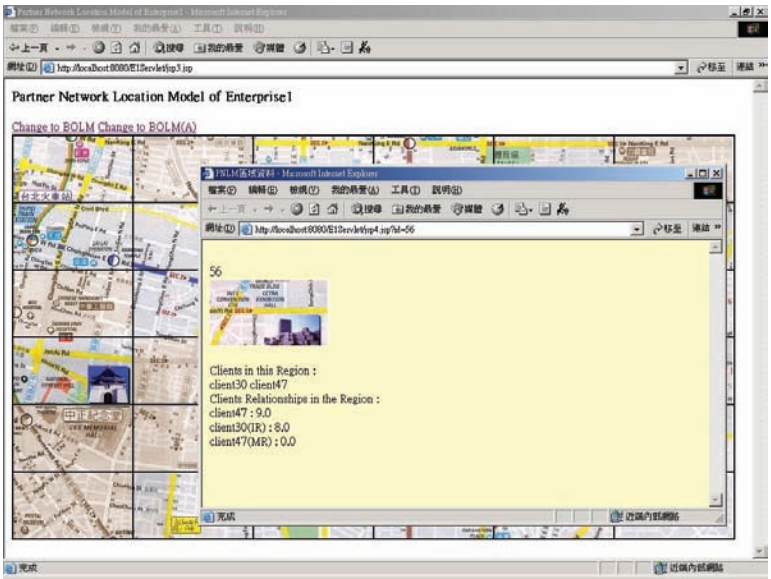
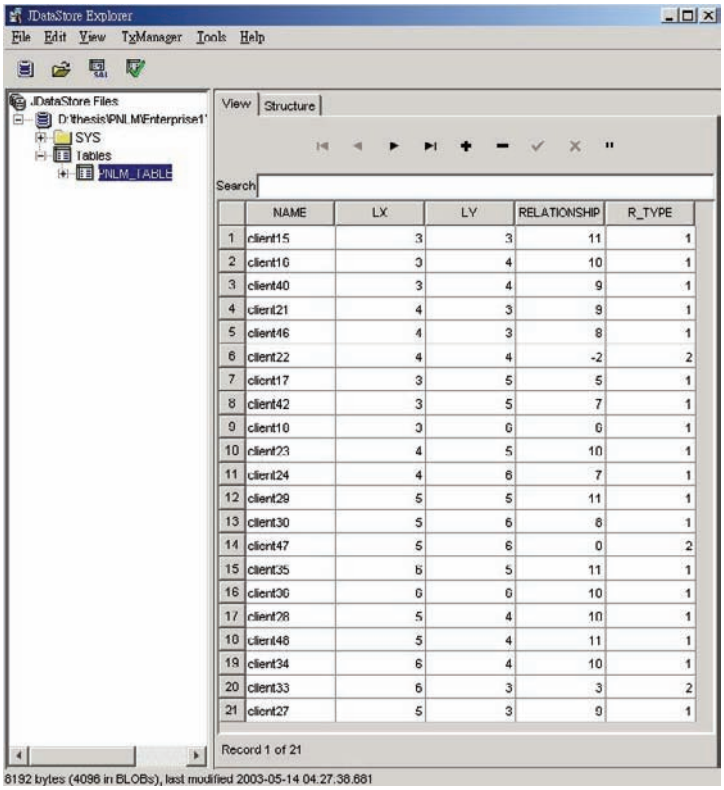
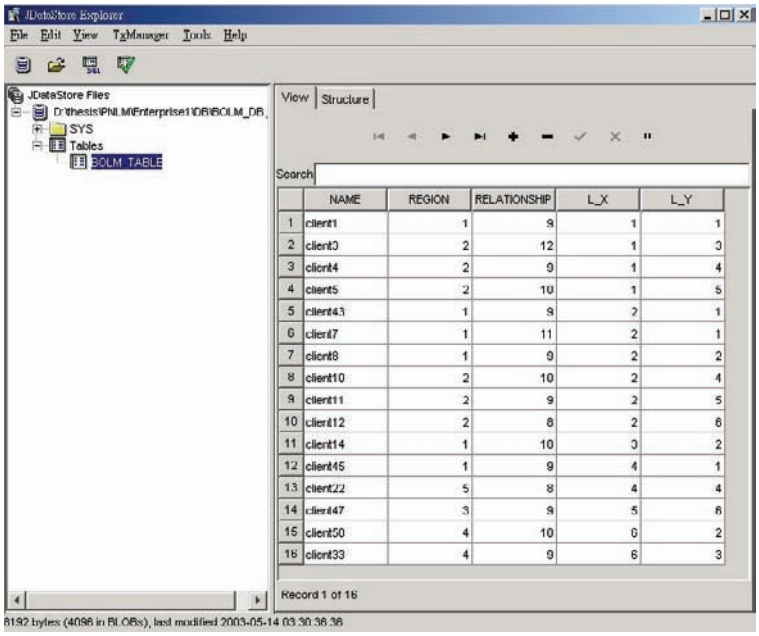


Figure 20. (a) shows the fragmented data associated with the PNLM of Enterprise1&2; (b) and (c) then show the data associated with the BOLMs of Enterprise1 and Enterprise2 respectively.

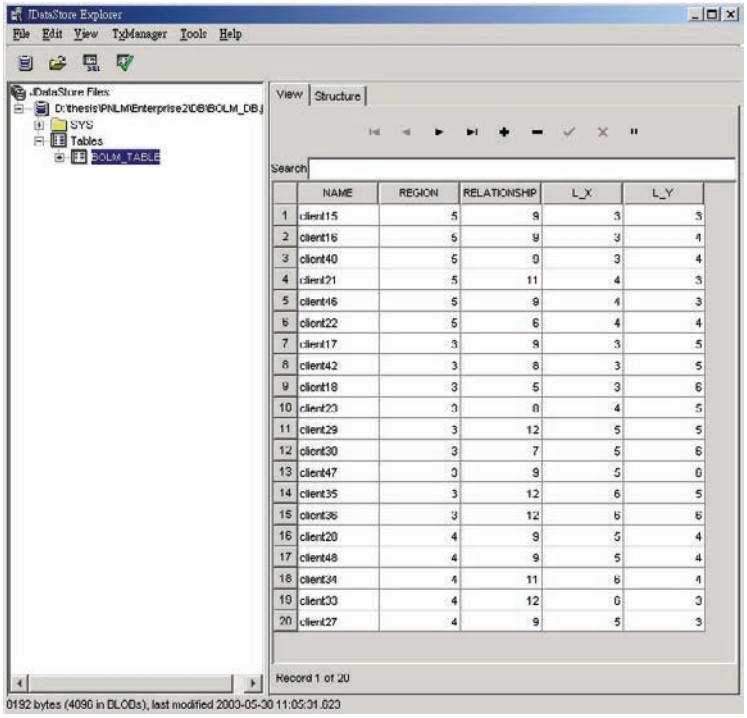


(a)



(b)

Figure 20. continued



(c)

Table 11. The (CAN, SIM) value vectors between Enterprise2-4

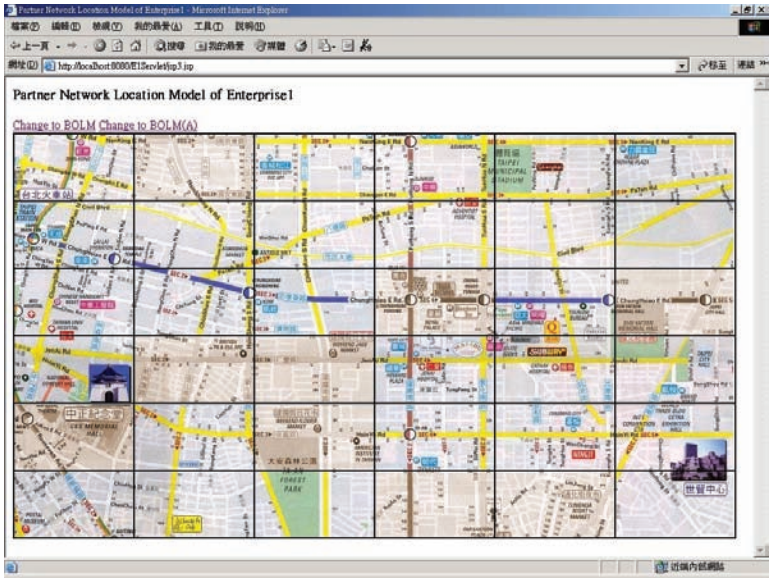
	(CAN)	(SIM)
Enterprise2	18	163 (75)*
Enterprise3	15	132 (43)*
Enterprise4	14	125 (13)*

*Note: * indicates the SIM value attained only from the clients in Region3*

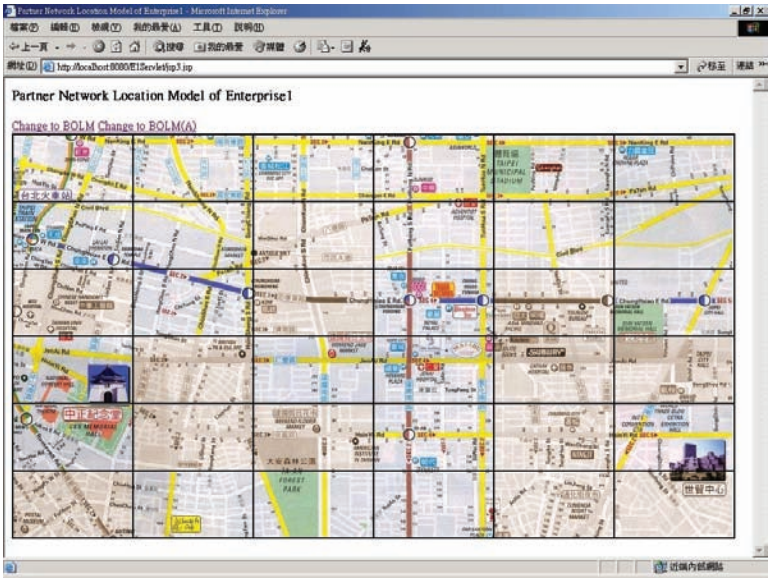
Discussion

The aforementioned evaluation results aim to justify the contributions of the DSLM framework by exemplifying certain integration of enterprise business models and the proposed location model, driving the enterprise LBS research a step further. We conclude our evaluation with a short discussion as follows:

Figure 21. The picturesque views of (a) PNLM of Enterprise1&3; (b) PNLM of Enterprise1&4



(a)



(b)

- Among the two aspired features for enterprise-based LBS ("*semantic*" and "*dynamic*"), BOLM enables to semantically and dynamically define for an enterprise its clients, client values, and client relationships in a location model, while PNLM empowers an enterprise to *merge* its location model with the other enterprise's due to their inter-enterprise cooperation. LMP, on the other hand, facilitates this inter-enterprise cooperation in a given industry. That is, DSLM bestows enterprise mobile workforces high-level location-sensitive decision information so as to properly serve the clients or justifiably negotiate contracts with other enterprises.
- The results of Section 4.1 show that *semantic* feature of the location model (i.e., the linking between the enterprise business models and the enterprise location model in BOLM) endows the enterprise higher values than those of *static* location models (e.g., First-In-First-Out and Far-Distance-Based), and is capable of adapting to the varying service needs of the clients.
- Section 4.2 evinces that the merits of the semantic feature in BOLM can be extended to PNLM. That is, PNLM fully (or partly) enables the valuation of inter-enterprise cooperation/inter-operability in terms of the myriad principles of enterprise cooperation (compatible, complementing, analytic, and feasible). However, the scenarios have not yet addressed the homogeneous principle (sharing the same competitive goal) that awaits further investigation.
- Scenario 1 of Section 4.2 demonstrates that LMP moderates the creation of PNLMs between enterprises, precipitating the diffusion of the merits of the extended semantic values.
- DSLM can be applied to myriad kinds of enterprises with appropriate ontology modeling: enterprises of operations sensible to locations (e.g., transportation, logistics, touring, etc.), enterprises of majority mobile workforces (e.g., insurance, estate agencies, etc.), and enterprises of clients spreading over various regions (e.g., newspaper, online merchants, etc.).
- DSLM primarily intends to bring about cooperation between enterprises. However, this framework can reverse its functions by serving as a tool for competitive analysis (that is, substituting clients data with competitors data in order to perform competitive analysis).
- The practical implication of DSLM is that BOLM, PNLM, and LMP case be used in location-based enterprise decision support for maximizing enterprise profits or engaging enterprise geographical expansion (or cooperation).

CONCLUSION

The contribution of dynamic semantic location modeling devised in this chapter is the first attempt in integrating enterprise business models with location models (that have been playing a very important role in mobile commerce). DSLM advances the former location modeling methods by embodying the dynamic and semantic features. The DSLM is a kind of symbolic model that includes business oriented location models (BOLM) as objects. BOLM describes different business units and their relationships. For location-based enterprise decision support, our chapter presents a framework of three different deployment of the DSLM (business-oriented location model, location model platform, and partner network location model).

Enterprises can build up their business-oriented location model first and then search for their potential partners in location model platform. Finally, if the advanced cooperation between two enterprises is possible, the partner network

location model can be constructed with their business-oriented location models. In short, the dynamic semantic location model provides certain solutions to enterprise-based LBS that take into account enterprise business models, bestowing enterprise mobile workforce location-sensitive decision information so as to properly serve the clients or justifiably negotiate contracts with other enterprises.

We have evaluated the partner network location model and location model platform and created a few innovative scenarios about the integration of enterprise business models and the proposed location model. These scenarios imply some suggestions about how the proposed framework could be utilized along with a variety of situations of enterprises having different perspectives and strategies in assessing their clients, partners, and business contexts. We hope our work can shed light on further integration of enterprise business models and location models for advanced mobile enterprise applications.

REFERENCES

Bolloju, N. (2003). Extended role of knowledge discovery techniques in enterprise decision support environments. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, Maui, HI.

Bouwman, H., Haaker, T., & Faber, E. (2005). Developing mobile services: Balancing customer and network value. In *Proceedings of the 2nd IEEE International Workshop on Mobile Commerce and Services (WMCS'05)*, Munich, Germany.

Bruce, B., Cross, M., Duncan, T., Hoey, C., & Wills, M. (2000). *e.Volution* (pp. 81-96). Prestoungrange University Press.

Domnitcheva, S. (2001). *Location modeling: State of the art and challenges.* Paper presented at the UbiComp Workshop on Location Modeling for Ubiquitous Computing, Atlanta, GA.

Doz, L. Y. (2002). *Managing partnerships and strategic alliances.* Insead. Retrieved August 22, 2007, from http://www.insead.edu/executives

Ericsson Enterprise. (2002). *The Path to the mobile enterprise*. Retrieved August 22, 2007, from http://www.ericsson.com/products/whitepapers_pdf/whitepaper_mobile_enterprise_rc.pdf

Example for Reinforcement Learning: Playing Checkers. (2001). Retrieved August 22, 2007, from http://www.cs.wustl.edu/~sg/CS527_SP02/lecture2.html.

Fetnet. http://enterprise.fetnet.net/event /Special_02.htm.

Grimes, S. (2005). Location, location, location. *intelligent enterprise*. Retrieved August 22, 2007, from http://www.intelligententerprise.com/toc/?day=01&month=09&year=2005

Hiramatsu, H. (2001). A spatial hypermedia framework for position-aware information delivery systems. *Lecture Notes in Computer Science, 2113*, 754-763.

Hynes, N., & Mollenkopf, D. (1998). *Strategic alliance formation: Developing a framework for research.* Paper presented at the Australia New Zealand Academy of Marketing Conference, Otago, New Zealand.

Jasper, R., & Uschold, M. (1999). *A framework for understanding and classifying ontology applications.* Paper presented at the 12th Workshop on Knowledge Acquisition, Modeling and Management, Banff, Canada.

Machiraju, V. (2001). *Service-oriented research opportunities in the in the world of appliances (Tech. Rep.).* HP Software Technology Lab.

Mitchell, T. M. (1997). *Machine learning* (pp. 10-11). McGraw-Hill.

Pradhan, S. (2002). *Semantic location*. Retrieved August 22, 2007, from http://cooltown.hp.com/dev/wpapers /semantic/semantic.asp

Rohs, M., & Roduner, C. (2006). Towards an enterprise location service. In *Proceedings of the International Symposium on Applications and the Internet Workshops*, Phoenix, AZ.

Varshney, U. (2000). Recent advances in wireless networking. *IEEE Computer, 33*(6), 100-103.

Ververidis, C., & Polyzos, G. (2002). Mobile marketing using location based services. In *Proceedings of the 1st International Conference on Mobile Business*, Athens, Greece.

vLeonhardt, U. (1998). *Supporting location: awareness in open distributed systems.* Unpublished doctoral thesis, Imperial College, Department of Computing, London.

Yuan, S. T., & Peng, K. H. (2004). Location based and customized voice information service for mobile community. *Information Systems Frontiers, 6*(4), 297-311.

Yuan, S. T., & Tsao, E. (2003). A recommendation mechanism for contextualized mobile advertising. *Expert Systems with Applications, 24*(4), 399-414.

ENDNOTES

1 The efforts require the capability of interfacing with geometric models when handling the ranges and the overlaps of the location objects.

2 The new modeling method is grounded on symbolic modeling (in which a location is considered as a set containing the objects residing in the designated location) so as to be appropriately extended as shown in the later sections.

3 In a BOLM, the information of the distance between a CU and a BU is captured as an attribute of the CU (whenever required) that subsequently enables location-based commerce as addressed in Section 1.

4 Averaged attribute values are used if there are multiple Source Enterprises.

5 Please note that for the same pair of BUs the MR measurement is different if the source BU exchanges with the target BU.

6 For simplicity, in Table 8 the IR values are presumed and the MR values are attained by respectively calculating the difference of the correspondent DR values. (The complete process of calculating these DR, IR, MR values involves the awareness of the attribute weights and the attribute values before the application of Algorithm 2 and Algorithm 3.)

7 Farther-distance service requests are presumed to be of higher values (than those of shorter-distance service requests) for a logistic delivery company (if its charges take into account the distances of the service requests).

8 The purpose of additionally tuning Request Revenue by 1.5 (3.5) is for making the comparison of the resulting values (for the six types of clients) more perceivable.

9 Without loss of generality, a request set of size 10 is used in our experiment settings (i.e., the size could be equal to any other number).

10 This tuning originates from an intuition that acceptance of requests implicitly encourages the occurrence of subsequent requests (in contrast to the situation that rejection of requests usually dismays the succeeding).

11 Due to the limited resources of the enterprise, there might be client requests that cannot be served and hence are not taken into account in the client reply set.

[12] Location represents the region number, and LX and LY indicate the coordinates of the 6*6 coordinated plane.

[13] In Figure 17, clients of the IR are indicated with the R_Type of value 1 (while clients of MR are indicated with R_Type of value 2) and thus the SIM value is the sum of these relationship measurements.

This work was previously published in Agent Systems in Electronic Business, edited by E. Li and S. Yuan, pp. 289-322, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Section VII
Critical Issues

This section addresses conceptual and theoretical issues related to the field of mobile computing, which include security issues in numerous facets of the discipline including mobile agents, mobile commerce, and mobile networks. Within these chapters, the reader is presented with analysis of the most current and relevant conceptual inquires within this growing field of study. Particular chapters also address quality of service issues in mobile networks, mobile ontologies and mobile web mining for marketing. Overall, contributions within this section ask unique, often theoretical questions related to the study of mobile computing and, more often than not, conclude that solutions are both numerous and contradictory.

Chapter 7.1
Mobile Code and Security Issues

E. S. Samundeeswari
Vellalar College for Women, India

F. Mary Magdalene Jane
P. S. G. R. Krishnammal, India

ABSTRACT

Over the years, computer systems have evolved from centralized monolithic computing devices supporting static applications, into client-server environments that allow complex forms of distributed computing. Throughout this evolution, limited forms of code mobility have existed. The explosion in the use of the World Wide Web, coupled with the rapid evolution of the platform-independent programming languages, has promoted the use of mobile code and, at the same time, raised some important security issues. This chapter introduces mobile code technology and discusses the related security issues. The first part of the chapter deals with the need for mobile codes and the various methods of categorising them. One method of categorising the mobile code is based on code mobility. Different forms of code mobility, like code on demand, remote evaluation, and mobile agents, are explained in detail. The other method is based on the type of code distributed. Various types of codes, like source code, intermediate code, platform-dependent binary code, and just-in-time compilation, are explained. Mobile agents, as autonomously migrating software entities, present great challenges to the design and implementation of security mechanisms. The second part of this chapter deals with the security issues. These issues are broadly divided into code-related issues and host-related issues. Techniques, like sandboxing, code signing, and proof-carrying code, are widely applied to protect the hosts. Execution tracing, mobile cryptography, obfuscated code, and cooperating agents are used to protect the code from harmful agents. The security mechanisms, like language support for safety, OS level security, and safety policies, are discussed in the last section. In order to make the mobile code approach practical, it is essential to understand mobile code technology. Advanced and innovative solutions are to be

developed to restrict the operations that mobile code can perform, but without unduly restricting its functionality. It is also necessary to develop formal, extremely easy-to-use safety measures.

INTRODUCTION

Mobile code computation is a new paradigm for structuring distributed systems. Mobile programs migrate from remote sites to a host, and interact with the resources and facilities local to that host. This new mode of distributed computation promises great opportunities for electronic commerce, mobile computing, and information harvesting. There has been a general consensus that security is the key to the success of mobile code computation.

Distributed applications involve the coordination of two or more computers geographically apart and connected by a physical network. Most distributed applications deploy the client/server paradigm. There are certain problems with the client/server paradigm, such as the requirement of a high-network bandwidth and continuous user-computer interactivity. Hence, the mobile code paradigm has been developed as an alternative approach for distributed application design.

In the client/server paradigm, programs cannot move across different machines and must run on the machines they reside on. The mobile-code paradigm, on the other hand, allows programs to be transferred among, and executed on, different computers. By allowing code to move between hosts, programs can interact on the same computer instead of over the network. Therefore, communication cost can be reduced. Besides, one form of mobile code is a program that can be designed to work on behalf of users autonomously. This autonomy allows users to delegate their tasks to the mobile code, and not to stay continuously in front of the computer terminal.

With the growth of distributed computer and telecommunications systems, there have been increasing demands to support the concept of ”mobile code,” sourced from remote, possibly untrustworthy systems, but executed locally.

MOBILE CODE

Mobile code consists of small pieces of software obtained from remote systems outside the enclave boundary, transferred across a network, and then downloaded and executed on a local system without explicit installation or execution by the recipient.

The mobile-code paradigm encompasses programs that can be executed on one or several hosts other than the one that they originate from. Mobility of such programs implies some built-in capability for each piece of code to travel smoothly from one host to another. A mobile code is associated with at least two parties: its producer and its consumer, the consumer being the host that runs the code.

Examples of mobile code include a Java script embedded within an HTML page, a visual basic script contained in a WORD document, an HTML help file, an ActiveX Control, a Java applet, a transparent browser plug-in or DLL, a new document viewer installed on demand, an explicitly downloaded executable binary, and so forth. Since mobile code runs in the execution context of the user that downloads the code, it can issue any system calls that the user is allowed to make, including deleting files, modifying configurations or registry entries, ending e-mails, or installing back-door programs in the home directory. The most common type of malicious mobile code is an e-mail attachment.

Mobile-code systems range from simple applets to intelligent software agents. These systems offer several advantages over the more traditional distributed computing approaches, like flexibility in software design beyond the well-established object-oriented paradigm and bandwidth optimization. As usual, increased flexibility comes

with a cost, which is increased vulnerability in the face of malicious intrusion scenarios akin to Internet. Possible vulnerabilities with mobile code fall in one of two categories: attacks performed by a mobile program against the remote host on which the program is executed, as with malicious applets or ActiveX programs; and the less-classical category of attacks due to the subversion of the mobile code and its data by the remote execution environment.

Advantages of Mobile Code

Here are some possible advantages of mobile code:

- Eliminates configuration and installation problems, and reduces software distribution costs of desktop applications
- The code is potentially portable to many platforms
- Enhances the scalability of client/server applications
- Achieves performance advantages
- Achieves interoperability of distributed applications

Categories of Mobile Code

One method of categorising the mobile code is based on code mobility (Ghezzi & Vigna, 1997). Different forms of code mobility are *code on demand*, *remote evaluation*, and *mobile agents. Code on demand* is the downloading of executable content in a client environment as the result of a client request to a server. In *remote evaluation*, the code is uploaded to a server, where this code is executed. Multihop migration of code across the network and autonomous execution on many different hosts is termed *mobile agent.*

Code on Demand

In the code on demand paradigm, the client component owns the resources needed for the execution of a service, but lacks the know-how needed to use them in performing the service. The corresponding code component can be retrieved from a remote server component, which acts as a code repository, and subsequently executed, thus providing enhanced flexibility by allowing the server to dynamically change the behavior of the client. This is the scheme typically employed by Web applets, or by the parameter-passing mechanism in Java/RMI.

Remote Evaluation

In the remote-evaluation paradigm, the client component owns the know-how about the service that must be executed, but lacks the resources needed to perform the service, which are owned by the server component. A sort of enhanced client-server interaction takes place, where the client sends a request to the server, but includes also the code component required to perform the service. After the code component is received on the server, the interaction proceeds as in the client-server paradigm, with the code component accessing the resources now colocated with it, and sending the results back to the client. This reduces network traffic by executing a computation close to the resources located at the server's side. A common example is SQL servers performing queries on a remote database.

Mobile Agents

In the mobile-agent paradigm, the mobile components explicitly relocate themselves across the network, preserving their execution state (or part thereof) across migrations. It is, therefore, associated with many security issues needed for "safe" execution. The mobile agents offer new possibilities for the e-commerce applications,

Table 1. Summary of mobile code techniques

Type of mobility	Category	Mobility of code	Resources	Processor
Weak	Code on demand	Remote to local (Pull)	Local side	Local side
	Remote evaluation	Local to remote (Push)	Remote side	Remote side
Strong	Mobile agent	Migration	Remote side	Agent's originator

*Where **Resource**s represent the information and other resources for code execution*
***Processor** is the abstract machine that holds the state of computation*

creating new types of electronic ventures from e-shops and e-auctions to virtual enterprises and e-marketplaces. The agent helps to automate many electronic commerce tasks such as simple information gathering tasks, and all tasks of commercial transactions, namely price negotiation, contract signing, and delivery of (electronic) goods and services. Such agents are developed for diverse business areas, for example, contract negotiations, service brokering, stock trading, and many others. Examples of systems supporting this type of mobility are Telescript (Telescript, 1995), Aglets (IBM Aglets, 2002), and JADE (Java Agent Development Framework, 2005).

The first two forms, code on demand and remote evaluation, can be classified as weak-mobility forms, as they involve the mobility of code only. Since the mobile agent involves the mobility of computation, it is commonly known as strong-mobility form.

The other method of categorizing "mobile code" technologies is based on the type of code distributed (Tennenhouse & Wetherall, 1996):

- Source code
- Intermediate code
- Platform-dependent binary code
- Just-in-time compilation

Source Code

The first approach is based on distributing the source for the "mobile code"used. This source will be parsed and executed by an interpreter on the user's system. The interpreter is responsible for examining the source to ensure it obeys the required syntactic and semantic restrictions of the language; and then for providing a safe execution "sand-box" environment. The safety of this approach relies on the correct specification and implementation of the interpreter.

The main advantages of the source code approach are the distribution of relatively small amounts of code; the fact that since the user has the full source, it is easier to check the code; and that it is easier for the interpreter to contain the execution environment. Disadvantages include the fact that it is slow, since the source must first be parsed; and that it is hard to expand the core functionality, since the interpreter's design limits this. Examples are programmable MUDs, JavaScript, and so forth.

Intermediate Code

A second approach to providing "mobile code" is to have the programs compiled to a platform-inde-

pendent intermediate code that is then distributed to the user's system. This intermediate code is executed by an interpreter on the user's system. Advantages are that it is faster to interpret than source, since no textual parsing is required, and the intermediate code is semantically much closer to machine code. The interpreter provides a safe execution "sand-box" and again, the safety of the system depends on the interpreter. The code, in general, is quite small, and the user's system can check the code to ensure it obeys the safety restrictions. Disadvantages of this approach are its moderate speed, since an interpreter is still being used, and the fact that less semantic information is available to assist in checking the code than if source was available. Java is a very good example for this category.

Native Binary Code

The third category of code distribution uses native binary code that is then executed on the user's system. This gives the maximum speed, but means that the code is platform-dependent. Safe execution of binary code requires the restricted use of an instruction set and the restricted address space access. Approaches to ensuring this can rely upon

- Traditional heavy address space protection that is costly in terms of system performance and support
- The verified use of a trusted compiler that guarantees to generate safe code that will not violate the security restrictions
- The use of "software fault isolation" technologies that augment the instruction stream, inserting additional checks to ensure safe execution.

A combination of verified use of a trusted compiler and the software fault isolation approach has created considerable interest, especially when used with a just-in-time compiler.

Just-in-Time Compilation

Just-in-time compilation (JIT) is an approach that combines the portability of intermediate or source code with the speed of binary code. The source or intermediate code is distributed, but is then compiled to binary on the user's system before being executed. If source is used, it is slower but easier to check. If intermediate code is used, then it is faster. Another advantage is that users can utilise their own trusted compiler to verify code, and insert the desired software fault isolation run-time checks. Individual procedures are translated on a call-by-call basis. This approach is being used with Java JIT compilers.

PROPERTIES OF MOBILE CODE

- Comes in a variety of forms
- Often runs unannounced and unbeknownst to the user
- Runs with the privilege of the user
- Distributed in executable form
- Run in multiple threads
- Can launch other programs

SECURITY ISSUES OF MOBILE CODE PARADIGMS

In this section, some possible security attacks to different mobile-code paradigms, and possible mechanisms against these attacks, are discussed.

A security attack is an action that compromises the security requirements of an application. Applications developed using different paradigms are subject to different attacks. In the conventional client/server model, the local computer is usually assumed to be fortress for code and data. Therefore, the sources of security attacks are outsiders of the local machine. The main possible attacks are *masquerading* (pretending the server or the

client), *eavesdropping* on the communication channel, and *forging messages* to the client or the server.

The security model of the client/server paradigm also applies to the *remote evaluation* and *code-on-demand* approaches, with the additional concern that the code-receiving side must make sure the code is not harmful to run. In remote evaluation, the code receiving side is the remote side, while it is the local side in code-on-demand. *Mobile agent*, on the other hand, is the most challenging area of mobile-code security, due to the autonomy of agents. Mobile-agent security is usually divided into two aspects: *host security* and *code security*. Host security (Loureiro, Molva, & Roudier, 2000) deals with the protection of hosts against malicious code/agent, whereas code security deals with the protection of code/agents against malicious hosts or other agents.

Host Security Against Malicious Code

In the interconnected world of computers, mobile code generated by a malicious outsider, has become an omnipresent and dangerous threat. Malicious code can infiltrate hosts using a variety of methods, such as attacks against known software flaws, hidden functionality in regular programs, and social engineering.

From the host perspective, a secure execution environment is necessary to protect itself from such types of code. The first step towards a secure environment is to simply limit the functionality of the execution environment in order to limit the vulnerabilities. Techniques for protection of hosts now evolve along two directions (1) executing mobile codes in a restricted environment, (2) a mobile code infrastructure that is enhanced with authentication, data integrity, and access control mechanisms. The following section details both the aspects.

Sandboxing

Sandboxing is a software technique used to protect hosts from malicious mobile code. In an execution environment, local code is executed with full permission, and has access to crucial system resources. On the other hand, mobile code is executed inside a restricted area called a "sandbox" that restricts the code to operating system functionality. A sandboxing mechanism enforces a fixed-security policy for the execution of the mobile code. The policy specifies the rules and restrictions that mobile code should conform to. A mechanism is said to be secure if it properly implements a policy that is free of flaws and inconsistencies.

To contain mobile code within a sandbox, extensive type checking is used. Also, memory accesses and jump addresses are checked at runtime. If these addresses do not fall within the sandbox, then they are redirected to a location within the sandbox. The error, however, is contained within the sandbox, and cannot affect the rest of the system. Sandboxing can also be used for restricting access to file systems, and limiting the ability to open network connections.

The most common implementation of sandboxing is in the Java interpreter inside Java-enabled Web browsers. A Java interpreter contains three main security components: classloader, verifier, and security manager . The classloader converts mobile code into data structures that can be added to the local class hierarchy. Thus, every remote class has a subtype of the classloader class associated with it. Before the mobile code is loaded, the verifier performs a set of security checks on it in order to guarantee that only legitimate Java code is executed. The mobile code should be a valid virtual machine code, and it should not overflow or underflow the stack, or use registers improperly. Additionally, remote classes cannot overwrite local names, and their operations are checked by the security manager before the execution.

Figure 1. Sandboxing technique

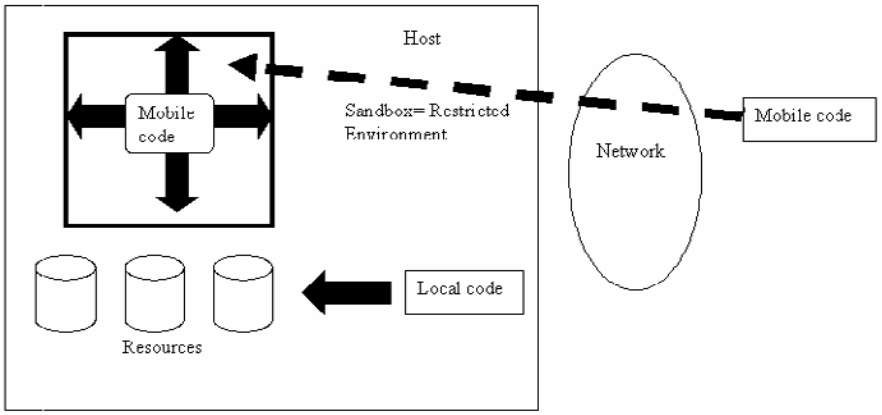
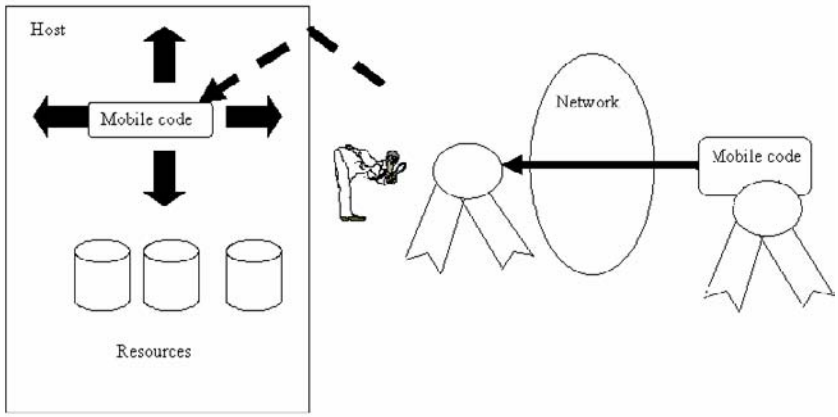


Figure 2. Code signing technique



The main problem with the sandbox is that any error in any security component can lead to a violation of the security policy. The sandbox also incurs a high runtime overhead. A downside of the sandboxing technique is that it increases the execution time of legitimate remote code.

Code Signing

In the "code signing" technique, a digitally signed piece of software identifies the producer who created and signed it. It enables the platform to verify that the code has not been modified since it was signed by the creator. Code signing makes use of a digital signature and one-way hash function where a private key is used to sign code, both ensuring transmission integrity and enabling policy defined by trust in the signer. Code signing enables the verification of the code producer's identity, but it does not guarantee that they are trustworthy.

The platform that runs mobile code maintains a list of trusted entities and checks the code against

the list. If the code producer is on the list, it is assumed that they are trustworthy and that the code is safe. The code is then treated as local code and is given full privileges; otherwise, the code will not run at all. An example is Microsoft's Authenticode system for ActiveX.

There are two main drawbacks of the code signing approach. First, this technique assumes that all the entities on the trusted list are trustworthy and that they are incorruptible. Mobile code from such a producer is granted full privileges. If the mobile code is malicious, it can use those privileges not only to directly cause harm to the executing platform, but also to open a door for other malicious agents by changing the acceptance policy on the platform. Moreover, the affects of the malicious agent attack may only occur later, which makes it impossible to establish a connection between the attack and the attacker. Such attacks are referred to as "delayed attacks." Secondly, this technique is overly restrictive towards agents that are coming from unrecognized entities, as they do not run at all.

Code Signing and Sandboxing Combined

This technique combines the advantages of both code signing and sandboxing. If the code consumer trusts the signer of the code, then the code will run as if it were local code, that is, with full privileges being granted to it. On the other hand, if the code consumer does not trust the signer of the code, then the code will run inside a sandbox. The main advantage of this approach is that it enables the execution of the mobile code produced by untrustworthy entities. However, this method still suffers from the same drawback as code signing, that is, malicious code that is deemed trustworthy can cause damage and even change the acceptance policy. The security policy is the set of rules for granting programs permission to access various platform resources. The "black-and-white" policy only allows the platform to label programs as completely trusted or untrusted. The combination of code signing and sandboxing implemented in JDK 1.2 incorporates fine-grained access control where it allows a user to assign any degree of partial trust to a code, rather than just "trusted" and "untrusted."

There is a whole spectrum of privileges that can be granted to the code. In JDK1.2, all code is subjected to the same security policy, regardless of being labelled as local or remote. The run-time system partitions code into individual groups, called protection domains, in such a way that all programs inside the same domain are granted the same set of permissions. The end-user can authorize certain protection domains to access the majority of resources that are available at the executing host, while other protection domains may be restricted to the sandbox environment. In between these two, there are different subsets of privileges that can be granted to different protection domains, based on whether they are local or remote, authorised or not, and even based on the key that is used for the signature.

Proof-Carrying Code

Proof-carrying code (PCC) (Proof-Carrying Code, 2002) strikes an effective balance between security and flexibility. The process, pioneered by Necula and Lee (1998), involves the code producer attaching additional data to a piece of code. This data can be interpreted as proof that a particular property holds for the piece of code.

In this technique, the code producer is required to provide a formal proof that the code complies with the security policy of the code consumer. The code producer sends the code, together with the formal safety proof, sometimes called machine-checkable proof, to the code consumer. Upon receipt, the code consumer checks and verifies the safety proof of the incoming code by using a simple and fast proof checker. Depending on the result of the proof validation process, the code is proclaimed safe, and consequently executed

without any further checking, or it is rejected. PCC guarantees the safety of the incoming code, providing that there is no flaw in the verification-condition generator, the logical axioms, the typing rules, and the proof checker.

PCC is considered to be "self-certifying" because no cryptography or trusted-third party is required. It involves low-cost static program checking, after which the program can be executed without any expensive run-time checking. In addition, PCC is considered "tamper-proof," as any modification done to the code or the proof will be detected. Other applications include active networks and extensible operating systems. Proof-carrying code also has some limitations that include the potential size of the proof and the time consumed in the proof-validation process.

Mobile Code Security against Malicious Host

While a mobile agent is roaming among host platforms, it typically carries information such as code, static data, data collected from other hosts that were visited, and the execution state of the mobile agent. The execution state is a dynamic data created during the execution of the agent at each host. Agents may be susceptible to observation of execution or any other information it possesses.

The possible attacks by the host platform on mobile agents are extracting sensitive information such as encryption keys, credit card information, corrupting or modifying the execution state and code information, and denial of service. The data collected by the agent from other hosts or from the host's own database is manipulated to report false information to the user. Similarly, the agent's code and execution sequence is manipulated to learn about the information the user is interested in, and make the agent perform something illegitimately. Denial of service includes terminating the agent without executing it, ignoring the agent's request for services and resources, providing insufficient resources, making it very difficult for the agent to complete execution in a timely fashion, or assigning continuous tasks to the agent so that it will never reach its goal. A malicious agent may assume the identity of another agent in order to gain access to platform resources and services, or simply to cause mischief or even serious damage to the platform. Likewise, a platform can claim the identity of another platform in order to gain access to the mobile agent data. This type of attack is known as masquerading.

It is intrinsically more difficult to protect the agents located on potentially untrusted hosts, since the environment has a total control over the mobile code (otherwise, protecting the host would be impossible). Three categories of solutions exist to protect agents (Chan & Anthony, 1999; Sanders & Tschudin, 1998a; Sanders & Tschudin, 1998b): agent tampering avoidance, detection, and prevention. In avoidance technique, a closed network is established by sending the agents only to trusted hosts, such as intraorganizational applications, or on a third-party-hosted network that is trusted by all parties involved. Such an arrangement is effective but obviously satisfies system openness. The attacks can be detected using techniques such as forward integrity and execution tracing. These techniques are not suitable for very critical actions, for which detection may be too late. The attacks can be prevented either by making the tampering difficult or expensive. This can be achieved either by digitally signing the agent state and the data, or encrypting them with a public key of the targeted host, or by obfuscated code. In cooperating agents technique, the agent code/state is duplicated to recover from an agent termination attack. These prevention techniques are not well developed and are of current research issue.

Tampering Detection Techniques

Execution tracing (Vigna, 1997) is a technique that enables the detection of any possible misbehaviour by a platform. It is based on cryptographic traces

that are collected during an agent's execution at different platforms and attached to the agent itself. Traces are the logs of actions performed by the agent during its lifetime, and can be checked by the agents' owner to see if it contains any unauthorized modifications. This technique has some limitations, such as the potential large size and number of logs to be retained, and the owner has to wait until it obtains suspicious results in order to run the verification process. Tracing is only triggered on suspicion that malicious tampering of an agent has occurred during its itinerary and is too complicated to be used for multithreaded agents. A variation of this technique is by assigning the trace verification process to a trusted third party, the verification server, instead of depending on the agent's owner. These techniques assume that all the involved parties own a public and private key that can be used for digital signatures to identify the involved parties. Another variation of this technique uses a list of secret keys provided by the agent's originator. For each platform in an agent's itinerary, there is an associated secret key. When an agent finishes an execution at a certain platform in its itinerary, it summarizes the results of its execution in a message for the home platform, which could be sent either immediately or later. The agent erases the used secret key of the current visited platform before its migration to the next platform. Destroying the secret key ensures the "forward integrity" of the encapsulation results. Forward integrity guarantees that no platform to be visited in the future is able to modify any results from the previously visited platform.

TAMPERING PREVENTION TECHNIQUES

Mobile Cryptography

This technique (Sanders & Tschudin, 1998a) is based on executing the agent in its encrypted form. It is not the code that is encrypted, but the function this code executes. The major challenge here is to find encryption schemes for expressing a program of arbitrary functions or login. An approach that uses the mobile cryptography is a time-limited blackbox (Hohl, 1998). It defines the blackbox as an agent that performs the same task as the original agent but has a different structure. The agent has the blackbox property if its code and data cannot be read or modified. The agent holds the blackbox property for a known time interval that should be sufficient to perform the required task. After this time the agent is invalidated, and the attacks have no effect. Various means of code obfuscation and authentication techniques are proposed to achieve this time-limited blackbox.

Obfuscated Code

Obfuscation (Motlekar, 2005) is a technique of enforcing the security policy by applying a behaviour-preserving transformation to the code before it is being despatched to different hosts. It aims to protect the code from being analysed and understood by the host; thereby, making the extraction and corruption of sensitive data, code, or state very difficult. Different obfuscating transformations are layout obfuscation — remove or modify some information in the code such as comments and debugging information; data obfuscation — modifying the data and data structures in the code without modifying the code itself; and control obfuscation — altering the control flow in the code without modifying the computing part of the code. Code mess up is a variation of this approach, where by the code is rendered to look illogically, using irrelevant variable names, having odd data representation, decomposing the variables bit-by-bit and reassembling them into the actual values during execution, adding a small amount of dead code that may appear to be active in the program. It is not sufficient to scramble the code only once, as the code may be reconstituted and comprehended by a malicious observer. The agent must have a new structure for each dispersal

from the home origin. Obfuscation concentrates on protecting the code from decompilers and debuggers. It could delay, but not prevent, the attacks on agent via reverse engineering.

Cooperating Agents

This technique distributes critical tasks of a single mobile agent between two cooperating agents. Each of the two cooperating agents executes the tasks in one of two disjoint sets of platforms. The cooperating agents share the same data and exchange information in a secret way. This technique reduces the possibility of the shared data being pilfered by a single host. Each agent records and verifies the route of its cooperating agent. When an agent travels from one platform to another, it uses an authenticated communication channel to pass information about its itinerary to its cooperating agent. The peer agent takes a suitable action when anything goes wrong. The drawbacks of this technique are the cost of setting up the authenticated communication channel for each migration; care should be taken to assign the two agents to disjoint platforms and never assigned to the same malicious host.

Security Mechanisms

Developing sound, reliable security mechanisms is a nontrivial task, and a history of vulnerable and/or incomplete implementations of these mechanisms led to the idea that mobile-code systems are inherently insecure, too complex, and very difficult to deploy. To overcome these problems, the mobile-code system must rely, as much as possible, on the security mechanisms already provided by the language used for developing, and by the underlying operating system. By doing this, it is possible to develop, with reduced effort, security services that rely on well-known, well-understood, and well-tested security mechanisms. Also, by describing the security of the mobile-code system in terms of the language and OS security mechanisms, system administrators can better evaluate the security implications of deploying the system.

Language Support for Safety

The features of the language needed to ensure that various code units do not interfere with each other, and with the system are given next.

- Heavy address space protection mechanisms
- Type-safe feature to ensure that arrays stay in bounds, pointers are always valid, and code cannot violate variable typing (such as placing code in a string and then executing it)
- Designing a modular system, separating interfaces from implementations in programs, and with appropriate layering of libraries and module groups, with particular care being taken at the interfaces between security boundaries.
- Replace general library routines that could compromise security with more specific, safer ones. For example a general file access routine can be replaced with one that can write files only in a temporary directory.
- Granting access to resources: Determining exactly which resources a particular code unit is to be granted access to. That is, there is a need for a security policy that determines what type access any "mobile code" unit has. This policy may be:

1. **Fixed for all "mobile code" units:** Very restrictive but easy, and the approach currently is used to handle applet security in Web browsers such as Netscape.
2. **User verifies each security-related access requests:** Relatively easy, but rapidly gets annoying, and eventually is self-defeating when users stop taking notice of the details of the requests. Whilst there is a place for

querying the user, it should be used exceedingly sparingly.

3. **Negotiate for each "mobile code" unit:** Much harder, as some basis is needed for negotiation, perhaps based on various profiles, but ultimately this is likely to be the best approach.

OS Level Security

The types of events to be monitored in association with the agent execution are very similar to those audited for the system's users. Moreover, the agents can be easily grouped and differentiated within the system. In addition to extensive authentication and authorization mechanisms, accounting and auditing mechanisms should be implemented.

In a system like "distributed agents on the go" (DAGO) (Felmetsger & Vigna, 2005), a mobile agent is viewed as an ordinary system's user who logs in to the host and uses some of the system's resources for its own needs. Every incoming mobile agent is given an individual account and a unique user identifier (UID) for the duration of its execution on a host. This approach allows the hosting OS to apply to mobile agents the same set of rules and policies that are applied by the OS to all of its users.

In Unix, a number of logging, auditing, and accounting mechanisms are available to monitor the action of its users and the status of its resources. These tools can work at the system call level and can be configured based on different types of events, such as opening and closing of files, reads and writes, programs executed, and so on. They also can allow one to specify groups of system objects to be monitored for certain activities, and can track system usage by recording the statistics about CPU and memory usage, I/O operations, running time, and other forms of system resource usage, along with the user IDs of the processes involved. These tools can be easily leveraged and extended to a multiagent environment.

A variety of customizable tools, such as SNARE — system intrusion analysis and reporting environment (SNARE, 2005), BSM — basic security module provide a greater degree of security assurance. SNARE is a dynamically loadable kernel nodule that can be used as a stand-alone auditing system or as a distributed tool. The tool can be configured to monitor events associated with certain groups of users, filter the monitored events with specific "search expressions," and submit reports in different formats and time frames. The type of events monitored can be either defined by a category (for example, system calls) or by an identifier (such as "denied access").

Safety Policies for Mobile Code Programs

A safety policy is a set of restrictions placed upon locally run untrusted code to ensure that the program does not behave in a manner that is detrimental to the system or to the system security. At the very least, a safety policy should guarantee the following fundamental safety properties (Muller, 2000):

- **Control flow safety:** The program should never jump to and start executing code that lies outside of the program's own code segment. All function calls should be to valid function entry points, and function returns should return to the location from where the function was called.
- **Memory safety:** The program should never be allowed to access random locations in memory. The program should only access memory in its own static data segment, live system heap memory that has been explicitly allocated to it, and valid stack frames.
- **Stack safety:** The program should only be allowed to access the top of the stack. Access to other areas of the stack should be completely restricted.

These three properties, combined, offer the minimum nontrivial level of security for mobile code. More complicated security policies are possible, depending on the application.

Trust

Security is based on the notion of trust. Basically, software can be divided into two categories, namely, software that is trusted and software that is not, separated by an imaginary trust boundary. All software on our side of the trust boundary is trusted and is known as the trusted code base.

All security implementations rely on some trusted code. As a result, a trust model of a particular implementation can be made. The trust model basically specifies which code is to be included in the trusted-code base and which code lies outside of the trust boundary.

At the very least, the trusted-code base should include the local operating system kernel, but can also include other items of trusted software, like trusted compilers or trusted program runtime environments (e.g., the Java interpreter). It is desirable, however, to keep the trusted-code base as small as possible to reduce the security vulnerabilities.

Performance and Security

Unfortunately, as it is in most applications, performance is sacrificed for increased security. It would, however, be profitable to have applications that are both secure and perform well at the same time. For this reason, there is much research concerned with resolving the conflict between these concepts in some way.

CONCLUSION

The purpose of this chapter is to raise readers' awareness of mobile code and various approaches to addressing security of mobile code and agents. All of the techniques discussed in this chapter offer different approaches to combating malicious mobile code. However, the best approach is probably a combination of security mechanisms. The sandbox and code signing approaches are already hybridized. Combining these with firewalling techniques, such as the playground, gives an extra layer of security. PCC is still very much in the research and development phase at present.

In order to make the mobile code approach practical, it is essential to develop advanced and innovative solutions to restrict the operations that mobile code can perform, but without unduly restricting its functionality. It is also necessary to develop formal, extremely easy–to-use safety languages to specify safety policy.

Organizations relying on the Internet face significant challenges to ensure that their networks operate safely, and that their systems continue to provide critical services, even in the face of attack. Even the strictest of security policies will not be able to prevent security breaches. Educating users in social-engineering attacks based around mobile code is also necessary.

REFERENCES

Alfalayleh, M., & Brankovic, L. (2004). *An overview of security issues and techniques in mobile agents.* Retrieved from http://sec.isi.salford.ac.uk/cms2004/Program/CMS2004final/p2a3.pdf

Brown, L. (1996). *Mobile code security* [Electronic version]. Retrieved from http://www.unsw.adfa.edu.au/~lpb/papers/mcode96.html

Chan, H. W., & Anthony. (1999). *Secure mobile agents: Techniques, modeling and application*. Retrieved from http://www.cse.cuhk.edu.hk/~lyu/student/mphil/anthony/term3.ppt

Felmetsger, V., & Vigna, G. (2005). *Exploiting OS-level mechanisms to implement mobile code security.* Retrieved from http://www.cs.ucsb.edu/

~vigna/pub/2005_felmetsger_vigna_ICECCS05.pdf

Ghezzi, C., & Vigna, G. (1997). Mobile code paradigms and technologies: A case study. In K. Rothermet & R. Popescu-Zeletin (Eds.), *Mobile agents, First International Workshop, MA'97, Proceedings* (LNCS 1219, pp. 39-49) Berlin, Germany: Springer.

Hefeeda, M., & Bharat, B. (n.d.) *On mobile code security.* Center of Education and Research in Information Assurance and Security, and Department of Computer Science, Purdue University, West Lafayette, IN. Retrieved from http://www.cs.sfu.ca/~mhefeeda/Papers/OnMobileCodeSecurity.pdf

Hohl, F. (1997). *An approach to solve the problem of malicious hosts.* Universität Stuttgart, Fakultät Informatik, Fakultätsbericht Nr. 1997/03. Retrieved from http://www.informatik.uni-stuttgart.de/cgi-bin/ncstrl_rep_view.pl?/inf/ftp/pub/library/ncstrl.ustuttgart_fi/TR-1997-03/TR-1997-03.bib

Hohl, F. (1998). *Time limited blackbox security: Protecting mobile agents from malicious hosts.* Retrieved from http://citeseer.ist.psu.edu/hohl-98time.html

Hohl, F. (1998). *Mobile agent security and reliability.* Proceedings of the Ninth International Symposium on Software Reliability Engineering (ISSRE '98).

Hohl, F. (1998). Time limited blackbox security: Protecting mobile agents from malicious hosts. *Mobile Agents and Security, 1419 of LNCS*. Springer-Verlag.

IBM Aglets. (2002). Retrieved from http://www.trl.ibm.com/aglets/

Jansen, W., & Karygiannis, T. (n.d.). *Mobile agent security* (NIST Special Publication 800-19) Retrieved from http://csrc.nist.gov/publications/nistpubs/800-19/sp800-19.pdf

Java Agent Development Framework. (2005). Retrieved from http://jade.tilab.com/

Karjoth, G., Lange, D. B., & Oshima, M. (1997). A security model for aglets. *IEEE Internet Computing, 1*(4), 68-77. [Electronic version]. Retrieved from http://www.ibm.com/java/education/aglets/

Loureiro, S., Molva, R., & Roudier, Y. (2000, February). *Mobile code security.* Proceedings of ISYPAR 2000 (4ème Ecole d'Informatique des Systems Parallèles et Répartis), Code Mobile, France. Retrieved from www.eurecom.fr/~nsteam/Papers/mcs5.pdf

Lucco, S., Sharp, O., & Wahbe, R. (1995). Omniware: A universal substrate for mobile code. In Fourth International World Wide Web Conference, MIT. [Electronic version] Retrieved from http://www.w3.org/pub/Conferences/WWW4/Papers/165/

McGraw, G., & Morrisett, G. (2000). *Attacking malicious code*. Retrieved from http://www.cs.cornell.edu/Info/People/jgm/lang-based-security/maliciouscode.pdf

Mobile Code and Mobile Code Security. (2005). Retrieved from http://www.cs.nyu.edu/~yingxu/privacy/0407/main.html

Mobile Code Security. (1996). [Electronic version] Retrieved from http://www.unsw.adfa.edu.au/~lpb/papers/mcode96.html

Mobile Code Security and Computing with Encrypted Functions [Electronic version] Retrieved from http://www.zurich.ibm.com/security/mobile

Motlekar, S. (2005). *Code obfuscation.* Retrieved from http://palisade.paladion.net/issues/2005Aug/code-obfuscation/

Muller, A. (2000). Mobile *code security: Taking the Trojans out of the Trojan horse*. Retrieved from www.cs.uct.ac.za/courses/CS400W/NIS/

papers00/amuller/essay1.htm

Necula, G. C., & Lee, P. (1998). Safe, untrusted agents using proof-carrying code. *Lecture Notes in Computer Science*, (1419). Springer-Verlag.

Oppliger, R. (2000). *Security technologies for the World Wide Web.* Computer Security Series. Artech House Publishers.

Proof-Carrying Code. (2002). Retrieved from http://raw.cs.berkeley.edu/pcc.html

Robust Obfuscation. (2005). Retrieved from http://www.cs.arizona.edu/~collberg/Research/Obfuscation/

Roger, A. G. (2001). *Malicious mobile code: Virus protection for Windows* [Electronic version]. O'Reilly & Associates.

Rubin, A. D., & Geer, D. E. (1998). Mobile code security. *IEEE Internet Computing.*

Sander, T., & Tschudin, C. (1998a). *Towards mobile cryptography.* Proceedings of the IEEE Symposium on Security and Privacy.

Sander, T., & Tschudin, C. (1998b). Protecting mobile agents against malicious hosts. [Electronic version] In G. Vigna (Ed.). Mobile agents and security, *Lecture Notes in Computer Science, 1419* (pp. 44-60). Retrieved from http://citeseer.ist.psu.edu/article/sander97protecting.html

SNARE — System iNtrusion Analysis and Reporting Environment (2005). [Electronic version] Retrieved from http://www.intersectalliance.com/projects/Snare

Telescript Language Reference. (1995). Retrieved from http://citeseer.ist.psu.edu/inc95telescript.html

Tennenhouse, D. L., & Wetherall, D. J. (1996) Towards an active network architecture. *Computer Communication Review.* Retrieved from http://www.tns.lcs.mit.edu/publications/ccr96.html

Vigna, G. (1997, June). Protecting mobile agents through tracing. *Proceedings of the 3rd ECOOP Workshop on Mobile Object Systems*, Jyvälskylä, Finland. Retrieved from http://www.cs.ucsb.edu/~vigna/listpub.html

This work was previously published in Web Services Security and E-Business, edited by C. Radhamani and G. Rao, pp. 75-92, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.2
Security of Mobile Code

Zbigniew Kotulski
Polish Academy of Sciences, Warsaw, Poland
Warsaw University of Technology, Poland

Aneta Zwierko
Warsaw University of Technology, Poland

ABSTRACT

The recent development in the mobile technology (mobile phones, middleware, wireless networks, etc.) created a need for new methods of protecting the code transmitted through the network. The oldest and the simplest mechanisms concentrate more on integrity of the code itself and on the detection of unauthorized manipulation. The newer solutions not only secure the compiled program, but also the data, that can be gathered during its "journey," and even the execution state. Some other approaches are based on prevention rather than detection. In this chapter we present a new idea of securing mobile agents. The proposed method protects all components of an agent: the code, the data, and the execution state. The proposal is based on a zero-knowledge proof system and a secure secret sharing scheme, two powerful cryptographic primitives. Next, the chapter includes security analysis of the new method and its comparison to other currently more widespread solutions. Finally, we propose a new direction of securing mobile agents by straightening the methods of protecting integrity of the mobile code with risk analysis and a reputation system that helps avoiding a high-risk behavior.

INTRODUCTION

A software agent is a program that can exercise an individual's or organization's authority, work autonomously toward a goal, and meet and interact with other agents (Jansen & Karygiannis, 1999). Agents can interact with each other to negotiate contracts and services, participate in auctions, or barter. Multi-agent systems have sophisticated applications, for example, as management systems for telecommunication networks or as artificial intelligence (AI)-based intrusion detection systems. Agents are commonly divided into two types:

- Stationary agents
- Mobile agents

The stationary agent resides at a single platform (host), the mobile one can move among different platforms (hosts) at different times.

The mobile agent systems offer new possibilities for the e-commerce applications: creating new types of electronic ventures from e-shops and e-auctions to virtual enterprises and e-marketplaces. Utilizing the agent system helps to automate many e-commerce tasks. Beyond simple information gathering tasks, mobile agents can take over all tasks of commercial transactions, namely, price negotiation, contract signing, and delivery of (electronic) goods and services. Such systems are developed for diverse business areas, for example, contract negotiations, service brokering, stock trading, and many others (Corradi, Cremonini, Montanari, & Stefanelli, 1999; Jansen & Karygiannis, 1999; Kulesza & Kotulski, 2003). Mobile agents can also be utilized in code-on-demand applications (Wang, Guan, & Chan, 2002). Mobile agent systems have advantages even over grid computing environments:

- Require less network bandwidth
- Increase asynchrony among clients and servers
- Dynamically update server interfaces
- Introduce concurrency

The benefits from utilizing the mobile agents in various business areas are great. However, this technology brings some serious security risks; one of the most important is the possibility of tampering with an agent. In mobile agent systems the agent's code and internal data autonomously migrate between hosts and can be easily changed during the transmission or at a malicious host site. The agent cannot itself prevent this, but different countermeasures can be utilized in order to detect any manipulation made by an unauthorized party. They can be integrated directly into the agent system, or only into the design of an agent to extend the capabilities of the underlying agent system.

Several degrees of agent's mobility exist, corresponding to possibilities of relocating code and state information, including the values of instance variables, the program counter, execution stack, and so forth. The mobile agent technologies can be divided in to two groups:

- **Weakly mobile:** Only the code is migrating; no execution state is sent along with an agent program
- **Strong mobile:** A running program is moving to another execution location (along with its particular state)

The protection of the integrity of the mobile agent is the most crucial requirement for the agent system. The agent's code and internal data autonomously migrate between hosts and can be easily changed during the transmission or at a malicious host site. A malicious platform may make subtle changes in the execution flow of the agent's code; thus, the changes in the computed results are difficult to detect. The agent cannot itself prevent this, but different countermeasures can be utilized in order to detect any manipulation made by an unauthorized party. They can be integrated directly into the agent system, or only into the design of an agent to extend the capabilities of the underlying agent system. However, the balance between the security level and solution implementation's cost, as well as performance impact, has to be preserved. Sometimes, some restrictions of agent's mobility may be necessary.

Accountability is also essential for the proper functioning of the agent system and establishing trust between the parties. Even an authenticated agent is still able to exhibit malicious behavior to the platform if such a behavior cannot later be detected and proved. Accountability is usually realized by maintaining an audit log of security-relevant events. Those logs must be protected from

unauthorized access and modification. Also the non-repudiability of logs is a huge concern. An important factor of accountability is authentication. Agents must be able to authenticate to platforms and other agents and vice versa. An agent may require different degrees of authentication depending on the level of sensitivity of the data.

The accountability requirement needs also to be balanced with an agent's need for privacy. The platform may be able to keep the agent's identity secret from other agents and still maintain a form of revocable anonymity where it can determine the agent's identity if necessary and legal. The security policies of agent platforms and their auditing requirements must be carefully balanced with agent's privacy requirements.

Threats to security generally fall into three main classes: (1) disclosure of information, (2) denial of service, and (3) corruption of information (Jansen, 1999). Threats in agent system can be categorized with regard to agents and platform relations (e.g., agent attacking an agent, etc.). Another taxonomy of attacks in agent system was proposed in Man and Wei (2001). The article describes two main categories of attacks: purposeful and frivolous. The first kind is carefully planned and designed and can be further classified by the nature of attack (read or non-read) and number of attackers (solo or collaborative). During the second kind of attacks, the attacker may not know the effect of his/her actions or gain an advantage. These attacks can be random or total. Another category of attacks is connected with traffic analysis (Kulesza, Kotulski, & Kulesza, 2006) or called *blocking attacks* (when a malicious platform refuses to migrate the agent), as described by Shao and Zhou (2006). In this chapter we will focus on the threats from an agent's perspective.

Among the mentioned threats, the most important are connected with the agent platform since the most difficult to ensure is the agent's code/state integrity. There are two main concepts for protecting mobile agent's integrity:

- Providing trusted environment for agent's execution
- Detection or prevention of tampering

The first group of methods is more concentrated on the whole agent system than on an agent in particular. These seem to be easier to design and implement but, as presented in Oppliger (2000), mostly lead to some problems. The assumption that an agent works only with a group of trusted hosts makes the agent less mobile than it was previously assumed. Also an agent may need different levels of trust (some information should be revealed to host while in another situation it should be kept secret). Sometimes, it is not clear in advance that the current host can be considered as trusted. A method to provide such an environment is special tamper-resistant hardware, but the cost of such a solution is usually very high.

The second group of methods provides the agents' manager with tools to detect that the agent's data or code has been modified, or an agent with a mechanism that prevents a successful, unauthorized manipulation. In this chapter we concentrate on the "built-in" solutions because they enable an agent to stay mobile in the strong sense and, moreover, provide the agent with mechanisms to detect or prevent tampering. Detection means that the technique is aimed at discovering unauthorized modification of the code or the state information. Prevention means that the technique is aimed at preventing changes of the code and the state information in any way. To be effective, detection techniques are more likely than prevention techniques to depend on legal or other social framework. The distinction between detection and prevention can be sometimes arbitrary, since prevention often involves detection (Jansen, 2000).

BACKGROUND

Many authors proposed methods for protecting integrity of the mobile code. The most interesting of them are presented in this section.

Time Limited Black-Box Security and Obfuscated Code

These methods are based on a *black-box* approach. The main idea of the black-box is to generate executable code from a given agent's specification that cannot be attacked by read (disclosure) or modification attacks. An agent is considered to be black-box if at any time the agent code cannot be attacked in the previous sense, and if only its input and output can be observed by the attacker. Since it is not possible to implement it today, the relaxation of this notion was introduced Hohl (1998): it is not assumed that *the black-box protection* holds forever, but only for a certain known time. According to this definition, an agent has the time-limited black-box property if for a certain known time it cannot be attacked in the aforementioned sense. The *time limited black-box* fulfills two black-box properties for this limited time:

- Code and data of the agent specification cannot be read
- Code and data of the agent specification cannot be modified

This scheme will not protect any data that is added later, although the currently existing variables will be changeable. Thus, it cannot protect the state of an agent, which can change between different hosts or any data, which the agent gathered.

In order to achieve the black-box property, several conversion algorithms were proposed. They are also called obfuscating or mess-up algorithms. These algorithms generate a new agent out of an original agent, which differs in code but produces the same results.

The *code obfuscation* methods make it more complicated to obtain the meaning from the code. To change a program code into a less easy "readable" form, they have to work in an automatic and parametric manner. The additional parameters should make possible that the same original program is transformed into different obfuscated programs. The difficulty is to transform the program in a way that the original (or a similar, easily understandable) program cannot be re-engineered automatically. Another problem is that it is quite difficult to measure the quality of obfuscation, as this not only depends on the used algorithm, but on the ability of the re-engineering as well. Some practical methods of code obfuscation are described by Low (1998) and general taxonomy proposed by Coilberg, Thomborson, and Low (1997).

Since an agent can become invalid before completing its computation, the obfuscated code is suitable for applications that do not convey information intended for long-lived concealment. Also, it is still possible for an attacker to read and manipulate data and code but, as a role of these elements cannot be determined, the results of this attack are random and have no meaning for the attacker.

Encrypted Functions

The encrypted functions (EF) method is one step forward in implementing the perfect black-box security. It has been proposed initially by Sander and Tschudin (1998). Since then other similar solutions were introduced (Alves-Foss, Harrison, & Lee, 2004; Burmester, Chrissikopoulos, & Kotzanikolaou, 2000) and the method is believed to be one of the canonical solutions for preserving agent's integrity (Jansen, 2000; Oppliger, 2000).

The goal of the EF, according to Jansen (2000), is to determine a method, which will enable the

mobile code to safely compute cryptographic primitives, such as digital signature, even though the code is executed in non-trusted computing environments and operates autonomously without interactions with the home platform. The approach is to enable the agent platform to execute a program assimilating an encrypted function without being able to extract the original form. This approach requires differentiation between a function and a program that implements the function.

The EF system is described as follows by Oppliger (2000):

A has an algorithm to compute function *f*. *B* has an input *x* and is willing to compute *f*(*x*) for *A*, but *A* wants *B* to learn nothing substantial about *f*. Moreover, *B* should not need interacting with *A* during the computation of *f*(*x*).

The function *f* can be, for example, a signature algorithm with an embedded key or an encryption algorithm containing the one. This would enable the agent to sign or encrypt data at the host without revealing its secret key.

Although the idea is straightforward, it is hard to find the appropriate encryption schemes that can transform arbitrary functions as shown. So far, the techniques to encrypt rationale functions and polynomials have been proposed. Also a solution based on the RSA cryptosystem was described (Burmester et al, 2000).

Cryptographic Traces

The articles by Vigna (1997, 1998) introduced cryptographic traces (also called execution traces) to provide a way to verify the correctness of the execution of an agent. The method is based on traces of the execution of an agent, which can be requested by the originator after the agent's termination and used as a basis for the execution verification. The technique requires each platform involved to create and retain a non-repudiation log or trace of the operations performed by the agent while resident there and to submit a cryptographic hash of the trace upon conclusion as a trace summary or fingerprint. The trace is composed of a sequence of statement identifiers and the platform signature information. The signature of the platform is needed only for those instructions that depend on interactions with the computational environment maintained by the platform. For instructions that rely only on the values of internal variables, the signature is not required and therefore is omitted.

This mechanism allows detecting attacks against code; state and control flow of mobile agents. This way, in the case of tampering, the agent's owner can prove that the claimed operations could never been performed by the agent. The technique also defines a secure protocol to convey agents and associated security-related information among the various parties involved, which may include a trusted third party to retain the sequence of trace summaries for the agent's entire itinerary. The approach has a number of drawbacks, the most obvious being the size and number of logs to be retained, and the fact that the detection process is triggered sporadically, based on suspicious results' observations or other factors.

Chained MAC Protocol

Different versions of chained message authentication code (MAC) protocol were described by Karjoth, Asokan, and Gulcu (1999) and Yee (1999). Some of them require existence of public key infrastructure, others are based on a single key. This protocol allows an agent to achieve strong forward integrity. To utilize this protocol, only the public key of the originator has to be known by all agent places. This can occur when the originator is a rather big company that is known by its smaller suppliers.

Assume that r_n is a random number that is generated by n^{th} host. This value will be used as a secret key in a MAC. The partial result o_n

(single piece of data, generated on *n* host), r_n and the identity of the next host are encrypted with the public key of the originator K_{i_0}, forming the encapsulated message O_n:

$$O_n = \{r_n, o_n, id(i_{n+1})\}K_{i_0}$$

A *chaining relation* is defined as follows (here *H* denotes a hash-function and *h* denotes the digest):

$$h_0 = \{r_0, o_0, id(i_1)\}K_{i_0}$$

and

$$h_{n+1} = H\{h_n, r_n, o_n, id(i_{n+1}))$$

When an agent is migrating from host i_n to i_{n+1}:

$$i_n \rightarrow i_n+1 : \{O_0, ..., O_n, h_{n+1}\}$$

Similar schemes are also called *partial results encapsulation* methods (Jansen, 2000).

Watermarking

Watermarking is mainly used to protect the copyrights for digital contents. A distributor or an owner of the content embeds a mark into a digital object, so its ownership can be proven. This mark is usually secret. Most methods exploit information redundancy and some of them can also be used to protect the mobile agent's data and code.

A method of watermarking of the mobile code was proposed by Esparza, Fernandez, Soriano, Munoz, and Forne (2003). A mark is embedded into the mobile agent by using software watermarking techniques. This mark is transferred to the agent's results during the execution. For the executing hosts, the mark is a normal part of results and is "invisible." If the owner of the agent detects that the mark has been changed (it is different than expected), he or she has proof that the malicious host was manipulating the agent's data or code. Figure 1 illustrates how the mark is appended to data during the mobile agent's computations on various hosts.

The paper by Esparza et al. (2003) presents three ways of embedding the watermark into the agent:

- Marking the code
- Marking the input data
- Marking the obfuscated code

The mark or marks are validated after the agent returns to its originator.

Possible attacks against this method include:

- **Eavesdropping:** If the data is not protected in any way (e.g., not encrypted) it can be read by every host.
- **Manipulation:** The malicious host can try to manipulate either the agent's code or data to change the results and still keep the proper mark.
- **Collusion:** A group of malicious hosts can cooperate to discover the mark by comparing the obtained results.

Fingerprinting

Software fingerprinting uses watermarking techniques in order to embed a different mark for each user. Software fingerprinting shares weaknesses with software watermarking: marks must be resilient to manipulation and "invisible" to observers.

The method for fingerprinting was proposed by Esparza et al. (2003). Contrary to the watermarking methods presented previously here, the embedded mark is different for each host. When the agent returns to the owner, all results are validated and the malicious host is directly traced (see Figure 2).

The article presents two ways of embedding the mark into the agent:

Figure 1. Example of watermarking

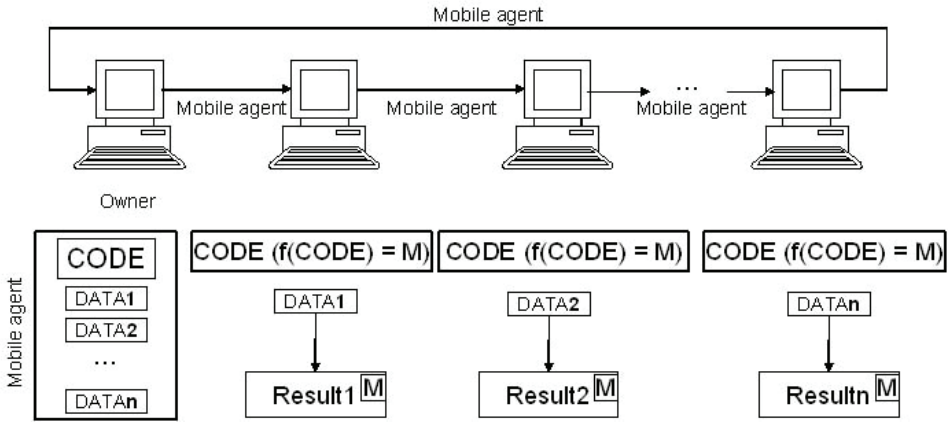
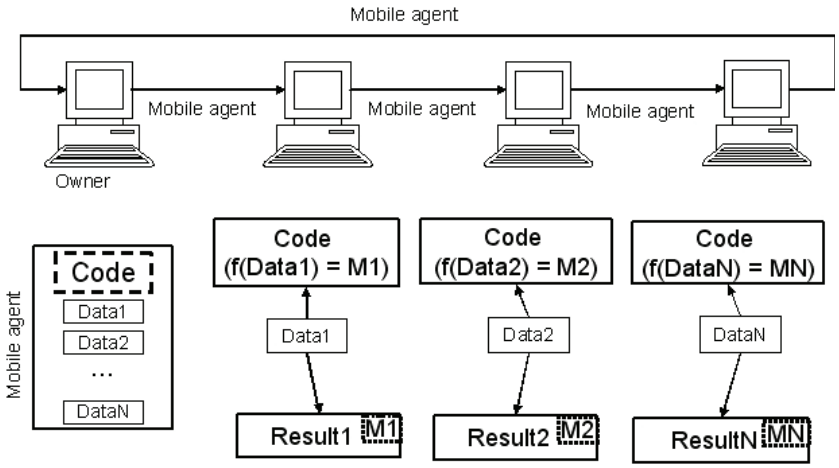


Figure 2. Example of fingerprinting



- **Marking the code:** In this case, malicious hosts have the possibility of comparing their different codes in order to locate their marks.
- **Marking the input data:** The data are usually different for each host, so it is harder to identify the mark.

The procedure is similar to the mobile agent watermarking approach. However, the owner must know each mark for each host and their location. One of the possibilities of reconstructing the marks is to catch the information about the previously chosen places in the results.

Possible attacks against this method include:

- **Eavesdropping:** If the data are not protected in any way (e.g., not encrypted) it can be read by every host.
- **Manipulation:** The malicious host can try to manipulate either the agent's code or data to change the results and still keep the proper mark.

- **Collusion:** Colluding hosts cannot extract any information about the mark comparing their data or results, because every host has a different input data and a different embedded mark.

The difference between mobile agent watermarking and fingerprinting is the fact that in the second case it is possible to detect collusion attacks performed by a group of dishonest hosts.

Publicly Verifiable Chained Digital Signatures

This protocol, proposed by Karjoth (1998) allows verification of the agent's chain of partial results not only by the originator, but also by every agent place. However, it is still vulnerable to interleaving attacks. This protocol makes it possible for every agent place, which receives an agent to verify that it has not been compromised. This saves computing power because if an agent has indeed been compromised, the agent place can reasonably refuse to execute the compromised agent.

Environmental Key Generation

This scheme allows an agent to take a predefined action when some environmental condition is true (Riordan & Schneier, 1998). The approach centers on constructing agents in such a way that upon encountering an environmental condition (e.g., via a matched search string), a key is generated, which is then used to cryptographically unlock some executable code. The environmental condition is hidden through either a one-way hash or public key encryption of the environmental trigger. This technique ensures that a platform or an observer of the agent cannot uncover the

Figure 3. Distributing ID and shares to hosts

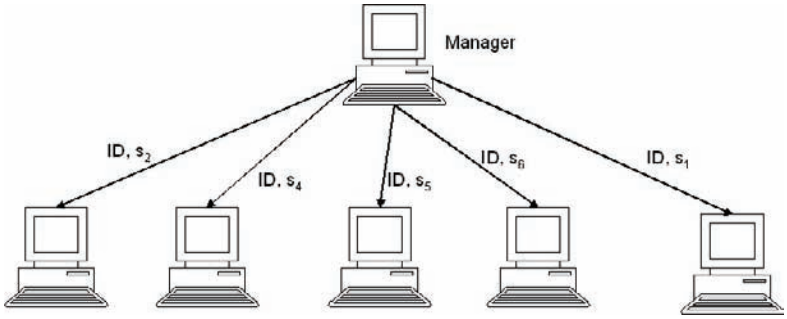
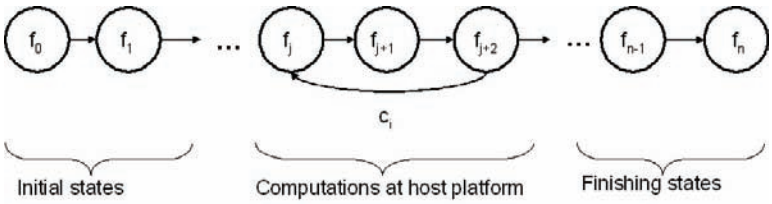


Figure 4. Mobile agent as an FSM



triggering message or response action by directly reading the agent's code.

Itinerary Recording with Replication and Voting

A faulty agent platform can behave similarly to a malicious one. Therefore, applying fault tolerant capabilities to this environment should help counter the effects of malicious platforms (Schneider, 1997). One such technique for ensuring that a mobile agent arrives safely at its destination is through the use of replication and voting. Rather than using a single copy of an agent to perform a computation, multiple copies are used. Although a malicious platform may corrupt a few copies of the agent, enough replicas avoid the encounter to successfully complete the computation. A slightly different method based on multiple copies of agent was proposed by Benachenhou and Pierre (2006). In this proposal, the copy of agent is executed on a trusted platform to validate results obtained on other platforms.

A METHOD BASED ON SECRETS AND PROOFS

In the proposed system we assume that there exist at least three parties:

- A manager
- An agent
- A host

The manager can be an originator of the agent. It plays a role of a verification instance in the scheme and creates initial countermeasures for the agent. The manager also plays a role of a trusted third party.

Outline of the Method

The zero-knowledge proof systems (Goldreich, 2002) enable the verifier to check validity of the assumption that the prover knows a secret. In our system the verifier would be the manager or owner of agents and, obviously, agents would be the provers. In the initial phase, the manager computes a set of secrets. The secrets are then composed into the agent, so that if the manager asks the agent to make some computations (denote them as a function f), the result of this would be a valid secret. This function should have the following property:

- If we have x_1 and $f(x_1)$ then it is computationally infeasible to find such x_2 that $f(x_1) = f(x_2)$

If the secret is kept within an agent, then also the host can use the zero-knowledge protocol to verify it. Every authorized change of agent's state results in such a change of the secret that the secret remains valid. On the other hand, every unauthorized change leads to loosing the secret, so at the moment of verification by host or manager, the agent is not able to prove possession of a valid secret. Since the host can monitor all agent's computations, the secret should not only change with agent's execution state, but should also be different for different hosts, so one host could only validate the secret prepared for operations that should be executed at this platform. In our system the host can tamper the agent and try to make such changes that so that he/she will be still able to obtain the proper secret, but the characteristics of function f will not allow doing this. Some possible candidates for the function f can be a hash function. Our approach is a detection rather than prevention (see Zwierko & Kotulski, 2007).

Specification of the Method

The Initial Phase

The initial phase has three steps:

1. The manager computes a set of so-called identities, denoted as ***ID***. It is public. For each identity, the manager computes appropriate secret, denoted as σ. The details for generating those values depend upon chosen zero knowledge system.
2. To compose σ into an agent, any secure secret sharing scheme (Pieprzyk, Hardjono, & Seberry, 2003) with threshold *t* can be used. The manager creates *n* shares, such that the reconstructed secret would be σ. The *t-1* shares are composed into an agent and the rest are distributed among the hosts via secure channels (this is illustrated in Figure 3).
3. The manager now needs to glue the shares into an agent in such a way, that when the agent is in a proper execution state, it is able to obtain from its code/state variables the correct shares. Since the agent is nothing more than a computer program, it can be described as a *finite state machine* (FSM). Assume, we have the agent of the form $<\Sigma, S, S_I, S_F, \delta>$, where:

    - Σ is the input alphabet
    - $S = \{f_0, \ldots, f_n\}$ is a set of all possible states
    - S_I is a subset of S with all initial states
    - S_F is a subset of S with all finishing states, possibly empty
    - $\delta: \Sigma \times S \rightarrow S$ is a state transition function.

Figure 4 shows an example of agent's FSM. It is obvious that only some execution states should be observed during the computation at the host platform (e.g., the ones connected with gathering and storing the data). If the state f_j is the first state of the agent's computations at the host platform, then it is natural that the shares should be generated only from this state. Additionally, some internal variables that differ for each host should be utilized to obtain different secrets for each host. Thus, to create agent's shares, f_j, $c_i \in \Sigma$, and the code should be used.

In other cases, where the pair f_j and c_i is not unique for each host, the previous states or other data should be used. It should be possible to obtain the proper shares for current host based on appropriate execution state and internal variables. If there is more than one unique combination of (f_j, c_i) for one host, then for each of them the host should obtain an ID and a share. The agent's code (in a certain form) should be a part of the data that are required to recreate the secret to enable detection of every unauthorized manipulation, which could be performed by previous host.

To create the shares from the mentioned data, the hash function or an encryption function with the manager's public key can be used.

The Validation Phase

1. The host, which wants to verify an agent's integrity, sends its share to the agent.
2. The agent creates the rest of the shares from its code and the execution state. It recreates the secret. The agent computes the secret σ and uses it for the rest of the scheme, which is a zero-knowledge identification protocol.
3. The agent and the host execute the selected zero-knowledge protocol, so that the host can confirm the correctness of σ.

The manager can compute many identities, which may be used with different execution states. In that situation the agent should first inform host which identity should be used, or the host can simply check the correctness of σ for all possible identities.

SECURITY AND SCALABILITY

Definitions and Notions

This section presents basic notions concerning agent's integrity that will be later used in description of the selected solutions. The integrity of an agent means that an unauthorized party cannot change its code or execution state, or such changes should be detectable (by an owner, a host or an agent platform, which want to interact with the agent). The authorized changes occur only when the agent has to migrate from one host to another. Next is a more formal definition:

Definition 1 (integrity of an agent). An agent's integrity is not compromised if no unauthorized modification can be made without the agent's owner noticing this modification.

The concept of forward integrity is also used for evaluation of many methods (Karjoth et al., 1999; Yee, 1999). This notion is used in a system where agent's data can be represented as a chain of partial results (a sequence of static pieces of data). Forward integrity can be divided into two types, which differ in their possibility to resist cooperating malicious hosts. The general goal is to protect the results within the chain of partial results from being modified. Given a sequence of partial results, the forward integrity is defined as follows:

Definition 2 (Karjoth et al., 1999; Yee, 1999). The agent posses the weak forward integrity feature if the integrity of each partial result m0, …, mn-1 is provided when in is the first malicious agent place on the itinerary.

Weak forward integrity is conceptually not resistant to cooperating malicious hosts and agent places that are visited twice. To really protect the integrity of partial result, we need a definition without constraints.

Definition 3 [strong forward integrity (Karjoth et al., 1999)]. The agent system preserves strong forward integrity of the agent if none of the agent's encapsulated messages mk, with k < n, can be modified without notifying the manager.

In this chapter we refer to forward integrity as to strong forward integrity (when applicable). To make notion of forward integrity more useful, we define also publicly verifiable forward integrity, which enables any host to detect compromised agents:

Definition 4. The agent posses the publicly verifiable forward integrity if every host in can verify that the agent's chain of partial results mi0 „…, min has not been compromised.

The other important notion concerning agent's integrity, a concept of black-box security (Hohl, 1998) was introduced in the Time Limited Black-Box Security and Obfuscated Code section.

Analysis

The proposed scheme should be used with more that one identity. This would make it very hard to manipulate the code and the data. The best approach is to use one secret for each host. We assume that the malicious host is able to read and manipulate an agent's data and code. He/she can try to obtain from an agent's execution state the proper shares. The host can also try to obtain a proper secret and manipulate the agent's state and variables in a way that the obtained secret would stay the same. But the host does not know other secrets that are composed into the agents; also he/she does not know more shares to recreate those secrets, so, any manipulation would be detected by the next host.

The protocol is not able to prevent any attacks that are aimed at destroying the agent's data or code, meaning that a malicious host can "invali-

date" any agent's data. But this is always a risk, since the host can simply delete an agent.

- **Weak forward integrity:** The proposed method posses the *weak forward integrity* property: the malicious host cannot efficiently modify previously generated results.
- **Strong forward integrity:** The protocol provides the agent also with *strong forward integrity*, because the host cannot change previously stored results (without knowledge of secrets created for other hosts). He/she cannot also modify the agent in a way that could be undetectable by the next host on the itinerary or by the owner.
- **Publicly verifiable forward integrity:** Each host can only verify if the agent's code or the execution state has not been changed. They cannot check wherever the data obtained on other platforms has not been modified. The agent's owner, who created all secrets, can only do this.
- **Black-box security:** The proposed system is not resistant to read attacks. A malicious host can modify the code or data, but it is detectable by agent's owner, so it is resistant to manipulation attack. The system does not have full black-box property.

Comparison with Other Methods

It is a difficult task to compare systems based on such different approaches as presented here. We decided to split comparison into two categories:

- **Practical evaluation:** If the method is hard or easy to implement:
  - **Hard:** No practical implementation exists at the moment
  - **Medium:** The method has been implemented, with much effort
  - **Easy:** The method is widely used and has been implemented for different purposes

  and what elements of an agent it protects:
- **Theoretical evaluation:** If the method satisfies the security definitions from the *Definitions and Notions* section.

The theoretical evaluation is quite hard, because some methods that have the black-box property do not "fit" other definitions. If the code or data cannot be read or manipulated (the ideal case), then how we can discuss if it can be verifiable, or, if it fulfills the forward integrity.

As for evaluation of the black-box property, it is very hard to provide the code that cannot be read. In all cases, marked by *, (see Table 2) the adversary can modify the agent but not in a way that owner or other host would not notice. This means that no efficient manipulation attack can be made, so one part of the black-box property is satisfied.

In # case the *publicly verifiable forward integrity* is satisfied only partially, because the agent's code can be verified but the data cannot.

Scalability

The initialization phase. The first phase is similar to the bootstrap phase of the system. The hosts and the manager create a static network. It is typical for agents' systems that the manager or the owner of an agent knows all hosts, so distribution of all IDs and shares is efficient. We can compare this to sending a single routing update for entire network as in OSPF protocol (the flooding). Whenever a new agent is added to the system, the same amount of information to all hosts has to be sent. Since the messages are not long (a single share and few IDs) and are generated only during creating a new agent, that amount of information should not be a problem. The sizes of parameters (keys lengths, number of puzzles, and number of shares) are appropriately adjusted to the agents' network size.

Table 1. Practical comparison of the integrity protection methods

Method	Implementation	Protects code	Protects data	Protects execution state
Encryption functions	Hard	Yes	Yes	No
Obfuscated code	Medium	Yes	No	No
Cryptographic traces	Hard	Yes	No	Yes
Watermarking	Easy	Yes	Yes	No
Fingerprinting	Easy	Yes	Yes	No
Zero knowledge proof	Easy	Yes	Yes	Yes

Table 2. Theoretical comparison of integrity protection methods

Method	Weak forward integrity	Strong forward integrity	Publicly verifiable forward integrity	*Black-box* property
Encryption functions	No	No	No	Yes
Obfuscated code	Yes	Yes	No	Partially*
Cryptographic traces	Yes	Yes	Yes	No
Watermarking	Yes	No	No	Partially*
Fingerprinting	Yes	Yes	No	Partially*
Zero knowledge proof	Yes	Yes	No#	Partially*

The operating phase. During the validation phase no additional communication between the manager and the hosts is required.

Modifications

A similar scenario can be used to provide integrity to the data obtained by the agent from different hosts. A malicious host could try to manipulate the data delivered to the agent by the previously visited hosts. To ensure that this is not possible, the agent can use the zero-knowledge protocol to protect the data. For each stored piece of data, the agent can create a unique "proof," utilizing the zero-knowledge protocol. Any third party, who does not possess σ, is not able to modify the proof. So the manager knowing σ can be sure that the data was not manipulated.

An area for development of the proposed integrity solution is to find the most appropriate function for composing secrets into hosts: The proposed solution fulfills the requirements, but some additional evaluation should be done. The next possibility for the future work would be to integrate the proposed solution to some agents' security architecture, possibly the one that would also provide an agent with strong authentication methods and anonymity (Zwierko & Kotulski, 2005). Then, such a complex system should be evaluated and implemented as a whole. A good example of such a system would be an agent-based electronic elections system for mobile devices, where the code integrity together with the anony-

mous authentication is crucial for correctness of the system (Zwierko & Kotulski, in press).

FUTURE TRENDS

In this chapter we presented methods of protection of mobile agents against attacks on their integrity. The methods offer protection on a certain level, but the agents' security can be significantly increased by avoiding risky behavior, especially visiting suspicious hosts. This can be done by using mechanisms built into individual agents or by distributed solutions based on cooperation of agents and hosts. The most promising solutions for improvement of the mobile code security can be based on risk analysis or on reputation systems. The first one needs some built-in analysis tools while the second one requires trust management infrastructure.

Risk analysis is one of the most powerful tools used in economics, industry, and software engineering (Tixier, Dusserre, Salvi, & Gaston, 2002). Most of the business enterprises carry out such an analysis for all transactions. The multi-agent or mobile agent system can be easily compared with such an economic-like scenario: There are a lot of parties making transactions with other parties. The risk analysis could be utilized to estimate how high is the probability that selected agent platform is going to harm the agent. The biggest advantage of this solution is lack of any form of cooperation between different managers: Everyone can make its own analysis based on gathered knowledge. However, the cooperation between different managers can benefit in better analysis.

Reputation systems (Sabater & Sierra, 2005; Zacharia & Maes, 2000) are well known and utilized in different applications, especially in peer-to-peer environments. They enable the detection of malicious parties based on their previous behavior, registered, valuated, and published. We can imagine an agent system where managers and owners of agents would also rate agent platforms based on their previous actions towards the agents. Of course, such a system still requires some integrity protection mechanisms, which could be used to verify if results obtained by the agent are correct. However, the applied mechanism can be rather simple, not as complicated as some presented methods, for example, EFs.

CONCLUDING REMARKS

Among security services for stored data protection two are the most important: availability and integrity. The data unavailable is useless for a potential user. Also, the data illegally defected or falsified is a worthless source of information. No other protection has sense if the data's content is destroyed. In the case of executables we face analogous problems. Except others, the executables must be available and protected against falsification (that is unauthorized changes of the designed functioning, internal state and the carried data). The problem of availability has been successfully solved by a concept of mobile agents that simply go to the destination place and work in there. However, this solution made the problem of integrity of the mobile code or mobile agent even more important than in the case of the stored data. The falsified mobile agent is not only useless. It can be even harmful as an active party making some unplanned actions. Therefore, preserving agents' integrity is a fundamental condition of their proper functioning.

In this chapter we made an overview of the existing protocols and methods for preserving the agent's integrity. The basic definitions and notions were introduced. The most important mechanisms were presented and discussed. We also proposed a new concept for detection of the tempering of an agent, based on a zero-knowledge proof system. The proposed scheme secures both, an agent's execution state and the internal data along with its code. For the practical implementation the system requires some additional research and development

work, but it looks to be a promising solution to the problem of providing an agent with effective and strong countermeasures against attacks on its integrity.

REFERENCES

Alves-Foss, J., Harrison, S., & Lee, H. (2004, January 5-8). The use of encrypted functions for mobile agent security. In *Proceedings of the 37th Hawaii International Conference on System Sciences*—Track 9 (pp. 90297b). US: IEEE Computer Society Press.

Benachenhou, L., & Pierre, S. (2006). Protection of a mobile agent with a reference clone. *Computer communications, 29*(2), 268-278.

Burmester, M., Chrissikopoulos, V., & Kotzanikolaou, P. (2000). Secure transactions with mobile agents in hostile environments. In E. Dawson, A. Clark, & C. Boyd (Eds.), Information security and privacy. *Proceedings of the 5th Australasian Conference ACISP* (LNCS 1841, pp. 289-297). Berlin, Germany: Springer.

Coilberg, Ch., Thomborson, C., & Low, D. (1997). *A taxonomy of obfuscating transformations* (Tech. Rep. No. 148). Australia: The University of Auckland.

Corradi, A., Cremonini, M., Montanari, R., & Stefanelli, C. (1999). Mobile agents integrity for electronic commerce applications. *Information Systems, 24*(6), 519-533.

Esparza, O., Fernandez, M., Soriano, M., Munoz, J. L., & Forne, J. (2003). Mobile agents watermarking and fingerprinting: Tracing malicious hosts. In V. Mařík, W. Retschitzegger, & O. Štěpánková (Eds.), *Proceedings of the Database and Expert Systems Applications (DEXA 2003)* (LNCS 2736, pp. 927-936). Berlin, Germany: Springer.

Goldreich, O. (2002). Zero-knowledge twenty years after its invention (E-print 186/2002). E-print, IACR.

Hohl, F. (1998). Time limited blackbox security: Protecting mobile agents from malicious hosts. In G. Vigna (Ed.), *Mobile agents and security* (LNCS 1419, pp. 92-113). Berlin, Germany: Springer.

Jansen, W. A. (2000). Countermeasures for mobile agent security. [Special issue]. *Computer Communications, 23*(17), 1667-1676.

Jansen, W. A., & Karygiannis, T. (1999). Mobile agents security (NIST Special Publication 800-19). Gaithersburg, MD: National Institute of Standards and Technology.

Karjoth, G., Asokan, N., & Gulcu, C. (1999). Protecting the computation results of free-roaming agents. In K. Rothermel & F. Hohl (Eds.), *Proceedings of the Second International Workshop on Mobile Agents (MA '98)* (LNCS 1477, pp. 195-207). Berlin, Germany: Springer.

Kulesza, K., & Kotulski, Z. (2003). Decision systems in distributed environments: Mobile agents and their role in modern e-commerce. In A. Lapinska (Ed.), *Proceedings of the Conference "Information in XXI Century Society"* (pp. 271-282). Olsztyn: Warmia-Mazury University Publishing.

Kulesza, K., Kotulski, Z., & Kulesza, K. (2006). On mobile agents resistant to traffic analysis. *Electronic Notes in Theoretical Computer Science, 142*, 181-193.

Low, D. (1998). Protecting Java code via code obfuscation. *Crossroads, 4*(3), 21-23.

Man, C., & Wei, V. (2001). A taxonomy for attacks on mobile agent. In *Proceedings of the International Conference on Trends in Communications, EUROCON'2001* (pp. 385-388). IEEE Computer Society Press.

Oppliger, R. (2000). *Security technologies for the World Wide Web*. Computer Security Series. Norwood, MA: Artech House Publishers.

Pieprzyk, J., Hardjono, T., & Seberry, J. (2003). *Fundamentals of computer security.* Berlin, Germany: Springer.

Riordan, J., & Schneier, B. (1998). Environmental key generation towards clueless agents. In G. Vinga (Ed.), *Mobile agents and security* (pp. 15-24). Berlin, Germany: Springer.

Sabater, J., & Sierra, C. (2005). Review on computational trust and reputation models. *Artificial Intelligence Review, 24* (1), 33-60.

Sander, T., & Tschudin, Ch. F. (1998, May 3-6). Towards mobile cryptography. In *Proceedings of the 1998 IEEE Symposium on Security and Privacy* (pp. 215-224). IEEE Computer Society Press.

Schneider, F. B. (1997). Towards fault-tolerant and secure agentry. In M. Mavronicolas (Ed.), *Proceedings 11th International Workshop on Distributed Algorithms* (pp. 1-14). Berlin, Germany: Springer.

Shao, M., & Zhou, J. (2006). Protecting mobile-agent data collection against blocking attacks. *Computer Standards & Interfaces, 28*(5), 600-611.

Tixier, J., Dusserre, G., Salvi, O., & Gaston, D. (2002). Review of 62 risk analysis methodologies of industrial plants. *Journal of Loss Prevention in the Process Industries, 15*(4), 291-303.

Vigna, G. (1997). Protecting mobile agents through tracing. In *Proceedings of the 3rd ECOOP Workshop on Mobile Object Systems.* Jyvälskylä, Finland.

Vigna, G. (1998). Cryptographic traces for mobile agents. In G. Vigna (Ed.), *Mobile agents and security* (LNCS 1419, pp. 137-153). Berlin, Germany: Springer.

Wang, T., Guan, S., & Chan, T. (2002). Integrity protection for code-on-demand mobile agents in e-commerce. *Journal of Systems and Software, 60*(3), 211-221.

Yee, B. S. (1999). A sanctuary for mobile agents. In J. Vitek & C. D. Jensen (Eds.), *Secure Internet programming: Security issues for mobile and distributed objects* (LNCS 1603, pp. 261-273). Berlin, Germany: Springer.

Zacharia, G., & Maes, P. (2000). Trust management through reputation mechanisms. *Applied Artificial Intelligence, 14*(9), 881-907.

Zwierko, A., & Kotulski, Z. (2005). Mobile agents: Preserving privacy and anonymity. In L. Bolc, Z. Michalewicz, & T. Nishida (Eds.), *Proceedings of IMTCI2004, International Workshop on Intelligent Media Technology for Communicative Intelligence* (LNAI 3490, pp. 246-258). Berlin, Germany: Springer.

Zwierko, A., & Kotulski, Z. (2007). Integrity of mobile agents: A new approach. *International Journal of Network Security, 2*(4), 201-211.

Zwierko, A., & Kotulski, Z. (2007). A lightweight e-voting system with distributed trust. *Electronic Notes in Theoretical Computer Science, 168*, 109-126.

KEY TERMS

Agent Platform (Host): Agent platform is a computer where an agent's code or program is executed. The software agent cannot perform its actions outside hosts. The host protects agents against external attacks.

Cryptographic Protocol: Cryptographic protocol is a sequence of steps performed by two or more parties to obtain a goal precisely according to assumed rules. To assure this purpose the parties use cryptographic services and techniques. They realize the protocol exchanging tokens.

Intelligent Software Agent: Intelligent software agent is an agent that uses artificial intelligence in the pursuit of its goals in contacts with hosts and other agents.

Mobile Agent: Mobile agent is an agent that can move among different platforms (hosts) at different times while the **stationary agent** resides permanently at a single platform (host).

Security Services: Security services guarantee protecting agents against attacks. During agent's transportation the code is protected as a usual file. At the host site, the agent is open for modifications and very specific methods must be applied for protection. For the agent's protection the following security services can be utilized:

- **Confidentiality:** Confidentiality is any private data stored on a platform or carried by an agent that must remain confidential. Mobile agents also need to keep their present location and the whole route confidential.
- **Integrity:** Integrity exists when the agent platform protects agents from unauthorized modification of their code, state, and data and ensure that only authorized agents or processes carry out any modification of the shared data.
- **Accountability:** Accountability exists when each agent on a given platform must be held accountable for its actions: must be uniquely identified, authenticated, and audited.
- **Availability:** Availability exists when every agent (local, remote) is able to access data and services on an agent platform, which responsible to provide them.
- **Anonymity:** Anonymity is when agents' actions and data are anonymous for hosts and other agents; still accountability should be enabled.

Software Agent: Software agent is a piece of code or computer program that can exercise an individual's or organization's authority, work autonomously at host toward a goal, and meet and interact with other agents.

Strong Mobility: Strong mobility of an agent means that a running program along with its particular (actual) state is moving from one host site to another.

Weak Mobility: Weak mobility of an agent means that only the agent's code is migrating and no execution state is sent along with an agent program.

This work was previously published in Handbook of Research on Wireless Security, edited by Y. Zhang, J. Zheng, and M. Ma, pp. 28-42, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.3
Security in Mobile Agent Systems

Chua Fang Fang
Multimedia University, Malaysia

G. Radhamani
Multimedia University, Malaysia

ABSTRACT

Agent technologies have grown rapidly in recent years as Internet usage has increased tremendously. Despite its numerous practical benefits and promises to provide an efficient way of mitigating complex distributed problems, mobile agent technology still lacks effective security measures, which severely restricts its scope of applicability. This chapter analyzes and synthesizes the different security threats and attacks that can possibly be imposed to mobile agent systems. The security solutions to resolve the problems and the research challenges in this field are presented.

INTRODUCTION

Software agent is a very generic term for a piece of software that can operate autonomously and that helps facilitate a certain task. Software agents can communicate and be intelligent in the way that they have the attributes of proactive/reactive, and have learning capabilities. In agent-based systems, humans delegate some of their decision-making processes to programs that are intelligent, mobile, or both (Harrison, Chess, & Kershenbaum, 1995). Software agents may be either stationary or mobile, such that stationary agents remain resident at a single platform while mobile agents are capable of suspending activity on one platform and moving to another, where they resume execution (Jansen, 2000). In most mobile intelligent agent systems, the software agent travels autonomously within the agent-enabled networks, executes itself in the agent execution environment, gathers related information, and makes its own decision on behalf of its owner.

SCOPE

Currently, distributed systems employ models in which processes are statically attached to hosts and communicate by asynchronous messages or synchronous remote procedure calls; mobile agent technology extends this model by including mobile processes (Farmer, Guttman, & Swarup, 1996a). Compared to the client/server model, the mobile agent paradigm offers great opportunities for performing various attacks because mobile agent systems provide a distributed computing infrastructure where applications belonging to different users can execute concurrently (Bellavista, Corradi, Federici, Montanari, & Tibaldi, 2003).

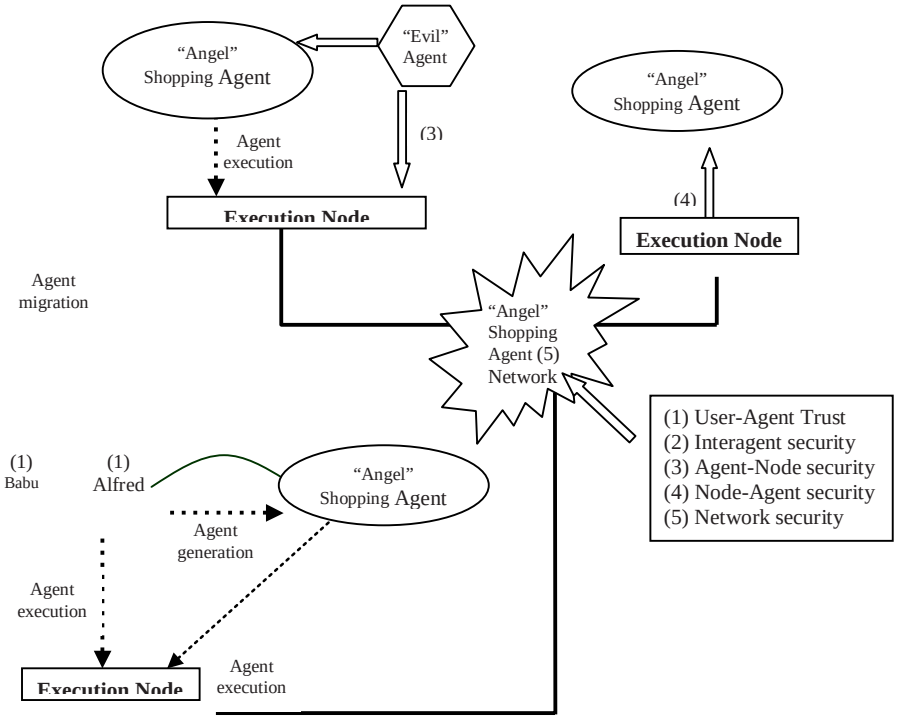
A mobile agent is an object that can migrate autonomously in a distributed system to perform tasks on behalf of its creator. It has the ability to move computations across the nodes of a wide-area network, which helps to achieve the deployment of services and applications in a more flexible, dynamic, and customizable way than the traditional client-server paradigm. For instance, if one needs to perform a specialized search of a large free-text database, it may be more efficient to move the program to the database server than to move large amounts of data to the client program. Security issues in regard to the protection of host resources, as well as the agent themselves, are extremely critical in such an environment. Apart from that, there is a greater chance for abuse or misuse, and it is difficult to identify a particular mobile process with a particular known principal and to depend on the reference monitor approach to enforce the security policy (Varadharajan, 2000).

PROBLEM STATEMENT

The general lack of security measures in existing mobile intelligent agent systems restricts their scope of applicability. According to Bellavista et al. (2003), the widespread acceptance and adoption of the mobile agent technology is currently delayed by several complex security problems that still need to be completely solved. Harrison et al. (1995) identifies security as a severe concern and regards it as the primary obstacle in adopting the mobile agent systems. Full-scale adoption of mobile agent technology in untrustworthy network environments, for example Internet, has been delayed by several security complexities. The security risks that can be encountered in mobile agent environments include malicious hosts, malicious agents, and malicious network entities. Without an appropriate security level for agents, mobile agent applications could only execute in trusted environments, and could not be deployed in the Internet scenario.

To illustrate the security requirements and issues raised by the mobile agent technology (Bellavista et al., 2003), consider the case of a shopping mobile agent that has to find the most convenient offer for a flight ticket. Suppose that Babu accesses a flight-ticket booking service (FBS) to search for and book the cheapest Rome-to-London flight ticket. Before starting an FBS provisioning session, the client requires Babu to authenticate. After a successful authentication, a middleware mobile proxy called Alfred is instantiated to represent Babu over the fixed network and to support Babu's shopping operations. A trusting relationship should be established between Babu and Alfred now that Alfred generates a shopping mobile agent and delegates it the flight searching and booking operations. The shopping agent could migrate among the various air-travel agencies' nodes to locally operate on needed resources. Once its tasks are completed, the shopping agent should be granted the same rights and submitted to the same restrictions as Alfred. In this scenario, several security issues arise and several attacks such as user-agent trust, interagent security, agent-node security, and so forth, are possible, as Figure 1 shows.

Figure 1. Security threats in mobile agent systems



Malicious Host

A malicious hosting node can launch several types of security attacks on the mobile agent and divert its intended execution towards a malicious goal, or alter its data or other information in order to benefit from the agent's mission (Sander & Tschudin, 1998). According to Jansen (2001), a receiving-agent platform can easily isolate and capture an agent and may attack it by extracting information, corrupting or modifying its code or state, denying requested services, or simply terminating it completely. An agent is very susceptible to the agent platform and may be corrupted merely by the platform responding falsely to requests for information or service, altering external communications, or delaying the agent until its task is no longer relevant.

In the case of the shopping agent scenario as mentioned (Mitchell, 2004), a malicious host could try to

- **Erase** all information previously collected by the agent so that the host is guaranteed at least to have the best current offer.
- **Change** the agent's route so that airlines with more favorable offers are not visited.
- **Terminate** the agent to ensure that no competitor gets the business either.
- Make the agent execute its commitment function, ensuring that the agent is committing to the offer given by the malicious

host. Besides this, the agent might be carrying information that needs to be kept secret from the airline (e.g., maximum price).

Integrity Attacks

Integrity of the mobile agent has been violated when tampering with the agent's code, state, or data. There are two subclasses of integrity attacks, namely **integrity interference** and **information modification** (Bierman & Cloete, 2002). Integrity interference occurs when the executing host interferes with the mobile agent's execution mission but does not alter any information related to the agent, whereas information modification includes several actions that the executing host can take against a mobile agent in an unauthorized way such as altering, manipulating, deleting the agent's code, data, status, and control flow. Modification of the agent by the platform is a particularly insidious form of attack, since it can radically change the agent's behavior or the accuracy of the computation (Jansen, 2001).

Availability Refusals

Availability refusal occurs when an authorized mobile agent is prevented from accessing objects or resources to which it should have legitimate access. It is a deliberate action performed by the executing nodes in order to obstruct the agent. There are three subclasses of availability refusal, namely **denial-of-service**, **delay-of-service,** and **transmission-refusal**.

- **Denial of service** occurs when the requested resources that the agent needs to accomplish its mission are denied. Nevertheless, it is also possible for a malicious host to bombard the agent with too much irrelevant information, so that the agent finds it impossible to complete its goals.
- **Delay of service** occurs when the host lets the mobile agent wait for the service and only provides the service or access to the required resources after a certain amount of time. This delay can have a negative effect on the actual purpose of the mobile agent.
- **Transmission refusal** occurs when a host with malicious intentions disregards the itinerary of the mobile agent and refuses to transmit the agent to the next host that is specified in the agent's itinerary.

Confidentiality Attacks

The privacy of the mobile agent is intruded when the assets of the mobile agent are illegally accessed or disposed by its host. The confidentiality attacks include **theft**, **eavesdropping,** and **reverse engineering** (Bierman & Cloete, 2002).

- **Eavesdropping** is an invasion of privacy that mostly occurs when the host spies on the agent and gathers information about the mobile agent's information or about the telecommunication between agents.
- **Theft** means that besides spying on the agent, the malicious host also removes the information from the agent. The malicious host may also "steal" the agent itself and use it for its own purposes, or simply kill it.
- **Reverse engineering** occurs when the malicious host captures the mobile agent and analyzes its data and state in order to manipulate future or existing agents. This kind of attack enables the host to construct its own similar agents, or update the profile of information to which the agent gets access.

Authentication Risks

The host may jeopardize the intended goal for the mobile agent by hiding its own identity or refusal to present its own credentials, for example, **masquerading** and **cloning**. Masquerading occurs if an executing host masks itself as one of the hosts

on the agent's itinerary when, in fact, it is not. Cloning happens when each agent carries its own credentials in order to gain authorized access to the services of its executing hosts.

Nonrepudiation

Interaction between the hosts can be very ad hoc due to the mobile agent's capability in moving autonomously in the network. The malicious host can deny the previous commitments or actions and cause dispute.

Malicious Agents

According to Schoeman and Cloete (2003), a host is faced with two potential threats from mobile agents, namely, a malicious agent that might be a virus or Trojan Horse vandalizing the host or a benign agent that might simply abuse the host's local resources. In an uncontrolled environment, mobile agents can potentially run indefinitely and consume the system level resources such as files, disk storage, I/O devices, and so forth, in their execution environment. An agent can interfere with other agents so that they cannot perform their tasks completely. Besides that, servers are exposed to the risk of system penetration by malicious agents, which may leak sensitive information. Agents may mount "denial-of-service" attacks on servers, whereby they hog server resources and prevent other agents from progressing. An attack made by a mobile agent is pretty annoying because the user may never know if the mobile agent has visited the host computer and (Ylitalo, 2000) has presented seven types of potential malicious agent attacks:

- **Damage and system modification** means a mobile agent can destroy or change resources and services by reconfiguring, modifying, or erasing them from memory or disk. Consequently, it inadvertently destroys all the other mobile agents executing there at the time.
- **Denial of service** means impeding the computer services to some resources or services. Executing mobile agent can overload a resource or service, for example, by constantly consuming network connections or blocking another process by overloading its buffers to create deadlock.
- **Breach and invasion of privacy or theft** means remove the data from the host or mobile agent illegally. A mobile agent may access and steal private information and uses covert channels to transmit data in a hidden way that violates a host's security policy.
- **Harassment and antagonism** means repeating the attacks to irritate people.
- **Social engineering** means using misinformation or coercion to manipulate people, hosts, or mobile agents.
- **Logic bomb** goes off when code, concealed within an apparently peaceful mobile agent, is triggered by a specific event, such as time, location, or the arrival of a specific person (Trojan horse program).
- **Compound attack** means using cooperating techniques whereby mobile agents can collaborate with each other in order to commit a series of attacks.

Malicious Network Entities

The network layer is responsible for the final encoding of the encrypted serialized agent object so that it can be transported by the underlying network to its next host (Schoeman & Cloete, 2003) and the network communication on the Internet is always insecure. Network entities outside the hosting node can launch attacks against a mobile agent in transit, interrupt it, and steal the encryption key and thus corrupt its integrity. Other entities both outside and inside the agent framework may attempt actions to disrupt, harm, or subvert the agent systems even when the lo-

cally active agents and the agent platform are well behaved. The obvious methods involve attacking the interagent and interplatform communications through masquerade or intercept. An attacking entity may also intercept agents or messages in transit and modify their contents, substitute other contents, or simply replay the transmission dialogue at a later time in an attempt to disrupt the synchronization or integrity of the agent framework (Jansen, 2001).

SECURITY GOALS/SOLUTIONS

The security infrastructure should have the ability to flexibly and dynamically offer different solutions to achieve different qualities of security service depending on application requirements. The mobile agent system must provide several types of security mechanisms for detecting and foiling the potential attacks that include confidentiality mechanisms, authentication mechanisms, and authorization mechanisms. Four types of countermeasures, namely measures based on trust, recording and tracking, cryptography, and time techniques to address malicious host problems were presented by Bierman and Cloete (2002).

Host's Security Mechanism (Protecting Host)

Yang et al. (Yang, Guo, & Liu, 2000) have suggested employing a number of security methods to ensure that an agent is suitable for execution. The suggestions are as follows:

Authentication

Authentication involves checking that the agent was sent from a trustworthy site. This can involve asking for the authentication details to be sent from the site where the mobile agent was launched or the site from which the agent last migrated. A mobile agent that fails authentication can be rejected from the site or can be allowed to execute as an anonymous agent within a very restricted environment. For authenticating incoming agents, agent principals can be associated with personal public/private keys and can be forced to digitally sign agents to ensure the correct identification of their responsible party. The public key-based **authentication** process safely verifies the correspondence between principal identities and keys and most authentication solutions based on public key cryptography delegate key lifecycle management to public key infrastructures (Bellavista et al., 2003).

Verification

Verification entails checking the code of a mobile agent to ensure that it does not perform any prohibited action. In order to protect the hosts, some formal techniques that can be used to develop the provably secure code are:

- **Proof carrying code:** Proof carrying code that forces agent code producer to formally prove that the mobile code has the safety properties required by the hosting-agent platform. The proof of the code correct behavior is transmitted to the hosting node that can validate the received node (Necula, 1997).
- **Path history logs:** Path history logs can be exploited to allow hosting platforms to decide whether to execute an incoming agent (Chess, Grosof, Harrison, Levine, Parris, & Tsudik, 1995). The authenticable record of the prior platforms visited by the agent is maintained so that a newly visited platform can determine whether to process the agent and the type of constraints to apply. Computing a path history requires each agent platform to add a signed entry to the agent path, indicating its identity and the identity of the next platform to visit, and to

supply the complete path history to the next platform.

- **State appraisal:** Another technique for detecting malicious agent logic uses a state appraisal function that becomes part of the agent code and guarantees that the agent state has not been tampered by malicious entities (Farmer, Guttman, & Swarup, 199b). The agent author produces the state appraisal function and it is signed together with the rest of the agent. The visited platform uses this function to verify that the agent is in a correct state and to determine the type of privileges to grant to the agent.

Authorization

After the authentication of an agent, some proper authorization must be realized (Vuong & Fu, 2001). **Authorization** determines the mobile agent's access permissions to the host resources. This indicates the amount of times a resource can be accessed or how much of a resource can be used, and the type of access the agent can perform (Yang et al., 2000). With an **authorization** language, a complete security policy can be implemented on a host, specifying which agents are allowed to do the operations and for resource usage control. Access control mechanisms can enforce the control of agent behavior at run time and can limit access to resources. For example, agents should run in a sandbox environment in which they have limited privileges, in which they are safely interpreted (Claessens, Preneel, & Vandewalle, 2003; Volpano & Smith, 1998). It is also ideally suited for situations where most of the code falls into one domain that is trusted, since modules in trusted domains incur no execution overhead.

Allocation

Allocation should prevent agents from flooding hosts and denying resources to other agents. A host has to allocate the available resources to the competing mobile agents and for some resources types, it may be possible to schedule requests in time such that all resources requests of authorized mobile agents can be satisfied eventually (Tshudin, 2000).

Payment for Services

Payment for services determines the mobile agent's ability or willingness to pay for services (Yang et al., 2000). This includes ensuring that a mobile agent can actually pay, that payment is effected correctly, and that the service paid for is satisfactory to the payee. Since the agent is consuming at least computational resources at the server and may in fact be performing transactions for goods, its liability must be limited, and this can also be done by the mechanism of payment for services.

Security Mechanism of Mobile Agents (Protecting Mobile Agents)

Bierman and Cloete (2002) presented four types of countermeasures to address the problem of malicious hosts in protecting the mobile agents. The first type of countermeasure refers to **trust-based computing,** where a trusted network environment is created in which a mobile agent roams freely and fearlessly without being threatened by a possible malicious host. A second type of countermeasure includes methods of **recording and tracking** that make use of the itinerary information of a mobile agent, either by manipulating the migration history or by keeping it hidden. The third type of solution includes **cryptographic** techniques that utilize encryption/decryption algorithms, private and public keys, digital signatures, digital timestamps, and hash functions to address different threat aspects. The forth type of countermeasure is based on **time techniques** to add restrictions on the lifetime of the mobile agent. On the other hand, similarly, Bellavista et al. (2003) explains

that the main issues to be addressed to protect agents against malicious hosts are agent **execution**, **secrecy,** and **integrity**.

Trust-Based Computing

Creating a trusted environment in which a mobile agent roams freely and fearlessly without being threatened by a possible malicious host can possibly alleviate most of the classes of threats. Protecting agent execution requires ensuring that agents are not hijacked to untrusted destinations that may present agents with a false environment, thus causing them to execute incorrectly, do not commit to unwilling actions, and do not suffer from premature termination or starvation due to unfair administrator's policies that fail to provide necessary system resources.

- **Tamper-resistant hardware:** Installing tamper-resistant hardware is a method well suited to implement the notion of trust in agent-to-host relationships. This method uses the concept of a secure coprocessor model, where physically secure hardware is added to conventional computing systems.
- **Trusted nodes:** Sensitive information can be prevented from being sent to untrusted hosts and certain misbehaviors of malicious hosts can be traced by introducing trusted nodes into the infrastructure to which mobile agents can migrate when required (Mitchell, 2004).
- **Detection objects:** Detection objects, such as dummy data items or attributes accompanying the mobile agent, are used to see if the host in question can be trusted. If the detection objects have not been modified, then reasonable confidence exists that legitimate data has not been corrupted also. Apparently, it is necessary that hosts are not aware of the inserted detection objects (Meadows, 1997).

Recording and Tracking

This type of countermeasure makes use of the itinerary information of a mobile agent, either by manipulating the migration history or by keeping it hidden.

- **Execution tracing:** To address the malicious host attacks, an **execution-tracing** mechanism is used. A host platform executing an agent creates a trace of an agent's execution that contains precisely the lines of code that were executed by the mobile agent and the external values that were read by the mobile agents (Tan & Moreau, 2002). When the mobile agent requests to move, a hash of this trace and of the agent's intermediate state are signed by the host platform. This guarantees nonrepudiation by providing evidence that a specific state of execution was achieved on the host platform prior to migration.
- **Path histories:** A record of all prior platforms visited by a mobile agent is maintained in this method. The computation of a path history requires that each host add a signed entry to the itinerary carried by the mobile agent. Ordille (1996) explains that this signed entry includes the identity of the host and the identity of the next host to be visited. A path history is a countermeasure that is strongly used in the malicious agent problem, where it is needed to maintain record of the agent's travels that can be substantiated.

Cryptographic

Techniques under this type of countermeasure, titled encryption/decryption algorithms, private and public keys, digital signatures, digital timestamps and hash functions, are used to address different threat aspects. Protecting agent integrity requires the identification of agent tampering,

either of its code or of its state, by malicious execution hosts (Bellavista et al., 2003).

- **Digital signature:** Yi et al. (Yi, Siew, & Syed, 2000) proposed a **digital signature** scheme in which users have a long-term key pair, but in which a message-dependent virtually certified one-time key pair is generated for each message that has to be signed. A private key that can only be used once would be an ideal solution for a mobile agent. The private key in this system is unfortunately message-related, which makes it unusable for a mobile agent that does not know the message to be signed in advance. According to Mitchell (2004), the simplest solution to tackle the malicious host problem is to use contractual means. Operators of agent platforms guarantee, via contractual agreements, to operate their environments securely and not to violate the privacy or the integrity of the agent, its data, and its computation.
- **Environmental key generation:** With environmental security measures, the execution of an agent is actually not kept private, but it is only performed when certain environmental conditions are met. **Environmental key generation** (Riordan & Schneier, 1998) is a concept in which cryptographic keys are constructed from certain environmental data. For example, an agent or part of it could be encrypted with such a key in order that it would only be decrypted and executed if this environmental data were present at the host. In theory, this could prevent agents from being executed on a malicious host; provided that the environmental conditions that identify whether a host is malicious can be defined.
- **Sliding encryption:** Young and Yung (1997) presented a special implementation of encryption, sliding encryption, that encrypts the mobile agent piecewise, which in turn yields small pieces of cipher text. The encryption is performed so that it is intractable to recover the plain text without the appropriate private key. Extra measures are employed so that it is extremely difficult to correlate the resulting cipher texts, thus making it possible to have mobile agents that are not easy to trace.
- **Proxy certificates:** Romao and Silva (1999) proposed **proxy certificates** in which instead of giving the mobile agent direct access to the user's private digital signature key, a new key pair is generated for the mobile agent. The key pair is certified by the user, thereby binding the user to that key pair; hence, proxy certificate, and as such to the transactions that the mobile agent will perform. The lifetime of the certificate is short and therefore revocation is not needed. It should be difficult for a malicious host to discover the private key before the certificate expires. Besides that, the proxy certificate can contain constraints that prevent the private key from being used for arbitrary transactions.
- **Blinded-key signature using RSA:** There are two encryption algorithms that are often used (Yang et al., 2000): secure key and public key. In secure key encryption algorithm (single key method), a common secure key used for encrypting/decrypting is shared by both sender and receiver. The typical algorithm of secure-key encryption methods is DES. In public key encryption algorithm, both parties create two particular keys, one public and the other secure. Sender encrypts the data using the public key of receiver, while receiver decrypts the very data using the secure key of its own. The typical algorithm of public-key encryption methods is RSA. It is obvious that RSA is more suitable for mobile agents, which run in an open environment.

Ferreira and Dahab (2002) presented an idea in which the private signature key is blinded. A blinded signature can be produced using this blinded-signature key. The blinding is claimed to be performed in such a way that only the resulting signature can be unblinded, but not the key. Mobile agents carry the blinded-signature key and a signed policy that defines the restrictions under which the signature key may be used. The blinding factor can be given to a third party or to the mobile agent. In the first case, the private key is cryptographically protected, as opposed to merely being obfuscated or distributed over multiple agents. The second case corresponds to the regular proxy certificate situation, where the host is able to obtain signatures on any message, but the signed policy will still determine which signatures should be considered valid.

Network Entities Security (Protecting Communication)

Security mechanisms can be included in the agent's transport protocols (Schoeman & Cloete, 2003). Secure socket layer (SSL) and transport layer security (TLS), although a bit heavyweight, can be used for securing transmission of data between two hosts. On the other hand, the key exchange protocol (KEP) offers a lightweight transport security mechanism that suits the notion of small transferable objects better. Protecting the communication can be achieved by setting up secure channels between the hosts. SSL is the most widely used protocol for secure network nowadays, which provides authentication and encryption services for TCP connections (Vuong & Fu, 2001). SSL provides encrypted communication so that eavesdropping attacks can be prevented. SSL also provides mutual authentication of both sides of the connection so that man-in-middle attacks can be prevented. SSL can be plugged into applications at the socket layer and the application does not need any special security knowledge or security-related code about SSL.

RELATED WORK (SECURITY ARCHITECTURE)

Secure Actigen System (SAS)

Many mobile agent systems have been built for both academic research and commercial purposes in recent years. The security system proposed by Vuong and Fu (2001), **secure actigen system (SAS)** uses a rich-security model that provides an identification capability to each principal and supports system resource access control to a very fine level of granularity. It offers some methods to detect if the behavior or data of an actigen agent is tampered.

Verifiable Distributed Oblivious Transfer (VDOT)

In mobile agent security, oblivious transfer (OT) from a trusted party can be used to protect the agent's privacy and the hosts' privacy. Zhong and Yang (2003) introduce a new cryptographic primitive called **verifiable distributed oblivious transfer (VDOT)** that allows the replacement of a single trusted party with a group of threshold-trusted servers. This design of VDOT uses two novel techniques: consistency verification of encrypted secret shares and consistency verification through rerandomization. CDOT protects the privacy of both the sender and the receiver against malicious attacks of the servers.

Concordia System

The agent platform protection is achieved through agent authentication and resource access control in the **Concordia system** (Wong, Paciorek, Walsh, Dicelie, Young, & Peet, 1997). Any Concordia agent has a unique identity associated with the identity of the user that has launched it, and the resource control is based on the Java 1.1 security model and relies on simple access control lists

that allow or deny access to resources on the basis only of agent identities.

Aglets System

The **aglets system** provides an aglet security manager to implement own security policies (Lange & Oshima, 1998). The behavior of the security manager cannot be changed directly, but via a GUI tool or directly editing policy files. In the aglet security model, agents can access resources depending on their associated principles.

Ajanta

The Ajanta security manager proposed by Tripathi (1999) is used only for mediating access to system-level resources. **Ajanta** protects hosting resources through an ad hoc security manager that uses identity-based access control lists to grant or deny agent access. For all application-defined resources, Ajanta uses a proxy-based mechanism where a proxy intercepts agent requests and denies or grants access based on its own security policy and on the agent's credentials.

The Secure and Open Mobile Agent (SOMA)

The secure and open mobile agent (Corradi, Montanari, & Stefanelli, 2001) developed at the University of Bologna, is another mobile agent system implemented in Java. A **SOMA** agent (a Java program) executes in an environment (the agent platform) called SOMA place, which represents physical machines, and the SOMA places can be grouped into domains that represent LANs. Places and domains provide two layers of abstraction that represent the Internet. SOMA takes security into consideration at a very early stage of its design; therefore, it provides a relatively rich and comprehensive solution for security problems. It uses a location-independent naming scheme for mobile agents' identities, which can be verified by the agent owner's digital signatures. The public keys of the agent owners are distributed by using X.509 certification infrastructure. Only the agents from the untrusted domains are subject to authentication checks and the agents from trusted domains will be trusted automatically.

RESEARCH CHALLENGES

The design challenges for interagent communication mechanisms arise due to the mobility of agents. There are several design choices such as connection-oriented communication such as TCP/IP, connectionless communication such as RPC or indirect communication. Security is an important concern in providing remote communication facilities to visiting agents, which provides a good research opportunity. Security and fault tolerance remain to be the most challenging problems in this field.

Most current security frameworks lack a clear separation between policies and security mechanisms and provide monolithic security solutions where applications cannot choose their suitable trade-off between security, scalability, and performance. A wider diffusion of the mobile agent technology is limited by the lack of an integrated and flexible security framework that is able to protect both execution sites and agents and that is capable of balancing application performance and security requirements. The interactions between the different entities in the framework need to be formalized so that specific security properties can be identified and maintained.

According to Montanari et al. (Montanari, Stefanelli, & Naranker, 2001), an approach that can provide the requested degree of flexibility and dynamicity in mobile agent-based applications is to integrate within mobile agent systems the solutions already proposed in the field of policy-driven management (Sloman, 1994). A primary advantage of this approach is the possibility of fully separating the control of agent behavior

from implementation details: policies are completely uncoupled from the automated managers in charge of their interpretation. Investigation needs to be carried out with regards to the other types of security techniques that can be employed in conjunction with execution tracing and the manner in which they can be integrated into the framework (Tan & Moreau, 2002).

Security policies may prohibit communication between two agents while any one of them is located at some untrusted host. The issue of the support that is needed for mutual authentication of mobile agents needs to be taken up in a wider context (Tripathi, Ahmed, & Karnik, 2000). There is a lack of experience with large-scale mobile agent-based applications. Most of the existing mobile agent applications are generally "small" in size, requiring at most a few tens of agents. Good program development and debugging tools can be an interesting line of research.

CONCLUSION

The revolution of the Internet enhances the rapid development of mobile agent technology, and mobile agent is potentially playing an important role in the future communication systems. There are a number of agent-based application domains for which basic and conventional security techniques should prove adequate (Jansen, 2001). Full-scale adoption of mobile agent technology in the Internet and standards definition for security in mobile agent frameworks can be achieved by effective and improved security mechanisms and strategies.

REFERENCES

Bellavista, P., Corradi, A., Federici,C., Montanari, R., & Tibaldi, D. (2003). *Security for mobile agents: Issues and challenges.* Retrieved April 20, 2005, from http://zeus.elet.polimi.it/is-manet/Documenti/pap-deis-10.pdf

Bierman, E., & Cloete, E. (2002). Classification of malicious host threats in mobile agent computing. In *Proceedings of SAICSIT* (pp. 141-148).

Chess, D., Grosof, B., Harrison, C., Levine, D., Parris, C., & Tsudik, G. (1995). Itinerant agents for mobile computing. *IEEE Personal Communications, 2*(5), 34-49.

Claessens, J., Preneel, B., & Vandewalle, J. (2003). (How) Can mobile agents do secure electronic transactions on untrusted hosts? A survey of the security issues and the current solutions. *ACM Transactions on Internet Technology, 3*(1), 28-48.

Corradi, A., Montanari, R., & Stefanelli, C. (2001). Security of mobile agents on the Internet. *Internet Research: Electronic Networking Applications and Policy, 11*(1), 84-95.

Farmer, W., Guttman, J., & Swarup, V. (1996a). Security for mobile agents: Issues and requirements. In *Proceedings of the 19th National Information Systems Security Conference*, Baltimore (pp. 591-597).

Farmer, W., Guttman, J., & Swarup, V. (1996b). Security for mobile agents: Authentication and state appraisal. In *4th European Symposium on Research in Computer Security*, Rome, Italy (pp. 118-130).

Ferreira, L., & Dahab, R. (2002). Blinded-key signatures: Securing private keys embedded in mobile agents. In *Proceedings of the 2002 ACM symposium on Applied Computing* (pp. 82-86).

Harrison, C. G., Chess, D. M., & Kershenbaum, A. (1995). *Mobile agents: Are they a good idea*? Technical Report, IBM Research Report, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY. Retrieved June 23, 2004, from http: //www.research.ibm.com/massive

Jansen, W. (2000). Countermeasures for mobile

agent security. *Computer Communications: Special Issue on Advances in Research and Application of Network Security* (pp.1667-1676).

Lange, D., & Oshima, M. (1998). *Programming and deploying Java mobile agents with aglets.* Menlo Park, CA: Addison Wesley.

Meadows, C. (1997). *Detecting attacks on mobile agents.* Foundations for Secure Mobile Code Workshop. Centre for High Assurance Computing Systems. Montery, CA: DAR A.

Mitchell, C. J. (2004). Cryptography for mobile security. Chapter 1 of *Security for Mobility* (pp. 3-10).

Montanari, R., Stefanelli, C., & Naranker, D. (2001). Flexible security policies for mobile agent systems. *Microprocessors and Microsystems* (pp. 93-99).

Necula, G. (1997). Proof carrying code. In *24th ACM Symposium on Principle of Programming Languages.* Paris: ACM Press.

Ordille, J. J. (1996). When agents roam, who can you trust? In *Proceedings of the First Conference on Emerging Technologies and Applications in Communications*, Portland, OR.

Riordan, J., & Schneier, B. (1998). Environmental key generation towards clueless agents. In G. Vigna (Ed.), *Mobile agents and security, Lecture Notes in Computer Science, 1419* (pp. 15-24). New York: Springer-Verlag.

Romao, A., & Silva, M. M. (1999). Proxy certificates: A mechanism for delegating digital signature power to mobile agents. In *Proceedings of the Workshop on Agents in Electronic Commerce* (pp. 131-140).

Sander, T., & Tschudin, C. (1998). Protecting mobile agents against malicious hosts. In *Mobile agents and security, Lecture Notes in Computer Science, 1419* (pp. 44-60). New York: Springer-Verlag.

Schoeman, M., & Cloete, E. (2003). Architectural components for the efficient design of mobile agent systems. In *Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement through Technology* (pp. 48-58).

Sloman, M. (1994). Policy driven management for distributed systems. *Plenum Press Journal of Network and Systems Management, 2*(4), 333-360.

Tan, H. K., & Moreau, L. (2002). Certificates for mobile code security. In *Proceedings of the 2002 ACM Symposium on Applied Computing* (pp. 76-81).

Tripathi, A. (1999). *Mobile agent programming in Ajanta*. 19th IEEE International Conference on Distributed Computing Systems Workshop (ICDCS'99), IEEE Computer Society Press, Austin, TX.

Tripathi, A., Ahmed, T., & Karnik, N. M. (2000). *Experiences and future challenges in mobile agent programming. Microprocessor and Microsystems.* Retrieved July 26, 2004, from http://www.cs.umn.edu/Ajanta/publications.html

Tshudin, C. (2000). Mobile agent security. In Matthias Klusch (Ed.), *Intelligent information agents: Agent based discovery and management on the internet* (pp. 431-446). Springer Verlag.

Varadharajan, V. (2000). Security enhanced mobile agents. In *Proceedings of the 7th ACM Conference on Computer and Communications Security* (pp. 200-209).

Volpano, D., & Smith, G. (1998). Language issues in mobile program security. In G. Vigna (Ed.), *Mobile Agents and Security, Lecture Notes in Computer Science, 1419* (pp. 25-43). New York: Springer-Verlag.

Vuong, S., & Fu, P. (2001). A security architecture and design for mobile intelligent agent systems.

ACM SIGAPP Applied Computing Review, 9(3), 21-30.

Wong, D., Paciorek, N., Walsh, T., Dicelie, J., Young, M., & Peet, B. (1997). Concordia: An infrastructure for collaborating mobile agents. *First International Workshop on Mobile Agents, LNCS 1219* (pp. 86-97). Berlin: Springer-Verlag.

Yang, K., Guo, X., & Liu, D. (2000). *Security in mobile agent systems: Problems and approaches, 34*(1), 21-28.

Yi, X., Siew, C. K., & Syed, M.R. (2000). Digital signature with one-time pair of keys. *Electron. Lett., 36*, 130-131.

Ylitalo, J. (2000). *Secure platforms for mobile agents.* Retrieved January 22, 2005, from http://www.hut.fi/~jylitalo/seminar99/

Young, A., & Yung, M. (1997). Sliding encryption: A cryptographic tool for mobile agents. In *Proceedings of the 4th International Workshop on Fast Software Encryption* (pp. 230-241).

Zhong, S., & Yang, R. (2003). Verifiable distributed oblivious transfer and mobile agent security. In *Proceedings of the 2003 Joint Workshop on Foundations of Mobile Computing* (pp. 12-21).

This work was previously published in Web Services Security and E-Business, edited by G. Radhamani and G. Rao, pp. 112-128, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.4
Security Issues and Possible Countermeasures for a Mobile Agent Based M-Commerce Application

Jyh-haw Yeh
Boise State University, USA

Wen-Chen Hu
University of North Dakota, USA

Chung-wei Lee
Auburn University, USA

ABSTRACT

With the advent of wireless and mobile networks, the Internet is rapidly evolving from a set of connected stationary machines to include mobile handheld devices. This creates new opportunities for customers to conduct business from any location at any time. However, the electronic commerce technologies currently used cannot be applied directly since most were developed based on fixed, wired networks. As a result, a new research area, mobile commerce, is now being developed to supplement existing electronic commerce capabilities. This chapter discusses the security issues related to this new field, along with possible countermeasures, and introduces a mobile agent based solution for mobile commerce.

INTRODUCTION

The Internet has been steadily growing at a rapid speed since its commercialization. The fast and convenient characteristics of the Internet attract a wide variety of users all over the world. Because of its ability to reach more potential customers, the Internet is changing the nature of business from a traditional model based on face-to-face negotiations to a more advanced model utiliz-

ing electronic commerce (e-commerce). People all over the world can sell, buy and trade goods online as long as they can access the Internet. As a result of recent advances in wireless and mobile network technology, accessing the Internet has become even more convenient. Users can now access the Internet with a handheld device from any location at any time they choose. This wireless technology evolution further broadens the scope of business from e-commerce to mobile commerce (m-commerce). Most major companies have foreseen this and devoted a significant effort to developing new m-commerce systems to facilitate this trend. However, the migration from e-commerce to m-commerce is not as easy as it first appears because all the existing e-commerce technologies were developed for wired networks, which are more reliable, more secure and faster than wireless and mobile networks. Therefore, without major revisions the current e-commerce technologies cannot be applied directly to m-commerce. This chapter addresses this issue by discussing possible solutions based on the use of mobile agent technology to overcome the underlying hardware limitations of m-commerce.

In order to fully deploy m-commerce for business, there are two levels of security requirements that must be satisfied. The lower level requirement is the need for a secure wireless infrastructure to protect each individual wireless communication and the higher level requirement is for a secure protocol with which to conduct mobile payment and business transactions, thus protecting the legitimate security concerns of the three parties involved, namely the customer, the merchant, and the bank. Wireless communication security is a serious problem for all wireless applications that must transmit data securely through an open airwave communication medium. IEEE 802.1x (IEEE, 2001) defines the standard for wireless authentication, key distribution, network monitoring, and similar issues. This standard uses EAP (Extensible Authentication Protocol) (Blunk & Vollbrecht, 1998) and its supported algorithms to authenticate exchanged messages. The algorithms supported by EAP are MD5 (Message Digest 5), TLS (Transport Layer Security) (Aboba & Simon, 1999; Dierks & Allen, 1999), TTLS (Tunneled TLS) (Funk & Blake-Wilson, 2002), LEAP (Lightweight EAP), and PEAP (Protected EAP) (Hakan, Josefsson, Zorn, Simon & Palekar, 2002). The security community has agreed that cryptography is the only solution to the problem of ensuring authenticity, privacy and integrity for communications through insecure media and many encryption algorithms have been developed over the past few decades. However, in a wireless environment with limited physical resources, most existing encryption algorithms are too computationally intensive. A lightweight encryption algorithm with an acceptable degree of security strength is a possible solution to this dilemma. Although the lower level security requirement, wireless communication security, is the topic of considerable ongoing research and is a vital preliminary to the deployment of all wireless applications, this chapter will instead focus on the higher level security requirement, mobile payment and transaction security.

A business transaction is likely to involve a secure negotiation made up of many back and forth messages. However, due to their limited bandwidth, mobile handheld devices cannot afford to receive and respond to those messages individually. To resolve this problem, the use of mobile software agent technology could provide a possible solution. The handheld device launches a smart mobile agent containing all the necessary negotiation and shopping logics to the Internet. The agent shops around and makes decisions based on the contained logics and returns only the final result to the customer via the handheld device. The handheld device verifies the result and performs the final transaction, that is, the actual purchase. In this way, the number of messages exchanged can be reduced considerably. Another advantage of using mobile agent technology is that it is not necessary for the handheld device to stay online

after launching the agent. The customer can disconnect the device from the network while the smart agent traverses the Internet, visiting Web sites and gathering information.

Mobile agent technology is still in its infancy, but it has attracted a great deal of research attention because of its potential utility. The major obstacle preventing the wider deployment of mobile agent technology is, again, the related security concerns. Without sufficient protection for both the mobile agents and the foreign host platforms they visit, malicious attacks may damage either the agents or their hosts. A contaminated agent could attack a host platform by planting a virus, consuming valuable resources, extracting secret data, and so forth. On the other hand, a malicious host may alter a visiting agent's shopping logics, or even kill the whole agent, to favor itself. In this chapter, these security threats and some possible countermeasures to protect the mobile agents will be discussed.

This chapter is structured as follows:

1. **Online Business Model** describes a generic business model and lists its security and resource concerns. E-commerce and m-commerce share many of the same security concerns, since both belong to this online business model. However, to satisfy the security requirements of their different underlying infrastructures, some resource concerns in m-commerce may become more important and present greater challenges than their e-commerce counterparts.
2. **E-Commerce Approach I: SET Protocol** presents the Secure Electronic Transaction protocol to illustrate how the security concerns can be satisfied.
3. **E-Commerce Approach II: Digital Cash** presents one of the existing digital cash systems that is currently used for e-commerce.
4. **Mobile Agent Technology** discusses the basic principles of mobile agent technology.
5. **The Use of Mobile Agents for Mobile Commerce** illustrates how the mobile agent technology can be applied for mobile commerce.
6. Finally, the **Conclusion** summarizes and concludes this chapter.

ONLINE BUSINESS MODEL

An online business transaction consists of two phases, shopping and purchase-payment. During the shopping phase, the customer may visit many online merchants searching for the best buy. Once a merchant has been selected, the customer may request a tamper-resistant quote from the merchant, which is a signed offer from the merchant listing the merchandise items and the offering prices. The format of a quote may look like Table 1.

The merchant's signature on the quote ensures that no other entity can modify the quote without being detected, thus guaranteeing the integrity of the quote. Once a merchant creates a quote and sends it to a customer, the merchant cannot repudiate it because no one except the merchant can generate a quote with the correct signature. Because the merchant's name is incorporated in the quote and its integrity is protected by their signature, the customer cannot maliciously present this quote to other merchants who may not want to sell the specified merchandise at the specified price. Similarly, as the customer's name is also included, a stolen quote would be useless.

After receiving a quote, the purchase-payment phase is initiated to perform the actual online purchase and payment. The customer prepares a purchase order and payment instructions based on the received quote, where

Table 1. The format of a quote from a merchant

Merchant Name	Customer Name	Merchandise	Merchant's Signature
Quantity	Unit Price	Expiration Date	

- *PO*: The purchase order includes the customer's name, the merchant's name, the merchandise items, the quantity and price of each item purchased, and the date.
- *PI*: The payment instruction consists of the customer's name, the merchant's name, the payment method such as the credit card number or the digital cash that is to be used, the total charge, and the date.

The customer initiates the purchase-payment phase by sending the prepared *PO* and *PI*, both encrypted, to the merchant. The merchant decrypts the *PO* to learn what items have been ordered, and then forwards the encrypted *PI* to the bank to ensure an authorized payment.

This online business model applies to both e-commerce and m-commerce since m-commerce is just an extension of e-commerce. However, due to the inherent physical limitations, additional challenges arise when conducting the two business phases in m-commerce. To better understand the challenges and their possible countermeasures, it is first necessary to clarify the resource and security concerns specific to m-commerce.

The two business phases present different resource concerns. The first phase is likely to generate many message round trips between a mobile device and online merchants, which will consume a lot of network bandwidth, while the second phase requires the mobile device to have high computational power in order to perform the many encryptions needed for a secure purchase and payment transaction. In a wireless environment, both of these resources are very precious and limited; existing e-commerce approaches could not be applied directly unless their resource consumption can be reduced considerably. Later in this chapter, the mobile agent technology will be introduced for this purpose.

To address the security concerns, generally speaking, a secure communication, depending on its application, must satisfy as many as possible of the following common security goals:

- *Authenticity*: The receiving end in a communication should make sure that the sender is really who it claims to be. For mutual authentication, both ends should authenticate each other.
- *Integrity*: It should not be possible to alter transmitted data without detection.
- *Confidentiality*: Only authorized entities should be able to see protected data.
- *Non-repudiation*: The recipient should have some sort of proof to show to a third party that the sender has really committed to an action in case the sender later repudiates the commitment.
- *Anonymity*: In some cases, an entity may want to initiate an activity without revealing his/her identity.

In a business transaction, because each of the three participants plays a different role, they will have different expectations and security concerns. The following list describes the main issues for the three participants:

- **Customer***:*
  1. *Authenticity*: The customer should be capable of authenticating the other two participants.
  2. *Integrity*: It should not be possible to alter purchase orders and payment instructions without detection.
  3. *Confidentiality*: The customer definitely does not want to reveal their credit card number to the merchant, and may also not want the card issuing bank to know the contents of the purchase order.
  4. *Non-repudiation*: The customer could use the received quote as a non-repudiation proof if the merchant refuses to sell the specified goods or services as previously agreed. Also, if the customer has been charged by the merchant before receiving the ordered goods or services, the customer should receive a payment receipt that can be presented as evidence if the merchant later refuses to deliver the order.
  5. *Anonymity*: For an online business transaction, a customer may want to hide his/her identity from the merchant and/or bank. Obviously the credit card system no longer works for such cases. As with the system of paying with cash used in the real world, the use of digital cash provides a possible solution and protects anonymity in the electronic world.

- **Merchant**:
  1. *Authenticity*: The merchant should be capable of authenticating the other two participants.
  2. *Integrity*: It should not be possible to alter purchase orders and payment instructions without detection.
  3. *Non-repudiation*: If the order has been delivered to the customer before payment, the merchant should receive a delivery receipt which can be presented as evidence if the customer later refuses to pay.

- **Bank**:
  1. *Authenticity*: The bank should be capable of authenticating the other two participants.
  2. *Integrity*: It should not be possible to alter purchase orders and payment instructions without detection.

With these resource and security concerns in mind, the following two sections will describe some existing e-commerce approaches to see how these concerns can be satisfied.

E-COMMERCE APPROACH I: SET PROTOCOL

The Secure Electronic Transaction (SET) Protocol (*http://www.setco.org*) was developed in the mid 90s in response to a call by two major credit card companies, Mastercard and Visa, for the establishment of an electronic commerce standard. The protocol extends the existing credit card system and allows people to use it securely over open media. As described in the previous section, the customer prefers to hide the credit card number from the merchant, as well as to hide the goods/service order from the bank. However, these two pieces of information need to be somehow linked together to prevent the merchant from maliciously attaching the payment information to a different order. The SET protocol uses dual signatures to solve this problem.

Protocol Description

In this protocol, a public hash function H and a public key cryptosystem are set up and used by the three business participants. Each of the three participants has his/her own public and private keys. Let E_C, E_M, E_B be the encryption or signa-

ture-verification functions for the customer, the merchant, and the bank, respectively. Similarly, let D_C, D_M, D_B be the decryption or signature functions for the three participants.

During the shopping phase, the customer shops around and requests a quote from the merchant who offers the best deal. After the quote is received, the customer prepares a purchase order, *PO*, and a payment instruction, *PI*, based on the quote received, and then activates the purchase-payment phase by performing the following actions:

1. Computes a *PIMD*, which is the message digest of an encrypted *PI*, that is:

   $PIMD = H(E_B(PI))$

2. Computes a *POMD*, which is the message digest of an encrypted *PO*, that is:

   $POMD = H(E_M(PO))$

3. Computes a *PIPOMD*, which is the message digest of the concatenated *PIMD* and *POMD*, that is:

   $PIPOMD = H(PIMD \mid\mid POMD)$

4. Generates a dual signature *DS*, which is the customer's signature on the *PIPOMD*, that is:

   $DS = D_C(PIPOMD)$

5. Sends the *PIMD*, $E_M(PO)$, $E_B(PI)$, and *DS* to the merchant.

There are thus four pieces of data sent to the merchant. However, only the clear text *PO* embedded in the cipher text can be retrieved by the merchant because it is encrypted by the merchant's public key. The clear text *PI* is encrypted by the bank's public key so that the merchant has no way to learn the credit card number inside the *PI*. Thus, the security goal of hiding the credit card number from the merchant is achieved. The merchant performs the following actions after receiving the message.

1. Computes *POMD* by applying the hash function to the received $E_M(PO)$, that is:

   $POMD = H(E_M(PO))$

2. Verifies the dual signature *DS* by computing the following two values:

   $H(PIMD \mid\mid POMD)$ and $E_C(DS)$

   If the two values are equal, the merchant has verified the customer's signature, and therefore authenticates the customer. Most importantly, the merchant is convinced that both the purchase order and the payment instruction were not forged during transmission and are really from the customer. Thus, the security goals of authentication and data integrity are achieved. The two values obtained are the *PIPOMD*.

3. Retrieves the purchase order *PO* by decrypting the received $E_M(PO)$, that is:

   $PO = D_M(E_M(PO))$

4. Computes $D_M(PIPOMD)$ to sign the *PIPOMD*, the value obtained at step 2.
5. Sends the *POMD*, $E_B(PI)$, $D_M(PIPOMD)$ and *DS* to the bank.

   Among the four data items sent to the bank, only the encrypted *PI* can be decrypted. The *PO* is embedded in the message digest *POMD* and therefore cannot be retrieved by the bank. Thus, the security goal of hiding the purchase order from the bank is achieved. Upon receiving the request from

the merchant, the bank performs the following actions.

1. Computes *PIMD* by applying the hash function to the received $E_B(PI)$, that is:

   $PIMD = H(E_B(PI))$

2. Verifies the dual signature *DS* by computing the following two values:

   $H(PIMD \mid\mid POMD)$ and $E_C(DS)$

   If the two values are equal, the bank has verified the customer's signature, and therefore authenticates the customer. This comparison also convinces the bank that the received *POMD*, $E_B(PI)$ and *DS* have not been modified and thus the security goal of data integrity is guaranteed. The two values obtained are the *PIPOMD*.

3. Uses the merchant's public key to verify the merchant's signature. That is, the bank computes:

   $E_M(D_M(PIPOMD))$

   and then compares the value to the *PIPOMD* obtained in the previous step. If the two values are equal, the bank is really communicating with the merchant as it claimed. Thus, the bank authenticates the merchant.

4. Retrieves the payment instruction *PI* by decrypting the received $E_B(PI)$, that is:

   $D_B(E_B(PI))$

5. Returns a digitally signed receipt to the merchant, guaranteeing payment.

After receiving the receipt from the bank, the merchant:

1. Verifies the bank's signature on the received receipt to authenticate the bank. That is, the merchant computes and compares the following two values:

   $D_M(PIPOMD)$ and *PIPOMD*

   If the two values are equal, the merchant successfully authenticates the bank and knows that the received receipt is indeed from the bank.

2. Returns the bank's receipt $D_B(PIPOMD)$, together with its own signed receipt $D_M(PIPOMD)$, to the customer.

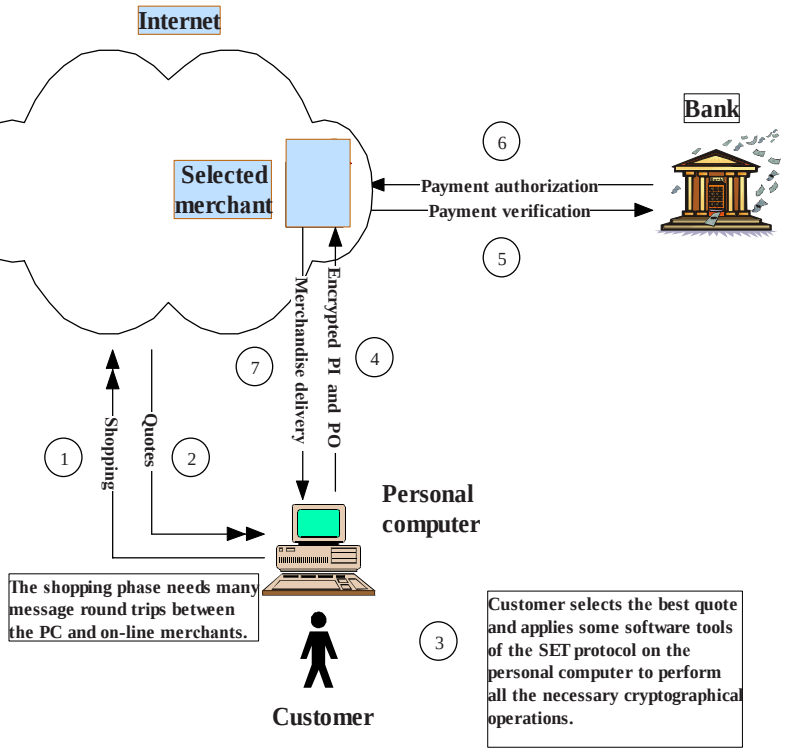
To complete the phase, the customer authenticates both the merchant and the bank by verifying the signatures on the two receipts received.

The protocol described in this section is not exactly the same as the SET protocol. Some modifications have been made, since the original protocol does not consider all the security concerns mentioned in the previous section. For example, the original SET protocol only deals with the purchase-payment phase. For a complete online business model, the authors believe that the shopping phase is also important and should be included. Moreover, in the modified SET protocol, some extra signatures on the *PIPOMD* are needed for mutual authentication purposes.

Online E-Business Using the SET Protocol

Figure 1 shows the basic sequence of events used to conduct an online e-business transaction using the SET protocol, although the figure ignores all the cryptographic details. The double arrowheads used in Steps 1 and 2 in Figure 1 represent the many back and forth messages exchanged between the personal computer and online merchants during the shopping phase. In Step 3, the customer selects the best merchant based on the received quotes and activates the purchase-payment phase

Figure 1. The sequence of events for an online e-business transaction using the SET protocol



by first performing the necessary cryptographic operations on his/her personal computer. For the remaining steps in the figure, the purchase order and payment instruction and the merchandise delivery will be securely performed by passing the encrypted/signed *PI*, *PO*, and *PIPOMD* between the three business participants, as described earlier in this section.

E-COMMERCE APPROACH II: DIGITAL CASH

Based on the use of digital cash to facilitate online business, Okamoto and Ohta (1992) identified six properties a digital cash system may have:

1. The digital cash can be sent securely through a computer network.
2. The digital cash cannot be "double spent"; that is, it cannot be copied and reused.
3. The anonymity of a digital cash spender (customer) should be preserved. If a business transaction uses digital cash, neither the merchant nor the bank should be able to identify the customer.
4. Business transactions using digital cash should not have to go through a central bank.
5. The digital cash can be transferred to others.
6. A piece of digital cash can be divided into smaller amounts.

The system developed by Okamoto and Ohta satisfies all these requirements; other digital cash systems only satisfy some. The most difficult part of developing a digital cash system is that

properties 2 and 3 above are in conflict with each other. Digital cash (a "coin") is an electronic object which is easily copied at essentially no cost. Therefore, the system must provide the business participants with some mechanism to detect a reproduced, or counterfeit, digital coin. Based on our current knowledge of the digital world, the most cost effective way to detect illegal electronic copies is by attaching a user's signature to each of the electronic coins. Any coin without a valid signature would be considered a counterfeit. However, using the existing digital signature schemes, such as DSS (FIPS, 1994) or RSA (Rivest, Shamir & Adleman, 1978), the anonymity of the coin spender cannot be preserved. In both DSS and RSA, the coin recipient must know who the coin spender is in order to identify his/her public key for signature verification.

A digital cash system developed by Brands (1994) uses a technique called "restricted blind signatures" to overcome the above problem. In this system, the customer's anonymity can be preserved if a digital coin is spent only once. However, if it is used twice, the customer can be identified by the bank. When receiving a digital coin, the merchant would first verify the validity of the coin and then request the customer to send proof that they legally possessed it. The purpose of requesting a proof is to prevent someone from stealing the coin and then trying to spend it. We will briefly describe this system below, but a more detailed treatment can be found in Brands (1994).

Initialization

The central authority and the three business participants need to perform the following steps to complete the initialization process:

- **The authority**:
  1. Picks two large prime numbers p and q, where $q = (p-1)/2$. Let g be the square of a primitive root mod p. This implies that $g^{d1} \equiv g^{d2} \pmod{p} \Leftrightarrow d_1 \equiv d_2 \pmod{q}$
  2. Chooses two secret random exponents d_1 and d_2. Let $g_1 = g^{d1} \pmod{p}$; $g_2 = g^{d2} \pmod{p}$ and then discards the two random exponents.
  3. Makes the three numbers g, g_1 and g_2 public.
  4. Chooses two public hash functions H_1 and H_2. The first hash function H_1 takes a tuple of 5 integers as input and outputs an integer mod q. The second hash function H_2 takes a tuple of 4 integers as input and outputs an integer mod q.

- **The bank**:
  1. Chooses its own secret identity number x.
  2. Computes three numbers h, h_1 and h_2 and makes them public, where $h \equiv g^x \pmod{p}$; $h_1 \equiv g_1^x \pmod{p}$; $h_2 \equiv g_2^x \pmod{p}$

- **The coin spender**:
  1. Chooses their own secret identity number u.
  2. Computes an account number C, where $C \equiv g_1^u \pmod{p}$
  3. Sends the number C to the bank, which stores C along with the coin spender's personal information such as name, address, and so forth.
  4. The bank sends back a value to the coin spender, where $z' \equiv (Cg_2)^x \pmod{p}$

- **The Merchant**:
  The merchant chooses an identity number m and registers it with the bank.

Creating a Coin

The coin spender requests digital coins through the bank by presenting its account identity C to the bank. A coin is a tuple of six numbers (D, E,

z, a, b, r) where the six numbers are constructed as follows:

1. After receiving the request from the coin spender, the bank picks a different random number v for each coin, and then computes $g_v \equiv g^v \pmod p$; $\alpha \equiv (Cg_2)^v \pmod p$. The bank sends both and to the coin spender. Note that each coin has a different pair of (g_v, α).
2. The coin spender picks a random secret tuple of five integers for each coin requested: (s, x_1, x_2, y_1, y_2).
3. The coin spender constructs the first five numbers of the tuple representing a coin as below.

   $D \equiv (Cg_2)^s \pmod p$; $E \equiv g_1^{x1} g_2^{x2} \pmod p$; $z \equiv z'^s \pmod p$;
   $a \equiv g_v^{y1} g_v^{y2} \pmod p$; $b \equiv \alpha^{sy1} D^{y2} \pmod p$

   $D = 1$ is prohibited. There are two possible cases for D to be 1. The first is if $s = 0$, then $D = 1$. Thus, the coin spender should not pick 0 for the random number s. The second is if $Cg_2 \equiv 1 \pmod p$, then $D = 1$. However, this case is highly unlikely to occur since it means that the coin spender has solved a difficult discrete logarithm problem by a lucky choice of u.
4. In order to construct the last (6th) number of the coin, the coin spender computes a value e and sends it to the bank, where:

   $e \equiv y_1^{-1} H_1(D, E, z, a, b) \pmod q$

5. Upon receiving e, the bank computes $e' \equiv (ex + v) \pmod q$ and sends it back to the coin spender.
6. The coin spender constructs r by computing $r \equiv (y_1 e' + y_2) \pmod q$

After this step, the coin construction is complete and the coin spender now owns the coin by knowing the magic six numbers. Finally, the bank deducts the amount of the coin from the spender's bank account to complete their withdrawal.

Spending the Coin

When the coin spender would like to spend a coin (D, E, z, a, b, r), he/she sends the tuple of six numbers to the merchant. The following procedure is then performed:

1. The merchant computes whether:

   $g^r \equiv a h^{H1(D, E, z, a, b)} \pmod p$; $D^r \equiv z^{H1(D, E, z, a, b)} b \pmod p$

   If both of the above hold, the merchant knows that the coin with the six numbers is constructed through the bank, and therefore is valid. However, to avoid double spending, more effort is necessary.
2. The merchant computes and sends a value $k = H_2(D, E, m, t)$ to the coin spender, where t is a timestamp of the transaction. Different transactions will thus have different values of k.
3. The coin spender computes and sends two numbers:

   $r_1 \equiv (kus + x_1) \pmod q$; $r_2 \equiv (ks + x_2) \pmod q$

   to the merchant.
4. The merchant computes whether

   $g_1^{y1} g_2^{y2} \equiv D^k E \pmod p$

   If the above checking procedure withstands this scrutiny, the coin is valid and the merchant accepts the coin. Note that a correct pair of (r_1, r_2) is a proof showing that the coin spender legally possesses the coin and has not stolen it from someone else.

Depositing the Coin in the Merchant's Bank Account

The merchant cashes the "coin" by depositing it to the bank. The merchant sends the coin (D, E, z, a, b, r), along with the triple (r_1, r_2, k), to the bank. The bank then performs the following two steps:

1. If the coin has been previously deposited, a fraud control procedure, discussed in the next section, will take over to deal with the fraudulent case. Otherwise, step 2 will be performed.
2. The bank checks whether:

$$g^r \equiv ah^{H1(D, E, z, a, b)} b \pmod{p}; \; D^r \equiv z^{H1(D, E, z, a, b)} \pmod{p}; \; g_1^{y1} g_2^{y2} \equiv D^k E \pmod{p}$$

If all three of the above are true, the coin is valid and the merchant's bank account is credited.

Double Spending

This subsection describes several possible fraudulent double spending cases and how the previously described digital cash system handles them.

1. The coin spender tries to spend the coin twice with two different merchants, M_1 and M_2. M_1 submits the coin with the triple (r_1, r_2, k) to the bank, but M_2 submits the coin along with a different triple (r_1', r_2', k'). The bank will detect the double deposits, and then initiate their fraud control procedure. The procedure will then be able to discover the malicious spender's secret identity, u, since:

$$r_1 - r_1' \equiv us(k - k') \pmod{q}; \; r_2 - r_2' \equiv s(k - k') \pmod{q}$$
$$\Rightarrow$$

The bank can then identify the coin spender by computing the spender's public identity $C \equiv g^u{}_1 \pmod{p}$.

2. The merchant tries to deposit the coin twice, once with the legitimate triple (r_1, r_2, k) and once with a forged triple (r_1', r_2', k'). Making up a valid forged triple is extremely difficult for the merchant since the merchant does not know the secret numbers u, s, x_1, and x_2, but must produce and such that:

$$g_1^{r1'} g_2^{r2'} \equiv D^{k'} E \pmod{p}$$

3. A malicious merchant *Devil* tries to deposit the coin to the bank, but also tries to use it to pay another merchant, *Angel*. *Angel* computes k', which has almost a zero chance of being equal to the original k. *Devil* doesn't know u, x_1, x_2 and s, but he must produce r_1' and r_2' such that:

$$g_1^{r1'} g_2^{r2'} \equiv D^{k'} E \pmod{p}$$

This is again a difficult discrete logarithm problem. Note that *Devil* cannot simply use the already known r_1 and r_2, since the merchant would detect that

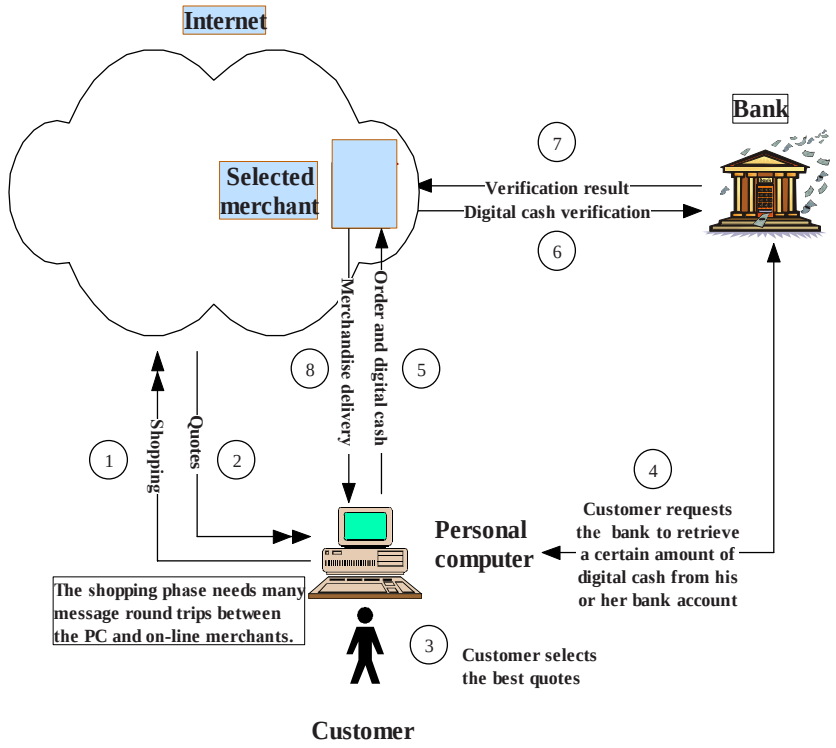
$$g_1^{r1} g_2^{r2} \neq D^{k'} E \pmod{p}.$$

Anonymity

To see how the above scheme preserves the anonymity of the coin spender, consider the following two cases:

1. Can the merchant by itself identify the coin spender? The answer is "no," since the coin spender need not provide any of his/her identities, neither u nor C, during the entire transaction with the merchant.
2. Is it possible for the merchant and the bank, acting together, to derive the spender's identity? Before answering this question,

Figure 2. The sequence of events for an online e-business transaction using a digital cash system



we would like to assume that banks are usually trustworthy, and thus this case is most likely not an issue. However, in certain rare situations, if the bank is malicious and tries to illegally identify the coin spender, the scheme described in this section also provides protection against it. The bank and the merchant together know about both the coin (D, E, z, a, b, r) and the triple (r_1, r_2, k). Since s, x_1, x_2, y_1, and y_2 are secret numbers and unknown to both the bank and merchant, the first five numbers D, E, z, a, b of the coin will just look like some random powers of g (mod p) . Therefore, the spender's identity C cannot be derived from those numbers. Note that when $e \equiv y_1^{-1} H_1(D, E, z, a, b) \pmod{q}$ is sent to the bank from the spender, the bank might calculate the value of H_1 and thus derive y_1. However, the bank has not actually seen the coin at the time of receiving the number e from the spender, and so cannot calculate the value of H_1. The bank could try to keep a list of all values of e it has received from the spenders and a list of all values of H_1 for all deposited coins, and then derive y_1 by trying all possible combinations of these two lists. Obviously, this approach requires a highly expensive and time-consuming exponential processing operation. For systems with millions of coins, this level of exhaustive matching is not practical.

Online E-Business Using Digital Cash

Figure 2 gives a basic model showing the sequence of events for conducting an online e-business transaction using a digital cash system. The shopping phase in this model, indicated by the double arrows in Steps 1 and 2 in Figure 2, is the same shopping scenario as that used in the SET protocol in Figure 1 and requires many back and forth message round trips between the personal computer and online merchants. However, these two models differ in their purchase-payment phases. Instead of providing the credit card number to the merchant, the personal computer in this model, on behalf of the customer, will withdraw an appropriate amount of digital cash from the bank and use the cash to make the purchase and payment. After receiving the digital cash from the customer, the merchant only forwards the received cash, without attaching any information about the customer, to the bank for verification. The bank is capable of verifying and authorizing the digital cash only by checking its own "blind signatures" on the cash, without the necessity of knowing the customer's identity. Thus, the anonymity of the customer can be preserved.

E-Commerce Approaches' Limitations

To conduct a business transaction using the existing e-commerce approaches, as described in the previous two sections, requires many message round trips and multiple cryptographic operations. If the underlying infrastructure is based on the use of wireless and mobile networks with limited resources, these approaches cannot be applied unless the resource consumption can be reduced significantly. The next two sections provide a possible solution that would reduce the necessary level of resource consumption for m-commerce by utilizing a new option, mobile agent technology.

MOBILE AGENT TECHNOLOGY

Mobile agent technology advances the distributed computing paradigm one step further to offer two extra properties: client customization and autonomy. End users are now able to virtually install new software in targeted foreign hosts by creating and launching a personalized mobile agent onto the Internet, thereby automatically accomplishing the assigned mission without the need for interactive guidance from the user. A mobile agent acts as a smart software agent that can be executed in foreign hosts on behalf of its owner. It can make decisions autonomously, based on the decision logics it contains. Once it has been launched, it is independent from its owner. During its life, it may visit many foreign hosts, communicate with other agents, and finally return to its owner with the results.

Several agent systems have been developed by both university and industrial research groups. Dartmouth College developed a mobile agent system, D'Agents (Gray, Kotz, Cybenko & Rus, 1998), which uses PKI for authentication, and applies the RSA algorithm to generate a public and private key pair. After a foreign host authenticates a visiting agent, the host assigns a set of access rights to the agent and sets up an appropriate execution environment. The resource access control within the host that interacts with the visiting agent is controlled by a stationary resource management agent who checks an access list each time an access request arrives. Ajanta is a Java-based mobile agent system developed at the University of Minnesota (Karnik & Tripathis, 1999). Here, an authentication server distributes a ticket to each of the registered clients. An agent acting on behalf of a client is authenticated by its possession of an appropriate ticket. Resource accesses are controlled by a security manager based on an access control list. Java Aglets (Lange & Oshima, 1998) are another Java-based mobile agent system developed at IBM's Tokyo Research Laboratory. The IBM Aglets Workbench consists

of a development kit for aglets and a host platform for aglet execution. Aglets may visit various hosts that are defined as a context in the IBM Aglets. The context owner must take steps to secure these hosts against malicious aglets. Other mobile agent systems include Ara (Peine & Stolpmann, 1997), Mole (Straser, Baumann & Hohl, 1996), and Telescript (White, 1994), the first two of which were developed as university projects and the third as a commercial product.

Sidestepping the lengthy standardization process needed for a new Internet application protocol, the customization feature of the mobile agent technology allows users to install new software into networks by simply launching appropriate agents. This great benefit of using mobile agents for applications is well understood. However, there is a major obstacle for widely deploying mobile agent technology. Until the security concerns can be resolved, the technology will not be able to reach its full potential. The concerns can be divided into four categories, as follows:

1. *Attacks on hosts by agents:* This type of attack was identified as soon as the mobile agent paradigm was proposed. Executing a program without knowing its real origin and purpose is extremely dangerous. Malicious codes can damage a computer in various ways, such as reading secret data without permission, exhausting resources by performing excessive amounts of computation or sending a huge number of messages, or changing the computer settings to make it behave abnormally. Trojan horses, viruses, and worms are well-known examples of malicious programs. In the mobile agent era, it is expected that attackers will have greater opportunities to implant such malicious codes. Fortunately, the countermeasures needed to resist this type of attack are relatively straightforward, being similar to the traditional protection techniques already employed in trusted systems. These techniques can be used to provide analogous protection to hosts in the mobile agent paradigm.
2. *Attacks on agents by rival agents:* An agent can launch an attack on a rival agent if the hosting environment does not provide sufficient protection. An agent can be malicious, eavesdropping on conversations between other agents and the host, launching a denial-of-service attack by sending messages to other agents repeatedly, or sending incorrect responses to requests it has received from other agents. A possible countermeasure is to allow the host to protect visiting agents against each other. Whenever an agent tries to access or communicate with a target agent, the host would consider the target agent as part of its own resources and provide the same level of protection as it does for its other resources.
3. *Attacks on agents by hosts:* A host can attack a visiting agent by changing the contained decision logic, spying on its accumulated data, or even killing the entire agent. In the mobile agent paradigm, there is an assumption that the host will provide appropriate resources for executing the mobile codes contained in a visiting agent. In other words, in order to execute the mobile codes, the host must have complete access rights and thus control of the agent. This leads to a serious vulnerability if the host itself is malicious. The possible countermeasures are trusted hardware (Chess, Grosof, Harrison, Levine, Parris & Tsudik, 1995), encrypted functions (Sander & Tschudin, 1998), time-limited blackbox protection (Hohl, 1998a), or a trusted virtual marketplace (Chavez & Maes, 1996; Collins, Youngdahl, Jamison, Mobasher & Gini, 1998; Tsvetovatyy & Gini, 1996). Trusted hardware consists of tamper-resistant hardware attached to each host, which can be used as a communication bridge between the host and the agent

so that a malicious host is unable to access the agent directly. Sander and Tschudin (1998) proposed the concept of encrypted functions. A function *f* is encrypted by users as *E(f)*, which is then executed by the host, without the host having access to *f*. This idea is a promising way to protect agents from malicious hosts. However, the actual implementation of this approach is not yet very clear. Time-limited blackbox protection is completely based on software. The agent code is obfuscated so that it is hard to analyze within a limited time period. However, the obfuscated code can be studied off-line by attackers. This off-line study may provide some hints that allow a faster analysis of future obfuscated mobile codes from the same source. The reason for protecting agents from hosts is because the hosts themselves may not be trustworthy. The trusted virtual marketplace approach is an attempt to provide a set of reliable hosts operated by trusted authorities. The marketplace not only guarantees the trustworthiness of all its hosts, but also needs to provide a good security mechanism to prevent attacks from other agents or outsiders. Within the marketplace, all agents can sell, buy, or trade goods without the fear of being attacked.

4. *Attacks on the agent system by other entities:* An agent system includes both mobile agents and host platforms. Other entities may attack the system by taking actions that disrupt, harm, or subvert the agent system. The mechanisms used to protect the hosts can be extended to protect the whole agent system by considering the visiting agents as part of the hosts' resources.

Mobile agents comprise a broad research area with two major categories: how to make mobile agent systems more secure and how to apply mobile agent technology to applications. This section described these security issues and their possible countermeasures and the next section will present ways to use mobile agents for m-commerce, illustrating how mobile agent technology is particularly suited to this application.

USE OF MOBILE AGENTS FOR MOBILE COMMERCE

A typical scenario applying mobile software agents for m-commerce would operate as follows. The mobile device launches a smart mobile agent containing all the necessary negotiation and shopping logics to the Internet. The agent shops around and makes decisions based on the contained logics and finally returns the best quote to the mobile device. As a result, during the shopping phase, once the agent has been launched only one message must be received and responded to by the mobile device. Another advantage of using mobile agent technology for m-commerce is the agent's real-time interaction capability. For many time-critical applications, the mobile agent can make decisions on the spot, without interactively asking for its owner's confirmation. Applications such as auctions or stock market transactions are typical time critical examples.

After the agent brings back a quote, the mobile device verifies the quote and performs the final purchase transaction. As discussed earlier in this chapter, the purchase-payment phase requires the business transaction initiator to perform a number of cryptographic operations. As an initiator, the mobile device usually lacks the computational power needed for these expensive operations. This will continue to pose a problem until lightweight encryption algorithms become available or until the hardware technology advances to provide sufficient computational power. However, an interim solution may be possible if each mobile access point is connected to a local auxiliary encryption server. The mobile device could make a request to the server for encryption service before triggering the final purchase-payment phase. However, this

approach is likely to increase the complexity of the protocol since it involves another entity. This server must also be trustworthy to avoid compromising the confidentiality of the customer.

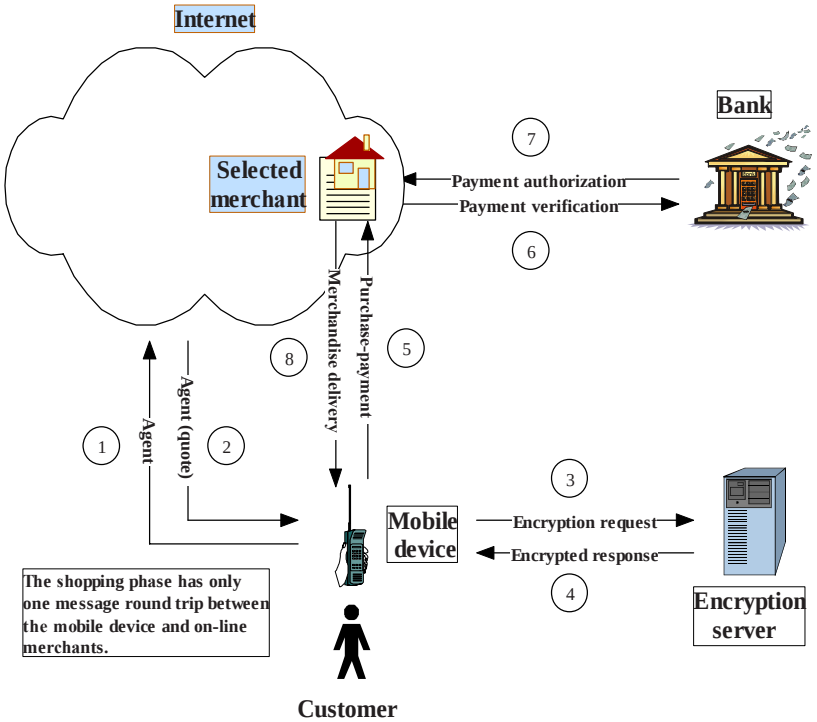
Online M-Business Using Mobile Agents

Figure 3 gives the sequence of events for an online mobile business transaction using mobile agent technology incorporating an encryption server.

In the figure, the shopping phase begins at Step 1 and ends at Step 2, in which the mobile agent shops around on the Internet for the best buy and brings back a quote from the selected merchant. The single arrowheads in Steps 1 and 2 in this figure indicate there is only one message round trip between the mobile device and online merchants during the shopping phase. To illustrate how a typical purchasing agent operates, we used an agent similar to the one used by Hohl (1998b). This agent consists of a code block and a data block as follows:

```
// CODE BLOCK
public void startAgent(){
1   if (merchantlist == null){
2         merchantlist = getTrader().getProviderOf("BuyFlowers");
3       go(merchantlist[1]);
4       break;
5   }
6  if(merchantlist[merchantlistindex].askprice(flowers) < bestprice){
7          bestprice = merchantlist[merchantlistindex].askprice(flowers);
8          bestmerchant = merchantlist[merchantlistindex];
```

Figure 3. The sequence of events for an online mobile business using mobile agents and an encryption server



```
9   }
10  if (merchantlistindex >= (merchantlist.length - 1)){
11      requestquote(bestmerchant, flowers);
12      go(home);
13  }
14  go(merchantlist[++merchantlistindex]);
15 }

// DATA BLOCK
address home = "PDA, sweet PDA";
float maximumprice = 20.00$;
good flowers = 10 red roses;
address merchantlist[] = empty list;
int merchantlistindex = 0;
float bestprice = 20.00$;
address bestmerchant = empty;
```

The purchasing agent visits a list of pre-selected online merchants to search for the lowest price of a bunch of flowers. This "lowest price" shopping strategy is encoded in the code block from line 6 to line 8. The data block specifies the agent owner's budget ($20), the merchandise to be purchased (10 red roses), the accumulated values of the agent's itinerary, and some other bookkeeping variables. Beginning at "home," the agent requests a list of online merchants to visit on line 2. Then the agent migrates to each of the merchants in the list. While visiting a merchant, the agent compares the merchant's offering price to the currently best known price, and then updates the "bestprice" and the "bestmerchant" variables if necessary. After all the listed merchants have been visited, the variable "bestmerchant" will contain the merchant who offered the best quote. Finally, line 11 in the agent's source code requests the best merchant to send an official signed quote to the agent or directly to the agent's home.

After receiving the official quote from the merchant selected, in order to activate the purchase-payment phase, the mobile device will request the encryption server to perform all the necessary cryptographic operations, as shown in Steps 3 and 4 in Figure 3. The necessary cryptographic operations were discussed in the sections "E-Commerce Approach I" and "E-Commerce Approach II". Finally, Steps 5 to 8 in Figure 3 perform the actual purchase and payment transaction by sending messages among the three business participants.

CONCLUSION

As wireless communication technology has advanced, new avenues of mobile commerce have become available. However, this opportunity to reach more customers through wireless channels and mobile devices has led to a higher risk for theft and fraud. Because of the portable features introduced for user convenience, mobile devices usually have a limited display size, limited input capability, limited computation power, limited power usage, and limited data transfer rate. The insecure broadcast medium and limited physical resources of mobile devices have made the development of security mechanisms even more challenging.

This chapter has discussed the common resource and security concerns for involved in conducting an online business. In spite of their different underlying communication infrastructures, both e-commerce and m-commerce face many of the same security concerns and thus share the same security requirements. To see how these security requirements are satisfied in e-commerce, this chapter described two existing approaches, SET protocol and digital cash. However, until the intensive resource consumption can be reduced, these existing approaches cannot be used directly for m-commerce. Fortunately, by utilizing the emerging mobile agent technology, the application of existing e-commerce methods for m-commerce becomes possible, especially for those methods that require many message round trips. This chapter also illustrated how to apply the mobile agent technology for m-commerce using an example.

REFERENCES

Aboba, B., & Simon, D. (1999). PPP EAP TLS Authentication Protocol. *IETF RFC 2716*.

Blunk, L., & Vollbrecht, J. (1998). PPP Extensible Authentication Protocol (EAP). *IETF RFC 2284*.

Brands, S. (1994). Untraceable off-line cash in wallets with observers. *Advances in Cryptology - CRYPTO'93.* Springer-Verlag.

Chavez, A., & Maes, P. (1996). Kasbah: An agent marketplace for buying and selling goods. *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM'96).*

Chess, D., Grosof, B., Harrison, C., Levine, D., Parris, C., & Tsudik, G. (1995). Internet agents for mobile computing. Technical Report, RC 20010. IBM T.J. Watson Research Center.

Collins, J., Youngdahl, B., Jamison, S., Mobasher, B., & Gini, M. (1998). A market architecture for multi-agent contracting. *Proceedings of the Second International Conference on Autonomous Agents.*

Dierks, T., & Allen, C. (1999). The TLS Protocol Version 1.0. *IETF RFC 2246*.

FIPS. (1994). Digital Signature Standard (DSS). *Federal Information Processing Standards Publication 186*.

Funk, P., & Blake-Wilson, S. (2002). EAP Tunneled TLS Authentication Protocol (EAP-TTLS). *IETF draft-ietf-pppext-eap-ttls-02.txt*

Gray, R., Kotz, D., Cybenko, G., & Rus, D. (1998). D'Agents: Security in a multiple-language, mobile-agent system. In G. Vigna (Eds.), *Mobile agents and security.* Springer-Verlag.

Hakan, A., Josefsson, S., Zorn, G., Simon, D., & Palekar, A. (2002). Protected EAP Protocol (PEAP). *IETF draft-josefsson-pppext-eap-tls-eap-05.txt*

Hohl, F. (1998a). Time limited blackbox security: Protecting mobile agents from malicious hosts. *Mobile agent security.* Springer-Verlag.

Hohl, F. (1998b). A model of attacks of malicious hosts against mobile agents. *Secure Internet mobile computation: Fourth Workshop on Mobile Object Systems (MOS'98).*

IEEE Standard for Local and Metropolitan Area Networks - Port-Based Network Access Control. (2001). *IEEE Std 802.1x-2001.*

Karnik, N., & Tripathis, A. (1999). Security in the Ajanta mobile agent system. Technical Report. Department of Computer Science, University of Minnesota.

Lange, D., & Oshima, M. (1998). *Programming and deploying JAVA mobile agents with aglets.* Addison-Wesley.

Okamoto, T., & Ohta, K. (1992). Universal electronic cash. *Advances in Cryptology - CRYPTO'91.* Springer-Verlag.

Peine, H., & Stolpmann, T. (1997). The architecture of the Ara platform for mobile agents. In Rothermel & Popescu-Zeletin (Eds.), *Mobile agents: 1st International Workshop MA'97.* Springer-Verlag.

Rivest, R., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM, 21*(2).

Sander, T., & Tschudin, C. (1998). Protecting mobile agents against malicious hosts. *Mobile agent security.* Springer-Verlag.

SET Secure Electronic Transaction Specification. *http://www.setco.org*

Straser, M., Baumann, J., & Hohl, F. (1996). A Java based mobile agent system. In M. Muhlauser

(Ed.), *Special issues in object-oriented programming: Workshop Reader of the 10th European Conference on Object-Oriented Programming ECOOP'96*.

Tsvetovatyy, M., & Gini, M. (1996). Toward a virtual marketplace: Architectures and strategies. *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM'96).*

White, J. (1994). *The foundation for the electronic marketplace*. Technical Report. General Magic, Inc.

This work was previously published in Advances in Security and Payment Methods for Mobile Commerce, edited by W. Hu, C. Lee, and W. Kou, pp. 140-163, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.5
XML Security with Binary XML for Mobile Web Services

Jaakko Kangasharju
Helsinki Institute for Information Technology, Finland

Tancred Lindholm
Helsinki Institute for Information Technology, Finland

Sasu Tarkoma
Helsinki Institute for Information Technology, Finland

ABSTRACT

In the wireless world, there has recently been much interest in alternate serialization formats for XML data, mostly driven by the weak capabilities of both devices and networks. However, it is difficult to make an alternate serialization format compatible with XML security features such as encryption and signing. We consider here ways to integrate an alternate format with security, and present a solution that we see as a viable alternative. In addition to this, we present extensive performance measurements, including ones on a mobile phone on the effect of an alternate format when using XML-based security. These measurements indicate that, in the wireless world, reducing message sizes is the most pressing concern, and that processing efficiency gains of an alternate format are a much smaller concern. We also make specific recommendations on security usage based on our measurements.

INTRODUCTION

In recent years, two developments in the computing landscape appear to be having a significant impact on the future. One of these is the rising popularity of XML (extensible markup language), which is now being used also for machine-to-machine messaging, most notably in the form of SOAP (World Wide Web Consortium [W3C], 2003a, 2003b). The other is the increasing number of available mobile devices with sophisticated networking capabilities, potentially heralding an age of truly pervasive, or ubiquitous, computing (Satyanarayanan, 2001; Weiser, 1993).

In a pervasive computing situation, a person carries a small computing device, such as a smart phone or a PDA (personal digital assistant). These kinds of devices have much less processing power available than typical personal computers. They are normally battery powered, meaning that the available energy should not be squandered, especially as battery capabilities tend to increase very slowly over time. Finally, their connection to other computers, including to the Internet, will often be on a low-bandwidth, high-latency wireless link, though in some places more powerful devices can take advantage of wireless LAN (local area network) hotspots that provide much better network connectivity.

There has been concern that XML is not suitable for use on mobile devices due to its verbosity and processing requirements. Because of this, there have been proposals to replace XML with an alternate binary XML format, which would be compatible with XML on some level but is purported to be more compact and more efficient to process. When communicating with existing systems on a fixed network, gateways can convert between this binary format and XML to permit piecewise introduction of the new format. A well-known gateway-based solution is the wireless application protocol (WAP; WAP Forum, 2001a) that includes one of the earliest binary formats for XML (W3C, 1999).

However, compatibility achieved through gateways breaks down in the case of security features such as encryption and digital signatures. If serialized content is encrypted, a gateway cannot convert it, so the ultimate recipient needs to be able to understand the used format. In the case of signatures, the signature will be computed over the serialized form, so again the recipient will need to be able to regenerate that version.

In this article, we explore the effect of a binary format in the context of XML security, in particular to determine what benefits, if any, such a format could bring. We focus on communication between a mobile device using a wireless link and a server in a fixed network. While direct peer-to-peer communication between mobile devices is also an important topic, the issues of compatibility arise more strongly in the client-server case due to the number of existing deployed systems.

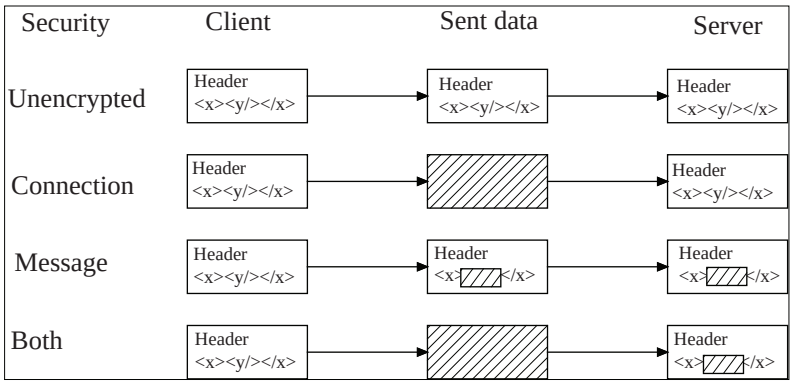
The main contributions of this article are a review of options for achieving compatibility between different formats and a comparison. We present extensive measurements, of both time and energy consumption, that were performed with real mobile devices over real networks. Finally, drawing on our measurements, we make recommendations for new features in XML security specifications that would support mobile devices better than is currently achievable.

We begin the article with usage scenarios supporting fine-grained XML security and an overview of the relevant specifications. We continue by presenting three different compatibility options to allow use of a binary format, and then show measurements using our proposed option. Next, we review related work, and finally conclude the article with specific recommendations and some view of the future.

XML SECURITY

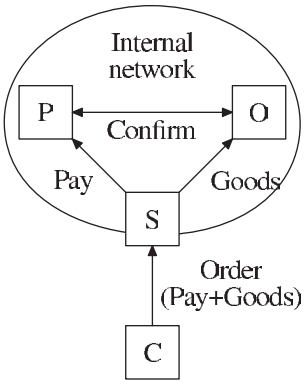
There are several existing ways to secure network traffic, many of which can be deployed immediately without needing to worry about interoperability at the application layer. On the network layer, it is possible to use IP (Internet protocol) security (Kent & Atkinson, 1998) for authentication and encryption. Transport-layer connections can be secured with SSL (secure sockets layer; Freier, Karlton, & Kocher, 1996), which provides authentication and a secure communication channel. The problems with these are that they only secure network traffic, so stored data need to be reencrypted and re-signed, and they lack the granularity to support some use cases that require multiple transport-layer connections.

Figure 1. Message flow and content with different kinds of security



The differences between connection-level security and message-level security are illustrated in Figure 1. Here we assume a message to consist of a protocol header and an XML document, and at the client end, we show the unencrypted form of the message in all cases. When connection-level security is used, the full message is encrypted in transit, and when message-level security is used, a part of the XML document is encrypted. The main difference in message-level security is that the message, as received by the server, still has the sensitive parts of the XML document encrypted.

Figure 2. The online retailer scenario



We continue by considering scenarios where this feature is an advantage to the overall system.

Usage Scenarios

We first consider the case of a user wishing to place an order with an online retailer. The order will include the user's identification, identification of the ordered goods, and payment information (such as a credit card number). The user will want to keep private, that is, encrypt, the payment information. The retailer wishes to authenticate the user to make sure no fraudulent orders are placed, which requires a digital signature. Authenticating the retailer to the user is better handled at the messaging protocol level and not at the message level.

Figure 2 shows a simple example of this scenario where the client (C) sends an order to the retailer's outward-facing system (S). S then further sends the payment information to its payment processor (P), and the list of ordered goods to its order processor (O). O also needs to confirm from P that the payment succeeded. Both O and P are located inside the retailer's secure internal network, and only S is accessible from outside.

With communication-level security, the decryption of the payment information and authentication of C both need to happen at S. In contrast, with message-level security, S only

needs to extract the relevant pieces of information from C's order and send them on to P and O. Therefore, S can be a simpler system, and since it does not perform security processing on the messages, compromising S is not sufficient for an external attacker to alter the orders or extract payment information. Furthermore, the message-level model is less coupled as payment information decryption and client authentication are separated into different components, P and O, respectively.

Workflow systems in business processing are another application area for fine-grained security. A workflow system consists of a number of message processors and communication channels between pairs of processors so that each message flows through certain processors. Typically, each processor is only interested in looking at specific parts of each message and not at the whole message. The combined actions of the processors then form the processing that is performed on a message.

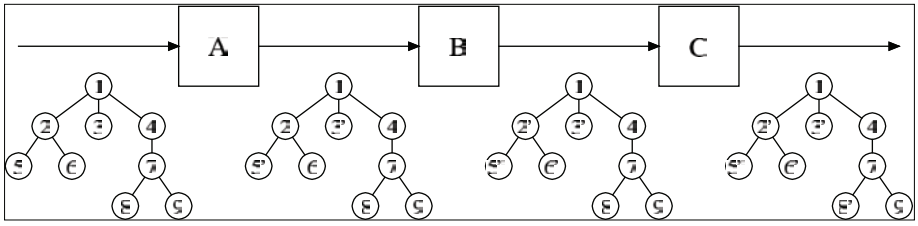
An example of a simple workflow system is shown in Figure 3, where there are three processors, A, B, and C, and an XML message (shown in tree form) is passed through the system. In the example, A needs to touch only Nodes 3 and 5, B only the subtree rooted at Node 2, and C only Node 8. In a fully XML-based system, such processing could be accomplished with XSLT (extensible stylesheet language transformations; W3C, 2007d) or XQuery (W3C, 2007c).

If this system were secured with connection-level security, each processor would have to decrypt, encrypt, verify, and sign the whole message. If the message were large, this could be a prohibitive additional cost. On the other hand, with message-level security, only the parts that are touched by each processor need to be processed, which should increase efficiency. Furthermore, the system is better compartmentalized as each encryption and signature can be targeted only to those processors that need it. Finally, each signature naturally retains the identity of its creator instead of that of the previous processor in the chain, which happens in connection-level security. This security processing can even be integrated with fully XML-based processing using XSLT (Chang & Hwang, 2004).

One way to provide fine-grained message-level security would be to use S/MIME (secure/multipurpose Internet mail extensions; Internet Engineering Task Force [IETF], 2004) by splitting the message into multiple parts, with each component to encrypt or sign being its own part, as is done with e-mail. Since much communication is moving toward XML, there would therefore need to be a way to represent XML documents as multipart MIME messages, as is done in XOP (W3C, 2005) to split Base64-encoded content out of an XML document to be transmitted in binary.

However, a solution for security based on S/MIME would require a subpart in the message for each piece of XML that is to be signed or

Figure 3. Example of an XML-based workflow system



encrypted, obscuring the content on the wire and increasing the message size due to the required MIME headers. Furthermore, due to potential security processing inside encrypted XML, this multipart solution would need to be integrated into XML processing, essentially forcing the integration of MIME into XML. Therefore, S/MIME is not very suitable for fine-grained XML security, and an XML-based solution is needed.

XML Security Standards

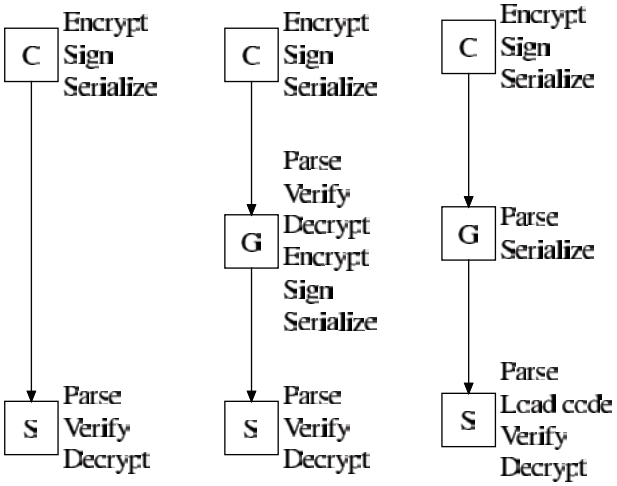
To solve the issue of fine-grained XML document security, W3C has produced specifications for XML signatures (W3C, 2002b) and XML encryption (W3C, 2002a). XML signatures are complemented by canonical XML (W3C, 2001), which specifies an algorithm to serialize an XML document so that equivalent XML documents produce the same byte sequence. This is necessary so that an XML document passed through processing can still have its signature verified.

In XML signatures, the content to be signed is marked with a reference. This reference also includes transformation methods, which are applied to the signed content to get the bytes to digest. These references are collected inside a single XML element, which is then canonicalized using canonical XML or the like; the resulting bytes are digested, and this digest is signed. If the signed data are in XML, one of the transformations applied to it will normally be a canonicalization.

The use of XML encryption results in an element that replaces the encrypted content. Such an element contains minimally an element containing the encrypted bytes. These bytes can be either embedded in the document or given as a URI (uniform resource identifier) reference. The encrypted element will also contain a type, which denotes, for example, that the encrypted content is an XML element. XML encryption also permits transformations to be applied, but these are performed only on referenced URIs to produce the actual encrypted bytes for the decrypter and not on decrypted data.

The Web services security specification from the Organization for the Advancement of Structured Information Standards (OASIS, 2004)

Figure 4. The processing by the client, gateway, and server in each compatibility model



(a) Dual format *(b) Trusted gateway* *(c) Thin gateway*

defines how XML encryption and signatures are used to secure SOAP messages. It defines a SOAP header that contains an XML signature element if the message contains signed content and an encrypted key if the message contains encrypted content. This latter element consists of a symmetric key encrypted with the recipient's public key and references to the XML content encrypted with that key.

COMPATIBILITY OPTIONS

For the purposes of this article, we assume that a Web service client resides on a mobile device and a server resides somewhere in a fixed network. The situation will be such that the server supports XML but potentially not any other formats. We also assume that the mobile user would prefer to use a binary format for its presumed compactness and processing efficiency. We consider three different ways to allow the mobile client and the server to communicate within these constraints, illustrated in Figure 4.

Dual Format

The simplest of these is the dual-format (Figure 4a) case when either the server also supports the client's binary format or the client also supports XML. In this case, the client will simply sign and encrypt according to the specifications, and the server understands this. However, when the client wishes to use a binary format that the server does not understand, this solution cannot be used.

Trusted Gateway

Currently it is possible for the mobile device to use a binary format if a gateway on the network side translates between this format and XML. Various gateways have been used for mobile device access in IP (Perkins, 1996), CORBA (common object request broker architecture; Object Management Group [OMG], 2005), and WAP (WAP Forum, 2001a). Of these, WAP is the only one that rewrites the actual messages, converting them between XML and the WAP binary format (W3C, 1999). The WAP gateway also includes the WTLS (wireless transport layer security) protocol (WAP Forum, 2001b), where the gateway reencrypts and re-signs all content passing between the mobile client and the network.

The WTLS solution can obviously be extended to handle XML security. Since the gateway already needs to handle translating between XML and the binary format, it can easily handle reencrypting and re-signing any element content as well. We call this model the trusted gateway (Figure 4b) because it requires the gateway to possess the private keys of the client and to sign and encrypt messages on its behalf. In essence, the communication path is split into two separate individually trusted paths with the gateway in between and no end-to-end security between the client and server.

The trusted-gateway model can be optimized by establishing a secure tunnel, for example, with SSL, between the client and the gateway. Then, the client will not need to actually perform XML encryption and signatures, but it will be enough to indicate which parts of the message are to be signed or encrypted by the trusted gateway. This is significant savings in processing time for the client since the secure tunnel, after establishment, requires only symmetric cryptography and not the more computationally intensive asymmetric cryptography.

Thin Gateway

Since the trusted-gateway model lacks true end-to-end security, we propose a model that we call the thin gateway (Figure 4c). In this model, the gateway is only responsible for conversions between XML and the binary format and will not do any security processing. This permits, among other things, a much larger selection of potential

gateways for clients since there is no additional trust involved.

The client behavior in the thin-gateway model is exactly the same as if it were communicating with a binary-aware server. Specifically, the canonicalization and transformation algorithms, as well as encrypted content types, are specified to be in the binary format. The gateway will only convert between the XML and binary formats and will not touch these values. Naturally, as the gateway does not possess the decryption keys, it can convert only the unencrypted data and has to pass the encrypted data along as the same byte sequence they were received.

On the server side, modifications are required only to the XML security processing. The XML security implementation needs to recognize the algorithms for the binary format and have these available (denoted by "Load code" in Figure 4c to indicate these algorithms may not always be integrated into the security implementation itself). This limits the recognition of the binary format to a single piece of code, which may even be provided by a separate entity, instead of requiring the binary format to be integrated into XML parsing and serializing.

The signature processing on the server side is essentially the same as with normal XML signatures. The only difference is that the final canonicalization algorithm that is applied to convert from the abstract infoset representation to bytes will produce the binary format. Similarly, any decrypted plaintext will be in the binary format, so the server will need to use a binary format parser to produce the SAX (simple application programming interface for XML) events or DOM (document object model) tree, or whichever application-level representation is expected.

Discussion

At the moment, the dual-format solution is only applicable by requiring the client to support XML since there is no widely accepted binary format. Therefore, this solution may not be acceptable to the mobile world, which does not consider XML's verbosity and processing requirements to be suitable. Even if agreement is reached on a binary format, support for it in the dual-format model will need to be implemented at the XML parser level, which may take some time. Therefore, gateway-based solutions will need to be at least considered.

Even if the server does not support a binary format, gateways still allow clients to use it for the unencrypted parts of a message. However, XML must be used for the encrypted contents as well as for computing the signature digests, but if the signed content is sent unencrypted, it can be serialized in binary up to the gateway. Since most traffic will likely be unencrypted even in the future, a simple gateway suffices for many applications. However, we specifically consider the security case here, so a simple gateway is not sufficient.

The obvious drawback of the trusted-gateway model is the requirement of trust. While, for example, the current mobile phone networks require placing some trust in the operator, it is still possible to engage in secure communication by performing security operations at the ends. The trusted-gateway model would effectively make the gateway owner a proxy for any secure communication initiated by the mobile client. Considering that more and more can be done through these kinds of systems, this appears unacceptable from the point of view of privacy and trust.

The main benefits of the thin-gateway model are that it does not require complete adoption of a binary format, that it permits a more flexible convergence in the binary format landscape, and that it does not require trusting the gateway. This last point is important as it means that the number of usable gateways can be significantly larger than with the trusted-gateway model.

A downside of the thin-gateway model compared to the trusted-gateway model is obviously that it requires modification to the server side.

However, as only the XML security implementation needs to be modified, the impact on the server code is smaller than if binary format support were required at the parsing level. Also, the security implementation will already need to process canonicalization and transformations generically, so adding this code is less of a burden than it would be to support an alternate message serialization format.

Still, there is the question of which binary formats are supported by the security implementation. If there is no standard, the expectation could be that implementations would be provided by third parties. Since the code in this context is security related, it would need to be carefully vetted and certified by a trusted entity. Furthermore, such a third-party implementation would need to be provided for several different Web service platforms. We do not consider this to be likely for generic binary formats, but rather expect at most one format to be widely supported.

We consider the thin-gateway model to be a reasonable alternative, especially because the same canonicalization and transformation algorithm specifications will need to be utilized with full binary support on the server, too. However, it does not seem feasible to support arbitrary binary formats. Rather, if a binary format is standardized, the thin-gateway model can be used as a stepping-stone toward full support of this format on the server.

EXPERIMENTATION RESULTS

We performed several experiments on XML security performance in the context of Web services security. Our measurements were intended to discover the effect of using a binary format instead of XML, especially with mobile phones. Furthermore, as battery life is a significant concern on mobile devices, we also measured battery consumption and show it broken down to its components so that we can determine the most fruitful avenues for improvements.

Experimentation Setup

Our experimentation platform consisted of three components: the client, gateway, and server. The server and the gateway were running on the same machine, which has a 1.5 GHz AMD Athlon XP processor, 512 MB of main memory, and the Debian GNU/Linux 3.1 operating system. For the client, we measured on two different machines. One was a desktop system with a 3 GHz Intel Pentium 4 processor, 1 GB of main memory, and Debian GNU/Linux 3.1. The other was a regular Nokia 7610 mobile phone that supports the second-generation GSM (global system for mobile communication) and GPRS (general packet radio service) networks. The desktop systems run Java 5.0 from Sun Microsystems, and the mobile phone supports Mobile Information Device Profile (MIDP) 2.0.

We used Axis, XML-Security, and WSS4J, all from the Apache project (http://www.apache.org), for the SOAP server and its Web Services security implementation. We implemented our thin-gateway model by extending XML-Security to recognize Java scheme URIs to indicate that the correct algorithm to use is the class given by the URI. We recognize that this is not the correct solution as allowing arbitrary Java classes to be loaded to perform security processing is an obvious weakness. However, if we wish to experiment with several alternate formats, this is a more extensible solution than hard-coding the binary processor classes. We also note that the thin-gateway model was extremely straightforward to implement, requiring only approximately 50 lines of new or changed code in four classes, which serves as partial validation of the feasibility of this model.

The client-side system was a simple one written by us, both to make it easy to switch XML serialization formats and to run the same system

Table 1. Formats used in the experiments

Format	Description
Xml	XML, security processing
Xebu	Xebu, security processing
Xmlunsec	XML, no security processing
Xebuunsec	Xebu, no security processing
Xmlssl	XML, no security processing, over SSL
Xebussl	Xebu, no security processing, over SSL

Table 2. Description of the experiments

Experiment	Description
Desktop	Messages with 20-200 elements at 20-element increments, 100 invocations, 20 replications
Phone	Messages with 2-20 elements at two-element increments, 10 invocations, 10 replications
Battery	Message with 10 elements, enough replications to completely drain the phone battery, measure number of invocations

on both clients. The cryptographic algorithms we used were 3DES for symmetric encryption, 1,024-bit RSA for asymmetric encryption, and SHA-1 for digests. On the server side, the implementations were the default ones shipped with Java, and on the client side, they were provided by the Bouncy Castle library (http://www.bouncycastle.org). We pregenerated the RSA keys for both the client and the server and hard-coded them on the client-side applications.

For the desktop client, we used a fixed two-hop network route with ICMP (ping) latency of approximately 0.25 ms, and for the phone client, a regular GPRS connection from a major provider with a 12-hop route and ICMP latency ranging between 600 ms and 1.3 s. (The latter figures were measured from a laptop computer using the same GPRS provider as in the actual experiments.) Both the magnitude and variation of the GPRS latency are what is expected in mobile phone networks. The maximum data rates of these networks were measured to be 100 Mbps for the fixed network and 32 Kbps for GPRS, but these are less important as the small data sizes mean that the TCP (transmission-control protocol) connections do not have time to achieve their steady-state behavior.

Experiments were performed with various formats, all described in Table 1. Xebu is our binary format (Kangasharju, Tarkoma, & Lindholm, 2005), which gives similar final sizes compared to other general-purpose binary formats. Our previous measurements indicate that the results reported below for the size and processing time are typical and not specific to this particular test data.

The actual scenario was a simple Web service invocation over HTTP (hypertext transfer protocol) containing a number of card elements, each containing four subelements (these elements are credit card descriptions, but their actual content is

less relevant to the measurements than their size, which is approximately 140 bytes in XML and 70 bytes in Xebu). The measurements reported below are all plotted against the number of card elements contained in a message and averaged for a single invocation. The sequence of elements was encrypted, and after this, the SOAP body was signed. The server responded with a similar message, that is, one containing the same elements with the same encryption and signing.

In the measurement application, a SOAP message is represented as an object that knows how to serialize itself as XML through a generic XML serialization API (application programming interface) that can support both normal XML and a binary format. The Web services security header is represented similarly as a component in the SOAP message object. The SOAP body is essentially an in-memory list of the SAX events constituting the body. The content to be encrypted or signed is indicated by qualified name, as is also the case with WS-Security.

We measured three different components of the total time. First, the system serializes the SOAP message into memory as bytes, including all security processing. The second component is opening an HTTP connection to the server, sending the message, and reading the response into memory. The server and gateway both measure the time they take in processing and include this in the response so that we can compute the time spent on communication alone. The third is parsing the server's response message, including decryption and signature verification. We also measured individual times for each security operation.

We summarize the experiments that we ran and their parameter variations in Table 2. In both the desktop and phone experiments, we began with some unmeasured invocations to eliminate any incidental startup costs. The required number of these was determined experimentally by increasing the number and observing when the measurements stabilized. The battery experiment was performed to determine the amount of energy consumed by computation and communication in this context.

Message Sizes

We first show the sizes of the messages and their relevant components. Table 3 shows the sizes of the request and response messages in bytes by giving the size of one element and the additional constant overhead in each message; that is, if a message contains *n* elements, its size is approximately *Over* + *n*×*Elem*. Table 4 gives, in similar format, the number of bytes that were actually encrypted or digested.

Table 3 shows that a secured Xebu message is between one third and two fifths of the corresponding XML message in size per element. The size of a secured Xebu message quickly becomes smaller than that of an unsecured XML message. The large overhead of the secured messages is due to the Web services security SOAP header. Finally, we note that since XML requires binary

Table 3. Message sizes in bytes, per element and constant overhead

	Request		Response	
Format	Elem	Over	Elem	Over
Xml	218	4743	163	6419
Xebu	69	1580	69	2598
Xmlunsec	151	676	121	386
Xebuunsec	69	49	70	43

Table 4. Encrypted and signed sizes in bytes, per element and constant overhead

Measurement	Encrypt	Decrypt	Sign	Verify
Xml elem	161	121	218	163
Xebu elem	69	69	69	69
Xml over	128	273	715	1502
Xebu over	15	16	111	257

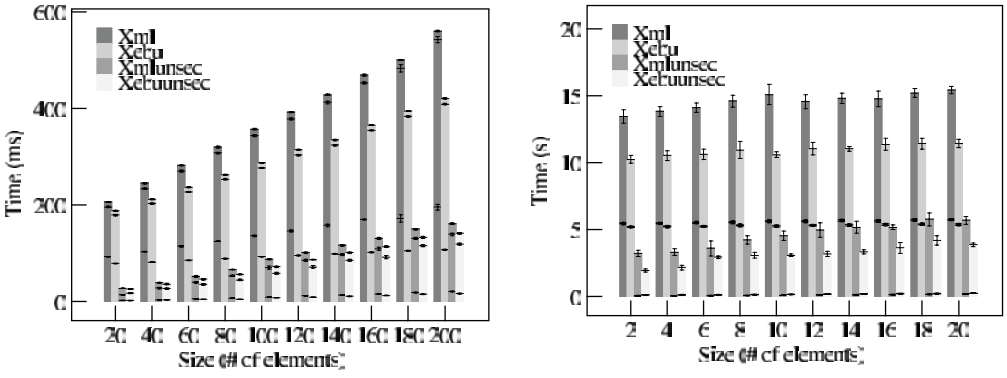
data (such as encrypted content) to be Base64 encoded, the signed content in Table 4 is one third larger than encrypted content, whereas with Xebu there is no difference.

Timing Measurements

Figure 5 shows the measured times for processing on the client, and on the gateway and server, and for communication from bottom to top. Error lines are marked at one standard deviation. Note that both figures have three lines marking processing times. However, in the phone experiment case, the time taken for remote processing at the gateway and the server is such a negligible part of the whole that its line is indistinguishable from the line drawn for local processing.

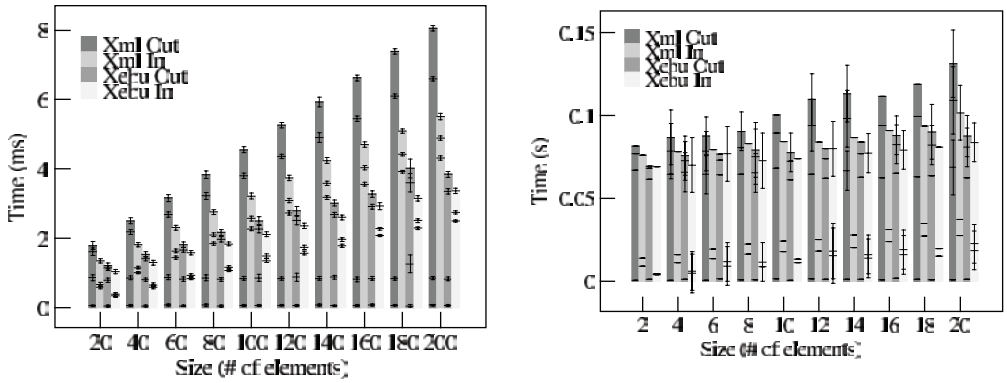
We can also see that the time taken for communication in the phone case is much higher for the security-enabled formats. By examining network packet dumps, we can see that the messages are sent in TCP segments of maximum size 1,348

Figure 5. Total times taken in the experiment for secured and unsecured formats (divisions are client, remote, network, from bottom to top)



Desktop *Phone*

Figure 6. Total times taken for the fast security operations (Out is key generation, key encryption, data encryption, digest computation, and In is data decryption, digest verification, signature verification, both from bottom to top)



Desktop *Phone*

Table 5. RSA private-key operation sizes and times

Measurement	Size (B)	
Signing size	35	
Keydec size	128	
	Desktop (ms)	Phone (s)
Xml Signing	35.70±0.01	2.19±0.01
Xml Keydec	35.52±0.04	2.26±0.00
Xebu Signing	35.71±0.03	2.20±0.01
Xebu Keydec	35.34±0.03	2.23±0.00

bytes. Since the sizes of the messages with security processing are so much larger, this generates additional round trips. With the aforementioned latency of the GPRS network, and recalling the slow-start algorithm of TCP, this becomes clearly visible in the timings.

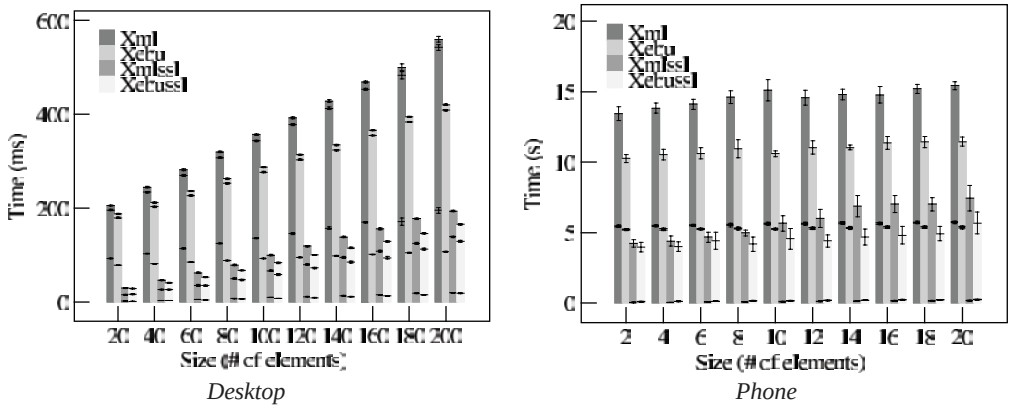
As with the size measurements, we also take a more detailed look into the security-enabled messages. Figure 6 shows the times taken processing encrypted data and message digests, excluding the expensive RSA private-key operations. The output bar shows symmetric key generation, key encryption, data encryption, and digest computation. The input bar shows data decryption, digest verification, and signature verification. Both of these sequences are from bottom to top.

For large messages, the dominant output component is data encryption because its processing requirements grow with message size. For the smaller messages of the phone experiment, the constant-time RSA public-key operation of symmetric key encryption dominates. For input processing, the most expensive one is again data decryption with the RSA public-key operation of signature verification dominating for smaller sizes.

There is much variation in many of the timings on the phone, which is due to the coarse granularity of the phone's internal clock. The actual measurement numbers show that the phone is capable of measuring time only in 15- or 16-ms increments. Since many of the operations only take a few such increments, we see both no variation and high variation, but very little low variation.

From the gathered data, we can estimate that for the same processing time as a single key encryption, we can encrypt symmetrically approximately 3.5 Kb of data on the desktop and 4.2 Kb on the phone. Similar numbers hold for the verification and data decryption pair of operations.

Figure 7. Total times taken in the experiment for secured XML and XML over SSL (divisions are client, remote, network, from bottom to top)



Desktop *Phone*

Finally, we show the time taken by RSA private-key operations, that is, signature computation and key decryption, in Table 5. These times do not depend on the message size at all. As we can see when comparing to Figure 6, these two operations consume several times the time spent on all other security-related processing. Extrapolating from Figure 6, we see that in the same time as one of these operations, we could potentially encrypt at least 250 Kb on the desktop.

Furthermore, we note that for the smallest messages on the phone, these operations consume up to 85% of the processing time on the client and almost 50% of the total invocation time. Therefore, it seems prudent to attempt to avoid repeated RSA private-key operations. However, there is no reason to avoid RSA public-key operations: As we saw earlier, the requirements of these are quite comparable to other processing.

Comparison with SSL

While SSL is not appropriate for all use cases, it is a widely deployed security solution and very suitable for many other cases. For this reason, we also ran our measurements using HTTP over SSL without XML-level security. This was done by replacing the HTTP URLs in Java's standard connection opening with HTTPS URLs. The algorithms used in SSL were forced to be the same as those with XML security, namely, 3DES, 1,024-bit RSA, and SHA-1.

We show the results of comparing SSL with our security-enabled formats in Figure 7 in the same way as the comparison with completely unsecured formats in Figure 5. To reiterate, each bar shows the time taken for local processing on the client, the time taken for processing at the gateway and the server, and the time taken for communication from bottom to top in this order.

Comparing to the regular HTTP case on the phone, we note that SSL adds approximately 2 s of processing to the invocation time. A network packet dump reveals this to be the SSL handshake cost, which takes two network round trips. Otherwise, the measurements remain the same. However, we also note that the first SSL handshake takes 6 s due to the key exchange; on later invocations, the session is reused, so the only overhead consists of the two network round trips of the abbreviated SSL handshake.

We noted when discussing the online retailer scenario that server authentication may be better handled at the transport level, so using both XML security and SSL is likely. In this case, we would expect the overhead of SSL to be essentially the same as with unsecured XML. The only addition would be for the larger messages, which require more processing when encrypted in SSL, but this is all symmetric cryptography, and as our timing measurements show, symmetric encryption is sufficiently low in processing costs that the overhead probably would not change significantly.

One thing to note is that the SSL experiment was somewhat problematic on the phone. An observation that we made concerned the unreliability of the connection: How many invocations are received by the gateway and not returned correctly to the client? For the regular HTTP case, this rate remained at a few tenths of a percent of the number of total invocations, but with SSL, we observed a rate of 6 to 7% of such dropped connections. The most probable cause is a too-stringent network time-out somewhere, but based on our past experience with phones, we cannot rule out the possibility of an unreliable SSL implementation on the mobile phone.

Battery Consumption Measurements

Our final experiment was the repeated invocation of a 10-element message from the phone until its battery ran out. The measurement in this case was the number of invocations performed at the server. We devised a method to use this number of invocations to determine the ratio of energy consumption between communication and com-

putation. To eliminate measurement errors as much as possible, we used only the WS-Security-based XML and Xebu formats to get measurements that were heavy on both communication and computation.

In addition to the Nokia 7610 that we used for the other experiments, we also repeated this experiment with a Nokia 9500 communicator. The 9500 is a much higher end device than the 7610: It has a slightly faster processor, more memory, a more capacious battery, and a more sophisticated Java virtual machine. Therefore, we would expect it to provide an interesting comparison. We used the GPRS network on the 9500 as well.

As a first step in computing the energy consumption ratio, let m be the amount of energy required to communicate 1 byte of data and let p be the amount of energy required to run 1 ms on the processor. If E is the total amount of energy on the phone, we get the formula:

$$E = i \cdot (s \cdot m + t \cdot p), \quad (1)$$

where s is the amount of data per invocation, t is the time spent in local processing of one invocation, and i is the total number of invocations achieved with a full battery.

By taking two measurements for the two different formats—s_1 and s_2 for size, t_1 and t_2 for processing time, and i_1 and i_2 for the number of invocations—and inserting them into Equation 1, we get a pair of linear equations that we can solve for m and p to get

$$m = E \cdot \frac{i_1 \cdot t_1 - i_2 \cdot t_2}{i_1 \cdot i_2 \cdot (s_2 \cdot t_1 - s_1 \cdot t_2)} \quad (2)$$

and

$$p = E \cdot \frac{i_1 \cdot s_1 - i_2 \cdot s_2}{i_1 \cdot i_2 \cdot (t_2 \cdot s_1 - t_1 \cdot s_2)}. \quad (3)$$

As E is here unknown to us, we can only compute the ratio between m and p. In Equations 2 and 3, the denominator is the determinant of the equation pair's matrix, so it cancels; we therefore get

$$r = \frac{m}{p} = \frac{i_2 \cdot t_2 - i_1 \cdot t_1}{i_1 \cdot s_1 - i_2 \cdot s_2}. \quad (4)$$

The numbers to insert into this equation are given in Table 6. The number of invocations is given as an interval between the minimum and maximum of our measurements. The sizes were measured from a network packet dump at the gateway so they are larger in proportion to the ones in Table 3. The times are the total processing times (i.e., total time minus communication time) of single invocations.

With these numbers, we can calculate r values for both devices from Equation 4, taking several different (i_1, i_2) pairs. The values, in the order of minimum, 25th percentile, median, 75th percentile, and maximum, are

0.47, 1.06, 1.44, 2.86, and 28.76
for the 7610, and
0.15, 0.15, 0.17, 0.19, and 0.21
for the 9500.

The calculated numbers for r give the number of milliseconds that the processor can run with the amount of energy it takes to communicate 1 byte. The Nokia 7610 has a 123 MHz ARM processor, and the Nokia 9500 a 150 MHz ARM processor. Therefore, taking the median value for both, the 7610 can run the processor for 177,000 cycles and the 9500 for 26,000 cycles.

One conclusion that we can draw from this is that the ratio of energy consumption for communication and computation is highly dependent on the device, and for different devices, even the order of magnitude can be different. Therefore, it

is not possible to make general statements of how much computation the transmission of 1 byte is worth. However, we do note that the equivalent of 1 byte is in both cases best measured in milliseconds instead of seconds or microseconds, so we might assume that to be true for a larger class of devices as well.

We note that there is a large variation in the measured values for the 7610 and hardly any variation for the 9500. We believe this to be an effect of the network conditions. Since we used a real network, the daily variation will affect the results, but the measurement on the 9500 took long enough that these variations happen for all measurement runs, whereas for the 7610, the time when a measurement run was performed may have a significant effect.

We believe, however, that it is useful to run measurements in real conditions, especially in cell-based networking where there is contention for the base stations. A private base station would give results for transmission that could never be achieved in actual conditions. Furthermore, by running measurements at a variety of times, we can get a better estimate of the range of possible network behaviors.

The method that we selected for measuring energy consumption is not very sophisticated, nor does it always produce very accurate results. Furthermore, it is not possible to measure absolute energy consumption, but only relative numbers for different types of energy-consuming operations. Finally, a single run will take a long time: On the 7610, draining the battery took 3 to 4 hours, and on the 9500, it took 8 to 9 hours.

However, our measurement method is much simpler than the proper alternative and does not require any additional equipment. Furthermore, as is evident from the measurements on the 9500, it can produce good accuracy in some cases. Therefore, we believe that in cases where absolute values do not matter, this technique may be useful.

RELATED WORK

Existing work on binary formats for XML data (Sandoz, Triglia, & Pericas-Geertsen, 2004; Schneider, 2003) led W3C to begin standardization in its Efficient XML Interchange (EXI) Working Group (http://www.w3.org/XML/EXI). As the measurements above indicate, the most pressing concern in wireless communication is reduction in size. The potential future format, now in working-draft stage (W3C, 2007a), can often achieve a size reduction of at least 50% for small messages, and well over 90% when a good schema is available (W3C, 2007b).

Gateway architectures have been very popular for adding support for mobile devices (OMG, 2005; Perkins, 1996), and protocol conversions have also been used to improve the performance of, for example, TCP (Kojo, Raatikainen, Liljeberg, Kiiskinen, & Alanko, 1997) and CORBA (DOLMEN, 1997). However, these architectures only consider underlying layers and do not address either the needs of content-based routing or the requirement for end-to-end security. As we noted, security is considered by WAP with the WTLS protocol (WAP Forum, 2001b), which is essentially our trusted-gateway model for SSL.

The other security solutions that we mentioned, namely, IP security, SSL, and S/MIME, have been

Table 6. Numbers of invocations and per-message communicated sizes and processing times in the battery experiment

Format	Invocations	Size (B)	Time (ms)
7610			
Xml	[842,979]	17385	5591.40±21.87
Xebu	[1444,1666]	8703	5225.04±8.72
9500			
Xml	[3084,3181]	17385	4481.30±35.72
Xebu	[4079,4153]	8703	4228.39±8.76

the targets of prior performance measurements on handheld devices as well (Argyroudis, Verma, Tewari, & O'Mahony, 2004). The processing requirement results of Argyroudis et al. appear to be in line with ours, but due to the use of wireless LAN and no mention of network latencies, total communication times are not directly comparable. This work also includes energy consumption measurements performed similarly to ours, but it does not consider how consumed energy is split between computing and communication.

The measurements that we performed were intended to reflect a single message exchange. For a longer term exchange of messages, it is beneficial to establish a security context, such as IP's security association or SSL's secure tunnel (we saw partial effects of this with the SSL session reuse). At the Web service level, such an establishment method is defined by WS-SecureConversation (IM, 2005).

The use of SSL in wireless communication was evaluated by Gupta and Gupta (2001), especially in contrast to the WTLS solution of WAP. The conclusion is that SSL can be implemented efficiently enough to be usable on mobile devices. Our measurements on SSL processing times essentially agree with this analysis.

The SSL protocol has been subjected to extensive measurements on energy consumption (Potlapally, Ravi, Raghunathan, & Jha, 2006). The tests of Potlapally et al. were performed on a PDA using wireless LAN. One of the findings is that the noncryptographic parts of the protocol, that is, mostly data transmission, consume 40 to 45% of the total energy, and for small data transmission, the cryptographic part is dominated by the asymmetric algorithms.

Comparing these figures on SSL to our results, we note that our messages were small enough that the asymmetric algorithms dominate. Estimating from our computed energy consumption ratios, we see that on the 7610, nearly 80% of the energy is spent on data transmission with XML and nearly 70% with Xebu. The 9500 is closer to a PDA in functionality than the 7610, and with it the amount spent on data transmission is 40% with XML and 25% with Xebu, which are closer to the values measured by Potlapally et al. (2006).

We note that existing detailed measurements of XML signatures (Shirasuna, Slominski, Fang, & Gannon, 2004) indicate that most of the time spent on signature processing is actually spent on canonicalization. As our experimental system was designed so that it wrote and read everything directly in canonical form, this effect is not visible in our timing measurements, but we could observe some of it on the server while running the experiments.

Shirasuna et al. (2004) also compare XML-based security solutions with SSL and note that SSL should be used if message-level security is not needed, a conclusion that agrees with ours. Also notable is the evaluation of WS-SecureConversation; the authors note that it provides a two-fold improvement for repeated messages and only a small overhead for single messages. This indicates that WS-SecureConversation is definitely a technology that is worth keeping in mind, but in the wireless communication context, the added overhead in message size may prove to be prohibitive.

An architecture for secure Web-services-based communication for pervasive computing is defined by Helander and Xiong (2005). This system uses Web services security for its security needs, and the authors' conclusion is that security interoperability is possible even with low-cost devices. Unlike the off-the-shelf components used in our measurements, this system's Web service implementation is a special-purpose one, written especially for the embedded devices they are targeting. Energy consumption is briefly considered and requirements for processing calculated, but there is no breakdown of costs.

RECOMMENDATIONS

The measurements shown in this article lead us to the bottlenecks in a secure Web services system for mobile devices. Based on these bottlenecks, we can note some recommendations on avoiding them. Some of these recommendations are usable even today with the existing standards, but others require modifications to existing practice.

First of all, we note that in our measurements, most of the processing time went to RSA private-key operations. In comparison, the time taken by RSA public-key operations is on par with the symmetric encryption of even a modest-sized message. The reason for this is that the public exponent of RSA is usually selected to be small (we used the common choice of $65537 = 2^{16}+1$), so the modular exponentiation does not take much time. In contrast, the private exponent, being the inverse of the public exponent, usually is of the same size as the modulus itself.

This property of RSA, of one operation being much faster than its opposite, is in contrast with other common asymmetric algorithms, DSA and Elgamal. In many cases, for example, in certificate verification, only RSA public-key operations are needed so we conclude that RSA is often the best algorithm to use despite its somewhat slower operation in its slower direction than the other algorithms.

Next, we move to extensions of existing standards. In light of our energy consumption measurements, we note that reducing the amount of data transmitted over the network can be worth a significant amount of computation. Therefore, it does not seem that in this specific case the purported processing efficiency gains of a binary format should matter much compared to compressed XML, especially if the latter produces smaller messages.

As an example, our previous measurements on the effectiveness of binary XML (Kangasharju et al., 2005) indicate that gzip on top of XML gives, for 3-Kb messages, a 75% reduction in size. These measurements were made on a desktop computer, so the additional time consumed cannot be directly translated, but based on the characteristics of the devices and other measurements, we estimate the additional processing to be between 10 and 100 ms on the 7610. In light of the ratios we computed for energy consumption, this is a massive benefit.

However, existing compression solutions are applied at the protocol level to the complete message. This means that compression is applied to the Base64-encoded form of the encrypted bytes, and should not be able to compress more than by the Base64 overhead of one fourth. This problem is naturally not specific to XML, and the common recommendation is to always compress messages before encryption (Schneier, 1990).

At the moment, XML encryption does not offer a method to encrypt compressed XML and have it be recognized as such by the receiver. This can be worked around by a simple extension to the EncryptedData element. Currently, it indicates with its Type attribute what kind of XML content has been encrypted. In our opinion, the simplest method to extend this to recognize compressed XML would be to add another attribute, perhaps ContentEncoding, that would work similarly to the Content-Encoding header of HTTP (Fielding et al., 1999), but would indicate how the XML fragment indicated by the Type attribute was encoded (compressed) before encryption. Our recent results (Kangasharju, 2007) demonstrate that large gains are possible with this simple extension.

Our final recommendation is a modification of the protocol itself and is specifically designed for the situation where the processing capabilities of the client and the server differ greatly. We assume that RSA is used per our earlier recommendation. In this case, there are two operations that require the expensive private-key use: signing a computed digest and decrypting a symmetric key received from the server.

To remove processing from signing, we suggest replacing the signature with HMAC (Bellare, Canetti, & Krawczyk, 1996). This requires a shared

secret between the client and the server, which can be established by the client with a securely generated random number encrypted for the server and signed by the client in the first message.

For decrypting the symmetric key, it is possible to use a similar method, that is, the client including an encrypted and signed random number in the message. Then, the server will use this number as the encryption key for the response message, and, as usual, include it encrypted with the client's public key. Now the client can simply encrypt the proposed key with its own public key and compare the encrypted keys for a match, thus replacing a private-key operation with a public-key one, which is much faster.

These suggestions have the drawbacks that the keys used in the operations will still need to be generated and signed by the client, and that the amount of traffic over the network is increased by the keys that are sent. For the former issue, we note that these keys can be generated beforehand when energy consumption is not an issue and stored securely on the client. Furthermore, the same keys can be used with the same server (but note that using the same key for two different servers is insecure).

For the latter, we note from our measurements that eliminating a single public-key operation saves 2.2 s on the 7610, which is equivalent to 1.5 Kb of data. However, on the 9500, the savings are 1.8 s, which is equivalent to only 300 bytes of data. In both cases, this is more than the 128-byte result of RSA, so our scheme is still more energy efficient than the standard. The results from the 9500 indicate that this is not necessarily true for all devices, but on the other hand, the large savings on the 7610 are definitely worth considering.

CONCLUSION

Summarizing our findings, we make the following main conclusions and recommendations:

- XML-level security is still a heavyweight operation and should only be used if the required security semantics demand it.
- SSL overhead is sufficiently small to make it fully usable in the wireless world.
- The compression of XML messages is vital for mobile devices, and specifications like XML encryption should be extended to integrate compression better.
- A characteristic of mobile Web services is a clear asymmetry in processing capabilities between clients and servers, and protocols may need to take this into account.

Based on our experience in the area, we believe that the adoption of an alternate serialization format for XML in the wireless world is very likely in the near future, especially if the W3C EXI effort makes progress. Since security is vitally important in the modern networked world, it must not be compromised or lessened. In our view, the thin-gateway model is a valid method for the initial inclusion of XML-based security in the case where an alternate format is adopted. Independently of that, however, we see the processing requirements of cryptography and energy requirements of large messages to be the major issues in this field.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for helpful comments. This research was performed in the Fuego Core research project, funded by the National Technology Agency of Finland, Nokia, and TeliaSonera.

REFERENCES

Argyroudis, P. G., Verma, R., Tewari, H., & O'Mahony, D. (2004). Performance analysis of

cryptographic protocols on handheld devices. In *Third IEEE International Symposium on Network Computing and Applications* (pp. 169-174).

Bellare, M., Canetti, R., & Krawczyk, H. (1996). Keying hash functions for message authentication. In *Advances in Cryptology: CRYPTO 1996* (LNCS 1109, pp. 1-15). Santa Barbara, CA: Springer-Verlag.

Chang, T.-K., & Hwang, G.-H. (2004). Using the extension function of XSLT and DSL to secure XML documents. In *18th International Conference on Advanced Information Networking and Applications* (pp. 556-561).

DOLMEN. (1997). *Bridging and wireless access for terminal mobility in CORBA* (Rep. No. LK-OMG01). Paper presented at the DOLMEN Consortium.

Fielding, R., Gettys, J., Mogul, J., Nielsen, H. F., Masinter, L., Leach, P., et al. (1999). *RFC 2616: Hypertext transfer protocol: HTTP/1.1.* Internet Engineering Task Force.

Freier, A. O., Karlton, P., & Kocher, P. C. (1996). *The SSL protocol version 3.0.* Netscape Communications.

Gupta, V., & Gupta, S. (2001). Securing the wireless Internet. *IEEE Communications Magazine, 39*(12), 68-74.

Helander, J., & Xiong, Y. (2005). Secure Web services for low-cost devices. In *Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing* (pp. 130-139).

IM. (2005). *Web services secure conversation language (WS-SecureConversation).* IBM, Microsoft, et al.

Internet Engineering Task Force (IETF). (2004). *RFC 3851: Secure/multipurpose Internet mail extensions (S/MIME) version 3.1 message specification.* Author.

Kangasharju, J. (2007). Efficient implementation of XML security for mobile devices. In *IEEE International Conference on Web Services* (pp. 134-141). Salt Lake City, UT: Institute of Electrical and Electronic Engineers.

Kangasharju, J., Tarkoma, S., & Lindholm, T. (2005). Xebu: A binary format with schema-based optimizations for XML data. In A. H. H. Ngu, M. Kitsuregawa, E. Neuhold, J.-Y. Chung, & Q. Z. Sheng (Eds.), *Sixth International Conference on Web Information Systems Engineering* (LNCS 3806, pp. 528-535). New York: Springer-Verlag.

Kent, S., & Atkinson, R. (1998). *RFC 2401: Security architecture for the Internet protocol.* Internet Engineering Task Force.

Kojo, M., Raatikainen, K., Liljeberg, M., Kiiskinen, J., & Alanko, T. (1997). An efficient transport service for slow wireless telephone links. *IEEE Journal on Selected Areas in Communication, 15*(7), 1337-1348.

Organization for the Advancement of Structured Information Standards (OASIS). (2004). *Web services security: SOAP message security 1.0.* Billerica, MA: Author.

Object Management Group (OMG). (2005). *Wireless access and terminal mobility in CORBA, version 1.2.* Needham, MA: Author.

Perkins, C. (1996). *RFC 2002: IP mobility support.* Internet Engineering Task Force.

Potlapally, N. R., Ravi, S., Raghunathan, A., & Jha, N. K. (2006). A study of the energy consumption characteristics of cryptographic algorithms and security protocols. *IEEE Transactions on Mobile Computing, 5*(2), 128-143.

Sandoz, P., Triglia, A., & Pericas-Geertsen, S. (2004). Fast infoset. *Sun Developer Network.*

Satyanarayanan, M. (2001). Pervasive computing: Vision and challenges. *IEEE Personal Communications, 8*(4), 10-17.

Schneider, J. (2003). Theory, benefits and require-

ments for efficient encoding of XML documents. In *W3C Workshop on Binary Interchange of XML Information Item Sets.* World Wide Web Consortium.

Schneier, B. (1990). *Applied cryptography* (2nd ed.). New York: John Wiley & Sons.

Shirasuna, S., Slominski, A., Fang, L., & Gannon, D. (2004). Performance comparison of security mechanisms for grid services. In R. Buyya (Ed.), *Fifth IEEE/ACM International Workshop on Grid Computing* (pp. 360-364).

WAP Forum. (2001a). *Wireless application protocol: Architecture specification.*

WAP Forum. (2001b). *Wireless transport layer security specification.*

World Wide Web Consortium (W3C). (1999). *WAP binary XML content format.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2001). *Canonical XML version 1.0.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2002a). *XML Encryption Syntax and Processing.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2002b). *XML signature syntax and processing.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2003a). *SOAP version 1.2 part 1: Messaging framework*. Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2003b). *SOAP version 1.2 part 2: Adjuncts.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2005). *XML-binary optimized packaging.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2007a). *Efficient XML interchange (EXI) format 1.0.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2007b). *Efficient XML interchange measurements note.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2007c). *XQuery 1.0: An XML query language.* Cambridge, MA: Author.

World Wide Web Consortium (W3C). (2007d). *XSL transformations (XSLT) version 2.0.* Cambridge, MA: Author.

Weiser, M. (1993). Some computer science issues in ubiquitous computing. *Communications of the ACM, 36*(7), 75-84.

This work was previously published in International Journal of Web Services Research, Vol. 5, Issue 3, edited by L. Zhang, pp. 1-19, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.6
Security Issues Concerning Mobile Commerce

Samuel Pierre
École Polytechnique de Montréal, Canada

INTRODUCTION

Electronic commerce or e-commerce can be briefly defined as a financial transaction or commercial information between two parties based on data transmitted over communication networks (Soriano & Ponce, 2002). It relies upon users' interventions to initiate a transaction and select the main steps of the process. Users' actions stem from a succession of virtual decisions. Indeed, when shopping with a virtual catalog, customers can select products which meet their needs, tastes, and respect their price range. Such decisions consistently require the users' input, thus costing them both time and money. These costs are even more exorbitant when a search is launched for an order that includes a variety of products from different sources which have different characteristics (price range, delivery dates, etc.). When transactions involve users who are moving or take place over mobile networks, this is referred to as *mobile electronic commerce*, a specific type of e-commerce.

Mobile electronic commerce (or m-commerce) refers to an ability to carry out wireless commercial transactions using mobile applications within mobile devices, such as mobile phones and personal digital assistants (PDAs). It is generally defined as the set of transactions or processes which can be carried out over a wireless mobile network. According to this definition, m-commerce constitutes a subset of all electronic commercial transactions (electronic commerce or e-commerce) from business to consumer (B2C) or business to business (B2B). Thus, short personal messages such as those from SMS (short messaging system) sent between two individuals do not fall into the category of m-commerce, whereas messages from a service provider to a salesperson or a consumer, or vice versa, do fit this very definition. M-commerce appears as an emerging manifestation of Internet electronic

commerce which meshes together concepts such as the Internet, mobile computing, and wireless telecommunications in order to provide an array of sophisticated services (m-services) to mobile users (Varshney, Vetter, & Kalakota, 2000; Veijalainen, Terziyan, & Tirri, 2003).

E-commerce includes an initial step where consumers search for a product they wish to purchase by virtually visiting several merchants. Once the product is found, negotiations can take place between the customer and the merchant (electronic negotiation or e-negotiation) (Paurobally, Turner, & Jennings, 2003). If an agreement is reached, the next step is the payment phase. At each step of the process, some problems arise, such as transaction security, confidence in the payment protocol, bandwidth limitations, quality of service, shipping delays, and so forth (Younas, Chao, & Anane, 2003; Zhang, Yuan, & Archer, 2002). The peak withdrawal periods have always presented a major challenge for certain types of distributed applications. The advent of m-commerce further highlights this problem. Indeed, in spite of rather optimistic predictions, m-commerce is plagued by several handicaps which hinder its commercial development, security being the main one.

Many market research studies, like those carried out by Strategy Analytics and the Gartner Group, predicted that by 2004 there would be over one billion wireless device users, some 600 million wireless Internet subscribers, a $200 billion m-commerce market, and 40% of consumer-to-business e-commerce will take place over Web-enabled phones (Gosh & Swaminatha, 2004). However, these business opportunities could be compromised by new security risks specific to the wireless medium and devices. As a result, the potential boom in the number of new m-commerce applications and markets can be achieved if and only if security and privacy can be integrated into online m-commerce applications.

This article analyzes some major security issues concerning mobile commerce. The next section presents background and related work, followed by a summary of some security issues and challenges. Future and emerging trends in secure m-commerce are then outlined, and the article is concluded.

BACKGROUND

While e-commerce systems are designed for purchases conducted on the wired Internet, m-commerce is extended to handle the mobility aspects related to the user equipment such as a mobile phone or a PDA. One of the main characteristics of an m-commerce system is the use of the Internet as the backbone and e-commerce with mobile terminals as user equipment. M-commerce applications can be as simple as a system to synchronize an address book or as complex as the system used to enable credit card transactions. They are deployed using mobile middleware which can be defined as a functional layer of software provided by application developers to link their e-commerce applications to an operating system and various mobile networks to allow their applications to bypass certain mobility issues.

Any party engaging in business needs a certain level of security. Security relies on a set of basic concepts and requirements such as: confidentiality, authentication, integrity, non-repudiation, and authorization. Confidentiality assures that the exchange of messages between parties over wireless access networks or global networks is not being monitored by non-authorized parties. Authentication ensures that the parties engaging in business are who they claim to be. Integrity allows users to verify whether modifications have occurred; however, it does not guarantee that information has not been altered. Non-repudiation certifies that the business transactions the parties engage in are legally binding. Authorization refers to a set of access rights assigned to an entity by a certification authority (CA). It does not guarantee that messages received do really come from a given counterpart; that is the task

of authentication.

In a wired network, the secure socket layer (SSL) protocol and the transport layer security (TLS) protocol, which are well-established security protocols, provide privacy and data integrity between two communicating applications. In fact, HTTP over TLS-SSL is used to secure transactions for security-sensitive applications like m-commerce. It is generally known that these protocols do not adapt well to wireless environments with reduced processing capability and low-bandwidth links. Indeed, wireless devices such as cellular phones and PDAs have limited storage and minimal computational capacity. As a result, security issues were not taken into account when they were designed.

The scheme devised during the wireless application protocol (WAP) forum, which has defined an entirely new suite of protocols, uses a WAP gateway or proxy between the wireless and wireline environments to ensure connection and security. The SSL and TLS ensure security within the Internet, while the wireless transport layer security (WTLS) protocol ensures secure channels between the client and the WAP gateway. Transactions between WTLS and TLS are executed by the WAP gateway. However, the use of the WAP proxy, which is also a point of failure, does not allow for end-to-end security. As a matter of fact, because there are storage and translation operations at the WAP proxy, it becomes a point of entry for attacks. A solution to strengthen this weakness was provided by Soriano and Ponce (2002). They suggested providing a secure end-to-end tunnel between an Internet server and a mobile user by implementing a TLS compatible security layer at the wireless application environment (WAE) layer on the client side, named WAE-Sec. WAE-Sec therefore prohibits translations by the WAP gateway and permits compatibility with the TLS protocol. Note, however, that this solution resembles the one proposed by Gupta et al. (2001).

On the other hand, Tang, Terziyan, and Veijalainen (2003) have defined other related security issues to m-commerce, namely hostility, information security, and vulnerability. Hostility means that dishonest customers who get fraudulent identities by stealing mobile devices can make illegal operations and, thus, should be quickly identifiable. Information is more vulnerable in wireless networks since other parties can easily intercept it. The solution is to encrypt data with adequate keys. Vulnerability arises from a malfunctioning of the mobile device itself or from the physical access of malicious persons to the terminals. To remedy these additional problems, Tang et al. (2003) suggested the use of a mixed personal identification number (PIN) storage scheme which let the PIN be partially stored on the mobile device while the remainder of the PIN is stored on the network. Researchers assume that the probability of discovering the PIN located at two different places does not depend on the length of the PIN nor on the fact that a single part was discovered. Thus, discovering the whole PIN will require digging and/or guessing for twice as long than if the PIN was located at a single place. The improvements brought about by this strategy have been shown using a probabilistic model, but its implementation has yet to be investigated.

A new protocol for m-commerce was proposed by Katsaros and Honary (2003). Fully applicable to third-generation mobile networks, this protocol is characterized by three novel properties, as opposed to the existing methods of m-commerce. In fact, it provides a simplified and secure transaction method, minimizes the number of entities involved in the transaction, and finally reduces the probability of security threats, thus reducing the risk of fraud. Unfortunately, this protocol does not solve certain security issues related to m-commerce.

SECURITY ISSUES AND CHALLENGES

Mobile commerce provides an exciting new set of capabilities which can lead to new services that enhance the end-user's experience. With these new business opportunities, the risk of new security threats also arises. New mobile devices such as PDAs and mobile phones enable easy access to the Internet and strongly contribute to the development of m-commerce services, while Smartcard platforms will enable operators and service providers to design and deploy new m-commerce services. Such technologies must guarantee a high level of security of customer information and transactions in order to be adopted and widely deployed. Thus, establishing security mechanisms which allow diverse mobile devices to support a secure m-commerce environment on a wireless Internet is a critical challenge.

There are a lot of other security issues and challenges related to m-commerce: security of the transactions, security of the payments, security of customer information, end-to-end security, authorization mechanisms, and so on. Providing security provisions for the m-commerce community is challenging due to the insecure air interface of wireless access networks, and limited computational capability of mobile devices and users' mobility (He & Zhang, 2003). The limited equipment resources require the e-payment protocol in the wireless Internet environment to be designed in consideration of the efficiency of the computing functions and the storage device. In this context, security issues, like those dealing with service and subscriber authorizations in enhanced prepaid implementations for m-commerce, must be addressed. In fact, client application and subscriber-level authentication and authorization are key mechanisms used to regulate access to and usage of content-based transactions in m-commerce. The objective is to provide an enriched rating engine and a highly configurable feature set for service and content charging on wireless networks (Cai et al., 2004).

Smartcard platforms will enable operators and service providers to design and deploy new m-commerce services. This development can only be achieved if a high level of security is guaranteed for the transactions and customer information (Renaudin et al., 2004). In this context, smartcard design is very challenging when it comes to providing the flexibility and the power required by the applications and services, while at the same time, guaranteeing the security of the transactions and the customer's privacy.

On the other hand, as the number of users of wired and wireless Internet services is increasing exponentially and m-commerce services are going to be activated, it is quite necessary to establish a wireless Internet public key infrastructure (PKI) service which accepts diverse mobile devices to support secure m-commerce environments on the wireless Internet. In this context, security/payment policy algorithms must be designed in order to dynamically adapt the level of security according to the domain-dependent properties and the independent properties to support secure m-commerce transactions and payment on wireless Internet (Kim et al., 2002).

An e-payment system for m-commerce uses existing wired systems as is. However, it implies certain security and inefficiency problems. In fact, the limited amount of equipment required by the e-payment protocol in the wireless Internet environment allows for the highest level of the efficiency pertaining to the computing function and the storage device. The issue is the basis of the design of an e-payment system for m-commerce that minimizes public key computing and guarantees anonymity concerning personal and purchasing information, as well as spatial storage efficiency (Kim, Kim, & Chung, 2003).

Another issue in m-commerce security concerns the increasing number of destructive messages with viruses that can harm mobile devices. Such an issue is truly critical in the context of

mobile applications which are generally deployed by small mobile devices with limited processing and storage capabilities.

Mobile commerce involves many risks related to security and privacy (Ghosh & Swaminatha, 2004). In fact, wireless devices introduce new security threats which are specific to their mobility and communication medium. Most Web sites are not currently configured to deal with the intermittent service failures which frequently occur during wireless connections. Furthermore, the most popular implementations of the WTLS protocol do not re-authenticate principles or double-check certificates once a connection has been established. As a result, attackers can take advantage of this vulnerability and compromise the integrity of the wireless networks which support the m-commerce applications.

The most significant security and privacy risks for wireless devices involved in m-commerce applications are: platform risks, software application risks, security risks of WML Script, among others (Ghosh & Swaminatha, 2004). Platform risks are related to the fact that many manufacturers have failed to include some basic operating system features necessary to enable some kinds of secure computing: memory protection for processes, protected kernel rings, file access control, authentication of principals to resources, biometric authentication, and so forth. Without a secure infrastructure provided by the platform and used by the device running m-commerce applications, it is difficult to achieve secure m-commerce.

Software application risks are related to the capability to design and develop secure wireless applications using good software engineering and assurance methods. One of the most important issues in this context is the ability to develop software for sending and executing mobile codes and agents to wireless devices, by taking into account the need to reduce the communication load on extremely bandwidth-limited wireless links.

Security risks of WML (wireless markup language) are related to the lack of access control for WML scripts, meaning that the type of attacks that can be launched using WML script is limited only by the imagination of malicious script writers. More generally, such risks are based on a fundamental lack of a model for secure computation (Ghosh & Swaminatha, 2004).

FUTURE TRENDS

Despite the differences between wired and wireless networks, both networks are vulnerable to the same kinds of attacks. Nevertheless, wireless networks, as the core infrastructure which supports m-commerce, are basically more exposed to security attacks due to the type of communication channel used.

Future trends in this field consist of considering hardware, software, and data as elements to be protected against security attacks in mobile environments. In particular, m-commerce deals with payments over the Internet, electronically sending both services and information, storage of consumer information on resources available from the Internet, as well as all other issues related to online shopping. Much research tackles security problems related to the overall process of m-commerce. Some of these problems will likely be solved in the near future on some levels by modifications to existing protocols. In particular, problems related to the wireless application protocol (WAP) are considered highly critical. In this context, data encryption over communication channels constitutes the strongest perceived security issue in the system.

Finally, other research directions address the setup of a trustworthy relationship with customers in order to deliver the service in due time. This also includes security issues related to the lack of anonymity and the possibility for an attacker to gain access to the users' account number and

their identities, particularly in the context of payments with credit cards.

CONCLUSION

This article analyzed some major security issues in mobile commerce. After a presentation of background, some security issues and challenges, then future and emerging trends in secure m-commerce were outlined. A set of privacy risks were also mentioned and their relationships to software development were outlined. In fact, the nature of the communication medium requires a degree of trust and cooperation between nodes in wireless networks. There is a certain risk that trust and cooperation are exploited by malicious entities to collect confidential information and disseminate false information. Other risks are related to the platform, the software application, and the WML scripts. The most significant risk to m-commerce systems is related to a malicious code which has the ability to undermine other security technologies as it resides on the device, thus having all of the owner's privileges. For all these reasons, encrypted communication protocols are necessary to provide confidentiality, authentication, integrity, non-repudiation, and authorization of services for m-commerce applications.

REFERENCES

Cai, Y., Kozik, J., Raether, H. L., Reid, J. B., Starner, G. H., Thadani, S., & Vemuri, K. V. (2004). Authorization mechanisms for mobile commerce implementations in enhanced prepaid solutions. *Bell Labs Technical Journal, 8*(4), 121-131.

Ghosh, A. K., & Swaminatha, T. M. (2004). Software security and privacy risks in mobile e-commerce. *Communications of the ACM, 44*(2), 51-57.

Gupta, V., & Gupta, S. (2003). Securing the wireless Internet. *IEEE Communications Magazine, 39*(12), 68-74.

He, L. S., & Zhang, N. (2003). An asymmetric authentication protocol for m-commerce applications. *Proceedings of the 8th IEEE Symposium on Computers and Communications* (vol. 1, pp. 244-250).

Katsaros, I., & Honary, B. (2003, June 25-27). Novel m-commerce security protocol for third generation mobile networks. *Proceedings of the 4th International Conference on 3G Mobile Communication Technologies,* London (3G 2003) (pp. 23-27).

Kim, M., Kim, H., & Chung, M. (2003). Design of a secure e/m-commerce application which integrates wired and wireless environments. *Proceedings of the 3rd IASTED International Conference on Wireless and Optical Communications* (pp. 259-264).

Kim, M. A., Lee, H. K., Kim, S. W., Lee, W. H., & Kang, E. K. (2002, June 29-July 1). Implementation of anonymity-based e-payment system for m-commerce. *Proceedings of the IEEE 2002 International Conference on Communications, Circuits and Systems* (vol. 1, pp. 363-366).

Paurobally, S., Turner, P. J., & Jennings, N. R. (2003, November). Automating negotiation for m-services. *IEEE Transactions on Systems, Man and Cybernetics, Part A, 33*(6), 709-724.

Renaudin, M., Bouesse, F., Proust, P., Tual, J. P., Sourgen, L., & Germain, F. (2004, February 16-20). High security smartcards. *Proceedings of the Europe Conference on Design, Automation and Test* (vol. 1, pp. 228-232).

Soriano, M., & Ponce, D. (2002, August). A security and usability proposal for mobile electronic commerce. *IEEE Communications Magazine, 40*(8), 62-67.

Tang J., Terziyan V., & Veijalainen J. (2003, April). Distributed PIN verification scheme for improv-

ing security of mobile devices. *Journal of Mobile Networks and Applications, 8*(2), 159-175.

Varshney, U., Vetter, R. J., & Kalakota, R. (2000, October). Mobile commerce: A new frontier. *Computer, 33*(10), 32-38.

Veijalainen, J., Terziyan, V., & Tirri, H. (2003, January 6-9). Transaction management for m-commerce at a mobile terminal. *Proceedings of the 36th Hawaii International Conference on Systems Sciences*, Big Island, HI (p. 10).

Younas, M., Chao, K. M., & Anane, R. (2003). M-commerce transaction management with multi-agent support. *Proceedings of 17th International Conference on Advanced Information Networking and Applications* (AINA 2003) (pp. 284-287).

Zhang, J. J., Yuan, Y., & Archer, N. (2002). Driving forces for m-commerce success. *Journal of Internet Commerce, 1*(3), 81-105.

KEY TERMS

Authentication: Technique by which a process verifies that its communication partner is who it is supposed to be and is not an imposter. It makes sure that the parties engaging in business are who they claim to be. Integrity allows the system to verify whether modifications have occurred; it does not ensure that information was not altered.

Authorization: One or many access rights assigned to an entity by a certification authority (CA). Authorization does not make sure that messages received really do come from a given counterpart.

Confidentiality: Assures that the exchange of messages between parties over wireless access networks or global networks is not being monitored by non- authorized parties.

Electronic Commerce (E-Commerce): Set of transactions or processes which can be carried out between two parties based on data transmitted over communication networks. E-commerce relies upon users' interventions to initiate a transaction and select the main steps of the process.

Integrity: Allows the system to verify whether modifications have occurred; it does not make sure that information was not altered.

Mobile Commerce (M-Commerce): Refers to an ability to carry out wireless commercial transactions using mobile applications within mobile devices, such as mobile phones and personal digital assistants (PDAs). It is generally defined as the set of transactions or processes which can be carried out over a wireless mobile network.

Mobile Middleware: The functional layer of software provided by application developers to link their e-commerce applications to an OS and various mobile networks to allow their applications to bypass certain mobility issues.

Non-Repudiation: Makes sure that the business transactions the parties engaged in are legally binding.

Public Key Infrastructure (PKI): Security mechanism based on public key cryptography used to provide end-to-end security required for the information, services, and means of access. The core component of a PKI is the certification authority (CA). This authority is trusted by the end entities in its administrative domain and is responsible for the status of the certificate it issues.

This work was previously published in Encyclopedia of Encyclopedia of E-Commerce, E-Government, and Mobile Commerce edited by M. Khosrow-Pour, pp. 968-972, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.7
Security Architectures of Mobile Computing

Kaj Grahn
Arcada Polytechnic, Finland

Göran Pulkkis
Arcada Polytechnic, Finland

Jonny Karlsson
Arcada Polytechnic, Finland

Dai Tran
Arcada Polytechnic, Finland

INTRODUCTION

Mobile Internet users expect the same network service quality as over a wire. Technologies, protocols, and standards supporting wired and wireless Internet are converging. Mobile devices are resource constrained due to size, power, and memory. The portability making these devices attractive also causes data exposure and network penetration risks.

Mobile devices can connect to many different wireless network types, such as cellular networks, personal area networks, wireless local area networks (WLANs), metropolitan area networks (MANs), and wide area networks (satellite-based WANs). Wireless network application examples are e-mailing, Web browsing, m-commerce, electronic payments, synchronization with a desktop computer, network monitoring/management, and reception of video/audio streams.

BACKGROUND

Major security threats for mobile computing devices are (Olzak, 2005):

- Theft/loss of the device and removable memory cards,
- Wireless connection vulnerabilities, and
- Malicious code.

Mobile computing devices are small, portable, and thus easily lost/stolen. Most mobile platforms only include support for simple software-based password login schemes. These schemes are easily bypassed by reading information from the device without login. Memory cards are also easily removed from the device.

Mobile devices support wireless network connections such as Bluetooth and WLAN. These connections are typically by default unprotected and thus exposed to eavesdropping, identity theft, and denial-of-service attacks.

Malware has constituted a growing threat for mobile devices since the first Symbian worm (Cabir) was detected in 2004. Mobile devices can be infected via MMS, Bluetooth, infrared, WLAN, downloading, and installing from the Web. Current malware is focused on Symbian OS and Windows-based devices. Malware may result in (Olzak, 2005):

- Loss of productivity,
- Exploitation of software vulnerabilities to gain access to resources and data,
- Destruction of information stored on a SIM (subscriber identity module) card, and
- Hi-jacking of airtime resulting in increased costs.

WIRELESS SECURITY PRINCIPLES

Security Policy

Examples of rules proposed for mobile device end users are:

- I agree to make sure my device is password protected and that latest security patches are installed.
- I agree to keep a firewall/anti-virus client with latest anti-virus signatures installed, and to use a remote access VPN client, if I will connect to the corporate network.
- I agree to use the security policies recommended by the corporate security team.

Examples of rules proposed for administrators of mobile devices in corporate use are:

- End-users get mobile network access after agreeing to the end-user rules of behavior.
- Handheld firewalls shall be configured to log security events and send alerts to *security-manager@company.com*.
- Handheld groups and Net groups shall have restricted access privileges and only to needed services.

Handheld security policies should be automated by restrictive configuration settings for handhelds, firewalls, VPNs, intrusion detection systems, and directory servers (Handheld Security, 2006).

Storage Protection

Mobile device storage protection is online integrity control of all stored program code and all data, optional confidentiality of stored user data, and protection against unauthorized tampering of stored content. Protection should include all removable storage modules used by the mobile device.

The integrity of the operating system code, the program code of installed applications, and system and user data can be verified by checksums, cyclic redundancy codes (CRCs), hashes, message authentication codes (MACs, HMACs), cryptographic signatures, and so forth. However, only hardware protection of verification keys needed by MACs, HMACs, and signatures provide strong protection against tampering attacks. Online integrity control of program and data files must be combined with online integrity control of the configuration of a mobile device for protection against malware intrusion attempts.

User data confidentiality can be granted by file encryption software. Such software also protects integrity of stored information, since successful decryption of an encrypted file is also an integrity proof.

Security Layers

Mobile computing security layers are based on the OSI (Open Systems Interconnection) Security Model. Defined security services are *authentication, access control, non-repudiation, data integrity, confidentiality, assurance/availability,* and *notarization/signature* (ISO/IEC 7498-1, 1994; ISO 7498-2, 1989).

Specific wireless security architecture issues include Mobile IP security features, and link-level and physical-level security protocols of wireless access technologies like WLAN, GPRS, and Bluetooth

Mobile IP security means that:

- A mobile node, which is a mobile device, has the same connectivity and security in a visited foreign network as in its home network; and
- The home network and visited foreign networks have protection against active/passive attacks.

These security goals require:

- That Mobile IP registration and location update messages have *data integrity protection, data origin authentication,* and *anti-replay protection*;
- *Access control* to foreign network resources used by visiting mobile nodes; and
- That IP packet redirecting tunnels provide *data integrity protection, data origin authentication,* and *data confidentiality.*

Moreover, mobile nodes should have *location privacy* and *anonymity* (Zao et al., 1999).

Replay prevention with timestamps or nonces for all mobile IP messages is specified in Perkins and Calhoun (2000). Other mobile IP security solutions are authentication schemes and protection of data communication (Calhoun et al., 2005; Barun & Danzeisen, 2001; Hwu, Chen, & Lin, 2006).

Identification Hardware

Identification hardware contains user information and cryptographic keys used to authenticate users to mobile devices, applications, networks, and network services.

The following identification hardware types are used:

- Subscriber identity module (SIM),
- Public key infrastructure SIM (PKI SIM),
- universal SIM (USIM), and
- IP multimedia services identity module (ISIM).

SIM

A basic SIM card is a smartcard securely storing a key (Ki) identifying a GSM network user. A SIM card is a microcomputer executing cryptographic operations with Ki. The SIM card also stores SMS (short message service) messages, MMS (multimedia messaging system) messages, and a phonebook. The use and content of a SIM card is PIN protected (Rankl & Effing, 2003).

PKI SIM

A PKI SIM card is a basic SIM card with added PKI functionality. An RSA co-processor is added for public key-based encryption and signing with private keys. The PKI SIM card stores private keys and certified public keys needed for digital signatures and encryption (Setec, 2006).

USIM

A USIM card is a SIM used in 3G mobile telephony networks. The physical size is the same as for a GSM SIM card, but hardware is different. USIM is actually an application running on a UICC (universal integrated circuit card) storing a pre-shared secret key (Lu, 2002).

ISIM

An ISIM card consists of an application (ISIM) residing on a UICC. ISIM provides secure authentication of handheld users to IMS (IP multimedia system) services (Dietze, 2005).

Wireless Security Protocols

Security protocols are—for wired networks—implemented by (Perelson & Botha, 2004): authentication services, confidentiality services, non-repudiation services, and authorization. Four wireless security protocol types are needed:

- Access control to mobile devices,
- Local access control to networks and network services,
- Remote access control to networks and network services, and
- Protection of data communication to/from mobile devices.

Different protocols are presented in Markovski and Gusev (2003).

Access Control to Mobile Devices

Access control must be implemented on a mobile device itself to prevent unwanted access to confidential data stored in the device (see Figure 1). Authentication confirms a claimed user identity.

PIN and Password Authentication

A PIN is four digits from a 10-digit (0-9) keypad. However, PINs are susceptible to shoulder surfing or to systematic trial-and-error attacks due to their limited length and alphabet. Passwords are more secure than PINs since their length and alphabet are larger (Jansen, 2003).

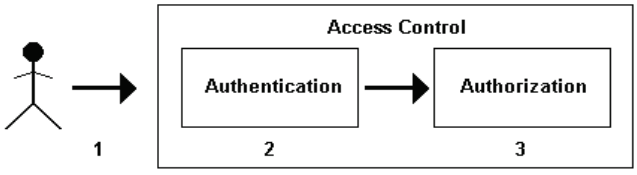
Visual and Graphical Login

Visual authentication means that a user must remember image sequences to authenticate to a mobile device. A picture password system can be designed to require a sequence of pictures or objects matching a certain criteria and not exactly the same pictures. For example, the user must find a certain number of objects with four sides. This makes the shoulder surfing quite difficult (Duncan, Akhtari, & Bradford, 2004).

Biometrics

Biometric user authentication is a hardware solution for examining one or more physical attributes of an authorized user. Biometric controls, such

Figure 1. The access control principle



1. The user presents an identity (e.g., password or biometric)
2. The user's identity is confirmed
3. An authenticated user is allowed access to a resource

as fingerprints, are becoming more common in handheld devices (Perelson & Botha, 2004).

Authorization

Usually mobile devices are personal, and authentication infers that the user is authorized. A corporate handheld device may however be used by several employees and may contain confidential company information. Needed user authorization features for such mobile devices include (Perelson & Botha, 2004):

- **File Masking:** Some files cannot be viewed by unauthorized users.
- **Access Control Lists:** User-related object permissions.
- **Role-Based Access Control:** User role-related permissions.

Local Network Access

Local network access protocols depend on the wireless access network type (WLAN, Bluetooth, Cellular Network, etc.). A WLAN is usually an access network to a LAN. Authentication for LAN resources is thus also needed unless WLAN authentication is integrated in a single-sign-on scheme. Local network access protocols are described in later sections.

Remote Network Access

Secure remote network access from a mobile device requires a VPN (virtual private network), which is a protected data path in an existing unsecured network to a private LAN. VPNs can be based on different protocols: IPSec (IP security), SSL (secure socket layer), or SSH (secure shell).

IPSec VPN

IPSec operates at the network layer of the OSI model. IPSec protocols are:

- ESP (encapsulating security payload) for authentication, data confidentiality, and message integrity;
- AH (authentication header) for authentication and message integrity; and
- IKE (Internet key exchange protocol) for encryption key exchange.

IPSec VPNs require VPN client software in mobile devices (Davis, 2001).

SSL VPN

The encrypted tunnel is established at the session layer of the OSI model. SSL VPN clients communicate with the VPN gateway using an SSL-supported application such as a Web browser or e-mail client. No separate VPN client software is therefore needed (Steinberg & Speed, 2005).

SSH

SSH (secure shell) is a protocol for login to and executing commands on a remote UNIX computer. SSH provides between two communicating hosts an encrypted communication channel, which can be used for port forwarding with VPN functionality (Barret et al., 2005).

Protection of Data Communication

Security protocols for protection of wireless data communication are integrated in protocols for local and remote access to networks/network services. In a cellular network a shared secret session key created by the authentication protocol is used for encryption/decryption of data communication. In a WLAN, the TKIP (temporal key integrity protocol) is integrated in the WPA security protocol, and AES (advanced encryption standard) is integrated in the WPA2 security protocol. The remote access protocols IPSec, SSL/TLS, and SSH also provide end-to-end protection of data communication with secure symmetric encryp-

tion algorithms and shared secret session keys created during authentication.

PLATFORMS FOR INTEGRATED ARCHITECTURES

Software signing and binary trust-models do not provide adequate protection against third-party programs. Fine-grained software authorization is emerging into mobile units. Typical examples include Java sandboxing and Symbian platform security. Software-based mobile platform examples are Java Mobile Environment, Symbian OS, Embedded Linux, Windows Mobile, Brew (Binary Runtime for Wireless), Blackberry OS, and Palm OS.

OS implementation vulnerabilities still remain a challenge. Integrated solutions have been proposed for executing trusted code and for secure boot. Standardization efforts are under development (e.g., Trusted Computing Group and Trusted Mobile Platform). There are different embedded on-chip security solutions, but mostly the security solution relies on combining hardware and software. Platform security examples are Texas Instruments OMAP™ Platform (Sundaresan, 2003) and Intel Wireless Trusted Platform (Intel Corporation, 2006b).

The TI platform relies on three layers of security: application layer security, operating system layer security, and on-chip hardware security. The main security features are:

- A *secure environment* provides secure execution of critical code and data by *secure mode*, *secure keys*, *secure ROM*, and *secure RAM*.
- *Secure boot/flash* prevents security attacks during device flashing/booting.
- *Run-time security* is included for security-critical tasks like encryption/decryption, authentication, and secure data management.
- A *hardware crypto engine* is also included for DES/3DES, SHA1/MD5, and RNG with two configuration modes: secure mode and user mode.

Intel platform building blocks are performance primitives (hardware) and cryptographic primitives (optimized software) for security services. Platform components include

- *Trusted boot ROM* integrity validation and booting to a correct configuration;
- *Wireless trusted module* processing secrets;
- *Security software stack* enabling access to platform resources through standard cryptographic APIs;
- *Protected storage* in system flash for secrets; and
- *Physical protection* by security hardware in a single device and discrete components in a single physical package.

WIRELESS APPLICATION SECURITY

The risks described above should be addressed in wireless application design. Wireless application security includes (Umar, 2004): application access control, client/server communications security, and anti-malware protection.

Application Access Control

Many mobile platforms lack support for individual user accounts and for operating system-level logon. Mobile applications handling confidential data should require user authentication before application access is granted. In case a mobile device is lost or stolen while the device user is logged in to an application, the application should also support "session timeout." This means that a limited inactive time is specified for an appli-

cation before re-authentication is required (Intel Corporation, 2006a).

Client/Server Communication Security

Typical wireless Internet connections are:

1. The wireless connection between a mobile device and an access device, and
2. The Internet connection between the mobile device and the Internet host/server via the access device.

Internet connection security should be provided at the application level.

For Web-based client/server applications, the SSL protocol provides encryption and signing of transmitted data. SSL application examples are:

- Web browsers for secure communications with Web servers,
- E-mail client software for secure reading of E-mail messages on e-mail servers, and
- SETs (secure electronic transactions) for secure financial transactions with credit cards.

For applications using customized protocols, security protocols are also customized. Alternatively, VPN techniques can be used.

Anti-Malware Protection

Most current mobile operating systems lack memory space protection. Malware can access and steal application data, such as credit card information stored in memory by wireless applications. Time and space for sensitive data in memory should be minimized (Intel Corporation, 2006a).

SECURITY OF MOBILE TECHNOLOGIES

A taxonomy of mobile technologies is:

- Wireless cellular networks (GSM, DECT, GPRS, and UMTS),
- Wireless long-range networks (WiMax, Satellite Communication Technology),
- Wireless local area networks (WLAN, ZigBee™), and
- Wireless short-range networks (Bluetooth, Wireless USB).

Wireless Cellular Networks

First Generation

First-generation cellular systems, such as AMPS (advanced mobile phone system) introduced in the early 1980s, use analog transmission and provide no security.

Second Generation

2G cellular systems, such as GSM (Global System for Mobile Communications) introduced in the late 1980s and DECT, use digital transmission.

GSM security is based on a unique IMSI (International Mobile Subscriber Identity) and a unique secret key (Ki) stored in the SIM card of each subscriber. The Ki is never transmitted over the network. Every GSM network has:

- *AUC (authentication center),* a protected database containing a copy of Ki;
- *HLR (home location register)* for subscriber information;
- *VLR (visitor location register)* for information of each mobile station currently located in the geographical area controlled by the *MSC (Mobile Station Controller)*; and

- *EIR (equipment identity register)* for lists of mobile stations on the network. Stations have unique IMEI (International Mobile Equipment Identity) numbers.

When a mobile station enters a GSM network for the first time, the IMEI is transmitted for determination in which AUC/HLR subscriber data is stored. The MSC/VLR of the visited network asks for and stores a security triplet (a unique random number RAND, a signed response SRES, a ciphering key Kc) from the AUC/HLR. SRES and Kc are calculated from RAND with Ki.

Subscriber authentication:

- RAND is sent to the mobile station.
- SRES' and Kc' are calculated from RAND with Ki.
- SRES' is sent back to MSC/VLR.
- Authenticated if SRES=SRES'.

Kc=Kc' is used for radio link encryption/decryption.

After the initial registration, IMSI is stored in the VLR. A TMSI (temporary mobile subscriber identity) is generated, transmitted back to the mobile station, stored in the SIM card, and used for future subscriber identification in the visited network.

DECT is a cellular system and a common standard for cordless telephony, messaging, and data transmission standardized by ETSI (European Telecommunications Standards Institute). DECT is similar to GSM, but cell ranges are shorter (DECT, 2006).

DECT uses several advanced digital radio techniques for efficient radio spectrum utilization. It enables high speech quality and security with low radio interference risks and low-power technology. Mobility management, responsible for DECT communication security, consists of procedures for *identity, authentication, location, access rights, key allocation, parameter retrieval,* and *ciphering* (Umar, 2004).

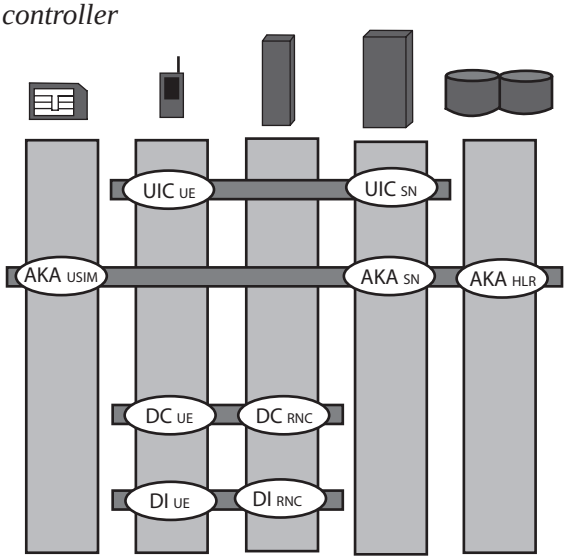
2.5 Generation

The GPRS (2.5G) infrastructure equals GSM. TSMI is replaced by P-TMSI (packet TMSI) and by P-TMSI signature as alternate identities. Mapping between IP addresses and IMSI is generated in the HLR GPRS Register. GPRS authentication is performed by SGSN (serving GPRS support node). As a consequence, user data and signaling are encrypted all the way from the mobile station to the SGSN. Tunneling, firewalls, and private IP techniques are used. IP addresses are assigned after authentication and encryption algorithm negotiations.

Third Generation

UMTS, Universal Mobile Telecommunications System, a standard for third-generation (3G) systems for mobile communication, referred to as International Mobile Telecommunications 2000 (IMT-2000) and initiated by the International Telecommunication Union (ITU), is presently being developed by the Third Generation Partnership Project (3GPP).

Figure 2. UMTS functional security architecture; UE is user equipment and RNC is radio network controller



The UMTS security architecture is based on 2G/2.5G security. Some GSM security features have been improved and some new features have been added. The UMTS security mechanisms are (see Figure 2): user identity confidentiality (UIC), authentication and key agreement (AKA), confidentiality of user and signaling data (DC), and integrity of signaling data (DI). See Lu (2002) for UMTS security details.

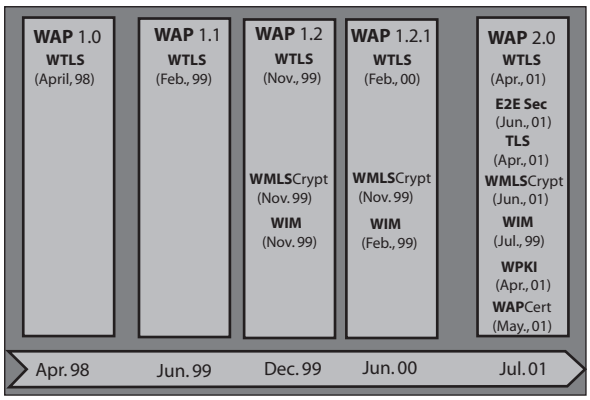
WAP

WAP (wireless application protocol) is an open mobile device application standard. WAP security protocols and specifications are being developed by the WAP Forum (Open Mobile Alliance, 2006). The evolution of WAP security specifications is shown in Figure 3.

WTLS/TLS/SSL

SSL/TLS are TCP-based security protocols for communication in client/server applications. WAP 2.0 adopts TLS as security protocol and supports the tunneling of SSL/TLS sessions through a WAP/WAP proxy. TLS/SSL in WAP 2.0 is a complement to the similar UDP-based WTLS protocol in earlier WAP versions. Server authentication and mutual authentication are options in WTLS/TLS/SSL-protected WAP applications.

Figure 3. The development of WAP security specifications



WMLScript Crypto Library, WIM, and WPKI

The lack of non-repudiation services and end-user authentication was addressed in WAP 1.2. The WMLScript (Wireless Markup Language Script) Crypto Library provides cryptographic functionality for WAP clients. WAP identity module (WIM) is used in WTLS and application-level security functions. A WIM stores and processes user authentication information, such as private keys. A WIM implementation example is a mobile phone S/WIM card (combined SIM and WIM). WPKI (wireless public key infrastructure) is a mobile environment PKI supported since WAP 2.0 (Open Mobile Alliance, 2006).

i-Mode

i-mode is a Japanese competitor to WAP for m-commerce. i-mode security features are:

- protection of the radio link between the i-mode handset and the base station,
- encryption/authentication of data transmitted between i-mode mobile devices and Web sites, and
- protection of private network links between the i-mode center and special service providers like banks.

The radio link is protected using SSL and other protocols, which are not public. Security of Web site connections and private network links are based on SSL. Mutual certificate authentication is supported (Umar, 2004).

Bluetooth

Bluetooth provides wireless short-distance transmission of data and voice signals between

Table 1. Bluetooth service levels

	Authorization	Authentication	Encryption
Trusted	Yes	Yes	Yes
Untrusted	No	Yes	Yes
Unknown	No	No	Yes

electronic devices. The specifications are defined by Bluetooth SIG (2006). The security is based on *authentication, authorization,* and *encryption.* The *security modes* are:

1. No security measures,
2. Security measures based on authorization, and
3. Authentication and encryption.

Authentication

Bluetooth device authentication is a unidirectional or mutual challenge/response process. Secret keys, called *link keys,* are generated either dynamically or by pairing. For dynamic link key generation, a passkey—the same passkey—must be entered in both connecting devices each time a connection is established. In pairing, a long-term stored link key is generated from a user-entered passkey, which can be automatically used in several connection sessions between the same devices.

Authorization

In authorization, a Bluetooth device determines whether or not another device is allowed access to a particular service. Levels of trust are *trusted*, *untrusted*, or *unknown*. Service levels are shown in Table 1.

Encryption

Bluetooth data transmission uses 128-bit encryption. Encrypted data can only be viewed by a device owning the proper decryption key. The encryption key is based on the link key.

ZigBee

ZigBee is a low-cost, low-power communications standard for wireless data communication in home and building automation. The ZigBee stack architecture is based on the standard OSI model. The IEEE 802.15.4-2003 standard defines the physical (PHY) layer and the medium access control (MAC) sub-layer. The ZigBee Alliance builds on this foundation by providing the network (NWK) layer and a framework for the application layer with: the application support sub-layer (APS), ZigBee device objects (ZDO), and manufacturer-defined application objects.

Security services are defined for key establishment, key transport, frame protection, and device management. The MAC, NWK, and APS layers are responsible for the secure transport of their respective frames. Data encryption uses the symmetric key 128-bit AES algorithm. Frame integrity is protected, since frames cannot be modified by parties without cryptographic keys. Replayed data frames are rejected by a frame freshness verification function of the NWK layer. Furthermore, the APS sub-layer establishes and maintains security relationships. ZDO manages the security policies and the security configuration of a device. Access control uses a list of trusted devices maintained by a ZDO (ZigBee Alliance, 2004).

WLAN

Broadband mobile communication is supported by a WLAN, which gives mobile users LAN connectivity through a high-speed radio link. Major WLAN security standards are (Pulkkis, Grahn, Karlsson, Martikainen, & Daniel, 2005): IEEE 802.11/WEP, WPA, and IEEE 802.11i.

WEP is not recommended due to security flaws. Data encryption is based on static encryption keys, and no user authentication mechanisms are speci-

fied. *WPA* addresses the WEP vulnerabilities and is based on IEEE 802.11i (see Figure 4).

The main features of WPA are:

- Temporal key integrity protocol (TKIP) to provide dynamical and automatically changed encryption keys, and
- IEEE 802.1X and EAP (extended authentication protocol) to provide strong user authentication.

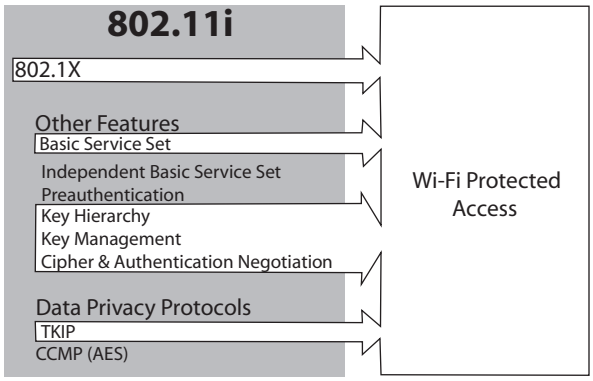
CCMP (cipher block chaining message authentication protocol) is an IEEE 802.11i protocol that uses the AES (advanced encryption standard) to provide stronger encryption than TKIP.

WiMax

WiMax is a new technology for wireless broadband Internet access. The MAC layer of the WiMax network stack has a security sub-layer with (Puthenkulam & Yin, 2005):

- A base station device and mobile user authentication capability based on the EAP protocol, X.509 certificates, and AAA servers (Radius, Diameter);
- Encryption key management using the privacy key management protocol (PKM) v2;
- AES-CCM authenticated encryption of all data communication—the Encryption Key Refresh Mechanism supports high data rates; and
- CMAC (cipher-based message authentication code) and HMAC (hash-based message authentication code), which handle control message integrity protection.

Figure 4. IEEE 802.11i features in WPA



Wireless USB

An USB wire provides two security services: (1) a wanted interconnection of two devices is created, and (2) all data in transit is protected from casual observation or malicious modification by external parties.

The goal of Wireless USB security is to provide analogous security services. Hosts and wirelessly connected devices are required to authenticate each other to avoid man-in-the-middle attacks. Data communication between a host and a wirelessly connected device is confidential and integrity-checked by AES-128/CCM encryption. Secret encryption keys are shared by mutually authenticated hosts and wirelessly connected devices (Wireless, 2005).

Satellite Communication Technology

A communication satellite permits two or more earth stations to send radio messages to each other over far distances. For satellite communication security it is necessary that earth stations have significant physical security, and RF (radio frequency) communication channels between satellites and earth stations are protected.

Satellite communications are normally secured by scrambling satellite signals using cryptography or transmitting same signals over several frequencies. The data bits are basically transmitted on different signals based on a secret scheme. The receiver of a signal must thus be aware of the secret scheme. Additional security protocols like IPSec can be used to encrypt radio messages. However, such protocols slow down

data transmission. The main challenge is thus to find a good balance between performance and security (Umar, 2004).

FUTURE TRENDS

Privacy, security, and trust issues are and will be of major importance. The growth of the Internet and m-commerce will dramatically increase the amount of personal and corporate information that can be captured or modified. In the near future ubiquitous computing systems will accentuate this trend. We can likewise expect an increase in privacy and security risks, not only with the emergence of mobile and wireless devices, but also with sensor-based systems, wireless networking, and embedded devices. Ubiquitous computing technologies will probably suffer from the same sorts of unforeseen vulnerabilities that met the Internet society.

CONCLUSION

Mobile terminals face security threats due to openness. Platforms are open for external software and content. Malicious software, like Trojan horses, viruses, and worms, has started to emerge. Fine-grained software authorization has been proposed. Downloaded software may then access particular resources only through user authorization. OS implementation vulnerability still remains a challenge because of difficulties in minimizing OS code running in privileged mode. Integrated hardware solutions may be the solution.

Wireless security architectures have many options, and many standards/protocols addressing wireless security are quite recent, especially standards/protocols based on public key cryptography. Therefore more practical experience from the use of these protocols/standards in mobile computing is needed for reliable estimation of the provided security.

REFERENCES

Barrett, J.D., Silvermann, E.R., & Byrnes, G.R. (2005). *SSH, the secure shell: The definitive guide* (2nd ed.). O'Reilly.

Barun, T., & Danzeisen, M. (2001). Secure mobile IP communication. *Proceedings of the IEEE 26th Annual Conference on Local Computer Networks* (pp. 586-593).

Bluetooth SIG. (2006). *The official Bluetooth wireless info site.* Retrieved August 8, 2006, from http://www.bluetooth.com

Calhoun, P., Johansson, T., Perkins, C., Hiller, T., & McCann, P. (2005, August). *Diameter mobile IPv4 application.* IETF, RFC 4004.

Davis, C. (2001). *IPSec: Securing VPNs.* New York: McGraw-Hill.

DECT Forum. (2006). Retrieved August 8, 2006, from http://www.dect.org

Dietze, C. (2005). The smart card in mobile communication: Enabler of next-generation (NG) services. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities.* Hershey, PA: IRM Press.

Duncan, M. V., Akhtari, M. S., & Bradford, P. G. (2004). Visual security for wireless handheld devices. *JOSHUA—Journal of Science & Health at the University of Alabama, 2.*

Handheld Security. (2006). *Laura Taylor, part I-V (2004-2005).* Retrieved August 8, 2006, from http://www.firewallguide.com/pda.htm

Hwu, J.-S., Chen, R.-J., & Lin, Y.-B. (2006). An efficient identity-based cryptosystem for end-to-end mobile security. *IEEE Transactions on Wireless Communication.*

Intel Corporation. (2006a). *Wireless application security: What's up with that?* Retrieved August 8, 2006, from http://www.intel.com/cd/ids/developer/asmo-na/eng/57399.htm?page=1

Intel Corporation. (2006b). *Intel wireless trusted platform: Security for mobile devices.* Retrieved August 8, 2006, from http://www.intel.com/design/pca/applicationsprocessors/whitepapers/300868.htm

ISO/IEC 7498-1. (1994). *Information technology—Open systems interconnection—Basic reference model: The basic model, 1994.*

ISO 7498-2. (1989). *Information processing systems—Open systems interconnection—Basic references model—Part 2: Security architecture, 1989.*

Jansen, W. A. (2003, May 12-15). Authenticating users on handheld devices. *Proceedings of the 15th Annual Canadian Information Technology Security Symposium (CITSS),* Ottawa, Canada. Retrieved August 8, 2006, from http://csrc.nist.gov/mobilesecurity/publications.html#MD

Lu, W.W. (2002). *Broadband wireless mobile, 3G and beyond.* New York: John Wiley & Sons.

Markovski, J., & Gusev, M. (2003, April). Application level security of mobile communications. *Proceedings of the 1st International Conference Mathematics and Informatics for Industry (MII 2003)* (pp. 309-317), Thessaloniki, Greece.

Olzak, T. (2005). *Wireless handheld device security.* Retrieved August 8, 2006, from http://www.securitydocs.com/pdf/3188.PDF

Open Mobile Alliance. (2006). *WAP forum*. Retrieved August 8, 2006, from http://www.wapforum.org/

Perelson, S., & Botha, R. (2004, July). An investigation into access control for mobile devices. In H. S. Venter, J. H. P. Eloff, L. Labuschagne, & M. M. Eloff (Eds.), *Proceedings of the ISSA 2004 Enabling Tomorrow Conference on Information Security,* South Africa.

Perkins, C., & Calhoun, P. (2000). *Mobile IPv4 challenge/response extensions*. IETF, RFC 3012.

Pulkkis, G., Grahn, K., Karlsson, J., Martikainen, M., & Daniel, D. E. (2005). Recent developments in WLAN security. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities.* Hershey, PA: IRM Press.

Puthenkulam, J., & Yin, H. (2005). *802.16e: A mobile broadband wireless standard.* Broadband Wireless Division, Mobility Group, Intel Corporation. Retrieved August 8, 2006, from http://www.ewh.ieee.org/r6/scv/comsoc/0512.zip

Rankl, W., & Effing, W. (2003). *Smart card handbook* (3rd ed.). New York: John Wiley & Sons.

Setec Portal. (2006). Retrieved August 8, 2006, from http://www.setec.fi

Steinberg, J., & Speed, T. (2005). *SSL VPN: Understanding, evaluating and planning secure, Web-based remote access.* Birmingham, UK: Packt Publishing.

Sundaresan, H. (2003). *OMAP™ platform security features.* Retrieved August 8, 2006, from http://focus.ti.com/pdfs/wtbu/omapplatformsecuritywp.pdf

Umar, A. (2004). *Mobile computing and wireless communications.* Middlesex, NJ: Nge Solutions.

Wireless Universal Serial Bus Specification. (2005, May 12). *Revision 1.0.* Retrieved August 8, 2006, from http://www.usb.org/developers/wusb/docs/WUSBSpec_r10.pdf

Zao, J., Kent, S., Gahm, J., Troxel, G., Condell, M., Helinek, P., Yuan, N., & Castineyra, I. (1999). A public-key based secure Mobile IP. *Wireless Networks, 5*(5), 393-390.

ZigBee Alliance. (2004, December 14). *ZigBee™ Specification v1.0.* Retrieved August 8, 2006, from http://www.zigbee.org

KEY TERMS

Bluetooth: A technology standard for wireless short distance communication.

DECT: A cellular system and a common standard for cordless telephony, messaging. and data transmission standardized by ETSI (European Telecommunications Standards Institute).

Mobile IP: Mobile Internet protocol for IP number preservation of a mobile computer.

USIM: A SIM used in 3G mobile telephone networks.

WiMax: A technology standard for wireless broadband Internet access.

ZigBee™: A low-cost, low-power communication standard for wireless data communication in home and building automation.

This work was previously published in Information Security Policies and Actions in Modern Integrated Systems, edited by M. Fugini; C. Bellettini, and J. Hsu, pp. 1-63, copyright 2004 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.8
Security Architectures for B3G Mobile Networks

Christoforos Ntantogian
University of Athens, Greece

Christos Xenakis
University of Piraeus, Greece

ABSTRACT

The integration of heterogeneous mobile/wireless networks using an IP-based core network materializes the beyond third generation (B3G) mobile networks. Along with a variety of new perspectives, the new network model raises new security concerns, mainly, because of the complexity of the deployed architecture and the heterogeneity of the employed technologies. In this chapter, we examine and analyze the security architectures and the related security protocols, which are employed in B3G networks focusing on their functionality and the supported security services. The objectives of these protocols are to protect the involved parties and the data exchanged among them. To achieve these, they employ mechanisms that provide mutual authentication as well as ensure the confidentiality and integrity of the data transferred over the wireless interface and specific parts of the core network. Finally, based on the analysis of the security mechanisms, we present a comparison of them that aims at highlighting the deployment advantages of each one and classifies the latter in terms of: (1) security, (2) mobility, and (3) reliability.

INTRODUCTION

The evolution and successful deployment of wireless LANs (WLANs) worldwide has yielded a demand to integrate them with third generation (3G) mobile networks. The key goal of this integration is to develop heterogeneous mobile data networks, named as beyond 3G (B3G) networks, capable of supporting ubiquitous computing. Currently, the network architecture (3rd Generation Partnership Project [3GPP] TS 23.234, 2006) that integrates 3G and WLAN specifies two different access scenarios: (1) the *WLAN Direct IP Access* and (2) the *WLAN 3GPP IP Access.* The first scenario provides to a user an IP connection to the public Internet or to an intranet via the WLAN access network (WLAN-AN), while the second allows a user to connect to packet switch (PS) based services (such

as wireless application protocol [WAP], mobile multimedia services [MMS], location-based services [LBS] etc.) or to the public Internet, through the 3G public land mobile network (PLMN).

Along with a variety of new perspectives, the new network model (3G-WLAN) raises new security concerns, mainly, because of the complexity of the deployed architecture and the heterogeneity of the employed technologies. In addition, new security vulnerabilities are emerging, which might be exploited by adversaries to perform malicious actions that result in fraud attacks, inappropriate resource management, and loss of revenue. Thus, the proper design and a comprehensive evaluation of the security mechanisms used in the 3G-WLAN network architecture is of vital importance for the effective integration of the different technologies in a secure manner.

In this chapter we examine and analyze the security architectures and the related security protocols, which are employed in B3G, focusing on their functionality and the supported security services for both WLAN Direct IP Access and 3GPP IP Access scenarios. Each access scenario (i.e., WLAN Direct Access and WLAN 3GPP IP Access) in B3G networks incorporates a specific security architecture, which aims at protecting the involved parties (i.e., the mobile users, the WLAN, and the 3G network) and the data exchanged among them. We elaborate on the various security protocols of the B3G security architectures that provide mutual authentication (i.e., user and network authentication) as well as confidentiality and integrity services to the data transferred over the air interface of the deployed WLANs and specific parts of the core network. Finally, based on the analysis of the two access scenarios and the security architecture that each one employs, we present a comparison of them. This comparison aims at highlighting the deployment advantages of each scenario and classifying them in terms of: (1) security, (2) mobility, and (3) reliability.

The rest of this chapter is organized as follows. The next section outlines the B3G network architectures and presents the WLAN Direct IP Access and the 3GPP IP Access scenarios. The third section elaborates on the B3G security architectures analyzing the related security protocols for each scenario. The fourth section compares the security architectures and consequently, the two access scenarios. Finally, the fifth section contains the conclusions.

BACKGROUND

The B3G Network Architecture

As shown in Figure 1, the B3G network architecture includes three individual networks: (I) the WLAN-AN, (II) the visited 3G PLMN, and (III) the home 3G PLMN. Note that Figure 1 illustrates the architecture for a general case where the WLAN is not directly connected to the user's home 3G PLMN. The WLAN-AN includes the wireless access points (APs), the network access servers (NAS), the authentication, authorization, accounting (AAA) proxy (Laat, Gross, Gommans, Vollbrecht, & Spence, 2000), and the WLAN-access gateway (WLAN-AG). The wireless APs provide connectivity to mobile users and act like AAA clients, which communicate with an AAA proxy via the Diameter (Calhoun, Loughney, Guttman, Zorn, & Arkko, 2003) or the Radius (Rigney, Rubens, Simpson, & Willens, 1997) protocol to convey user subscription and authentication information. The AAA proxy relays AAA information between the WLAN and the home 3G PLMN. The NAS allows only legitimate users to have access to the public Internet, and finally, the WLAN-AG is a gateway to 3G PLMN networks. It is assumed that WLAN is based on the IEEE 802.11 standard (IEEE std 802.11, 1999).

On the other hand, the visited 3G PLMN includes an AAA proxy that forwards AAA information to the AAA server (located in the home 3G PLMN), and a wireless access gateway (WAG), which is a data gateway that routes users' data to

the home 3G PLMN. On the other hand, the home 3G PLMN includes the AAA server, the packed data gateway (PDG) and the core network elements of the universal mobile telecommunications system (UMTS), such as the home subscriber service (HSS) or the home location register (HLR), the Gateway GPRS support node (GGSN) and the Serving GPRS support node (SGSN). The AAA server retrieves authentication information from the HSS/HLR and validates authentication credentials provided by users. The PDG routes user data traffic between a user and an external packet data network, which is selected based on the 3G PS-services requested by the user. The latter identifies these services by means of a WLAN-access point name (W-APN), which represents a reference point to the external IP network that supports the PS services to be accessed by the user.

As mentioned previously, the integrated architecture of B3G networks specifies two different network access scenarios: (1) the WLAN direct IP access and (2) the WLAN 3GPP IP Access. The first scenario provides to a user connection to the public Internet or to an intranet via the WLAN-AN. In this scenario both the user and the network are authenticated to each other using the extensible authentication protocol method for GSM subscriber identity modules (EAP-SIM) (Haverinen & Saloway, 2006) or the Extensible Authentication Protocol-Authentication and Key Agreement (EAP-AKA) (Arkko & Haverinen, 2006) protocol. Moreover, in this scenario, the confidentiality and

Figure 1. The B3G network architecture

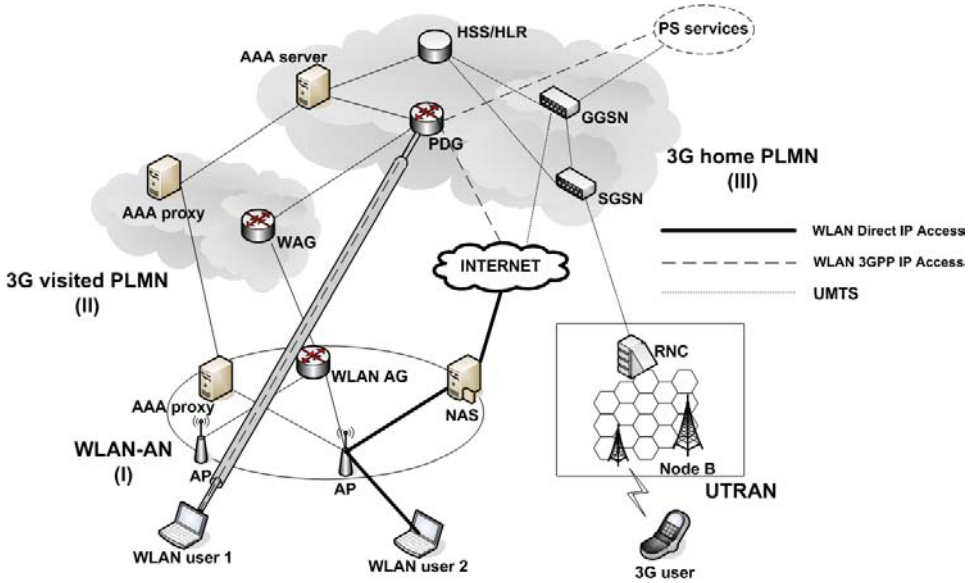


Table 1. 3G-WLAN interworking security mechanisms

Security	*WLAN Direct IP Access*	*3GPP IP Access*
Authentication	EAP-SIM or EAP-AKA	IKEv2 with EAP-SIM or EAP-AKA
Data protection	CCMP or TKIP protocol	IPsec based VPN tunnel using the ESP protocol

CCMP = Counter-Mode/CBC-Mac Protocol TKIP = Temporal Key Integrity Protocol

integrity of users data transferred over the air interface is ensured by the 802.11i security framework (IEEE std 802.11i, 2004). On the other hand, the WLAN 3GPP IP Access scenario allows a WLAN user to connect to the PS services (like WAP, MMS, LBS, etc.) or to the public Internet through the 3G PLMN. In this scenario, the user is authenticated to the 3G PLMN using the EAP-SIM or alternatively the EAP-AKA protocol encapsulated within IKEv2 (Kaufman, 2005) messages. The execution of IKEv2 is also used for the establishment of an IP security (Ipsec)-based virtual private network (VPN) (Kent & Atkinson, 1998a) tunnel between the user and the PDG that provides confidentiality and integrity services to the data exchanged between them (see Figure 1). Table 1 summarizes the security protocols employed in each access scenario.

SECURITY ARCHITECTURES FOR B3G NETWORKS

Each network access scenario (i.e., WLAN direct access and WLAN 3GPP IP access) in B3G networks incorporates a specific security architecture, which aims at protecting the involved parties (i.e., the mobile users, the WLAN, and the 3G network) and the data exchanged among them. These architectures (3GPP TS 23.234, 2006) consist of various security protocols that provide mutual authentication (i.e., user and network authentication) as well as confidentiality and integrity services to the data sent over the air interface of the deployed WLANs and specific parts of the core network. In the following, the security architectures and the involved security protocols, which are employed in B3G networks, are presented and analyzed focusing on their functionality and the supported security services.

WLAN Direct IP Access Scenario

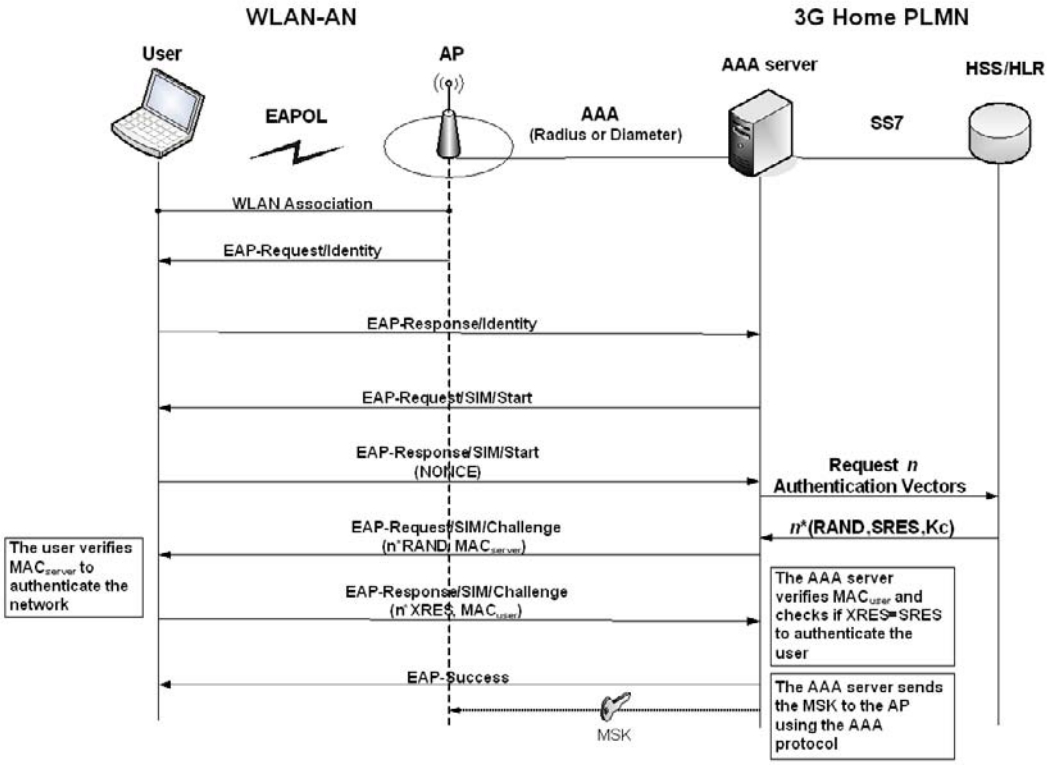
In the WLAN Direct IP Access scenario, both the user and the network are authenticated to each other using EAP-SIM or EAP-AKA, which are based on the 802.1X port access control (IEEE std 802.1X, 2001). After a successful authentication, the user obtains an IP address from the WLAN-AN and then, he/she gets access to the public Internet or an intranet, depending on the requested service. In this scenario, the confidentiality and integrity of user's data conveyed over the air interface of WLAN (IEEE std 802.11, 1999) are ensured by 802.11i (IEEE std 802.11i, 2004), which is analyzed next.

Authentication in WLAN Direct IP Access

The specific security protocol (i.e., EAP-AKA or EAP-SIM) that will be used for mutual authentication between the user and the network depends on the user's subscription. If the user possesses a UMTS subscribers identity module (USIM) card (3GPP TS 22.100, 2001), then, the EAP-AKA protocol is employed. Otherwise, EAP-SIM is used in cases that the user has a SIM-card (European Telecommunications Standards Institute [ETSI] TS 100 922, 1999) of global system for mobile communications (GSM)/general packet radio service (GPRS) (3GPP TS 0.3.6, 2002). When the AAA server receives the user's identity, it fetches from the HSS/HLR the user's profile in order to determine the employed authentication protocol that will be employed (i.e., EAP-SIM or EAP-AKA). In the following, we analyze the functionality of these two protocols focusing on the security services that each one provides.

EAP-SIM. EAP-SIM (Haverinen & Saloway, 2006) provides mutual authentication in a network environment that integrates 3G and WLANs using the credentials included in a SIM-card of a GSM/GPRS subscription. It involves a user, an AAA client (which is actually a wireless AP), and an AAA server that obtains authentication information (i.e., authentication triplets) from the HSS/HLR of the

Figure 2. The EAP-SIM authentication and session key agreement procedure



network where the user is subscribed (see Figure 2). EAP-SIM incorporates two basic enhancements that eliminate known security weaknesses of the authentication and key agreement procedure of GSM/GPRS (Haverinen & Saloway, 2006). First, the keys used in EAP-SIM are enhanced to have 128-bits security, in contrast to the 64-bit security of the original GSM/GPRS keys. Second, EAP-SIM supports mutual authentication, in contrast to the GSM/GPRS authentication, which performs only user to network authentication.

For the generation of stronger keys, the EAP-SIM protocol combines *n* (*n*=2 or *n*=3) individual random challenge (RANDs) that result in the derivation of *n* session keys, Kc. These keys are combined with a random number (NONCE payload), the user identity and other context-related information in order to generate the master key (*MK*) of the EAP-SIM protocol, as shown in the following formula:

$$MK = SHA1(Identity | n*Kc | NONCE | Version\ List | Selected\ Version)^{1}, \quad (1)$$

where SHA1 is a hash function (Eastlake & Jones, 2001). In the sequel, the produced key *MK* is fed into a pseudo random function (prf) that generates other keys used in EAP-SIM. From these keys the most important are: (1) the master session key (*MSK*), which is used in 802.11i to generate the encryption keys, as described later on, and (2) the *K_auth* key, which is used in EAP-SIM for the generation of keyed message authentication codes (MACs) for authentication purposes.

Figure 2 shows the message exchange of EAP-SIM between the user and the AAA server, which

is analyzed next. Note that the user communicates with the wireless AP via the EAP over LAN (EAPOL) protocol (IEEE std 802.1X, 2004).

- First, the user associates with the wireless AP and the latter sends an EAP-Request/Identity message to the user asking for his/her identity.
- The user responds with a message (EAP-Response/Identity) that includes his/her identity in the format of network access identifier (NAI) (Aboba & Beadles, 1999). This identity can be either the International Mobile Subscriber Identity (IMSI), or a temporary identity (i.e., pseudonym).
- Knowing the user's identity, the AAA server issues an EAP-Request/SIM/Start message, which actually starts the authentication procedure.
- The user sends back an EAP-Response/SIM/Start message that includes a nonce parameter (NONCE), which is the user's challenge to the network.
- Upon receiving this message, the AAA server communicates with HSS/HLR and obtains *n* (*n*=2 or *n*=3) authentication triplets (RAND, SRES, Kc) for the specific user (the holder of the SIM-card). The generation of the GSM authentication triplets is based on a permanent, pre-shared (between the user and the network) secret key, *Ki*, which is assigned to the user when the latter is subscribed to the GSM/GPRS network.
- Then, the AAA server sends to the user an EAP-Request/SIM/Challenge message, which contains the *n* RANDs and the MAC_{server} of the message payload, which is calculated using the *K_auth* key as follows:

$$MAC_{server}=HMAC\_SHA1_{K\_auth}(EAP\text{-}Request/SIM/Challenge(n*RAND)|\ NONCE)^{2}, \quad (2)$$

where *NONCE* is the nonce sent by the user to the AAA server, and HMAC-SHA1 (Krawczyk, Bellare, & Canetti, 1997) is the MAC algorithm that generates the keyed hash value. Before the calculation of the MAC_{server} value, the AAA server must first generate the *MK* key (see Eq. 1), and, subsequently, the *K_auth* and *MSK* keys.

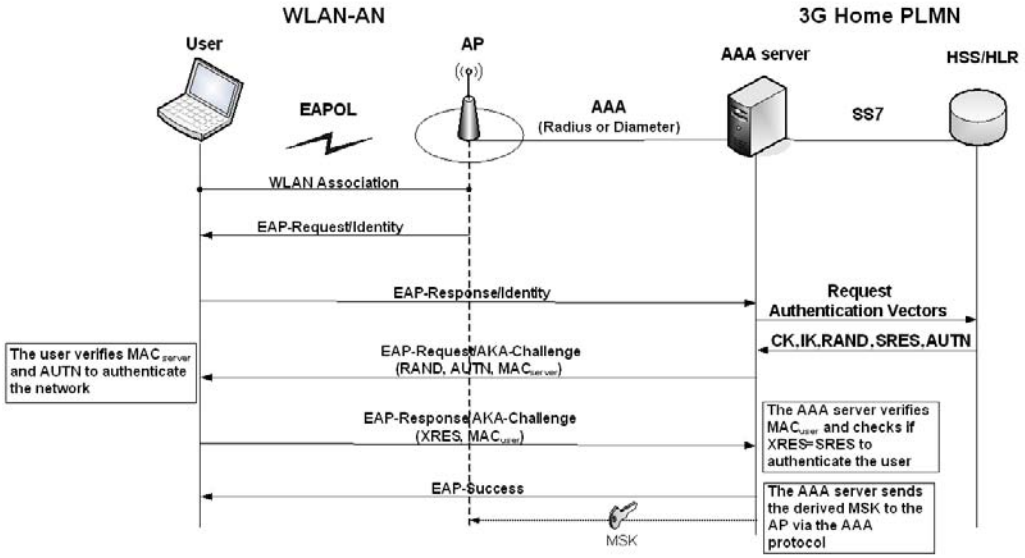
- Upon receiving the EAP-Request/SIM/Challenge, the user executes the GSM/GPRS authentication algorithms *n* times (one for each received RAND), in order to produce the *n* Kc keys and the *n* expected response (XRES) values. In the sequel, using the produced *n* *Kc* keys he/she generates the *MK* (see Eq. 1), and, consequently, the *K_auth* and the *MSK* keys, similarly, to the AAA server.
- Next, the user verifies the MAC_{server} using the *K_auth* key, and if this check is successful, then, the network is authenticated to the user, and the latter conveys to the AAA server the generated *n* XRES values within a EAP-Response/SIM/Challenge message. This message also includes the MAC_{user} value generated as follows:

$$MAC_{user}=HMAC\text{-}SHA1_{K\_auth}(EAP\text{-}Response/SIM/Challenge(n*XRES)\ |\ n*XRES)^{3}, \quad (3)$$

- Upon receiving this message, the AAA server examines whether the produced MAC_{user} is valid and if the *n* XRES values are equal to the *n* SRES values received from HSS/HLR for authentication. If these checks are successful, the AAA server sends an EAP-Success message to the user indicating the successful completion of the authentication procedure. In addition, the AAA server sends to the wireless AP the session key *MSK* within an AAA message (e.g., Radius or Diameter).

At this point, both the user and the network are mutually authenticated, and the user and the wireless AP share the key *MSK*, which is used for encryption purposes in the employed 802.11i security framework (see the *Data protection-802.11i*

Figure 3. The EAP-AKA authentication procedure and session key agreement



standard section).

EAP-AKA.EAP-AKA (Arkko & Haverinen, 2006) is an alternative to the EAP-SIM authentication protocol that uses a USIM-card and the UMTS AKA procedure. It involves the same network components with EAP-SIM (i.e., a user, an AAA client and an AAA server) and uses the same protocols for communication between them (i.e., EAPOL, Radius, Diameter, etc.). In the following, the EAP-AKA message exchange is analyzed:

- Likewise EAP-SIM, in the first two messages in the EAP-AKA negotiation (see Figure 3) the wireless AP requests for the user's identity (EAP request/identity message), and the latter replies by sending an EAP response/identity message, which contains his/her permanent IMSI or a temporary identity in an NAI format.
- After obtaining the user's identity, the AAA-server checks whether it possesses a 3G authentication vector, stored from a previous authentication with the specific user. If not, the AAA server sends the users IMSI to the HSS/HLR. The latter generates *n* 3G authentication vectors for the specific user by using the UMTS permanent secret key, *K*, which is assigned to the user when he/she is subscribed to the network, and sends it to the AAA-server. Note that an authentication vector includes a RAND, the authentication token (AUTN), the XRES, the encryption key (*CK*), and the integrity key (*IK*) (Xenakis & Merakos, 2004).
- In the sequel, the AAA server selects one out of *n* obtained authentication vectors to proceed with the EAP-AKA authentication procedure and stores the remaining *n*-1 for future use. From the selected authentication vector, the AAA server uses the keys *CK* and *IK* and the identity of the user to compute the *MK* of EAP-AKA as shown in the following formula:

$$MK = SHA1(Identity|IK|CK), \quad (4)$$

MK is used as a keying material to generate the *MSK* and the *K_auth* key. The AAA server uses

the *K_auth* key to calculate a keyed MAC_{server} (see Eq. 5), which verifies the integrity of the next EAP-AKA message (EAP-Request/AKA-Challenge).

$$MAC_{server}=HMAC\text{-}SHA1_{K\_auth}(EAP\text{-}Request/AKA/Challenge(RAND,AUTN)), \quad (5)$$

- The AAA server sends this message (EAP-Request/AKA-Challenge) to the user that contains the RAND, AUTN, and MAC_{server} payload. After receiving this information message, the user executes the UMTS-AKA algorithms and verifies the AUTN payload (Xenakis & Merakos, 2004). In the sequel, he/she generates the *IK* and *CK* keys and uses these two keys, as shown in Equation 4, to calculate the key *MK*. Subsequently, he/she uses *MK* to calculate the key *MSK* and the key *K_auth*, in order to verify the received MAC_{server} value.
- If these verifications (i.e., AUTN, MAC_{server}) are successful, the user computes the user's response to the challenge, noted as XRES payload, and sends an EAP-Response/AKA-Challenge message to the AAA server that includes the XRES and a new MAC_{user} value, which covers the whole EAP message and it is calculated using the *K_auth* key as follows:

$$MAC_{user}=HMAC\text{-}SHA1_{K\_auth}(EAP\text{-}Response/AKA/Challenge(n*XRES)), \quad (6)$$

- Upon receiving the EAP-Response/AKA-Challenge message the AAA server verifies the received MAC_{user} value and checks if the received user's response to the challenge (XRES) matches with the response (i.e., SRES) received from the HLR/HSS.
- If all these checks are successful, the AAA server sends an EAP-Success message along with the key *MSK* to the wireless AP. The latter stores the key and forwards the EAP-Success message to the user.

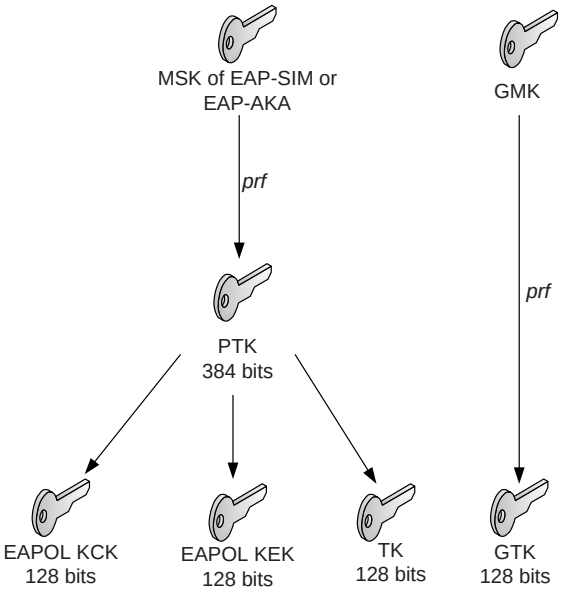
Finalizing the EAP-AKA protocol, both the user and the network have been authenticated to each other, and the user and the wireless AP share the key *MSK*, which is used in the security framework of 802.11i for generating the session encryption keys, as described in the next section.

Data Protection-802.11i Standard

As mentioned previously, 802.11i is employed to provide confidentiality and integrity services to users' data conveyed over the radio interface of the deployed WLAN in the WLAN Direct IP Access scenario. The 802.11i standard was developed to enhance the security services provided in WLANs. Its design was motivated by the fact that the wired equivalent privacy (WEP) protocol, due to its security flaws, could not fulfil the security requirements of WLANs (Borisov, Goldberg, & Wagner, 2001). The design goal of 802.11i is twofold: (1) to provide session key management by specifying a four-way handshake and group key handshake procedures, and (2) to enhance the confidentiality and integrity services provided to users' data by incorporating two security protocols (1) the counter-mode/CBC-MAC protocol (CCMP), which employs the advanced encryption standard (AES), and (2) the temporal key integrity protocol (TKIP), which uses the same encryption (RC4) with the WEP protocol. In the following, we analyze the four-way and group key handshake procedures of 802.11i and we present the functional details of the CCMP protocol. Since the TKIP protocol is considered to be a short term solution and it is merely a software enhancement of WEP, we do not elaborate further on it.

Four-way and group key handshakes. After a successful completion of the authentication procedure of EAP-SIM or EAP-AKA, the user and the AP perform the four-way and group key handshakes of 802.11i (IEEE std 802.11i, 2004) in order to generate the session keys. In the four-way handshake, both the user and the AP derive the

Figure 4. The CCMP protocol key hierarchy



pairwise transient key (*PTK*) from the *MSK* key that was generated in EAP-SIM or EAP-AKA to protect the four-way handshake messages and the unicast messages. In addition, the AP delivers to the user a group temporal key (*GTK*), which is used to protect broadcast/multicast messages. The *GTK* key is generated from the group master key (*GMK*), which is stored and maintained in the AP. The group key handshake is executed whenever the AP wants to deliver a new *GTK* key to the connected users. Note that all the messages exchanged during the four-way and the group key handshakes comply with the EAPOL-Key message format (IEEE std 802.1X, 2004).

As its name implies, the 802.11i four-way handshake consists of a total of four EAPOL-Key messages, which are analyzed next. Each of these messages includes key information (key_info payload), such as key identity, key replay counter, and so forth.

- At the beginning of the four-way handshake, the AP sends an EAPOL-Key message to the user that includes the A_{nonce}, which is a random number used as input for the generation of the *PTK* key, as described later on.
- Upon receiving the first EAPOL-Key message, the user generates a new random number called S_{nonce}. Then, he/she calculates the 384-bits *PTK* key using the first 265 bits of the *MSK* key (*MSK* was generated during the authentication procedure of EAP-SIM or EAP-AKA as described in the *Authentication in the WLAN Direct AP Access* section), the user's address, the AP's address, the S_{nonce} value, and the A_{nonce} value, as follows:

$$PTK = prf(MSK, \text{“}Pairwise\ key\ expansion\text{”}, Min(AP\ address, user's\ address) \mid Max(AP\ address, user's\ address) \mid Min(A_{nonce}, S_{nonce}) \mid Max(A_{nonce}, S_{nonce})), \quad (7)$$

where *prf* is a pseudo random function, “*Pairwise key expansion*” is a set of characters, and, finally, the Min and Max functions provide the minimum and maximum value, respectively, between two inputs. In the sequel, the generated *PTK* key is partitioned to derive three other keys: (1) a

128-bits key confirmation key (*KCK*) that provides integrity services to EAPOL-Key messages, (2) a 128-bits key encryption key (*KEK*) used to encrypt the *GTK* key as described next, and, (3) a 128-bits temporal key (*TK*) used for user's data encryption (see Figure 4).

- After the calculation of these keys, the user forwards to the AP the second EAPOL-Key message (step 2-Figure 5) that includes the S_{nonce}, the user's Robust Security Network Information Element (RSN IE) payload, which denotes the set of authentication and cipher algorithms that the user supports, and a message integrity code (MIC), which is a cryptographic digest used to provide integrity services to the messages of the four-way handshake and it is computed as follows:

$$MIC = HASH_{KCK}(EAPOL\text{-}Key\ message), \quad (8)$$

where $HASH_{KCK}$ denotes a hash function (i.e., HMAC-MD5 or HMAC-SHA-128) that uses the *KCK* key to generate the cryptographic hash value over the second EAPOL-Key message.

- Upon receiving this message, the AP calculates the key *PTK* and the related keys (i.e., *KCK*, *KEK*, and *TK* keys), (the same with the user), and, then, verifies the integrity of the message (producing the MIC value). Next, it generates the 128-bits *GTK* key from the *GMK* key as follows:

$$GTK = prf(GMK, \text{"Group key expansion"} \mid AP\ address \mid G_{nonce}), \quad (9)$$

where G_{nonce} is a random number generated from the AP to derive the *GTK* key

- In the sequel, the AP replies to the user by sending the third EAPOL-Key message (step 3), which includes the A_{nonce} value (the same

Figure 5. The four-way and group key handshakes of 802.11i

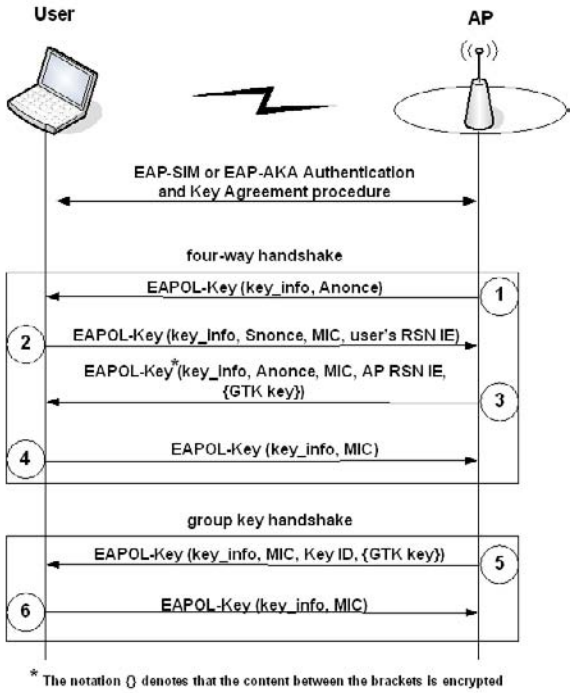
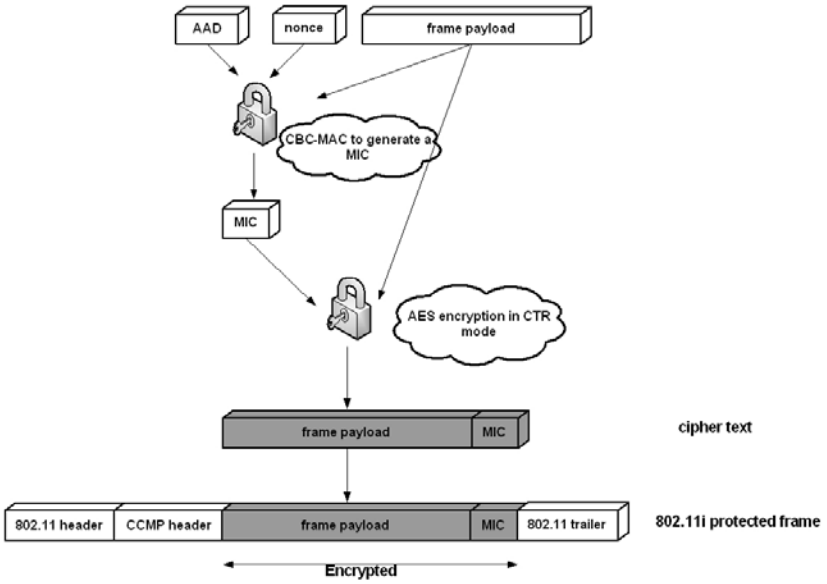


Figure 6. The CCMP protocol



with the first EAPOL-Key message), an MIC over the third EAPOL-Key message, the AP's RSN IE, and the *GTK* key, which is used to protect the broadcast/multicast messages and it is conveyed encrypted using the *KEK* key, as follows:

$$Encrypted\ GTK = ENC_{KEK}(GTK), \quad (10)$$

where ENC_{KEK} denotes the encryption algorithm (i.e., AES or RC4), which uses the *KEK* key to encrypt the *GTK* key.

- By receiving this message, the user checks whether the MIC is valid and compares his/her RSN IE with the AP's RSN IE ensuring that they support the same cryptographic algorithms. If all these checks are correct, the user decrypts the *GTK* key using the *KEK* key and sends to the AP the last message of the four-way handshake (step 4), which includes an MIC payload over the fourth EAPOL-Key message, to acknowledge to the AP that he/she has installed the *PTK* key and the related keys (i.e., *KEK*, *KCK*, and *TK* keys), as well as the *GTK* key.
- Once the AP receives the fourth EAPOL-Key message, it verifies the MIC as previously. If this final check is successful, the four-way handshake is completed successfully, and both the user and the AP share: (1) the *TK* key to encrypt/decrypt unicast messages, and (2) the *GTK* key to encrypt/decrypt broadcast/multicast messages.

In case that the AP wants to provide a new *GTK* key to the connected users, it executes the group key handshake, as shown in Figure 5.

- The AP first generates a fresh *GTK* key from the *GMK* key and sends an EAPOL-Key message that includes an MIC value and the new *GTK* key to the users. Note that MIC is computed over the body of this EAPOL-Key message using the *KCK* key, and the *GTK* key is conveyed encrypted using the *KEK* key. Recall that both the user and the AP

share the *KEK* and *KCK* keys, which were generated in the four-way handshake.

- Upon receiving the previous message, the user employs the *KCK* key to verify whether the MIC is valid and then, he/she decrypts the *GTK* key using the *KEK* key. Finally, he/she replies to the AP with an EAPOL-Key message, which includes an MIC that acknowledges to the AP that he/she has installed the *GTK* key.
- Once the AP receives this message, it verifies the MIC. If this final verification is successful, then, the group key handshake is completed successfully and the user can encrypt broadcast/multicast messages using the new *GTK* key.

CCMP Protocol. 802.11i incorporates the CCMP protocol to provide confidentiality and integrity services to users' data conveyed over the radio interface of WLANs. The CCMP protocol combines the AES encryption algorithm in CounTeR mode (CTR-AES) to provide data confidentiality and the Cipher Block Chaining Message Authentication Code (CBC-MAC) protocol to compute an MIC over the transmitted user's data that provides message integrity (Whiting, Housley, & Ferguson, 2003).

The operation of the CCMP protocol can be divided into three distinct phases. In phase 1, the CCMP protocol constructs an additional authentication data (AAD) value from constant fields of the 802.11 frame header (IEEE std 802.11, 1999). In addition, it creates a nonce value from the priority field of the 802.11 frame header and from the packet number (PN) parameter, which is a 48-bit counter incremented for each 802.11i protected frame. In phase 2, the CCMP protocol computes an MIC value over the 802.11 frame header, the AAD, the nonce, and the 802.11 frame

Figure 7. 3GPP IP access authentication procedure

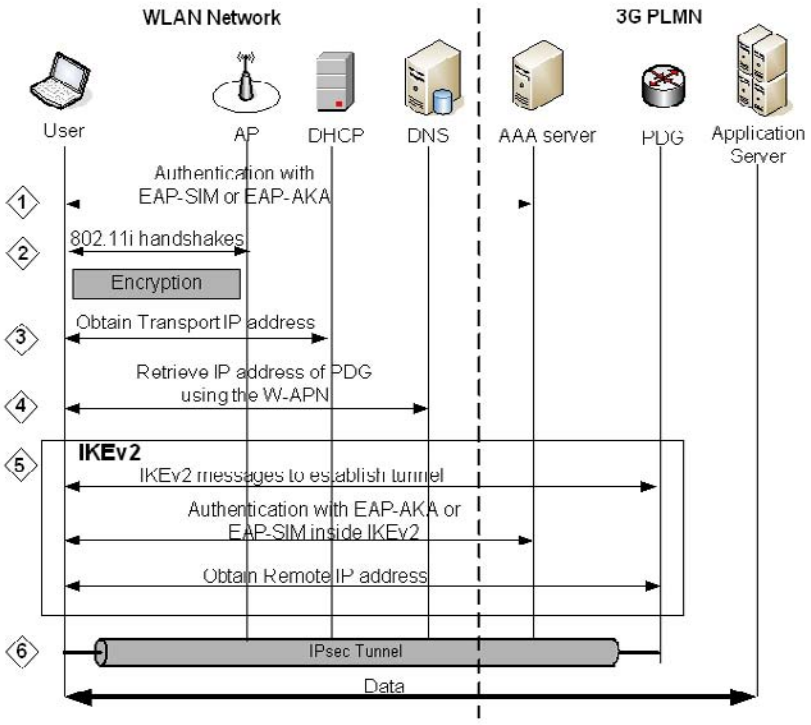
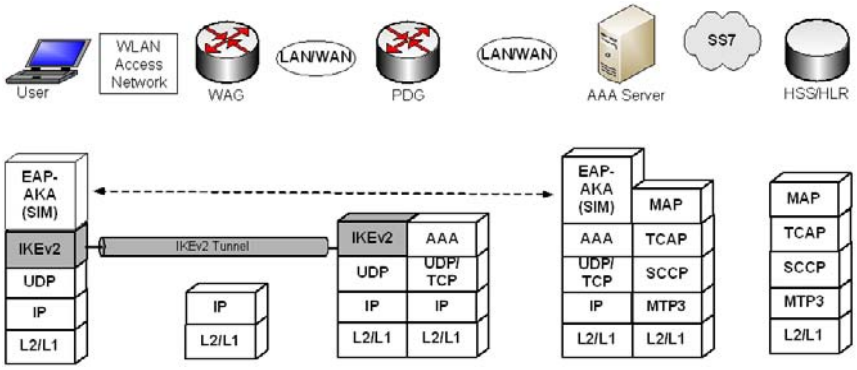


Figure 8. 3GPP IP access authentication protocol stack



payload using the CBC-MAC algorithm and the *TK* key (or the *GTK* key for broadcast/mulitcast communication). Recall that the *TK* key is part of the *PTK* key that is generated in the four-way handshake. In the sequel, CCMP forms the cipher text of the 802.11 frame payload and the produced MIC, using the CTR-AES encryption algorithm and the *TK* key (or the *GTK* key). Finally, in phase 3, the CCMP protocol constructs the 802.11i frame from the concatenation of: (1) the 802.11 header, (2) the CCMP header, which is created from the PN parameter and the identity of the encryption key, (3) the cipher text, and (4) the 802.11 trailer, which is the frame check sequence (FCS) (see Figure 6). The receiver of the 802.11i frame must verify that the PN parameter is fresh and the MIC value is valid. If these checks are successful, then, the receiver decrypts the 802.11i frame payload using the *TK* key (or the *GTK* key).

WLAN 3GPP IP Access

In contrast to the WLAN Direct IP Access scenario, in which a user gets access to the public Internet, directly, through the WLAN-AN, the WLAN 3GPP IP Access scenario provides to the WLAN user access to the PS services or the Internet through the 3G PLMN. Before getting access to them, the user must perform the six (6) discrete steps, presented in Figure 7 and described as follows:

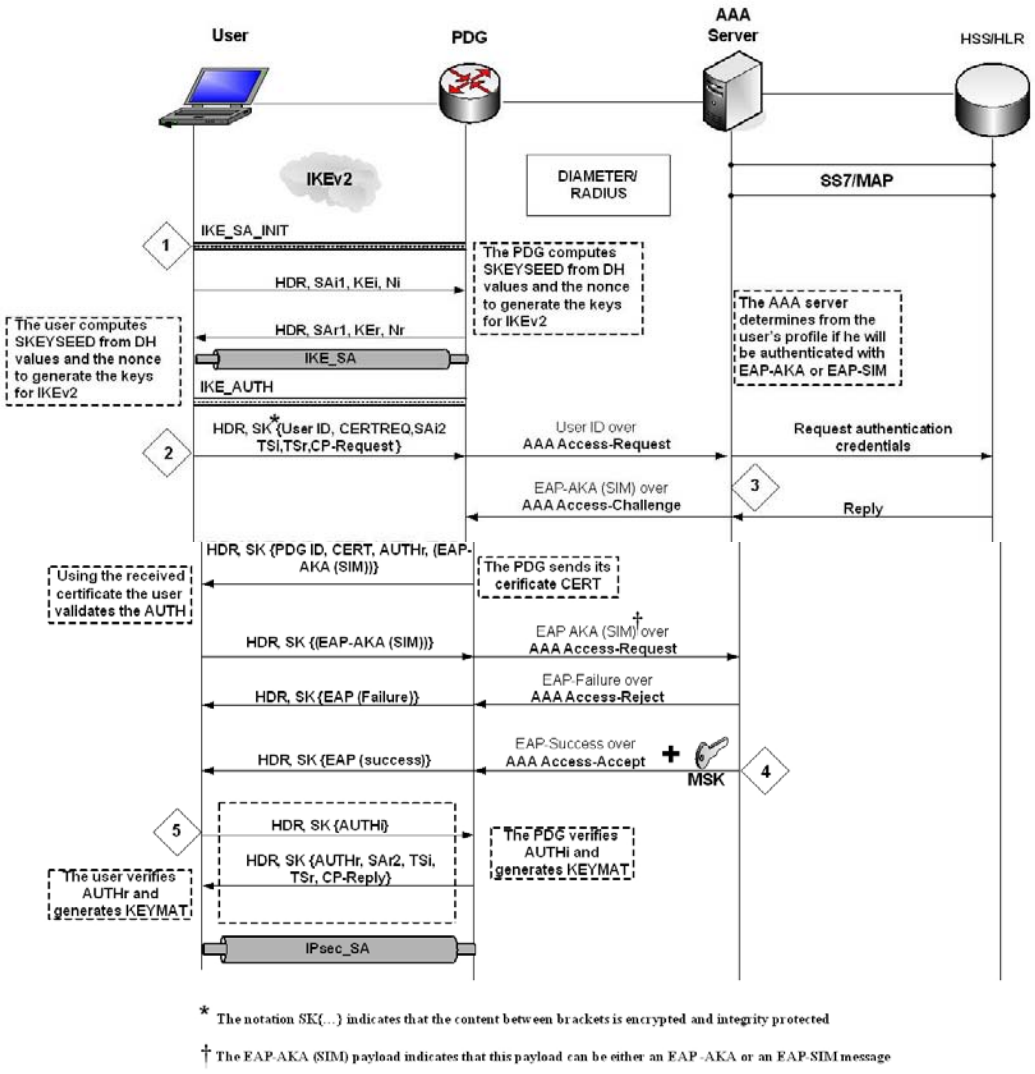
1. **Initial authentication.** The user and the network are authenticated to each other using either the EAP-SIM or EAP-AKA protocol. This authentication step enables the user to obtain a local IP address, called transport IP address, which is used for access to the WLAN environment and the PDG. Note that this initial authentication can be omitted, if the PDG trusts the WLAN network and its users.
2. After the EAP-SIM or EAP-AKA execution, the four-way handshake and optionally the group key handshake follow to provide the 802.11i session keys. Then, the communication between the user and the wireless AP is encrypted using the CCMP or alternatively the TKIP protocol.
3. After the completion of the initial authentication step and the 802.11i handshakes, the user communicates with the Dynamic Host Configuration Protocol (DHCP) server to obtain the transport IP address. This local address is used by the user to execute the IKEv2 in step 4.
4. The user retrieves the IP address of the PDG using the W-APN identity and the domain name system (DNS) protocol. Thus, both the user and the PDG participate in a second

authentication step that combines IKEv2 and EAP-SIM or EAP-AKA.

5. **Second authentication.** The user and the PDG execute the IKEv2 negotiation protocol, which encapsulates either EAP-SIM or EAP-AKA for authentication of the negotiating peers. After authentication completion, the user obtains a global IP address, called remote IP address, which is used for access to the PS services and the public Internet via the 3G PLMN. In addition, the execution of IKEv2 results in the establishment of a pair of IPsec security associations (SAs) between the user and the PDG, which are used for the deployment of an IPsec-based VPN.
6. The deployed IPsec based VPN protects user's data exchanged between the user and the PDG (in both directions) ensuring data origin authentication, data confidentiality and message integrity.

Figure 8 presents the protocol stack used in

Figure 9. The execution of IKEv2 based on EAP-SIM or EAP-AKA



* The notation SK{...} indicates that the content between brackets is encrypted and integrity protected

† The EAP-AKA (SIM) payload indicates that this payload can be either an EAP-AKA or an EAP-SIM message

the 3GPP IP Access scenario for each entity that participates in the authentication procedure. The main authentication protocol is EAP-SIM or EAP-AKA, which is executed between the user and the AAA server. The user encapsulates EAP-SIM or EAP-AKA messages within IKEv2 and conveys them to the PDG. The latter acting as an AAA client transfers the EAP-SIM or EAP-AKA messages to the AAA server using an AAA protocol. Note that the AAA protocol can be either RADIUS, which runs over the user datagram protocol (UDP) or Diameter, which runs typically over the TCP protocol. The AAA server also includes the mobile application part (MAP) protocol stack to be able to communicate with the HSS/HLR and obtain authentication triplets and authorization information.

From the previous steps that a user has to perform to get access to the PS services or the public Internet in the WLAN 3GPP IP Access scenario, the initial authentication using either EAP-SIM or EAP-AKA (step 1) and the 802.11i handshakes (step 2) are the same with these of the WLAN Direct IP Access scenario, which has been analyzed in the *Authentication in the WLAN Direct IP Access* and *Data protection-802.11i standard* sections. Moreover, the acquisition of a local IP address (step 3) and the retrieval of the PDG address (step 4) do not present any significant interest from a security point of view. Thus, in the following sections we analyze the second authentication step (step 5), which includes a combined execution of IKEv2 with EAP-SIM or EAP-AKA, and the deployment of a bidirectional VPN that protects data exchanged.

Authentication in WLAN 3GPP IP Access

IKEv2 (Kaufman, 2005) is a simplified redesign of IKE (Harkins & Carrel, 1998) that allows two peers to authenticate each other (i.e., mutual authentication) and derive keys for secure communication with IPsec. The exchanged messages within IKEv2 are protected ensuring confidentiality and integrity, while the peers are authenticated using certificates, pre-shared keys, or the EAP protocol. In the context of WLAN 3GPP IP Access scenario, the user and the PDG execute IKEv2. The authentication of the user is based on EAP-SIM or EAP-AKA, while the authentication of the PDG is based on certificates.

The IKEv2 protocol is executed in two sequential phases (i.e., phase 1 and phase 2). In phase 1, the user and the PDG establish two distinct SAs: (1) a bidirectional IKE_SA that protects the messages of phase 2, and (2) an one-way IPsec_SA that protects user's data. During phase 2, the user and the PDG using the established IKE_SA can securely negotiate a second IPsec_SA that is employed for the establishment of a bidirectional IPsec based VPN tunnel between them.

The IKEv2 phase 1 negotiation between the user and the PDG is executed in two sub-phases: (1) the IKE_SA_INIT, and (2) the IKE_AUTH exchange, as shown in Figure 9. The IKE_SA_INIT exchange (noted as step 1 in Figure 9) consists of a single request and reply messages, which negotiate cryptographic algorithms, exchange nonces, and do a Diffie-Hellman exchange. In the context of this sub-phase, four cryptographic algorithms are negotiated: (1) an encryption algorithm, (2) an integrity protection algorithm, (3) a Diffie-Hellman group, and (4) a prf. The latter prf is employed for the construction of keying material for all of the cryptographic algorithms used. After the execution of the IKE_SA_INIT, an IKE_SA is established that protects the IKE_AUTH exchange. The second sub-phase (i.e., IKE_AUTH) authenticates the previous messages; exchanges identities and certificates; encapsulates EAP-SIM or alternatively EAP-AKA messages; and establishes an IPsec_SA (step 2-5 in Figure 9). All the messages of IKEv2 include a header payload (HDR), which contains a security parameter index (SPI), a version number, and security-related flags. The SPI is a value chosen by the user and the PDG to identify a unique SA. In the following, the IKEv2 negotiation is

analyzed:

- At the beginning of the IKEv2 negotiation (step 1 in Figure 9), the user sends to the PDG the SAi1, which denotes the set of cryptographic algorithms for the IKE_SA that he/she supports, the KEi that is the Diffie-Hellman value, and an Ni value that represents the nonce. The nonce (i.e., a random number at least 128 bits) is used as input to the cryptographic functions employed by IKEv2 to ensure liveliness of the keying material and protect against replay attacks.
- The PDG answers with a message that contains its choice from the set of cryptographic algorithms for the IKE SA (SAr1), its value to complete the Diffie-Hellman exchange (KEr) and its nonce (Nr). At this point, both the user and the PDG can calculate the SKEYSEED value as follows:

$$SKEYSEED = prf((Ni \mid Nr), g \wedge ir)^4, \tag{11}$$

  where prf is the pseudo random function negotiated in the previous messages, and g^ir is the shared secret key that derives from the Diffie-Hellman exchange. The SKEYSEED value is used to calculate various secret keys. The most important are: the *SK_d* used for providing the keying material for the IPsec SA; *SK_ei* and *SK_ai* used for encrypting and providing integrity services, respectively, to the IKEv2 messages from the user to the PDG (IKE_SA); and, finally, *SK_er* and *SK_ar* that provide security services in the opposite direction (IKE_SA).

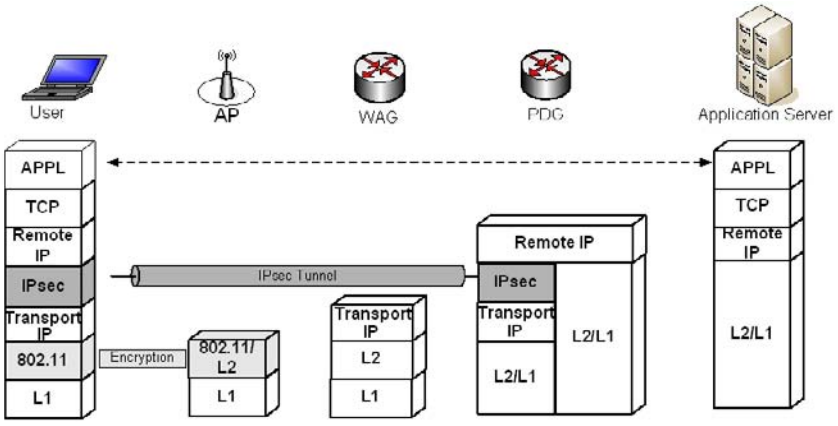
Finalizing the IKE_SA_INIT exchange, the IKE_AUTH exchange can start. It is worth noting that from this point all the payloads of the following IKEv2 messages, excluding the message header (HDR payload), are encrypted and integrity protected using the IKE_SA (see step 2 in Figure 9).

- The IKE_AUTH exchange of messages starts when the user sends to the PDG a message that includes his/her identity (IDi), which could be in an NAI format, the CERTREQ payload (optionally), which is a list of the certificate authorities (CA) whose public keys the user trusts, and the traffic selectors (TSi and TSr), which allow the peers to identify the packet flows that require processing by IPsec. In addition, in the same message the user must include the Configuration Payload Request (CP-Request), which is used to obtain a remote IP address from the PDG and get access to the 3G-PLMN.
- After receiving this information, the PDG forwards to the AAA server the user identity (IDi) including a parameter, which indicates that the authentication is being performed for VPN (tunnel) establishment. This will facilitate the AAA server to distinguish between authentications for WLAN access and authentications for VPN setup.
- Upon receiving the IDi, the AAA server fetches the user's profile and authentication credentials (GSM triplets if authentication is based on EAP-SIM, or 3G authentication vectors if authentication is based on EAP-AKA) from HSS/HLR (if these are not available in the AAA server in advance).
- Based on the user's profile, the AAA server initiates an EAP-AKA (if the user possesses a USIM card) or an EAP-SIM authentication (if the user possesses a GSM/GPRS SIM card) by sending to the PDG the first message of the related procedure (i.e., EAP-SIM or EAP-AKA) included in a AAA protocol (i.e., Radius or Diameter) (step 3 in Figure 9). Note that since there is no functional difference between the EAP-SIM and the EAP-AKA authentication when these protocols are encapsulated in IKEv2, we present them in a generic way. Thus, we introduce

the EAP-AKA (SIM) payload notation (see Figure 9) to indicate that this payload can be an EAP-SIM or an EAP-AKA message.

- Upon receiving the first EAP-AKA (SIM) message, the PDG encapsulate it within an IKEv2 message and forwards the encapsulated message to the user. Except for the EAP-AKA (SIM) payload, this message also includes the PDG's identity, which identifies the provided 3G services (W-APN) (see the *Background* section), the PDG's certificate (CERT), and the AUTHr field. The latter contains signed data used by the user to authenticate the PDG. Similarly to the previous messages, the payload of this IKEv2 message, except for the message header, is encrypted using the IKE_SA.
- Upon receiving the EAP-AKA (SIM) payload, the user verifies the AUTHr field by using the public key of the PDG included in the certificate field (CERT), and answers by sending an EAP-AKA (SIM) response message encapsulated again within an IKEv2 message. From this point, the IKEv2 messages contain only EAP-AKA (SIM) payloads, which are encrypted and integrity protected as described previously.
- The EAP-SIM or EAP-AKA exchange continues, normally, until an EAP-SUCCESS message (or an EAP-FAILURE in case of a failure) is sent from the AAA server to the PDG, which ends the EAP-AKA or the EAP-SIM dialogue. Together with the EAP-SUCCESS message, the key *MSK* is sent from the AAA server to the PDG via the AAA protocol, as shown in Figure 9 (step 4).
- After finishing the EAP-AKA or EAP-SIM dialogue, the last step (step 5) of IKEv2 reauthenticates the peers, in order to establish an IPsec_SA. This authentication step is necessary in order defeat man-in-the-middle attacks, which might take place because the authentication protocol (e.g., EAP-SIM or EAP-AKA) runs inside the secure protocol (e.g., IKEv2). This combination creates a security hole since the initiator and the responder have no way to verify that their peer in the authentication procedure is the entity at the other end of the outer protocol (Asokan, Niemi, & Nyberg, 2002). Thus, in order to prevent possible attacks against IKEv2 (i.e., man-in-the-middle attacks), both the user and the PDG have to calculate the AUTHi and the AUTHr payloads, respectively, using the *MSK* key that was generated from the EAP-SIM or EAP-AKA protocol. Then, both the user and the PDG send each other the AUTHi and AUTHr payloads to achieve a security bind-

Figure 10. 3GPP IP access data plane



ing between the inner protocol (EAP-SIM or EAP-AKA) and the outer protocol (IKEv2). Note that the PDG together with the AUTHr payload sends also its traffic selector payloads (TSi and TSr), the SAr2 payload, which contains the chosen cryptographic suit for the IPsec_SA and the assigned user's remote IP address in the Configuration Payload Reply (CP-REPLY) payload.

After the establishment of the IPsec_SA the keying material (*KEYMAT*) for this SA is calculated as follows:

$$KEYMAT = prf(SK\_d, Ni \mid Nr), \quad (12)$$

where Ni and Nr are the nonces from the IKE_SA_INIT exchange, and *SK_d* is the key that is calculated from the SKEYSEED value (see eq. 11). The *KEYMAT* is used to extract the keys that the IPsec protocol uses for security purposes. Note that the deployed IPsec_SA protects the one-way communication between the user and the PDG. For bidirectional secure communication, one more SA needs to be established between them (the user and the PDG) by executing the IKEv2 phase 2 over the established IKE_SA.

Data Protection

After the completion of the authentication procedure and the execution of IKEv2 between the PDG and the user, a pair of IPsec_SAs has been established between these two nodes. This pair deploys a bidirectional VPN between them that allows for secure data exchange over the underlying network path. At the same time, the user has been subscribed to the 3G PLMN network for charging and billing purposes using either the EAP-AKA or EAP-SIM protocol.

The deployed VPN runs on top of the wireless link and extends from the user's computer to the PDG, which is located in the user's home 3G PLMN (see Figure 1 and 10). It is based on IPsec (Kent & Atkinson, 1998a), which is a developing standard for providing security at the network layer. IPsec provides two choices of security service through two distinct security protocols: the Authentication Header (AH) protocol (Kent & Atkinson, 1998c), and the encapsulating security payload (ESP) protocol (Kent & Atkinson, 1998b). The AH protocol provides support for connectionless integrity, data origin authentication, and protection against replays, but it does not support confidentiality. The ESP protocol supports confidentiality, connectionless integrity, anti-replay protection, and optional data origin authentication. Both AH and ESP support two modes of operation: transport and tunnel. The transport mode of operation provides end-to-end protection between the communicating end points by encrypting the IP packet payload. The tunnel mode encrypts the entire IP packet (both IP header and payload) and encapsulates the encrypted original IP packet in the payload of a new IP packet.

In the deployed VPN of the WLAN 3GPP IP Access scenario, IPsec employs the ESP protocol and is configured to operate in the tunnel mode. Thus, VPN provides confidentiality, integrity, data origin authentication, and anti-reply protection services protecting the payload and the header of the exchanged IP packets. From the two IP addresses (i.e., transport and remote IP address) of each authenticated user, the remote IP address serves as the inner IP address, which is protected by IPsec, and the transport IP address serves as the IP address of the new packets, which encapsulate the original IP packets and carry them between the user and the PDG (see Figure 10). Thus, an adversary can not disclose, fabricate unnoticed, or perform traffic analysis to the data exchanged between the user and the PDG. Finally, IPsec can use different cryptographic algorithms (i.e., DES, 3DES, AES, etc.) depending on the level of security required by the two peers and the data that they exchange.

COMPARISON OF THE SCENARIOS

Based on the presentation of the two access scenarios (i.e., WLAN Direct IP Access and 3GPP IP Access) that integrate B3G networks and the analysis of the security measures that each one employs, this section provides a brief comparison of them. The comparison aims at highlighting the deployment advantages of each scenario and classifies them in terms of: (1) security, (2) mobility, and (3) reliability.

Regarding the provided security services, both scenarios support mutual authentication. In the WLAN Direct IP Access scenario, the authentication procedure employs either EAP-SIM or EAP-AKA, depending on the user's subscription. However, both protocols present the same security weaknesses, which can be exploited by adversaries to perform several attacks such as identity spoofing, denial of service (DoS) attacks, replay attacks, and so forth (Arkko & Haverinen, 2006; Haverinen & Saloway, 2006). On the other hand, the authentication procedure of the 3GPP IP Access scenario is more secured, since it combines the aforementioned protocols (i.e., EAP-SIM and EAP-AKA) with IKEv2. Specifically, the PDG is authenticated using its certificate, and the user is authenticated using EAP-SIM or EAP-AKA. It is worth noting that since the EAP-SIM and EAP-AKA messages are encapsulated in protected IKEv2 messages, the identified security weaknesses associated with them are eliminated.

Regarding confidentiality and data integrity services, both scenarios protect sensitive data conveyed over the air interface. More specifically, in the WLAN Direct IP Access scenario, high level security services are provided only in cases that the CCMP security protocol is applied, since it incorporates the strong AES encryption algorithm. A downside of applying CCMP is that it requires hardware changes to the wireless APs, which might be replaced. In the WLAN 3GPP IP Access scenario, data encryption is applied at the layer 2 (using WEP, TKIP, or CCMP) and layer 3 (using IPsec), simultaneously (see Figure 10). This duplicate encryption provides advanced security services to the data conveyed over the WLAN radio interface, but at the same time it may cause bandwidth consumption, longer delays, and energy consumption issues at the level of mobile devices.

Another deployment feature, which can be used for comparing the two scenarios, has to do with mobility. The WLAN Direct IP Access scenario may support user mobility by employing one of the mobility protocols, proposed for seamless mobility in wireless networks (Saha, Mukherjee, Misra, & Chakraborty, 2004). On the other hand, in the WLAN 3GPP IP Access scenario, the established VPN between a user and the PDG adds an extra layer of complexity to the associated mobility management protocols of this scenario. This complexity arises from the fact that as the mobile user moves from one access network to another and his/her IP address changes, the mobility protocols must incorporate mechanisms that maintain, dynamically, the established VPN, enabling the notion of mobile VPN. An attempt to address this problem can be found in Dutta et al., 2004) that designs and implements a secure universal mobility architecture, which incorporates standard mobility management protocols, such as mobile IP for achieving mobile VPN deployment.

Finally, the deployed IPsec-based VPNs between the users and the PDG in the 3GPP IP Access scenario may raise reliability issues. Reliability is perceived as the ability to use VPN services at all times, and it is highly related to the network connectivity and the capacity of the underlying technology to provide VPN services. In the 3GPP IP Access scenario, all data traffic passes through the VPN tunnels that are extend from the users to the PDG. The number of the deployed VPNs can grow significantly, due to the fact that each user can establish multiple VPNs at the same time to access different services. Thus, the PDG must be able to

support a large number of simultaneous VPNs in order to provide reliable security services.

CONCLUSION

This chapter has analyzed the security architectures employed in the interworking model that integrates 3G and WLANs, materializing B3G networks. The integrated architecture of B3G networks specifies two different network access scenarios: (1) the WLAN Direct IP Access, and (2) the WLAN 3GPP IP Access. The first scenario provides to a user connection to the public Internet or to an intranet via the WLAN-AN. In this scenario both the user and the network are authenticated to each other using EAP-SIM or EAP-AKA, depending on the user's subscription. Moreover, the confidentiality and integrity of the user's data transferred over the air interface are ensured by the 802.11i security framework. On the other hand, the WLAN 3GPP IP Access scenario allows a user to connect to the PS services (like WAP, MMS, LBS, etc.) or to the public Internet through the 3G PLMN. In this scenario, the user is authenticated to the 3G PLMN using EAP-SIM or alternatively EAP-AKA encapsulated within IKEv2, while the network is authenticated to the user using its certificate. In addition, the execution of IKEv2 is used for the establishment of an IPsec-based VPN between the user and the network that provides extra confidentiality and integrity services to the data exchanged between them.

ACKNOWLEDGMENT

Work supported by the project CASCADAS (IST-027807) funded by the FET Program of the European Commission.

REFERENCES

3rd Generation Partnership Project (3GPP) TS 22.100. (v3.7.0). (2001). *UMTS Phase 1 Release '99.* Sophia Antipolis Cedex, France: Author.

3rd Generation Partnership Project (3GPP) TS 0.3.6. (V7.9.0). (2002). *GPRS service description, Stage 2.* Sophia Antipolis Cedex, France: Author.

3rd Generation Partnership Project (3GPP) TS 23.234 (v7.3.0). (2006). *3GPP system to WLAN interworking. System description. Release 7.* Sophia Antipolis Cedex, France: Author.

3rd Generation Partnership Project (3GPP) TS 33.234 (v7.2.0). (2006). *3G security and WLAN interworking security. System description. Release 7.* Sophia Antipolis Cedex, France: Author.

Aboba, B., & Beadles, M. (1999). *The network access identifier* (RFC 2486). Retrieved from http://tools.ietf.org/html/rfc2486

Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., & Levkowetz, H. (2004). *The extensible authentication protocol* (RFC 3748). Retrieved from http://www.ietf.org/rfc/rfc3748.txt

Arkko, J., & Haverinen, H. (2006). *EAP-AKA authentication* (RFC 4187). Retrieved from http://www.rfc-editor.org/rfc/rfc4187.txt

Asokan, N., Niemi, V., & Nyberg, K. (2002). *Man-in-the-middle in tunneled authentication protocols.* Cryptology ePrint Archive, Report 2002/163. Retrieved from http://eprint.iacr.org/2002/163

Borisov, N., Goldberg, I., & Wagner, D. (2001, July). *Intercepting mobile communications: The insecurity of 802.11.* Paper presented at the 7th ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM), Rome, Italy.

Calhoun, P., Loughney, J., Guttman, E., Zorn, G., & Arkko, J. (2003). *Diameter base protocol* (RFC 3588). Retrieved from http://www.rfc-editor.org/rfc/rfc3588.txt

Dutta, A., Zhang, T., Madhani, S., Taniuchi, K.,

Fujimoto, K., Katsube, Y., et al. (2004, October). Secure universal mobility for wireless Internet. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots (WMASH)*, Philadelphia, PA.

Eastlake, D., & Jones, P. (2001). *US secure hash algorithm 1 (SHA1)* (RFC 3174). Retrieved from http://www.ietf.org/rfc/rfc3174.txt

Eronen, P. (2006). *IKEv2 mobility and multihoming protocol (MOBIKE)* (RFC 4555). Retrieved from http://www.ietf.org/rfc/rfc4555.txt

European Telecommunications Standards Institute (ETSI) TS 100 922 (v7.1.1). (1999). *Subscriber identity modules (SIM) functional characteristics.*

Harkins, D., & Carrel, D. (1998). *The Internet key exchange (IKE)* (RFC 2409). Retrieved from http://faqs.org/rfcs/rfc2409.html

Haverinen, H., & Saloway, J. (2006). *EAP-SIM authentication* (RFC 4186). Retrieved from http://www.ietf.org/rfc/rfc4186.txt

IEEE std 802.11 (1999). *Wireless LAN medium access control (MAC) and physical layer (PHY) specifications.*

IEEE std 802.11i. (2004). *Wireless medium access control (MAC) and physical layer (PHY) specifications: Medium access control (MAC) security enhancements.*

IEEE std 802.1X. (2004). *Port based access control.*

Kaufman, C. (2005). *The Internet key exchange (IKEv2) protocol* (RFC 4306). Retrieved from http://www.rfc-editor.org/rfc/rfc4306.txt

Kent, S., & Atkinson, R. (1998a). *Security architecture for Internet protocol* (RFC 2401). Retrieved from http://www.faqs.org/rfcs/rfc2401.html

Kent, S., & Atkinson, R. (1998b). *IP encapsulating security payload (ESP)* (RFC 2406). Retrieved from http://www.faqs.org/rfcs/rfc2406.html

Kent, S., & Atkinson, R. (1998c). *IP authentication header* (RFC 2402). Retrieved from http://www.rfc-editor.org/rfc/rfc2402.txt

Kivinen, T., & Tschofenig, H. (2006). *Design of the Mobike protocol* (RFC 4621). Retrieved from http://www.ietf.org/rfc/rfc4621.txt

Krawczyk, H., Bellare, M., & Canetti, R. (1997). *HMAC: Keyed-hashing for message authentication* (RFC 2104). Retrieved from http://www.faqs.org/rfcs/rfc2104.html

Laat, C., Gross, G., Gommans, L., Vollbrecht, J., & Spence, D. (2000). *Generic AAA architecture* (RFC 2903). Retrieved from http://isc.faqs.org/rfcs/rfc2903.html

Rigney, C., Rubens, A., Simpson, W., & Willens, S. (1997). *Remote authentication dial in user services (RADIUS)* (RFC 2138). Retrieved from http://tools.ietf.org/html/rfc2138

Saha, D., Mukherjee, A., Misra, I. S., & Chakraborty, M. (2004). Mobility support in IP: A survey of related protocols. *IEEE Network, 18*(6), 34-40.

Whiting, D., Housley, R., & Ferguson, N. (2003). *Counter with CBC MAC (CCM)* (RFC 3610). Retrieved from http://www.ietf.org/rfc/rfc3610.txt

Xenakis, C., & Merakos, L. (2004). Security in third generation mobile networks. *Computer Communications, 27*(7), 638-650.

KEY TERMS

Authentication, Authorization, and Accounting (AAA): AAA is a security framework which provides authentication, authorization, and accounting services. The two most prominent AAA protocols are Radius and Diameter.

Beyond Third Generation (B3G): B3G is the integration of heterogeneous mobile networks

through an IP-based common core network.

Counter-Mode/CBC-MAC Protocol (CCMP): CCMP is a security protocol defined in 802.11i, which employs the AES encryption to provide confidentiality and data integrity services.

Extensible Authentication Protocol (EAP): EAP is a security framework used to provide a plethora of authentications options, called EAP methods.

Extensible Authentication Protocol-Authentication and Key Agreement (EAP-AKA): EAP-AKA is an EAP method based on UMTS authentication of USIM cards.

Extensible Authentication Protocol method for GSM Subscriber Identity Modules (EAP-SIM): EAP-SIM is an EAP method based on GSM authentication of SIM cards.

802.11i: 802.11i is a security framework that incorporates the four-way handshake and group-key handshake for session key management and specifies the TKIP and CCMP security protocols to provide confidentiality and integrity services in 802.11 WLAN.

IKEv2: IKEv2 is a security association (SA) negotiation protocol used to establish an IPsec-based VPN tunnel between two entities.

IP security (IPsec): IPsec is a security protocol used to provide VPN services.

ENDNOTES

1 (| means string concatenetation and the notation *n**Kc denotes the *n* Kc keys concatenated)

2 (The notation *n**RAND denotes the *n* RAND values concatenated)

3 (The notation *n**XRES denotes the *n* XRES values concatenated)

4 | means string concatenation

This work was previously published in Handbook of Research on Wireless Security, edited by Y. Zhang; J. Zheng; M. Ma, pp. 297-317, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.9
Privacy and Anonymity in Mobile Ad Hoc Networks

Christer Andersson
Combitech, Sweden

Leonardo A. Martucci
Karlstad University, Sweden

Simone Fischer-Hübner
Karlstad University, Sweden

ABSTRACT

Providing privacy is often considered a keystone factor for the ultimate take up and success of mobile ad hoc networking. Privacy can best be protected by enabling anonymous communication and, therefore, this chapter surveys existing anonymous communication mechanisms for mobile ad hoc networks. On the basis of the survey, we conclude that many open research challenges remain regarding anonymity provisioning in mobile ad hoc networks. Finally, we also discuss the notorious Sybil attack in the context of anonymous communication and mobile ad hoc networks.

INTRODUCTION

The quest for privacy in today's increasingly pervasive information society remains a fundamental research challenge. In the traditional (wired) Internet, one essential means for protecting privacy is *anonymous communication*. Being anonymous usually implies that a user remains unlinkable to a set of items of interest (e.g., communication partners, messages) from an attacker's perspective (Pfitzmann & Hansen, 2006). The capabilities of the attacker are usually modeled by an *attacker model*, which can, for instance, include a rogue communication partner or an observer tapping the communication lines. Further, more advanced

applications can be deployed on top of anonymous communication mechanisms, to, for instance, enable pseudonymous applications.

This chapter investigates how anonymous communication can be enabled in *mobile ad hoc networks* (Corson & Macker, 1999); networks constituted by mobile platforms that establish on-the-fly wireless connections among themselves and ephemera networks without central entities to control it. They are of great importance as they constitute a basic core functionality needed for deploying *ubiquitous computing.* In short, ubiquitous computing would allow for computational environments providing information instantaneously through "invisible interfaces," thus allowing unlimited spreading and sharing of information. If realized, ubiquitous computing could offer an invaluable support for many aspects of our society and its institutions. However, if privacy aspects are neglected, there is a great likelihood that the end product will resemble an Orwellian nightmare.

In this chapter, we study how privacy and anonymity issues are tackled today in mobile ad hoc networks by surveying existing anonymous communication mechanisms adapted for mobile ad hoc networks[1]. Only recently, a number of such proposals have been suggested. In the survey, we evaluate some of these approaches against a set of general requirements (Andersson, Martucci, & Fischer-Hübner, 2005), which assess to which degree these approaches are suitable for mobile ad hoc networks. We also discuss Sybil attacks (Douceur, 2002) in the context of anonymous communication and mobile ad hoc networks.

This chapter is structured as follows. First, an introduction to privacy, anonymity, and anonymity metrics is provided in "Background." Then, existing approaches for enabling anonymity in ad hoc networks are described in "Anonymous Communication in Mobile Ad Hoc Networks." In "Survey of Anonymous Communication Mechanisms for Ad Hoc Networks" these approaches are evaluated against the aforementioned requirements. Then, Sybil attacks in the context of anonymous communication and mobile ad hoc networks are discussed in "Future Trends." Finally, conclusions are drawn in "Conclusions."

BACKGROUND

In this section, the concepts of privacy and anonymity and their relation are introduced. Methods for quantifying anonymity are also discussed.

Definitions of Anonymity and Related Concepts

Pfitzmann and Hansen (2006) define *anonymity* as "the state of being not identifiable within a set of subjects, the *anonymity set*" (p. 6). The

Figure 1. Unlinkability between a user in the anonymity set and an item of interest

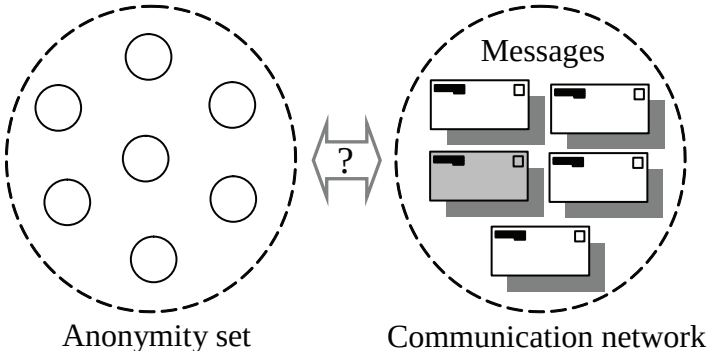
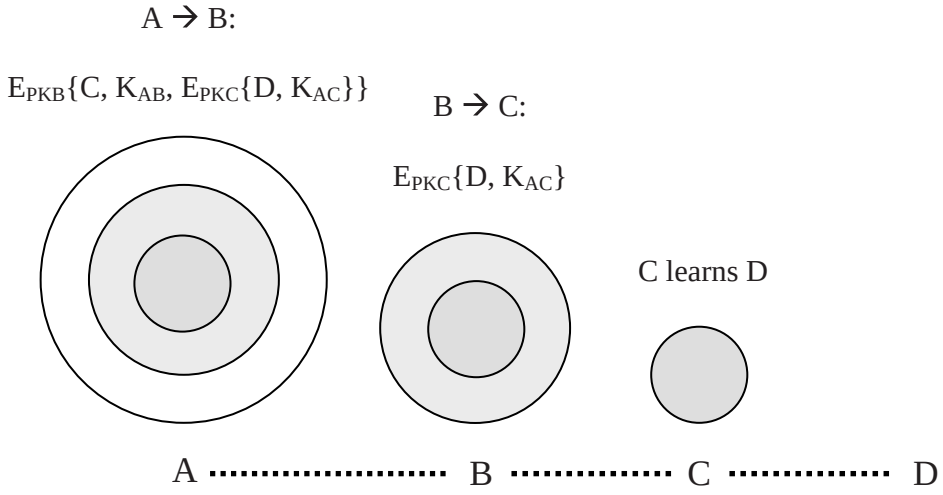


Figure 2. Setting a path between A and D (through B and C) using layered encryption; PK_B and PK_C are the public keys of B and C. K_{AB} and K_{AC} are shared symmetric keys. D is an external receiver



anonymity set includes all possible subjects in a given scenario, such as possible senders of a message.

Related to anonymity is *unlinkability*, where unlinkability of two or more items of interest (IOIs, e.g., subjects, messages, events, actions, etc.) means that within the system (comprising these and possibly other items), from the attacker's perspective, these items of interest are no more and no less related after his observation than they are related concerning his a-priori knowledge. (Pfitzmann & Hansen, 2006, p. 8)

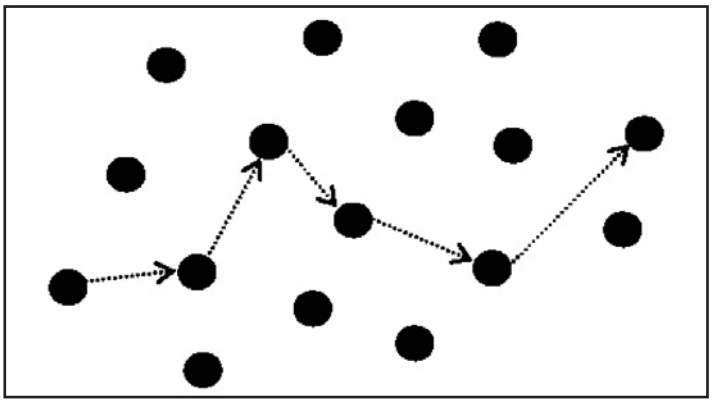
Anonymity can be defined in terms of unlinkability: *sender anonymity* entails that a message cannot be linked to the sender, while *receiver anonymity* implies that a message cannot be linked to the receiver (see Figure 1).

In traditional networks, such as the Internet, anonymous communication is often realized by *anonymous overlay networks*, which establish *virtual paths* consisting of one or more intermediary nodes, along which packets are transmitted. Using methods described below, the anonymous overlay network constructs the paths in such a manner that the correlation between the sender and receiver, and possibly also the identity of the sender and/or the receiver, is hidden.

A classic method enabling anonymity, where the sender determines the full path, is *layered encryption*[2]: a message is wrapped into several encryption layers. As the message propagates the network, these layers are sequentially decrypted by each successive node in the path, until the receiver decrypts the final layer. Each layer usually includes the identity of the next node in the path and a symmetric key shared with the initiating node (see Figure 2). In this way, expensive public key encryption is only used for constructing the path; for data delivery symmetric encryption is used. Messages encrypted in layers are often denoted *message onions*. Layered encryption enables anonymity as intermediary nodes do not know whether their predecessor and successor nodes are the sender or receiver, respectively.

An alternative approach, first applied in Crowds (Reiter & Rubin, 1997), is to let the sender select its successor randomly, which in turn flips a biased coin to decide whether it should end the path and connect to the receiver, or extend the path to a random node. The flipping of the biased coin is repeated until a node decides to connect to the receiver (see Figure 3). In this approach, link-to-link encryption between intermediary hops

Figure 3. "Crowds-like" path setting between the sender and receiver



in the path is usually combined with end-to-end encryption. This approach enables sender anonymity towards network nodes and the receiver, as neither of these nodes can deduce if the previous node in the path is the sender.

Another method specifically tailored for providing receiver anonymity is *invisible implicit addressing* (Pfitzmann & Waidner, 1987). Invisible implicit addressing hides the identity of the receiver by first encrypting a message (or a part of it) with the receiver's public key (or a shared symmetric key). Instead of sending the message directly to the receiver, the message is then *broadcasted* to all nodes in the network, which all must try to decrypt the message. However, only the intended receiver will be able to successfully decrypt the message.

On the Relation between Privacy and Anonymity

Privacy is recognized either explicitly or implicitly as a fundamental human right by most constitutions of democratic societies. Privacy can be defined as the right to *informational self-determination*, that is, individuals must be able to determine for themselves when, how, to what extent, and for what purpose personal information about them is communicated to others.

In Europe, the right for privacy of individuals is protected by the by a legal framework mainly consisting of the EU Data Protection Directive 95/46/EC, which defines general privacy requirements, and the E-Communications Privacy Directive 2002/58/EC, which specifically applies for personal data processing within the electronic communication sector.

An important privacy principle is *data minimization*, stating that the collection and processing of personal data should be minimized. Clearly, the less personal data are collected or processed, the less the right to informational self-determination is affected. Art. 6 (1) of the EU Data Protection Directive 95/46/EC embodies the principle of data minimization by stating that personal data should be limited to data that are adequate, relevant, and not excessive, and by requiring that data should only be kept in a form that permits identification of data subjects for no longer than it is necessary for the purpose for which the data were collected or for which they are further processed. Consequently, technical tools such as privacy-enhancing technologies should be available to contribute to the effective implementation of these requirements by providing anonymity and/or pseudonymity for the users and other concerned individuals.

More specific legal requirements for anonymization can also be found in the E-Communi-

cations Privacy Directive 2002/58/EC: Pursuant to Art.9 of the Directive: location data may only be processed when they are made anonymous, or with the consent of the user or subscriber to the extent and for the duration necessary for the provision of a value-added service.

On Measuring Anonymity

This section discusses *anonymity metrics*, which quantify the degree of anonymity in a given scenario in the following manner. First, the given attacker model, together with the properties of the anonymous communication mechanism, are passed as input to the anonymity metric. Then, the metric determines the degree of anonymity based using for example, analysis or by simulation, depending on the metric at hand. In Table 1, we summarize the most common anonymity metrics.

Although the metrics listed above differs in many respects, the main parameters contributing to the degree of anonymity in all metrics are *size of anonymity set* (anonymity set size and *k*-anonymity), *probability distributions* (entropy-based metric by Diaz et al.), and both (entropy-based metric by Serjantov and Danezis and the Crowds-based metric).

Anonymous Communication in Mobile Ad Hoc Networks

In *proactive* routing protocols (Perkins, 2001), each node always maintains routes to all other nodes, including nodes to which no packets are being sent. Standard proactive protocols do not enable anonymity as all nodes know significant amounts of information about other nodes.

In *reactive* routing protocols (Perkins, 2001), routes between nodes are established on demand, meaning that less packets are circulated in the network, for example, for status sensing. Also standard reactive routing protocols fail to enable anonymity. As a proof of concept, consider the reactive protocols dynamic source routing (DSR) (Johnson & Maltz, 1996) and ad hoc on-demand

Table 1. A summary of anonymity metrics

Anonymity set size
A classic indicator of anonymity is the size of the anonymity set. This metric is appropriate for mechanisms in which all users are equally likely to be the sender of a particular message, as in the DC-networks (Chaum, 1988) or Crowds, regarding the Web server (Reiter & Rubin, 1997).
K-anonymity
If a mechanism provides *k*-anonymity (Sweeney, 2002), *k* constitutes a lower bound of the anonymity set size *n*. For example, $k = 3$ implies that an attacker cannot exclude more than $(n - 3)$ users from the anonymity set.
Crowds-based metric
In the Crowds-based metric[3] (Reiter & Rubin, 1997), anonymity is measured on a continuum, including the points *possible innocence* (the probability that a user is not the sender is not negligible), *probable innocence* (the probability that a user is a sender $\geq 1/2$), and *beyond suspicion* (the user is not more likely than any other user to be the sender). The analysis is based on the communication patterns in Crowds, and the result is a probability depending on the anonymity set size and the number of corrupted users.
Entropy-based metrics
In entropy-based metrics (Diaz, Seys, Claessens, & Preneel, 2002; Serjantov & Danezis, 2002), each user is first assigned with a probability of being the sender of a message. The entropy regarding which user sent the message is then calculated using Shannon's theories (Shannon, 1948). The resulting degree is system-wide and may change depending on, for example, changes in the attacker's knowledge. Diaz et al. solely bases their analysis on the probability distributions (equally distributed probabilities → max degree of anonymity), while in Serjantov and Danezis metric, a large anonymity set contribute positively to the degree of anonymity.

distance vector routing (AODV) (Perkins & Royer, 1999).

- In DSR, during route discovery[4] the route request (RREQ) includes the IP addresses of the sender and receiver in plain. The IPs are also disclosed by the route reply (RREP) message. During data transfer, the path between the sender and receiver is included in plain in the packet headers.
- Also in AODV, the RREQ and RREP messages disclose the sender and receiver IP addresses. Also, routing data at each node in an active path discloses the receiver IP.

This situation applies for virtually any standard routing protocol. So far, two methods for enabling anonymous communication in mobile ad hoc networks have been proposed: *anonymous routing protocols* and *anonymous overlay networks.* They are explained in the next sections.

Anonymous Routing Protocols

An anonymous routing protocol replaces the standard routing protocol with a protocol preserving anonymity (see Figure 4). Anonymous routing protocols normally include building blocks for *anonymous neighborhood authentication, anonymous route discovery,* and *anonymous data transfer.* The first phase is not always included; instead many approaches assume that other mechanisms offer this service.

During anonymous neighborhood authentication, nodes establish trust relationships with their *neighbors* (i.e., nodes within one-hop distance). "Trust" implies that the nodes prove mutual possession of some valid identifiers, such as certificates, pseudonyms, public/private key-pairs, or combinations thereof.

The task of anonymous route discovery is to establish an anonymous path between the sender and receiver. Sender anonymity is often achieved through layered encryption. Sometimes, receiver anonymity is enabled by invisible implicit addressing, meaning in this context that a challenge is included in the RREQ that only the receiver can decrypt[5].

The main disadvantage with invisible implicit addressing is that all nodes receiving the RREQ must try to decrypt the challenge, resulting in considerable overhead (especially as the RREQ reaches all nodes). When the RREP is propagated back to the sender on the path created by the corresponding RREQ message, *visible implicit addressing* (Pfitzmann & Waidner, 1987) is often used to hinder nodes other than the sender from

Figure 4. Anonymous routing protocol

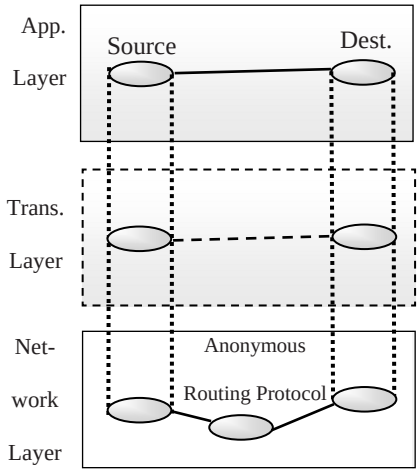
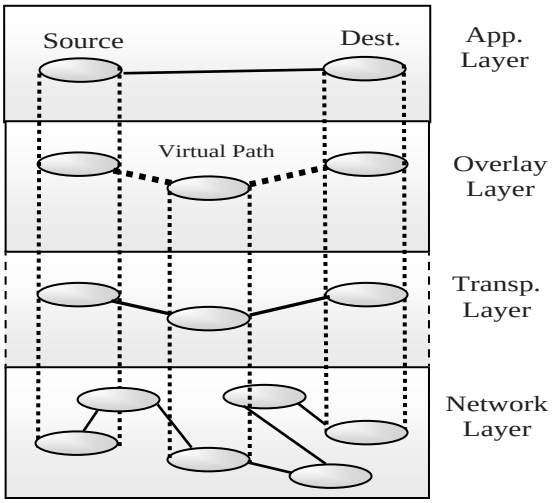


Figure 5. Anonymous overlay network



matching RREP messages with corresponding RREQ messages. This is often enabled by including sequence numbers in the RREP and RREQ so that only the sender can conclude that the sequence number of a given RREP corresponds to an earlier sent out RREQ.

During anonymous data transfer, data messages are sent along the paths created during route discovery. Only protocols that use *source routing* can apply layered encryption, as the sender in this case needs to decide the full path. Else, link-to-link encryption, possibly combined with end-to-end encryption, is normally used.

Anonymous Overlay Networks

In mobile ad hoc networks, anonymous overlay networks are normally deployed above the routing or transport layer (see Figure 5), where they can use services from the standard routing protocol (e.g., finding a route to the next node in the path) or the transport layer (e.g., reliable data delivery).

Anonymous overlay networks can be divided into the following phases: *group buildup, path construction,* and *data transfer.*

During group buildup, the user base of the overlay network is populated. One strategy for group buildup is to assign this task to one or more *directory servers,* where a set of nodes (or at least one node) must act as a directory server (Martucci, Andersson, & Fischer-Hübner, 2006). Similarly as in anonymous routing protocols, virtual path setting and data transfer are either based on layered encryption, or link-to-link encryption combined with end-to-end encryption.

Comparison between Anonymous Routing Protocols and Anonymous Overlay Networks

In Table 2 we summarize the respective pros and cons with anonymous routing protocols and anonymous overlay networks.

Survey of Anonymous Communication Mechanisms for Ad Hoc Networks

The survey is divided into two parts: one part for anonymous routing protocols and one for anonymous overlay networks[6]. Before the survey,

Table 2. Pros and cons with anonymous routing protocols and anonymous overlay networks

Advantages with Anonymous Routing Protocols
They make it possible to control already on the routing level what information is being disclosed during routing. Yet, this does not exclude the possibility that additional efforts may be needed in upper layers. Also, most approaches use the shortest path between the sender and receiver.
Disadvantages with Anonymous Routing Protocols
The replacement of the standard routing protocol; this will likely decrease the user base, which degrades anonymity according to many metrics. Besides, nodes may be exposed if a connection-oriented transport layer is used above the anonymous routing protocol, as they establish direct connections between nodes.
Advantages with Anonymous Overlay Networks
Flexibility; an anonymous overlay network is independent of the routing protocol and, further, compatible with applications expecting services from for example, a reliable transport layer.
Disadvantages with Anonymous Overlay Networks
The performance can be expected to be slightly worse as messages are detoured through a set of overlay nodes, instead of being transmitted on the shortest route between the sender and recipient.

however, we list the evaluation criteria against which the mechanisms included in the survey are evaluated.

Evaluation Criteria

Six requirements were defined by Andersson et al. (2005) that an anonymous overlay network should meet to be suitable for mobile ad hoc networks. These requirements are general enough to be suitable for providing the criteria against which the mechanisms surveyed in this chapter are evaluated. They are listed below:

R1. The anonymous communication mechanism must scale well. It should perform well also with a large number of participants.

R2. The anonymous communication mechanism must provide strong anonymity properties. We examine how the studied approaches resist an attacker model including a global observer[7], path insiders, other network nodes, and the receiver.

R3. The anonymous communication mechanism must be fair regarding the distribution of workload among the nodes. The workload should be equally distributed (and nodes should not be forced to spend a lot of resources on behalf of others). Else, incentives should be given for accepting a higher workload.

R4. The anonymous communication mechanism must provide acceptable performance. It should be lightweight (e.g., generate few messages and avoid public key operations). We evaluate whether the studied approaches presents arguments indicating a good performance. We also evaluate whether there are strong assumptions that could hamper performance.

R5. The anonymous communication mechanism must employ a peer-to-peer paradigm (P2P) model. There should be no dependence on central hardware/services, or at least, it should be minimized. We also study whether there are some implicit requirements for centralized services that are hidden by strong assumptions.

R6. The anonymous communication mechanism must handle a dynamic topology. It must tolerate that nodes are frequently entering or leaving the network.

In the survey, we grade the approaches according to which degree they satisfy these requirements: ●●● = the requirement is satisfied to a high degree; ●● = … is satisfied to a medium degree; ● = … is satisfied to a low degree; and ○ = … is violated. Regarding the grading of R2, the approaches are graded according to which degree they provide anonymity against each item in the assumed attacker model (see R2).

Survey of Anonymous Routing Protocols

In this section, we survey a variety of prominent anonymous routing protocols proposed in recent years. The ratings of the mechanisms are listed in table-form in the next section.

Anonymous Dynamic Source Routing Protocol (AnonDSR)

AnonDSR (Song, Korba, & Yee, 2005) is a source routing protocol using invisible implicit addressing for route discovery. The RREP is created as a message onion. Both the sender and recipient know the intermediary nodes in the path. Data messages are sent as message onions on bidirectional paths. AnonDSR includes a security parameter establishment (SPE) protocol for exchanging security parameters prior to route discovery, which contains a major flaw (see R2).

R1. As the SPE protocol is used to establish shared secrets between sender and receivers, the issues regarding it (see R2) may hamper scalability.

R2. The SPE protocol broadcasts the IDs of the senders and receivers in plain. If used, AnonDSR provides merely confidentiality. If not used, AnonDSR provides sender and receiver anonymity against observers, path insiders, and network nodes.
AnonDSR changes the message appearance at intermediary hops. Yet, a global observer may correlate the RREQ sizes or trace data flows in the network.

R3. During route discovery, nodes spend energy to assess whether they are the intended receiver. Intermediary nodes must perform public key encryptions.

R4. The range of the nodes and the network size is not specified in the performance simulation of AnonDSR, and only route discovery is evaluated while data transfer and node mobility are not considered. Also, as implicit addressing with public key cryptography is used, AnonDSR cannot be expected to provide high performance.

R5. No special nodes needed, and thus AnonDSR adheres well to the P2P paradigm.

R6. AnonDSR does not support rebuilding of broken paths. Also, the insecurities in the SPE protocol may cause problems for new nodes joining the network that wish to establish security parameters with existing nodes.

Secure Distributed Anonymous Routing Protocol (SDAR)

SDAR (Boukerche, El-Khatib, Xu, & Korba, 2004) is a source routing protocol enabling a system for managing trust: nodes associate their neighbors with a trust level based on past behavior. Invisible implicit addressing is used to hide the receiver identity in the RREQ. The RREP and data messages are sent as message onions.

R1. SDAR can be expected to scale badly as every node in the network must perform three public key operations per received RREQ message.

R2. SDAR offers sender and receiver anonymity against observers and other network nodes.
SDAR alters messages appearance and applied padding to thwart global observers. Still, only nodes assumed to forward RREQ/RREP packets do so, others drop them.

R3. It is not specified whether the certificate authority (CA) is a central service or distributed among the nodes. When processing RREQ packets, all nodes must perform one public key encryption, one public key decryption, and one signature generation.

R4. There are serious performance issues in SDAR. For instance, every node must perform must perform three public key operations for each RREQ it forwards.

R5. The existence of a CA (or similar) is assumed for distributing public keys. It is not specified how it would be implemented.

R6. We predict that the trust management system in SDAR would suffer in a dynamic topology; it would be difficult for nodes to be highly trusted as they would be ● punished for leaving the network in the midst of a communication. Also, path rebuilding in case of broken paths is not considered.

MASK

MASK (Zhang, Liu, & Lou, 2005) does not use source routing. Prior to route discovery, MASK performs anonymous neighborhood authentication, and nodes know each other by temporal pseudonyms. For performance reasons, MASK avoids invisible implicit addressing during route discovery; instead, the receiver identity is disclosed in the RREQ. After route discovery,

a sender may have multiple active paths to the receiver. End-to-end and/or link-to-link encryption is employed during data transfer, depending on the application at hand.

R1. MASK can be expected to scale well as it avoids the usage of implicit addressing. Yet, an increased node density (i.e., more neighbor nodes) may degrade performance during anonymous neighborhood authentication.
R2. MASK offer sender anonymity against path insiders, network nodes, and observers, but no receiver anonymity. MASK uses altered message appearance, random choice of paths, and per-hop message delay to harden traffic analysis during low traffic. No node forwards RREQ/RREP messages more than once.
R3 The avoidance of implicit addressing bears a positive impact on fairness.
R4. Simulation results indicate that MASK provides good performance. However, the mutual authentication between neighboring nodes was shown to be the most costly operation and in scenarios where the transmission range is small compared to the network size,
this may affect performance negatively.
R5. A trusted authority (TA) is used during the bootstrapping phase of the network.
R6. Broken paths are handled by broadcasting error packets in case of a broken path. Still, the tight synchronization scheme between neighboring nodes may lead to problems in some situations where neighboring nodes leave and join often.

Anonymous On-Demand Routing (ANODR)

ANODR (Kong, Hong, Sanadidi, & Gerla, 2005) is a source routing protocol aiming to protect privacy by avoiding persistent identifiers. Invisible implicit addressing based on symmetric encryption is used to hide the receiver identity during route discovery. The RREP is created as a message onion. During data transfer, it is not specified whether or not the data payload is encrypted.

R1. It is unclear how senders and receivers share symmetric keys. Given that they share a key, to solve the challenge in the RREQ, the receiver
may have to try all keys shared with other nodes (see R4). Further, other network nodes must try all their shared keys to conclude that they are not the intended receiver.
R2. ANODR offers sender and receiver anonymity against observers, path insiders, and networks nodes. Senders and receivers are not mutually anonymous. ANODR uses traffic mixing to thwart observers, where messages are independently and randomly delayed. Yet, traffic patters are leaked as only nodes assumed to forward the RREP does so. Further, as the payload of data messages is not altered at intermediary hops, it is trivial for a global observer to trace data traffic.
R3. Each node must spend considerable resources when forwarding RREQ packets.
R4. There are serious performance issues in ANODR (see R1).
Although ANODR has performed reasonably well in a simulation scenario, problems can be expected in a real world scenario.
R5. No special nodes are needed, and thus ANODR adheres well to the P2P paradigm.
R6. ANODR supports path rebuilding in case of broken paths. However, it is unclear how new nodes should share symmetric keys with old nodes

Discount Anonymous On-Demand Routing (Discount ANODR)

Discount ANODR (Yang, Jakobsson, & Wetzel, 2006) is a low-latency source routing protocol that

avoids invisible implicit addressing. A random time to live counter is used for RREQ/RREP messages to confuse observers (implemented by flipping a biased coin). Data are sent as message onions along unidirectional paths (i.e., a new path must be build for the reply).

R1. Discount ANODR can be expected to scale well. However, the bias of the coin flipping may have to be adapted if the geographical size of the network increases.

R2. Discount ANODR provides sender anonymity against local observers, as the coin flipping and random padding during route discovery confuse observers to a certain degree. No receiver anonymity.
Data messages are padded with random bits.

R3. There are no special nodes and no public encryption on behalf of other nodes.

R4. Discount ANODR avoids public key encryption and invisible implicating addressing.
The coin flipping may degrade performance as nodes on the shortest path may drop the RREQ, resulting in nonoptimal paths. Also, RREP packets can be lost for the same reason. Unidirectional paths also hamper performance.

R5. The nodes have to collectively administrate two values determining the bias of the coins deciding whether a node should forward a RREQ and a RREP, respectively.

R6. Discount ANODR rebuilds broken paths, but does not discuss how to collectively adapt the bias of the coin flipping when the network characteristics change.

Anonymous Routing Protocol for Mobile Ad Hoc Networks (ARM)

ARM (Seys & Preneel, 2006) aims to foil global observers by using random time-to-live values and padding for all messages. Senders and receivers share one-time pseudonyms. Invisible implicit addressing hides the receiver by including the secret pseudonym in the RREQ. The RREP is created as a message onion. Link-to-link encryption is used for data transfer.

R1. As a tight synchronization scheme is used between sender and recipients, it is assumed that senders shares keys and pseudonyms with a limited set of receivers.

R2. ARM offers sender and receiver anonymity against networks nodes, path insiders, and observers. Senders and receivers have an a-priori relationship. In ARM, data messages have a uniform size, RREQ/RREP messages are randomly padded, and RREQ/RREP/ data messages are propagated using random time-to-live values. The effectiveness of this limited dummy traffic is not formally proven.

R3. While no nodes perform public key operations, the amount of nodes forwarding RREQ/ RREP and data messages increases due to the random time-to-life values.

R4. If assuming a static environment, there are no conclusive arguments orthogonal to performance. However, all nodes in ARM generate overhead traffic.
ARM has not yet been simulated to assess the performance.

R5. There are no special nodes in ARM. In a real world scenario, central infrastructure may be required to realize the assumption that each node should possess a unique identifier; it is unclear how this would clash with the P2P paradigm.

R6. The assumption that each node establishes a broadcast key with its neighbors is problematic when considering dynamic topologies. Further, ARM does not consider path rebuilding in case of broken paths.

Distributed Anonymous Secure Routing Protocol (ASRP)

ASRP (Cheng & Agrawal, 2006) is a routing protocol not based on source routing where nodes are known by dynamic random pseudonyms. Invisible implicit addressing (based on public encryption) is used for both RREQ and RREP packets. Data messages are link-to-link and end-to-end encrypted. It is not specified whether the paths are bidirectional or unidirectional.

R1. All nodes in the network must perform two public key operations per RREQ (one private key decryption and one public key generation). This hampers scalability as the more nodes in the network, the more generated RREQ packets.

R2. Senders and receivers are not mutually anonymous as they have an a-priori relationship. Anonymity is offered against path insiders and network nodes, and ASRP alters message appearance and maintains a uniform message size to confuse attackers.

R3. All nodes spend significant resources when forwarding RREQ and RREP packets. For the RREQ, see R1. For propagation of RREP packets, all nodes on the path must perform three public key operations (one private key decryption and two public key encryptions).

R4. The performance of ASRP has not been simulated. Route discovery can be expected to offer a low performance, as public key encryption is extensively used.

R5. No special nodes are needed, and thus ASRP adheres to the P2P paradigm.

R6. Path rebuilding in case of broken paths is not considered. This means that the expensive route discovery process has to be initiated for each case of path failure.

Privacy Preserving Routing (PPR)

PPR (Capkun, Hubaux, & Jakobsson, 2004) is a proactive protocol for communication between ad hoc networks interconnected by fixed access points (AP). Nodes know each other by temporal pseudonyms. In the sender network, nodes maintain the shortest path to the AP. In the receiver's network, the AP maintain the shortest paths to the nodes. Routing consists of three parts: *uplink* (distance vector protocol), *inter-station*, and *downlink* (source routing). In uplink, a sender sends a message that reaches the AP as a message onion. In downlink, the receiver's AP send an onion to the receiver.

R1. The AP and the CA are the major points of workload aggregation in PPR, but as these are centrally offered services, PPR can be expected to scale well.

R2. PPR offers sender and receiver anonymity against observers, network nodes, and path insiders. There are no countermeasures against global observers in the senders or receivers networks, except message alteration at intermediary hops. Anonymity is quantified using the entropy-based anonymity metric (see section "On Measuring Anonymity"). There is no anonymity against the AP.

R3. Nodes do not perform special roles or execute public key operations on behalf of others.

R4. Public key encryption is only used for establishing trust relationships among neighboring nodes. The performance of PPR has not yet been simulated.

R5. PPR violates the P2P model as the existence of a CA and several AP is assumed.

R6. The existence of the AP facilitate the handling of trust and security issues in a dynamic topology. The uplink protocol is the most vulnerable part regarding routing, but it can be expected to handle dynamic topologies well.

Table 3. Summary of survey results (except R2)

Requirement	*ARM*	*AnonDSR*	*ANODR*	*SDAR*	*Discount ANODR*	*ASRP*	*MASK*	*PPR*
R1: Scalability	●	●●	●	●	●●●	●●	●●	●●●
R3: Fairness	●●	●●	●	●●	●●●	●	●●●	●●
R4: Performance	●●	●	●	●	●●	●	●●	●●
R5: P2P	●●	●●●	●●●	●●	●●	●●●	●●	○
R6: Dyn. Top.	●	●	●●	●	●●	●	●●	●●●

Table 4. Summary of anonymity requirement R2

Attacker model	*ARM*	*AnonDSR*	*ANODR*	*SDAR*	*Discount ANODR*	*ASRP*	*MASK*	*PPR*[8]
Sender – observer	●●	●	●	●	●	●	●●	●
Send. – path insider	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
Sender – net. node	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●
Sender – receiver	○	○	○	○	○	○	○	○
Rec. – observer	●●	●	●	●	○	●	○	●
Rec. - path insider	●●●	●●●	●●●	●●●	○	●●●	○	●●●
Rec. – net. node	●●●	●●●	●●●	●●●	○	●●●	○	●●●

Summary of Survey Results for Anonymous Routing Protocols

The survey results for all requirements (except R2) are summarized in Table 3. The survey results for R2 are summarized in Table 4.

SURVEY OF ANONYMOUS OVERLAY NETWORKS

In this section, we study two anonymous overlay networks for ad hoc networks: Chameleon (Martucci et al., 2006) and MRA (Jiang, Vaidya, & Zhao, 2004).

Chameleon

Chameleon can be described as a variant of Crowds adapted for mobile ad hoc networks. In Chameleon, the nodes share the responsibility of being directory servers during group buildup. Node authentication is based on certificates (the existence of a TCP (transmission control protocol)/SSL (secure socket layer) layer is assumed). Data messages are end-to-end and link-to-link encrypted or only link-to-link encrypted.

R1. The load on each node is approximately constant as the size of the network grows. However, if too few directory servers are used, this may put a limit on scalability.

R2. Chameleon offers sender anonymity against receivers and sender and receiver anonymity against local observers and malicious nodes.
The degree of anonymity is quantified by the Crowds-based metric (see "On Measuring Anonymity").

R3. A small subset of the nodes must act as directory servers. It is suggested that nodes take turns in acting as the directory servers.

R4. Chameleon is based on light-weight encryption.
However, the performance of Chameleon has not yet been assessed through simulation.

R5. Chameleon generally follows the P2P paradigm. However, nodes are assumed to possess certificates obtained in advance and the global probability deciding the expected path length has to be administrated collectively by the nodes

R6. Chameleon repairs broken paths at the point of breach, rather that rebuilding the whole path. Without redundancy, vanishing directory servers may be a problem.

Mix Route Algorithm (MRA)

MRA applies traffic mixing[9] (Chaum, 1981) in a mobile ad hoc scenario. A subset of the nodes acts as mixes, which constitute the virtual paths. Each node assigns a mix as its *dominator mix*. A RREQ is sent to the receiver via the sender's dominator mix, triggering the receiver to register at its dominator mix with a DREG (dominator registration) message. Each mix periodically broadcasts RUPD (route update) messages containing its registered receivers and a path field, which is updated as the RUPD propagates through the network. When it reaches the sender, it contains the path to the receiver.

R1. Scalability may be hampered if the mix set is static in a growing network.

R2. As the min path length is one, a mix may learn the identity of both the sender and receiver. The first mix always learns the sender ID.
Receiver anonymity is in doubt as all mixes broadcast information in the network about which receivers it is currently providing services for (i.e., the RUPD messages).

R3. Incentives for the costly operating of mixes are left as a future research problem.

R4. MRA is based on pubic-key cryptography. Basing MRA on symmetric cryptography is left as future research. Results from a performance simulation are presented, but only different mix settings are compared.

R5. No central services are needed. Still, establishing trust between mixes and other nodes are left as future research. This may require aid from external trusted nodes.

R6. If the sender or dominator mix move, the sender may have to switch dominator mix. If the mix set is small, problems may arise regarding the mix advertisement as nodes only retransmit advertisement messages from their dominator mixes.

Summary of Survey Results for Anonymous Overlay Networks

The results from the survey are summarized in Table 5.

DISCUSSION

From the survey, we can make the following observations:

1. **It is difficult to protect against a global eavesdropper.** None of the studied approaches implement powerful and proven countermeasures against global observers. We believe that it is an open research problem regarding how to enable such countermeasures while at the same time offering an acceptable level of performance in mobile ad hoc networks[12].
2. **It is difficult to implement invisible implicit addressing efficiently.** There is a clear trade-off between on the one hand enabling receiver anonymity by using invisible implicit addressing and on the other hand satisfying the fairness, dynamic, and scalability requirements. The proposals using invisible implicit addressing either use costly public key cryptography (e.g., AnonDSR, SDAR, ASRP) or avoid public key operations at the cost of including strong assumptions regarding in beforehand mutual distribution of secrets (e.g., ARM, ANODR).

3. **It is straightforward to hide the identity of the sender from other network nodes.** This is probably because most of the approaches use classical techniques for hiding the identity of the sender, such as layered encryption, that have been used before in other contexts.
4. **No anonymous routing protocol implements sender anonymity towards the receiver.** Hiding the sender identity during route discovery would require a mechanism for hiding the propagation of the RREP messages similar (and equally costly as) to the invisible implicit addressing schemes used for hiding the propagation of the RREQ messages.

FUTURE TRENDS

A *Sybil attack* (Douceur, 2002) implies one attacker forging multiple identifiers in the network to control an unbalanced portion of the network. Sybil attacks can undermine security in, for instance, mobile ad hoc networks based on reputation schemes or threshold cryptography (Piro, Shields, & Levine, 2006). Douceur has showed that *preventing* Sybil attacks is practically impossible as it requires a TTP (trusted third party) to manually assert that each identity corresponds to only one logical entity in the network. Yet, during the years, and recently also for mobile ad hoc networks, many approaches for *detecting* Sybil attacks have been proposed. In this section, we discuss why Sybil attacks threaten anonymity in ad hoc networks, and discuss some proposed countermeasures.

The Sybil Attack in Mobile Ad Hoc Networks

Mobile ad hoc networks are highly susceptible to Sybil attacks because of, for instance, the lack of reliable network or data link identifiers, and the absence of a trusted entity capable of vouching for the one-to-one binding between physical devices and logical network identifiers. This may give the impression that ad hoc nodes are naturally anonymous as nodes could confuse observers by regularly changing their {IP, MAC} pairs. Although this may prevent long-term tracking, other problems may arise. For instance, when there is a need to identify a node offering a specific service, a rouge node could easily impersonate this service. The absence of reliable network identifiers may also disrupt routing, as a rouge user could announce false information using multiple {IP, MAC} pairs. Also, as senders and receivers establish direct connections, they are still vulnerable to traffic analysis and physical layer oriented attacks (Capkun et al., 2004).

However, the Sybil attack also poses a threat against anonymous routing protocols and anonymous overlay networks. For both approaches, the anonymity set denotes the user base. There are some differences though. In an anonymous overlay network, the anonymity set is used as a pool of nodes serving as an input parameter to the path

Table 5. Summary of survey results (left) and summary of anonymity requirement R2 (right)

Requirement	*Chameleon*	*MRA*
R1: Scalability	●●	●●
R3: Fairness	●●	●
R4: Performance	●●	●
R5: P2P	●●	●●
R6: Dyn. Top.	●●	●●

Attacker model	*Chameleon*	*MRA*
Sender – observer	●	●●
Send. – path insider	●●●	●●●/○[10]
Sender – net. node	●●●	●●●
Sender – receiver	●●●	○
Rec. – observer	●	●
Rec. - path insider	○	●/○[11]
Rec. – net. node	●●●	●

creation algorithm. Polluting the anonymity set with many Sybil identities might yield a path only containing Sybil identities. If this happens, the attacker can easily break anonymity by linking the sender to the receiver. In an anonymous routing protocol, however, each node only stepwise extends the path to another node within a single-hop distance, until the receiver is reached. Thus, the locations of the nodes play a more important role here, and as all Sybil identities share the same location, it is difficult for the attacker to force the creation of paths in which it controls all nodes.

Thus, the Sybil attack poses a greater threat to anonymity in anonymous overlay networks compared to anonymous routing, although it still poses a great threat to other security properties for anonymous routing.

Mechanisms for Detecting the Sybil Attack in mobile Ad Hoc Networks

In this section, we describe two recent proposals for thwarting Sybil attacks in mobile ad hoc networks.

- The fact that Sybil nodes in mobile ad hoc networks naturally travel together in clusters can be used for detecting Sybil attacks (Piro et al., 2006). Piro et al. propose a detection mechanism in which each node records all encountered {IP, MAC} pairs. If a user repeatedly observes a set of {IP, MAC} pairs sharing the same location, there is an increased likelihood that these {IP, MAC} pairs represent Sybil nodes. One drawback with this strategy is that it is unclear how to prevent a detected attacker from generating new {IP, MAC} pairs and relaunch a new attack later, as there is no underlying long-term identity that can be blocked from the system.
- Another strategy is to cryptographically guarantee a one-to-one mapping between all temporal network identifiers seen in a particular network and corresponding certified long-term identifiers (Martucci et al., 2008). To tailor this approach for ad hoc networks, the nodes must be able to assert the validity of the temporal identifiers without having to interact with the TTP. Further, to protect privacy, only the TTP should be able to link a temporal identifier to the corresponding long-term identifier and there should be unlinkability between temporal identifiers used in different contexts. The fact that you need reliable identifiers to protect against the Sybil attack and to provide reliable anonymous communication has been labeled as the *identity-anonymity paradox* (Martucci et al., 2006).

CONCLUSION

In mobile ad hoc networks, anonymous communication can either be enabled by anonymous routing protocols or anonymous overlay networks. Currently, anonymous routing is the most popular approach, although future requirements, such as flexibility regarding the applications, may raise the need for anonymous overlay networks.

We evaluated commonly proposed anonymous routing protocols and anonymous overlay networks for mobile ad hoc networks against a set of evaluation criteria and showed that a number of research challenges remain. For instance, it is difficult to offer receiver anonymity without using a complex and performance-hampering invisible implicit addressing scheme, and it is further difficult to protect against global observers.

Finally, we introduced Sybil attacks, a notorious threat to all computer networks, including mobile ad hoc networks. We expect that the area of enabling reliable identifiers in a privacy-friendly manner is an interesting future research area.

REFERENCES

Andersson, C., Martucci, L. A., & Fischer-Hübner, S. (2005). Requirements for privacy: Enhancements in mobile ad hoc networks. In *Proceedings of the 3rd German Workshop on Ad Hoc Networks (WMAN 2005)* (pp. 344-348). Gesellschaft für Informatik (GI).

Boukerche, A., El-Khatib, K., Xu, L., & Korba, L. (2004). A novel solution for achieving anonymity in wireless ad hoc networks. In *Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems* (pp. 30-38).

Capkun, S., Hubaux, J. P., & Jakobsson, M. (2004). *Secure and privacy-preserving communication in hybrid ad hoc networks* (EPFL-IC Tech. Rep. No. IC/2004/10). Lausanne, Switzerland: Laboratory for Computer Communications and Applications (LCA)/Swiss Federal Institute of Technology Lausanne (EPFL).

Chaum, D. (1981). David Chaum: Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM, 24*(2), 84-88.

Cheng, Y., & Agrawal, D. P. (2006). *Distributed anonymous security routing protocol in wireless mobile ad hoc networks.* Paper presented at the OPNETWORK 2005.

Corson, M. S., & Macker, J. (1999). *Mobile ad hoc networking (MANET): Routing protocol performance issues and evaluation considerations* (RFC-2501), Internet RFC/STD/FYI/BCP Archives.

Dıaz, C., Seys, S., Claessens, J., & Preneel, B. (2002). Towards measuring anonymity. In *Proceedings of the Workshop on Privacy Enhancing Technologies (PET 2002)* (LNCS 2482). Springer-Verlag.

Douceur, J. R. (2002). The Sybil attack. In P. Druschel, F. Kaashoek, & A. Rowstron (Eds.), *Peer-to-peer Systems: Proceedings of the 1st International Peer-to-Peer Systems Workshop (IPTPS)* (pp. 251-260). Springer-Verlag.

Goldschlag, D. M., Reed, M. G., & Syverson, P. F. (1996). Hiding routing information. *Information hiding* (LLNCS 1174, pp. 137-150). Springer-Verlag.

Jiang, S., Vaidya, N. H., & Zhao, W. (2004). A mix route algorithm for mix-net in wireless mobile ad hoc networks. In *Proceedings of the 1st IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS 2004).*

Johnson, D. B., & Maltz, D. A. (1996). Dynamic source routing in ad hoc wireless networks. In *Computer Communications Review: Proceedings of the ACM SIGCOMM'96 Conference on Communications Architectures, Protocols and Applications.*

Kong, J., Hong, X., Sanadidi, M. Y., & Gerla, M. (2005). Mobility changes anonymity: Mobile ad hoc networks need efficient anonymous routing. In *Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC 2005).*

Levine, B. N, Shields, C., & Margolin, N. B. (2006). *A survey of solutions to the Sybil attack* (Tech. Rep. 2006-052). Amherst, MA: University of Massachusetts Amherst.

Martucci, L. A., Andersson, C., & Fischer-Hübner, S. (2006). Chameleon and the identity-anonymity paradox: Anonymity in mobile ad hoc networks. *In Short-Paper Proceedings of the 1st International Workshop on Security (IWSEC 2006)* (pp. 123-134).

Martucci, L., Kohlweiss, M., Andersson, C.,& Panchenko, A. (2008). Self-certified Sybil-free pseudonyms. In *1st ACM Conference on Wireless Network Security* (WiSec 2008).

Perkins, C. E. (2001). *Ad hoc networking.* Addison-Wesley Professional.

Perkins, C. E., & Royer, E. M. (1999). Ad-hoc on demand distance vector routing. In *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications (WMCSA '99).*

Pfitzmann, A., & Hansen, M. (2006) *Anonymity, unlinkability, unobservability, pseudonymity, and identity management: A consolidated proposal for terminology v0.27.* Retrieved April 25, 2007, from http:// dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.28.doc

Pfitzmann, A., & Waidner, M. (1987). Networks without user observability. *Computers and Security, 6*(2), 158-166.

Piro, C., Shields, C., & Levine, N. L. (2006). Detecting the Sybil attack in mobile ad hoc networks. In *Proceedings of the IEEE/ACM International Conference on Security and Privacy in Communication Networks (SecureComm).*

Reiter, M., & Rubin, A. (1997). *Crowds: Anonymity for Web transactions.* Technical report No. 97-15, DIMACS (pp. 97-115).

Serjantov, A., & Danezis, G. (2002). Towards and information theoretic metric for anonymity. In *Proceedings of the Workshop on Privacy Enhancing Technologies (PET 2002)* (LNCS 2482) .Springer-Verlag.

Seys, S., & Preneel, B. (2006). ARM: Anonymous routing protocol for mobile ad hoc networks. In *Proceedings of International Workshop on Pervasive Computing and Ad Hoc Communications (PCAC '06).*

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27,* 379-423.

Song, R., Korba, L., & Yee, G. (2005). AnonDSR: Efficient anonymous dynamic source routing for mobile ad-hoc networks. In *Proceedings of the 2005 ACM Workshop on Security of Ad Hoc and sensor Networks (SASN 2005)* (pp. 32-42). Alexandria.

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), 557-570.

Yang, L., Jakobsson M., & Wetzel, S. (2006). Discount anonymous on demand routing for mobile ad hoc networks. In *Proceedings of SecureComm 2006,* Baltimore, MD.

Zhang, Y., Liu, W., & Lou, W. (2005). Anonymous communication in mobile ad hoc networks. In *Proceedings of the 24th Annual Joint Conference of the IEEE Communication Society (INFOCOM 2005),* Miami.

KEY TERMS

Anonymity: The state of being not identifiable within a set of subjects.

Anonymity Metrics: Metrics for quantifying the degree of anonymity in a scenario.

Mobile Ad Hoc Network: Networks constituted of mobile devise which may function without the help of central infrastructure or services.

Privacy: The right to informational self-determination, that is, individuals must be able to determine for themselves when, how, to what extent, and for what purpose personal information about them is communicated to others.

Receiver Anonymity: Implies that a message cannot be linked to the receiver.

Sender Anonymity: Means that a message cannot be linked to the sender.

Unlinkability: If two items are unlinkable, they are no more or less related after an attacker's observation than they are related concerning the attacker's a-priori knowledge.

END NOTES

[1] As devices in ad hoc networks are responsible for their own services, including security and routing, protocols for anonymous communication for wired networks are not suitable for ad hoc networks, not even those based on the peer-to-peer paradigm (P2P) (Andersson et al., 2005).

[2] This method is sometimes also called telescope encryption. A public key based version of the method was initially introduced by Chaum (1981). Onion Routing, which only uses public key encryption for setting the path, and then relies on symmetric encryption, was later proposed by Goldschlag, Reed, and Syverson (1996).

[3] The Crowds-based metric was developed for Crowds, but has since been used in other contexts.

[4] This denotes the process of setting a path between the sender and a receiver. First, the sender floods a route request (RREQ) into the network, which triggers the sending of a route reply (RREP) from the receiver to the sender. During the propagation of the RREQ and RREP, respectively, the path is interactively formed.

[5] In the context of mobile ad hoc networks, this method is often referred to as a global trapdoor.

[6] In the survey, we omit approaches relying on the existence of either a positioning device (e.g., GPS) in the mobile devices or a location server in the mobile ad hoc network.

[7] A global observer is an observer that is capable of observing all networks traffic in the whole network.

[8] Note that no anonymity is provided against the access points (not included in attacker model).

[9] Batching and reordering traffic to hide the correlation between incoming and outgoing traffic.

[10] No sender anonymity if path length is one.

[11] No receiver anonymity against last mix on the path.

[12] It is commonly believed that omnipresent protection against a global observer can only be achieved if all nodes transmit a constant flow of traffic, requiring massive usage of dummy traffic.

This work was previously published in Handbook of Research on Wireless Security, edited by Y. Zhang; J. Zheng; M. Ma, pp. 431-448, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.10
Integrity Protection of Mobile Agent Data

Sheng-Uei Guan
Brunel University, UK

INTRODUCTION

One hindrance to the widespread adoption of mobile-agent technology is the lack of security. Security will be the issue that has to be addressed carefully if mobile agents are to be used in the field of electronic commerce. SAFER (secure agent fabrication, evolution and roaming) is a mobile-agent framework that is specially designed for the purpose of electronic commerce (Guan & Hua, 2003; Guan, Zhu, & Maung, 2004; Zhu, Guan, Yang, & Ko, 2000). Security has been a prime concern from the first day of our research (Guan & Yang, 2002; Yang & Guan, 2000). By building strong and efficient security mechanisms, SAFER aims to provide a trustworthy framework for mobile agents to assist users in conducting mobile or electronic-commerce transactions.

Agent integrity is one such area crucial to the success of agent technology (Wang, Guan, & Chan, 2002). Despite the various attempts in the literature, there is no satisfactory solution to the problem of data integrity so far. Some of the common weaknesses of the current schemes are vulnerabilities to revisit attacks, when an agent visits two or more collaborating malicious hosts during one roaming session, and illegal modification (deletion or insertion) of agent data. The agent monitoring protocol (AMP; Chionh, Guan, & Yang, 2001), an earlier proposal under SAFER to address agent data integrity, does address some of the weaknesses in the current literature. Unfortunately, the extensive use of PKI (public-key infrastructure) technology introduces too much overhead to the protocol. Also, AMP requires the agent to deposit its data collected to the agent owner or butler before it roams to another host. While this is a viable and secure approach, the proposed approach, Secure Agent Data Integrity Shield (SADIS), will provide an alternative by allowing the agent to carry the data by itself without depositing them (or the data hash) onto the butler.

Besides addressing the common vulnerabilities of current literature (revisit attacks and data-modification attacks), SADIS also strives to achieve maximum efficiency without compromising security. It minimizes the use of PKI technol-

ogy and relies on symmetric key encryption as much as possible. Moreover, the data encryption key and the communication session key are both derivable from a key seed that is unique to the agent's roaming session in the current host. As a result, the butler can derive the communication session key and data encryption key directly. Another feature in SADIS is strong security.

Most of the existing research works focus on detecting integrity compromise (Esparza, Muñoz, Soriano, & Fomé, 2006) or bypassing integrity attacks by requiring the existence of a cooperating agent that is carried out within a trusted platform (Ouardani, Pierre, & Boucheneb, 2006). However, these works neglect the need to identify the malicious host. With SADIS, the agent butler will not only be able to detect any compromise to data integrity, but will identify the malicious host effectively.

BACKGROUND

Agent data integrity has been a topic of active research in the literature for a while. SADIS addresses the problem of data integrity protection via a combination of techniques discussed by Borselius (2002): execution tracing, encrypted payload, environmental key generation, and undetachable signature.

One of the recent active research works is the security architecture by Borselius, Hur, Kaprynski, and Mitchell (2002). Their architecture aims at defining a complete security architecture designed for mobile-agent systems. It categorizes security services into the following: agent management and control, agent communications service, agent security service, agent mobility service, and agent logging service. SADIS addresses the agent communication service as well as agent security services (integrity protection), while previous research on SAFER addresses agent mobility service.

While many of the security services are still under active research, the security mechanisms for protecting agents against malicious hosts were described by Borselius, Mitchell, and Wilson (2001). The paper proposes a threshold scheme to protect mobile agents. Under the mechanism, a group of agents is dispatched to carry out the task, with each agent carrying a vote. Each agent is allowed to contact a merchant independently and gathers a bid based on the given criteria. Each agent votes for the best bid (under a trading scenario) independently. If more than n out of m ($m > n$) agents vote for the transaction, the agent owner will agree to the transaction.

Such a mode of agent execution effectively simplifies agent roaming by allowing one agent to visit one merchant only. While the approach avoids the potential danger of having the agent compromised by the subsequent host, it does not employ a mechanism to protect the agent against the current host. Most important of all, the threshold mechanism's security is based on the probability that no more than n hosts out of m are malicious. In another words, the security is established based on probability. Different from this approach, SADIS's security is completely based on its own merits without making any assumption about probability of hosts being benign or malicious. This is because the author believes that in an e-commerce environment, security should not have any dependency on probability.

Other than the research by Borselius (2002), Borselius et al. (2002), and Borselius et al. (2001), there are related research works in the area. One such research work on agent protection is SOMA (Secure and Open Mobile Agent) developed by Corradi, Cremonini, Montanari, and Stefanelli (1999). It is a Java-based mobile-agent framework that provides for scalability, openness, and security on the Internet. One of the research focuses of SOMA is to protect the mobile agent's data integrity. To achieve this, SOMA makes use of two mechanisms: the multihop (MH) protocol and trusted third party (TTP) protocol. The MH

protocol works as follows. At each intermediate site the mobile agent collects some data and appends them to the previous ones collected. Each site must provide a short proof of the agent computation, which is stored in the agent. Each proof is cryptographically linked with the ones computed at the previous sites. There is a chaining relation between proofs. When the agent moves back to the sender, the integrity of the chained cryptographic proofs is verified, allowing the sender to detect any integrity violation.

The advantage of the MH protocol is that it does not require any trusted third party or even the agent butler for its operation. This is a highly desirable feature for an agent integrity protection protocol. Unfortunately, the MH protocol does not hold well against revisit attacks when the agent visits two or more collaborating malicious hosts during one roaming session (Chionh et al., 2001). Roth (2001) provides more detailed descriptions on potential flaws of the MH protocol.

Another agent system that addresses data integrity is Ajanta (Tripathi, 2002). Ajanta is a platform for agent-based application on the Internet developed in the University of Minnesota. It makes use of an append-only container for agent data integrity protection. The main objective is to allow a host to append new data to the container but prevents anyone from modifying the previous data without being detected. Similar to the MH protocol, such an append-only container suffers from revisit attacks.

From these attacks in existing research, the importance of protecting agent itinerary is obvious. In SADIS, the agent's itinerary is implicitly updated in the agent butler during key seed negotiation. This prevents any party from modifying the itinerary recorded on the butler and guards against all itinerary-related attacks.

There is one recent research work on agent data integrity protection called the One-Time Key Generation System (OKGS) being studied in Kwang-Ju Institute of Science and Technology, South Korea (Park, Lee, & Lee, 2002). OKGS does protect the agent data against a number of attack scenarios under revisit attacks, such as data-insertion attacks and data-modification attacks, to a certain extent. However, it does not protect the agent against deletion attacks as two collaborating malicious hosts can easily remove roaming records in between them.

Inspired by OKGS's innovative one-time encryption key concept, SADIS will extend this property to the communication between agent and butler as well. Not only the data encryption key is one-time, but the communication session key as well. Using efficient hash calculations, the dynamic communication session key can be derived separately by the agent butler and the agent with minimum overhead. Despite the fact that all keys are derived from the same session-based key seed, SADIS also ensures that there is little correlation between these keys. As a result, even if some of the keys are compromised, the key seed will still remain secret.

PROTECTION OF AGENT DATA INTEGRITY

SADIS is designed based on the SAFER framework. The proposal itself is based on a number of assumptions that were implemented under SAFER. First, entities in SAFER, including agents, butlers, and hosts, should have globally unique identification number (IDs). These IDs will be used to uniquely identify each entity. Second, each agent butler and host should have a digital certificate that is issued by a trusted certificate authority (CA) under SAFER. Each entity with a digital certificate will be able to use the private key of its certificate to perform digital signatures and, if necessary, encryption. Third, while the host may be malicious, the execution environment of mobile agents should be secure and the execution integrity of the agent should be maintained. This assumption is made because protecting the agent's execution environment is

a completely separate area of research that is independent of this chapter. Without a secure execution environment and execution integrity, none of the agent data protection scheme will be effective. The last assumption is that entities involved are respecting and cooperating with the SADIS protocol. Finally, SADIS does not require the agent to have a predetermined itinerary. The agent is able to decide which host is the next destination independently.

Key Seed Negotiation Protocol

When an agent first leaves the butler, the butler will generate a random initial key seed, encrypt it with the destination host's public key, and deposit it into the agent before sending the agent to the destination host. It should be noted that agent transmission is protected by the supervised agent transport protocol (Guan & Yang, 2002). Otherwise, a malicious host ("man in the middle") can perform an attack by replacing the encrypted key seed with a new key seed and encrypt it with the destination's public key. In this case, the agent and the destination host will not know the key seed has been manipulated. When the agent starts to communicate with the butler using the wrong key seed, the malicious host can intercept all the messages and reencrypt them with the correct key derived from the correct key seed and forward them to the agent butler. In this way, a malicious host can compromise the whole protocol.

The key seed carried by the agent is session based; it is valid until the agent leaves the current host. When the agent decides to leave the current host, it must determine the destination host and start the key seed negotiation process with the agent butler.

The key seed negotiation process is based on the Diffie-Hellman (DH) key exchange protocol (Diffie & Hellman, 1976) with a variation. The agent will first generate a private DH parameter *a* and its corresponding public parameter *x*. The value *x*, together with the ID of the destination host, will be encrypted using a communication session key and sent to the agent butler.

The agent butler will decrypt the message using the same communication session key (derivation of communication session key will be discussed later in the section). It, too, will generate its own DH private parameter *b* and its corresponding public parameter *y*. With the private parameter *b* and the public parameter *x* from the agent, the butler can derive the new key seed and use it for communications with the agent in the new host. Instead of sending the public parameter *y* to the agent as in normal DH key exchange, the agent butler will encrypt the value *y*, host ID, agent ID, and current time stamp with the destination host's public key to get message *M*. Message *M* will be sent to the agent after encrypting with the communication session key.

$$M = E(y + \text{host ID} + \text{agent ID} + \text{time stamp}, H_{pubKey})$$

At the same time, the agent butler updates the agent's itinerary and sends it to the agent. When the agent receives the double-encrypted DH public parameter *y*, it can decrypt with the communication session key.

Subsequently, the agent will store *M* into its data segment and requests the current host to send itself to the destination host using the agent transport protocol (Guan & Yang, 2002).

On arriving at the destination host, the agent will be activated. Before it resumes normal operation, the agent will request the new host to decrypt message *M*. If the host is the right destination host, it will be able to use the private key to decrypt message *M* and thus obtain the DH public parameter *y*. As a result, the decryption of message *M* not only completes the key seed negotiation process, but also serves as a means to authenticate the destination host. Once the message *M* is decrypted, the host will verify that the agent ID in the decrypted message matches the incoming agent, and the host ID in the decrypted message matches that of the current host.

With the plain value of *y*, the agent can derive the key seed by using its previously generated private parameter *a*. With the new key seed derived, the key seed negotiation process is completed. The agent can resume normal operation in the new host.

Whenever the agent and the butler need to communicate with each other, the sender will first derive a communication session key using the key seed and use this communication session key to encrypt the message. The receiver can make use of the same formula to derive the communication session key from the same key seed to decrypt the message.

The communication session key K_{CSK} is derived using the formula below.

K_{CSK} = Hash(key_seed + host ID + seqNo)

The sequence number is a running number that starts with 1 for each agent roaming session. Whenever the agent reaches a new host, the sequence number will be reset to 1. Given the varying communication session keys, if one of the messages is somehow lost without being detected, the butler and agent will not be able to communicate afterward. As a result, SADIS makes use of TCP/IP (transmission-control protocol/Internet protocol) as a communication mechanism so that any loss of messages can be immediately detected by the sender. In the case of an unsuccessful message, the sender will send ping messages to the recipient in plain format until the recipient or the communication channel recovers. Once the communication is reestablished, the sender will resend the previous message (encrypted using the same communication session key).

When the host provides information to the agent, the agent will encrypt the information with a data encryption key K_{DEK}. The data encryption key is derived as follows.

K_{DEK} = Hash(key_seed + host ID)

Data Integrity Protection Protocol

The key seed negotiation protocol lays the necessary foundation for integrity protection by establishing a session-based key seed between the agent and its butler. Agent data integrity is protected through the use of this key seed and the digital certificates of the hosts. Our data integrity protection protocol is comprised of two parts: chained signature generation and data integrity verification. Chained signature generation is performed before the agent leaves the current host. The agent gathers data provided by the current host d_i and construct D_i as follows.

$$D_i = E(d_i + ID_{host} + ID_{agent} + time\ stamp, k_{DEK})$$

or

$$D_i = d_i + ID_{host} + ID_{agent} + time\ stamp$$

The inclusion of the host ID, agent ID, and time stamp is to protect the data from possible replay attacks, especially when the information is not encrypted with the data encryption key. For example, if the agent ID is not included in the message, a malicious host can potentially replace the data provided for one agent with that provided for a bogus agent. Similarly, if the time stamp is not included in the message, earlier data provided to the same agent can be used at a later time to replace current data provided to the agent from the same host. The inclusion of the IDs of the parties involved and a time stamp essentially creates an unambiguous memorandum between the agent and the host.

After constructing D_i, the agent will request the host to perform a signature on the following:

ci = Sig(Di + ci-1 + IDhost + IDagent + time stamp, kpriv),

where c_0 is the digital signature on the agent code by its butler.

There are some advantages with the use of chained digital signature compared to the conventional signature approach. In the scenario where a malicious host attempts to modify the data from an innocent host *i* and somehow manages to produce a valid digital signature c_i, the data integrity would have been broken if the digital signatures were independent and not chained to each other. The independent digital signature also opens the window for host *i* to modify data provided to the agent at a later time (one such scenario is the agent visits one of the host's collaborating partners later). Regardless of the message format used, so long as the messages are independent of each other, host *i* will have no problem reproducing a valid signature for the modified message. In this way, data integrity can be compromised. With chained digital signature, even if the malicious host (or host *i* itself) produces a valid digital signature after modifying the data, the new signature c_i' is unlikely to be the same as c_i. If the new signature is different from the original signature, as the previous signature is provided as input to the next signature, the subsequent signature verification will fail, thus detecting compromise to data integrity. The inclusion of the host ID, agent ID, and time stamp prevents anyone from performing a replay attack.

When the agent reaches a new destination, the host must perform an integrity check on the incoming agent. In the design of SADIS, even if the new destination host does not perform an immediate integrity check on the incoming agent, any compromise to the data integrity can still be detected when the agent returns to the butler. The drawback, however, is that the identity of the malicious host may not be established. One design focus of SADIS is not only to detect data integrity compromise, but more importantly, to identify malicious hosts. To achieve malicious-host identification, it is an obligation for all hosts to verify the incoming agent's data integrity before activating the agent for execution. In the event of data integrity verification failure, the previous host will be identified as the malicious host.

FUTURE TRENDS

Besides agent data integrity and agent transport security, there are other security concerns to be addressed in SAFER. One such concern is a mechanism to assess the agent's accumulated risk level as it roams. There have been some considerations for using the agent battery concept to address this during the earlier stages of research. Furthermore, in order to establish the identity of different agents from different agent communities, a certain level of certification by trusted third parties or agent passports are required (Guan, Wang, & Ong, 2003). More research can be conducted in these areas.

CONCLUSION

In this chapter, a new data integrity protection protocol, SADIS, is proposed under the SAFER research initiative. Besides being secure against a variety of attacks and robust against vulnerabilities pointed out in related work in the literature, the research objectives of SADIS include efficiency. This is reflected in the minimized use of PKI operations and reduced message exchanges between the agent and the butler. The introduction of a variation to DH key exchange and evolving communication session keys further strengthened the security of the design. Unlike solutions suggested in some existing literature, the data integrity protection protocol aims not only to detect data integrity compromise, but more importantly, to identify the malicious host.

With security, efficiency, and effectiveness as its main design focus, SADIS works with other security mechanisms under SAFER (e.g., agent transport protocol) to provide mobile agents with a secure platform.

FUTURE RESEARCH DIRECTIONS

Recently there have been active research activities on the use of intelligent agents to mine user preferences: so-called personalization agents. Such agents, when equipped with inference engines, would be able to derive personal interests when observing Web or mobile-user interactions or click streams during online transactions. They would carry sensitive, personal data that should not be disclosed to outsiders. The protection of data in such agents is crucial. The migration of such agents or personalized data may be necessary when the service platform consists of multiple servers. For now, such agents usually reside on the server side, where strict security may already be in place. In the near future, such agents could be deployed on the client side, with a different name such as personal secretary, personal agent, and so forth. Such an agent may be dispatched by the user to run errands such as product brokering, information collection, or even transaction negotiation. An agent that carries user preference data is therefore vulnerable to attacks due to the fact that it has sensitive data inside. Protection of data in such type of agents would then be necessary.

REFERENCES

Borselius, N. (2002). Mobile agent security. *Electronics & Communication Engineering Journal, 14*(5), 211-218.

Borselius, N., Hur, N., Kaprynski, M., & Mitchell, C. J. (2002). A security architecture for agent-based mobile systems. *Proceedings of the Third International Conference on Mobile Communications Technologies* (pp. 312-318).

Borselius, N., Mitchell, C. J., & Wilson, A. T. (2001). On mobile agent based transactions in moderately hostile environments. *Advances in Network and Distributed Systems Security: Proceedings of the IFIP TC11 WG11.4 First Annual Working Conference on Network Security* (pp. 173-186).

Chionh, H. B., Guan, S.-U., & Yang, Y. (2001). Ensuring the protection of mobile agent integrity: The design of an agent monitoring protocol. *Proceedings of the IASTED International Conference on Advances in Communications* (pp. 96-99).

Corradi, A., Cremonini, M., Montanari, R., & Stefanelli, C. (1999). Mobile agents and security: Protocols for integrity. *Proceedings of the Second IFIP WG 6.1 International Working Conference on Distributed Applications and Interoperable Systems (DAIS'99).*

Diffie, W., & Hellman, M. E. (1976). New directions in cryptography. *IEEE Transactions on Information Theory, 22*, 644-654.

Esparza, O., Muñoz, J. L., Soriano, M., & Forné, J. (2006). Secure brokerage mechanisms for mobile electronic commerce. *Computer Communications, 29*(12), 2308-2321.

Guan, S.-U., & Hua, F. (2003). A multi-agent architecture for electronic payment. *International Journal of Information Technology and Decision Making (IJITDM), 2*(3), 497-522.

Guan, S.-U., Wang, T., & Ong, S.-H. (2003). Migration control for mobile agents based on passport and visa. *Future Generation Computer Systems, 19*(2), 173-186.

Guan, S.-U., & Yang, Y. (2002). SAFE: Secure agent roaming for e-commerce. *Computer & Industrial Engineering Journal, 42*, 481-493.

Guan, S.-U., Zhu, F., & Maung, M. T. (2004). A factory-based approach to support e-commerce agent fabrication. *Electronic Commerce and Research Applications, 3*(1), 39-53.

Ouardani, A., Pierre, S., & Boucheneb, H. (2006). A security protocol for mobile agents based upon the cooperation of sedentary agents. *Journal of Network and Computer Applications.*

Park, J. Y., Lee, D. I., & Lee, H. H. (2002). One-time key generation system for agent data protection. *IEICE Transactions on Information and Systems* (pp. 535-545).

Roth, V. (2001). On the robustness of some cryptographic protocols for mobile agent protection. *Mobile Agents 2001 (MA'01)* (pp. 1-14).

Tripathi, A. R. (2002). Design of the Ajanta system for mobile agent programming. *Journal of Systems and Software, 62*(2), 123-140.

Wang, T., Guan, S.-U., & Chan, T. K. (2002). Integrity protection for code-on-demand mobile agents in e-commerce. *Journal of Systems and Software, 60*(3), 211-221.

Yang, Y., & Guan, S.-U. (2000). Intelligent mobile agents for e-commerce: Security issues and agent transport. In *Electronic commerce: Opportunities and challenges.* Idea Group Publishing.

Zhu, F., Guan, S.-U., Yang, Y., & Ko, C. C. (2000). SAFER e-commerce: Secure agent fabrication, evolution and roaming for e-commerce. In *Electronic commerce: Opportunities and challenges.* Idea Group Publishing.

FURTHER READING

Boll, S. (n.d.). Modular content personalization service architecture for e-commerce applications. *Proceedings of the Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS 2002)* (pp. 213-220).

Evans, A., Fernandez, M., Vallet, D., & Castells, P. (2006). Adaptive multimedia access: From user needs to semantic personalization. *Proceedings of the 2006 IEEE International Symposium on Circuits and Systems (ISCAS 2006).*

Guan, S.-U., Tan, S. L., & Hua, F. (2004). A modularized electronic payment system for agent-based e-commerce. *Journal of Research and Practice in Information Technology, 36*(2), 67-87.

Guan, S.-U., & Yang, Y. (1999). *SAFE: Secure-roaming agent for e-commerce.* 26th International Conference on Computers & Industrial Engineering, Australia.

Guan, S.-U., & Zhu, F. (2002). Agent fabrication and its implementation for agent-based electronic commerce. *International Journal of Information Technology and Decision Making (IJITDM), 1*(3), 473-489.

Guan, S.-U., Zhu, F. M., & Ko, C. C. (2000). Agent fabrication and authorization in agent-based electronic commerce. *Proceedings of International ICSC Symposium on Multi-Agents and Mobile Agents in Virtual Organizations and E-Commerce* (pp. 528-534).

Gunupudi, V., & Tate, S. R. (2004). Performance evaluation of data integrity mechanisms for mobile agents**.** *Proceedings of the International Conference on Information Technology: Coding and Computing, ITCC 2004, 1*, 62-69.

Jorstad, I., van Thanh, D., & Dustdar, S. (2005). The personalization of mobile services. *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, (WiMob 2005), 4*, 59-65.

Koutrika, G., & Ioannidis, Y. (2004). Personalization of queries in database systems. *Proceedings of the 20th International Conference on Data Engineering* (pp. 597-608).

Panayiotou, C., Andreou, M., Samaras, G., & Pitsillides, A. (2005). Time based personalization for the moving user. *International Conference on Mobile Business (ICMB 2005)* (pp. 128-136).

Park, J. Y., Lee, D. I., & Lee, H. H. (2001). Data protection in mobile agents: One-time key based approach. *Proceedings of the 5th International Symposium on Autonomous Decentralized Systems* (pp. 411-418).

Poh, T. K., & Guan, S.-U. (2000). Internet-enabled smart card agent environment and applications. In S. M. Rahman & M. Raisinghani (Eds.), *Electronic commerce: Opportunities and challenges.* Idea Group Publishing.

Sim, L. W., & Guan, S.-U. (2002). An agent-based architecture for product selection and evaluation under e-commerce. In S. Nansi (Ed.), *Architectural issues of Web-enabled electronic business* (pp. 333-346). Idea Group Publishing.

Specht, G., & Kahabka, T. (2000). Information filtering and personalisation in databases using Gaussian curves. *2000 International Database Engineering and Applications Symposium* (pp. 16-24).

Tam, K. Y., & Ho, S. Y. (2003). Web personalization: Is it effective? *IT Professional*, *5*(5), 53-57.

Tan, X., Yao, M., & Xu, M. (2006). An effective technique for personalization recommendation based on access sequential patterns. *IEEE Asia-Pacific Conference on Services Computing, APSCC '06* (pp. 42-46).

Treiblmaier, H., Madlberger, M., Knotzer, N., & Pollach, I. (2004). Evaluating personalization and customization from an ethical point of view: An empirical study. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences.*

Tseng, B. L., Lin, C.-Y., & Smith, J. R. (2002). Video personalization and summarization system. *2002 IEEE Workshop on Multimedia Signal Processing* (pp. 424-427).

Wang, Y., Kobsa, A., van der Hoek, A., & White, J. (2006). PLA-based runtime dynamism in support of privacy-enhanced Web personalization. *10th International Software Product Line Conference*.

Wang, Y. H., Wang, C. L., & Liao, C. H. (2004). Mobile agent protection and verification in the Internet environment. *The Fourth International Conference on Computer and Information Technology* (pp. 482-487).

Wu, D., Im, I., Tremaine, M., Instone, K., & Turoff, M. (2003). A framework for classifying personalization scheme used on e-commerce Websites. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences* (pp. 12-23).

Yang, Y. (2006). Provisioning of personalized pervasive services: Daidalos personalization functions. *2006 1st International Symposium on Pervasive Computing and Applications* (pp. 110-115).

Yang, Y., & Guan, S. U. (2000). Intelligent mobile agents for e-commerce: Security issues and agent transport. In S. M. Rahman & M. Raisinghani (Ed.), *Electronic commerce: Opportunities and challenges.* Idea Group Publishing.

Yee, G. (2006). Personalized security for e-services. *The First International Conference on Availability, Reliability and Security (ARES 2006).*

Yu, P. S. (1999). Data mining and personalization technologies. *Proceedings of the 6th International Conference on Database Systems for Advanced Applications* (pp. 6-13).

Zhao, Y., Yao, Y., & Zhong, N. (2005). Multilevel Web personalization. *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 649-652).

TERMS AND DEFINITIONS

Agent: An agent is a piece of software that acts to accomplish tasks on behalf of its user.

Cryptography: Cryptography is the art of protecting information by transforming it (encrypting it) into an unreadable format, called cipher text. Only those who possess a secret key can decipher (or decrypt) the message into plain text.

Flexibility: Flexibility is the ease with which a system or component can be modified for use in applications or environments other than those for which it was specifically designed.

Integrity: Integrity regards the protection of data or program code from being modified by unauthorized parties.

Mobile Agent: Also called a roaming agent, it is an agent that can move from machine to machine for the purpose of data collection or code execution.

Protocol: A protocol is a convention or standard that controls or enables the connection, communication, and data transfer between two computing endpoints. Protocols may be implemented by hardware, software, or a combination of the two. At the lowest level, a protocol defines a hardware connection.

Security: Security involves the effort to create a secure computing platform designed so that agents (users or programs) can only perform actions that have been allowed.

This work was previously published in Handbook of Research on Public Information Technology, edited by G. Garson; M. Khosrow-Pour, pp. 423-462, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.11
Key Distribution and Management for Mobile Applications

György Kálmán
University Graduate Center – UniK, Norway

Josef Noll
University Graduate Center – UniK, Norway

ABSTRACT

This chapter deals with challenges raised by securing transport, service access, user privacy, and accounting in wireless environments. Key generation, delivery, and revocation possibilities are discussed and recent solutions are shown. Special focus is on efficiency and adaptation to the mobile environment. Device domains in personal area networks and home networks are introduced to provide personal digital rights management (DRM) solutions. The value of smart cards and other security tokens are shown and a secure and convenient transmission method is recommended based on the mobile phone and near-field communication technology.

A PROBLEM OF MEDIA ACCESS

On the dawn of ubiquitous network access, data protection is becoming more and more important. While in the past network connectivity was mainly provided by wired connections, which is still considered the most secure access method, current and future users are moving towards wireless access and only the backbone stays connected by wires. In a wired environment, eavesdropping is existent, but not as spread and also not easy to implement. While methods exist to receive electromagnetic radiation from unshielded twisted pair (UTP) cables, a quite good protection can be achieved already by transport layer encryption or deploying shielded twisted pair (STP) or even fibre.

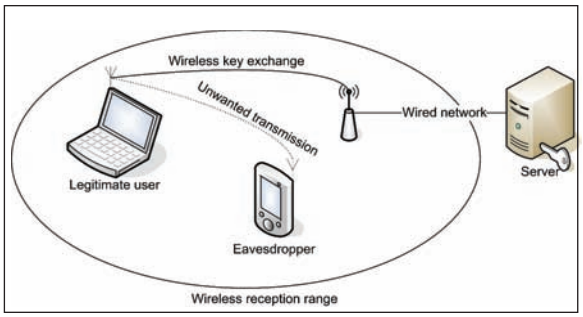
New technologies emerged in the wireless world, and especially the IEEE 802.11 family has drastically changed the way users connect to networks. The most basic requirements for new devices are the capability of supporting wireless service access. The mobile world introduced general packet radio service (GPRS) and third generation (3G) mobile systems provide permanent IP connectivity and provide together with Wi-Fi access points continuous wireless connec-

tivity. Besides communications devices such as laptops, phones, also cars, machines, and home appliances nowadays come with wireless/mobile connectivity.

Protecting user data is of key importance for all communications, and especially for wireless communications, where eavesdropping, man-in-the-middle, and other attacks are much easier. With a simple wireless LAN (WLAN) card and corresponding software it is possible to catch, analyse, and potentially decrypt wireless traffic. The implementation of the first WLAN encryption standard wired equivalent privacy (WEP) had serious weaknesses. Encryption keys can be obtained through a laptop in promiscuous mode in less than a minute, and this can happen through a hidden attacker somewhere in the surrounding. Data protection is even worse in places with public access and on factory default WLAN access points without activated encryption. Standard Internet protocols as simple mail transport protocol (SMTP) messages are not encoded, thus all user data are transmitted in plaintext. Thus, sending an e-mail over an open access point has the same effect as broadcasting the content. With default firewall settings an intruder has access to local files, since the local subnet is usually placed inside the trusted zone. These examples emphasise that wireless links need some kind of traffic encryption.

When the first widespread digital cellular network was developed around 1985, standardisation of the global system for mobile communication (GSM) introduced the A5 cryptographic algorithms, which can nowadays be cracked in real-time (A5/2) or near real-time (A5/1). A further security threat is the lack of mutual authentication between the terminal and the network. Only the terminal is authenticated, the user has to trust the network unconditionally. In universal mobile telecommunications system (UMTS), strong encryption is applied on the radio part of the transmission and provides adequate security for current demands, but does not secure the transmission over the backbone. UTMS provides mutual authentication through an advanced mechanism for authentication and session key distribution, named authentication and key agreement (AKA).

Figure 1. A basic problem of broadcast environment



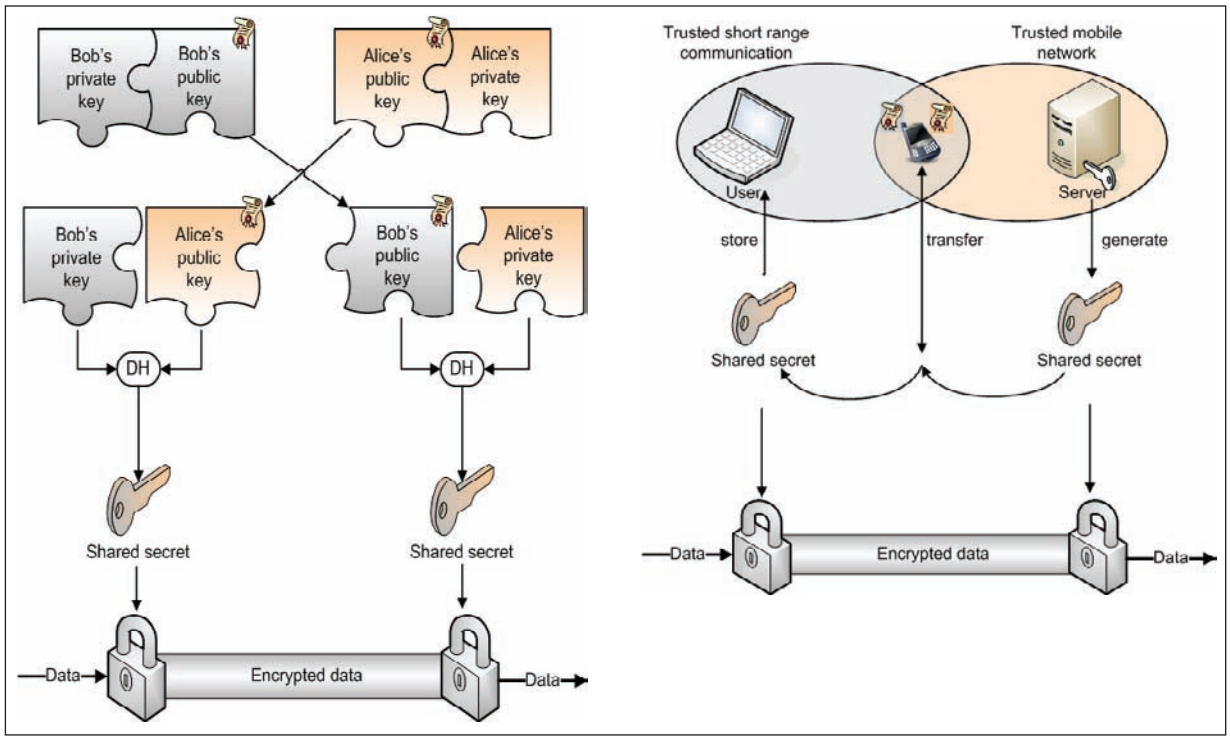
A LONG WAY TO SECURE COMMUNICATION

Applying some kind of cryptography does not imply a secured access. Communicating parties must negotiate the key used for encrypting the data. It should be obvious that the encryption key used for the communication session (session key) cannot be sent over the air in plaintext (see Figure 1).

In order to enable encryption even for the first message, several solutions exist. The simplest one, as used in cellular networks is a preshared key supplied to the mobile terminal on forehand. This key can be used later for initialising of the security infrastructure and can act as a master key in future authentications.

In more dynamic systems the use of preshared keys can be cumbersome. Most of WLAN encryption methods support this kind of key distribution. The key is taken to the new unit with some kind of out of band method, for example with an external unit, as indicated in Figure 2. Practically all private and many corporate WLANs use static keys, allowing an eavesdropper to catch huge amounts

Figure 2. (a) Diffie-Hellmann key exchange and (b) out-of-band key delivery



of traffic and thus enable easy decryption of the content. This implies that a system with just a secured access medium can be easily compromised. Non-aging keys can compromise even the strongest encryption, thus it is recommended to renew the keys from time to time.

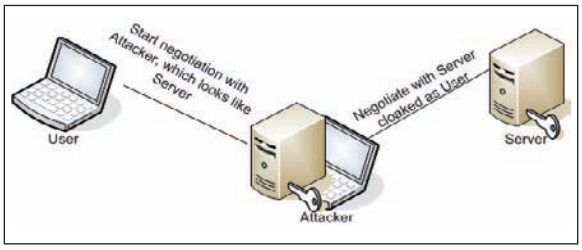
Outside the telecom world it is harder to distribute keys on forehand, so key exchange protocols emerged, which offer protection from the first message and do not need any preshared secret. The most widespread protocol is the Diffie-Hellman (DH) key exchange of Figure 2, which allows two parties that have no prior knowledge of each other to jointly establish a shared secret key over an insecure communications channel.

This protocol does not authenticate the nodes to each other, but enables the exchange data, which can be decoded only by the two parties. Malicious attackers may start a man-in-the-middle attack (see Figure 4). Since this problem is well-known, several modifications enable identity based DH, for example Boneh, Goh, and Boyen (2005) showed a hierarchical identity based encryption method, which is operating in fact as a public key system, where the public key is a used chosen string.

Public key infrastructure (PKI) can help defending corresponding parties against man-in-the-middle attacks. Public key cryptography is based on the non polynomial (NP) time problems, for example of factorisation or elliptic curves.

Two keys, a public and a private are generated. The public key can be sent in plaintext, because messages encrypted with the public key can only be decoded by the private key and vice versa. The two way nature of public keys makes it possible to authenticate users to each other, since signatures generated with the public key can be checked with the public key. Message authenticity can be guaranteed. Still, the identity of the node is not proven. The signature proves only that the message was encoded by the node,

Figure 3. Principle of a man-in-the-middle attack



which has a public key of the entity we may want to communicate with.

Identity can be ensured by using certificates. Certificate authorities (CA) store public keys and after checking the owner's identity out of band, prove their identity by signing the public key and user information with their own keys. This method is required for financial transactions and business and government operations. Without a CA, the public keys can be gathered into a PKI, which provides an exchange service. Here, most commonly, a method called web of trust is used. A number of nodes, who think that the key is authentic, submit their opinion by creating a signature. The solution enables community or personal key management, with a considerable level of authenticity protection.

While public keys can be sent, private keys must be kept secret. Although they are protected usually with an additional password, this is the weakest point in the system. If the user saves a key in a program in order to enter the key automatically, security provided by the system is equal to the security of the program's agent application. Private firewalls and operating system policies usually will not stop a good equipped intruder.

Another security issue for terminals is the lack of tamper resistant storage. Usage of smart cards is a solution to this issue, but introduces additional hardware requirements. The lack of secure storage is getting much attention in DRM schemes. Most DRM schemes use a software-based method, but also hardware-assisted ones have lately been introduced.

All these authentication methods, secure storage and rights management support secure data exchange, but they do not protect the privacy of user credentials, preferences, and profiles. Ad hoc networks, like personal area networks (PANs), which move around and are dynamically configured open for intrusion attacks on the privacy.

Thus, protection of user credentials in wireless environments is one of the focal points of current research. Before addressing privacy, we will first summarise issues in key management protocols.

FROM KEY EXCHANGE TO ACCESS CONTROL INFRASTRUCTURE

Mobility and wireless access introduced new problems in network and user management, as compared to fixed network installations with, for example, port-based access restrictions. The network operators want to protect the network against malicious intruders, charge the correct user for the use, and provide easy and open access to their valued services.

The first step to get access to an encrypted network is to negotiate the first session key. This has been solved in coordinated networks like mobile networks through pre-shared keys. Authentication and access control is provided by central entities to ensure operations.

In computer networks, which are not controlled in such way and usually not backed-up by a central authorisation, authentication, and accounting (AAA), different methods have been created for connection control. The basic method is still to negotiate encryption keys based on a preshared secret. Typical preshared keys are a password for hash calculation, one time password sent via cell phone or keys given on an USB stick.

There are several solutions to protect the data transmitted over a wireless link. In private networks, security based on preshared keys is a

Figure 4. TLS key negotiation

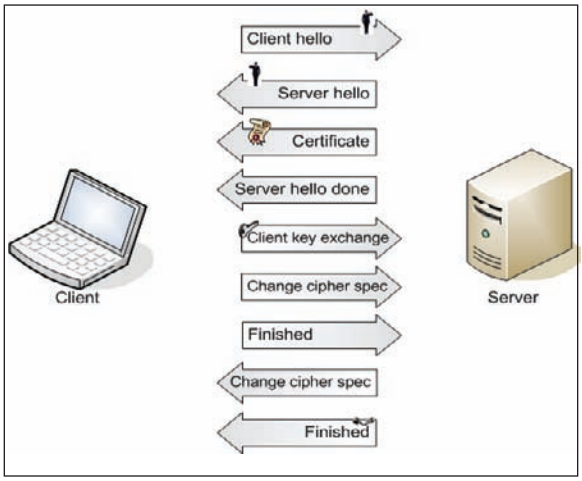
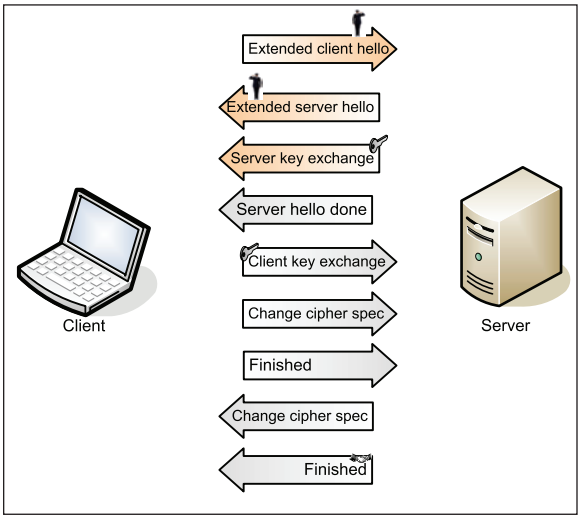


Figure 5. TLS-KEM key negotiation



working solution. In corporate or public networks, a more robust solution is needed. The most promising way is to integrate session key negotiation into the AAA process. Since providers or companies have to identify the connected user, they rely on an AAA infrastructure and have an encryption of user credentials as compulsory policy. A certificate-based medium access control and AAA system is advised, where AAA messages can carry also the certificates needed to secure the message exchange.

As public key operations induce a lot of network traffic, the negotiated session keys have to be used in the most efficient way. Encryption protocols designed for wired environments, like transport layer security (TLS) do not consider problems associated with the broadcast transmissions and limitations of mobile devices. In a wired, or at least fixed environment, computational cost of key negotiations is usually neglected. For example TLS is using several public key operations to negotiate a session key. This can be a problem for mobile devices, since computational cost is much higher in asymmetric encryption. The standard TLS suite uses lots of cryptographic operations and generates a too large message load on wireless links (see Figure 5).

If a mobile device wants to execute mutual authentication with a service provider, with certificate exchanges, it can lead to big amounts of data transferred over the radio interface beside the high computing power needs.

In environments with limited resources, authentication and identity management based on preshared keys is still the most effective solution. Badra and Hajjeh (2006) propose an extension to TLS, which enables the use of preshared secrets instead the use of asymmetric encryption. This is in line with the efforts to keep resource needs at the required minimum level in mobile devices. A preshared key solution was also proposed by the 3rd Generation Partnership Projects (3GPP, 2004) and (3GPP2, 2007) as an authentication method for wireless LAN interworking. The problem with the proposed solution is preshared keys does not provide adequate secrecy nor identity protection in Internet connections. To deal with this problem, the TLS-key exchange method (TLS-KEM) provides identity protection, minimal resource need, and full compatibility with the original protocol suite as seen in Figure 6.

In direct comparison, the public key based TLS needs a lot more computing, data traffic, and deployment effort.

In UMTS networks, an array of authentication keys is sent to the mobile in authentication vectors. In the computer world a good solution would be using hash functions to calculate new session keys, as these consume low power and require little computing.

A moving terminal can experience a communication problem, as the overhead caused by key negotiation might extend the connection time to a network node. A preserved session key for use in the new network is a potential solution in a mobile environment, as it speeds up the node's authentication. Lee and Chung (2006) recommend a scheme, which enables to reuse of session keys. Based on the AAA infrastructure, it is possible to forward the key to the new corresponding AAA server on a protected network and use it for authentication without compromising system security. This can reduce the delay for connecting, and also reduces the possibility of authentication failure. Since the old session key can be used for authenticating the node towards the new AAA server, connection to the home AAA is not needed any more. The messages are exchanged as follows (Lee & Chung, 2006): when sending the authorisation request to the new network, the node also includes the old network address it had. The foreign agent connects to the new local AAA server and sends an authentication request. The new AAA server connects to the old one sending a message to identify the user. The old AAA authenticates the message by checking the hash value included, and generates a nonce for the terminal and the foreign agent. The server composes an AAA-terminal answer, which is composed from a plain nonce, an encrypted nonce using the key shared between the old foreign agent and the terminal. Then the whole message is signed and encrypted with the key used between the two AAA servers. When the new AAA receives it, decrypts and sends the message to the new foreign agent. Based on the plain nonce, the agent generates the key and sends down the reply, which includes also the nonce encrypted by the old AAA. After the authentication of the user towards the network, the user can start using services.

Key distribution and efficiency in e-commerce applications is another important aspect. The network's AAA usually does not exchange information with third parties or can not use the authentication data of the network access because of privacy issues. Current security demands require mutual identification of communicating parties in an e-commerce application. This can easily lead to compromising the customer to companies (for example in a GSM network, the user has to trust the network unconditionally). If the user can also check the identity of the service provider, at least man-in-the-middle attacks are locked out.

When a user starts a new session with a service provider, this session should be based on a new key set. The session key has to be independent from the previous one in means of traceability and user identity should not be deductible from the session key, thus ensuring user privacy. For mutual identification, a key exchange method is proposed by Kwak, Oh, and Won (2006), which uses hash values to reduce resource need. The key calculation is based on random values generated by the parties, which ensures key freshness.

The use of hash functions is recommended in mobile environments, providing better performances for public key based mechanisms (Lim, Lim, & Chung, 2006). Mobile IPv4 uses symmetric keys and hashes by default. Since symmetric keys are hard to manage, a certificate-based key exchange was recommended, but this demands more resources. To lower the resource demand, a composite architecture was recommended (Sufatrio, 1999). The procedure uses certificates only in places where the terminal does not require processing of the public key algorithm and does not require storage of the certificate.

The result of the comparison shows that hash is by far the most efficient method in terms of key generation, but suffers from management difficulties. Lim et al. (2006) also demonstrates

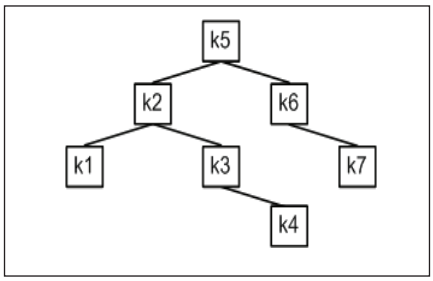
that a pure certificate-based authentication is unsuitable for mobile environments. Partial use of certificates and identity-based authentication with extensive use of hash functions can be a potential way ahead.

AUTHENTICATION OF DEVICE GROUPS

In a ubiquitous environment, moving networks appear. PANs and ad hoc connections based on various preferences emerge and fall apart. These devices communicate with each other and have usually very limited capabilities in terms of computing power and energy reserves. In order to provide secure communication between any part of the network, hierarchical key management methods emerged (Kim, Ahn, & Oh, 2006). Here a single trusted server is used to manage the group key. These entities are usually storing the keys in a binary tree, where nodes are the leaves.

Public key operations are usually required when a terminal wants to connect to a group for the first time. A group management system needs frequent key generation rounds, because it has to ensure forward and backward secrecy. Strict key management policies ensure that no new node is capable of decoding former traffic and none of the old nodes have the possibility to decrypt current traffic. To adjust resource usage to mobile environment, a management scheme which uses mainly simple operations like XOR and hash is advisable (Kim et al., 2006). As the key in the root of the binary tree is used to authenticate the whole group, keys need to be regenerated when a node leaves the network. This procedure is starting from the parent of the former node and goes up to the root. Then the management unit sends out the new keys in one message. Building a tree from keys ensures fast searches and a simple, clean structure. In addition, all keys in the internal nodes are group keys for the leaves under them. So a subset of devices can be easily addressed.

Figure 6. Keys in a binary tree



The root unit has to compute these keys in acceptable time, requiring a more complex architecture. In PANs this is usually not a problem, but when a member of a larger subnet is leaving, calculations could be more demanding. A standard group key handling method is the Tree-based Group Diffie-Hellman (TGDH), where management steps assume that all nodes have the same processing capabilities. To ensure maximal efficiency, the highest performance unit shall be the one in the root of the tree (Hong & Lopez-Benitez, 2006). When node computing capabilities are showing big differences, the overhead caused by tree transformations does not represent a drawback.

Another significant group of devices that need encryption can be found in home networks, where the focus is on management of content and personal data.

SECURE HOME NETWORK AND RIGHTS MANAGEMENT

Deployment of wired or wireless home networks happens in roughly 80% of all households with broadband access (Noll, Ribeiro, & Thorsteinsson, 2005). Network-capable multimedia devices, media players, game consoles, and digital set-top boxes are widespread and part of the digital entertainment era. Content is stored within this network, and provided through the Internet to

other users. Since the birth of peer-to-peer (P2P) networks, such technologies are in the crosshair of content providers. Recently, some software developers and a few musicians started using the torrent network for cost effective delivery of their content. A digital rights management method designed for such network is still missing.

Current right protection solutions are not compatible with each other and the user friendliness is also varying. The basic problem is, that just a very few devices are equipped with tamper resistant storage and integrated cryptographic capabilities. Beside software solutions, which are meant as weak solutions, hardware-based encryption can severely limit the lawful use of digital content. Recent lawsuits related to Sony's rootkit protection mechanism also reveals that customer rights of usage is considered to be more important than the legitimate wish of content providers to protect the content.

Trusted platform modules (TPM) are the most likely candidate for content protection in hardware-based solutions. While providing encryption capabilities, it is very likely that these components will be used to dispose the users' right to decide over the user's own resources.

The current discussions on DRM for audio content are regarded as minor when compared to high definition (HD) content protection. Even the connection to the screen has to use strong encryption, which has to exceed GSM/UMTS encryption in order to be acceptable for content providers. Enforcing a digital, end-to-end encrypted stream means that a HD-TV purchased at the end of 2006 may not work with the new encryption standards for HD. There is no current solution for computers to legally play full resolution HD. By the end of 2006 it was announced, that a workaround is arising to deal with the advanced content protection system of HD.

A more discrete, but not intrusive business model discussion for digital content management is presented in order to visualise the requirements of this market. Apple's FairPlay enables making backup copies of audio tracks, which is permitted by law in several European countries, and copy of content between the user's iPod players. This solution is considered being to open for some content providers, and the distribution is limited to a server-client infrastructure. For HD content with high bandwidth needs such a server-client infrastructure is not advisable, both from a server and network point of view. The ever growing size of P2P networks form a perfect infrastructure to deliver content with high bandwidth need practically without substantial transmission costs. P2P networks are usually run without any DRM support. An additional infrastructure supporting DRM in a P2P network used to transmit content will enable high volume distribution of digital content (Pfeifer, Savage, Brazil, & Downes, 2006). If seamless license delivery and user privacy could be guaranteed, such a network could be the foundation of a low cost content delivery scheme.

While the usage of P2P networks is an excellent idea, the recommended solution proposed by Nützel and Beyer (2006) is similar to the Sony's rootkit solution: It bypasses the user control and is thus not acceptable. While the primary goal is to secure content, the software used in such solutions acts like hidden Trojans and opens backdoors not only for the content providers, but also other hackers.

Content usage across platforms is not supported yet, as a common standard does not exist. Pfeifer et al. (2006) suggests a common management platform for DRM keys with an XML-based, standard MPEG-REL framework. Users will also produce content with digital protection, in order to ensure that personal pictures cannot be distributed electronically. Social networks and groups of interest, as well as distribution of content in PANs is a challenge for DRM development. Zou, Thukral, and Ramamurthy (2006) and Popescu, Crispo, Tanenbaum, and Kamperman (2004) propose a key delivery architecture for device groups, which could be extended by a local license manager.

The central key management unit could distribute licenses seamlessly to the device, which wants to get access, without invading user experience.

Kálmán and Noll (2006) recommend a phone-based solution. This represents a good trade-off between user experience and content protection. The phone is practically always online, most of them have Bluetooth or other short range radio transmitters, so licenses can be transmitted on demand. Since the phone has a screen and a keyboard, it is possible to request authorisation from the user before every significant message exchange, so the user can control the way licenses are distributes.

If we look aside the issues related to business aspects, computational issues still remain. Highly secure DRM entities will use asymmetric encryption and certificates. Sur and Rhee (2006) recommend a device authentication architecture, which eliminates traditional public key operations except the ones on the coordinator device. This is achieved by using hash chains including the permission, for example, a device can get keys to play a designated audio track ten times or permission to use five daily permits on demand. Such schemes allow end devices to be simpler and lower network communication overhead.

If a central device is not appreciated, a composite key management scheme may be used. The parties in the PAN will form a web of trust like in a confidentiality scheme, for example, pretty good privacy (PGP). In this web, the main key is split between nodes and cooperation is needed for significant operations. This means that if the scheme is operating on a (k, n) basis, k-1 nodes can be lost before the system needs to be generate a new key. Fu, He, and Li (2006) mention the problem of the PAN's ad hoc nature as the biggest problem. Since this scheme selects n nodes randomly, the ones that are moving between networks fast can cause instability in the system. Also, the resource need of this proposal is quite high on all nodes present.

When a scheme is enabling off-line use of license keys, attention should be given to problems arising from leaving or compromised nodes. Identity-based schemes become popular recently because of their efficiency in key distribution. The main drawback is that these proposals do not provide a solution for revocation and key renewal. Hoeper and Gong (2006) propose a solution based on a heuristic (z, m) method. The solution is similar to the threshold scheme shown before, but enables key revocation. If *z* nodes are accusing one node to be compromised, based on their own opinion, the node is forced to negotiate a new key. If a node reaches a threshold in number of regenerations in a time period, it could be locked out, since most likely an intruder is trying to get into the system or the internal security of the node is not good enough. The assumptions about the system are strongly limiting the effectiveness of the solution. The most stringent assumption is that they require to nodes to be in promiscuous mode. This can lead to serious energy problems. Another requirement is that there has to be a unit for out-of-band key distribution. This unit could be the cellular phone.

SMART CARDS AND CELLULAR OPERATORS

The use of smart cards has its roots in the basic problem of security infrastructures: even the most well designed system is vulnerable to weak passwords. A card, which represents a physical entity, can be much easier protected compared to a theoretical possession of a password. Smart cards integrate tamper resistant storage and cryptographic functions. They are usually initialised with a preshared key and creating a hash chain, where values can be used as authentication tokens.

The remote authentication server is using the same function to calculate the next member. The encryption key is the selection of a collision resis-

tant hash function. While the tokens they provide are quite secure, a problem with smart cards is that they represent a new unit that has to be present in order to enable secure communication, and user terminals must be equipped with suitable readers. The additional hardware does not only cause interoperability problems, but is usually slow, as a measurement conducted shows (Badra & Hajjeh, 2006). This becomes eminent when high traffic is associated with asymmetric encryption; sending a "hello" message with standard TLS to the smart card needed 10 seconds. In contrast, the modified TLS-KEM needed 1.5 s.

A user-friendly, seamless key delivery system can be created with the help of cellular operators and SIM cards with enhanced encryption capabilities. The SIM and USIM modules used in GSM/UMTS are quite capable smart cards. They offer protected storage with the possibility of over the air key management, good user interface, and standard architecture. Danzeisen, Braun, Rodellar, and Winiker (2006) shows the possible use of the mobile operator as trusted third party for exchanging encryption keys out of band for other networks.

Delivery of the mobile phone key to a different device can be problematic, since most devices do not have a SIM reader, or it is inconvenient to move the SIM card from the mobile phone to another device. New developments in near field communication may overcome this and enable short range secure key transfer.

BREAKING THE LAST CENTIMETRE BOUNDARY

Frequency of authentication request is a key factor in user acceptance. If a system asks permanently for new passwords or new values from the smart card hash chain, it will not be accepted by the user. On the other hand, if a device gets stolen and it asks for a password only when it is switched on, then a malicious person can impersonate the user for a long time. A potential solution is to create a wearable token with some kind of wireless transmission technology and define the device behaviour such that if the token is not accessible, it should disable itself in the very moment of notification.

Since the main challenge is not securing data transfer between the terminal and the network, but to authenticate the current user of the terminal, a personal token has to be presented. As proposed by Kálmán and Noll (2007), the mobile phone can be a perfect personal authentication token if it is extended by a wireless protocol for key distribution.

With the capabilities of user interaction, network control of the mobile phone, it can be ensured that critical operations will need user presence by requiring PINs or passwords. Possible candidates for key exchange are Bluetooth (BT), radio frequency identification (RFID), and Near Field Communications (NFC). NFC is a successor of RFID technology in very short range transmissions. BT is close to the usability limit, since its transmit range reaches several meters. But the two later ones are promising candidates. Depending on the frequency, general RFID has a range of several meters while NFC operates in the 0-10 cm range. NFC is recommended, as the range alone limits the possibilities of eavesdroppers and intruders who want to impersonate the token while it is absent. The use of repeaters in the case of NFC, a so-called wormhole attack as described by Nicholson, Corner, and Noble (2006), looks not feasible because of the tight net of repeaters required. Also, the capability of user interaction provides an additional level of security.

Mobile phones with integrated NFC functionality are already available and serve as user authentication devices. To use these devices as tokens for other terminals, they have to be placed very close to each other. This prevents accidental use in most cases. To check presence of the token, heartbeat messages might be introduced. By design, this solution is very capable of distribut-

ing preshared keys for other devices out of band. Meaning, the phone can get the keys from the cellular network from an identity provider and send it down to the appropriate device by asking the user to put the devices close to each other for a second or two.

Transmission of the key must be done only when needed, so the programmable chip on the phones has to be in a secured state by default and only activated by the user's interaction. Protection of RFID tags is shown by Rieback, Gaydadjiev, Crispo, Hofman, and Tanenbaum (2006), where a proprietary hardware solution is presented. In case of a phone-based NFC key transmission, additional active devices might be unnecessary to use, but for general privacy protection, IDs with RFID extensions must be treated with care.

Transmission of certificates would not need additional encryption over the NFC interface, while other keys may require a preshared key between the phone and the terminals, which can be done via a wired method or by the phone provider. Most providers have at least one secret key stored on phones and a public key connected to that one. Based on this, DH key exchange would be possible between terminals and the phone using the cellular network as a gateway. An NFC-enabled phone could be the central element of a home DRM service, as it is online, capable of over the air downloads, and still able to ensure user control.

ON THE DAWN ON PERSONAL CONTENT MANAGEMENT

From the viewpoint of secure data transmission and user authentication, access and distribution of digital content can be ensured. Open issues remain for moving PANs and devices with limited capability. Focus nowadays is on protecting the user's privacy. As usage of digital devices with personal information was limited, user privacy was not of primary concern for a long time. Since PANs and home networks hold a large amount of critical personal data, this has to change (Jeong, Chung, & Choo, 2006; Ren, Lou, Kim, & Deng, 2006).

In a ubiquitous environment users want to access their content wherever they are. This has to be enabled in a secure manner. With upcoming social services, also fine grained access control methods have to be deployed inside the personal infrastructure. The focus of DRM research has to shift towards the end user, who will also require the right to protect himself/herself and his/her content with the same strength as companies do.

Extending the phone's functions may be problematic because of energy consumption and limited computing power. This could be easily solved by the technology itself, since a new generation of mobile terminals is arriving every half year. The capacity and functionalities of the SIM cards will be extended, the newest 3GPP proposals are predicting high capacity and extended cryptographic possibilities.

Regarding legal aspects, extending the SIM possibilities may cause some concern, since the SIM cards are currently owned by the network operators.

CONCLUSION

Transport encryption and authentication of devices has been the subject of research for a long time and resulted in sufficient secure solutions with current technologies. The focus in recent proposals is on the limited possibilities of mobile terminals and adoption of encryption technologies for mobile and wireless links.

Distributing keys between nodes is solved, except for the first step, which usually requires out-of-band transmissions. A solution for this initial key distribution might be the mobile phone with its integrated smart card and already existing communication possibility. As phones come with NFC, they may act as contact-less cards to distribute keys between devices.

While device authentication is handled sufficiently, user identity is hard to prove. A knowledge-based password or PIN request is not a user-friendly solution. Current proposals tend to be insecure when performing the trade-off between user experience and security.

Focus on research should be paid towards personal area and home networks. These networks hold most of the user's personal private data and content, either purchased or created by the user. Currently no standard solution exists for managing content rights or for access control of own content.

REFERENCES

3rd Generation Partnership Projects (3GPP). (2004, July). *Technical standardization groups-system and architecture (TSG-SA) working group 3 (Security) meeting, 3GPP2 security—Report to 3GPP, S3-040588*. Retrieved December 20, 2006, from www.3gpp.org/ftp/TSG_SA/WG3_Security/TSGS3_34_Acapulco/Docs/PDF/S3-040588.pdf

3rd Generation Partnership Projects (3GPP)2. (2007). *TSG-X/TIA TR-45.6, 3GPP2 system to wireless local area network interworking to be published as 3GPP2 X.S0028*. Retrieved December 22, 2006

Badra, M., & Hajjeh, I. (2006). Key-exchange authentication using shared secrets. *IEEE Computer Magazine*, *39*(3), 58-66.

Boneh, D., Goh, E.-J., & Boyen, X. (2005). Hierarchical identity based encryption with constant size ciphertext. In *Proceedings of Eurocrypt '05.*

Danzeisen, M., Braun, T., Rodellar, D., & Winiker, S. (2006). Heterogeneous communications enabled by cellular operators. *IEEE Vehicular Technology Magazine*, *1*(1), 23-30.

Fathi, H., Shin, S., Kobara, K., Chakraborty, S. S., Imai, H., & Prasad, R. (2006). LR-AKE-based AAA for network mobility (NEMO) over wireless links. *IEEE Selected Areas in Communications, 24*(9), 1725-1737.

Fu, Y., He, J., & Li, G. (2006). A composite key management scheme for mobile ad hoc networks. In *On the move to meaningful Internet systems, OTM 2006 Workshops* (LNCS 4277).

Hoeper, K., & Gong, G. (2006). Key revocation for identity-based schemes in mobile ad hoc networks, ad-hoc, mobile, and wireless networks (LNCS 4104).

Hong, S., & Lopez-Benitez, N. (2006). Enhanced group key generation algorithm. In *Network 10th IEEE/IFIP Operations and Management Symposium, NOMS 2006* (pp 1-4).

Jeong, J., Chung, M. Y., Choo, H. (2006). Secure user authentication mechanism in digital home network environments. In *Embedded and Ubiquitous Computing* (LNCS 4096).

Kálmán, Gy., & Noll, J. (2006). *SIM as a key of user identification: Enabling seamless user identity management in communication networks.* Paper presented at the WWRF meeting #17.

Kálmán, Gy., & Noll, J. (2007). SIM as secure key storage in communication networks. In *The International Conference on Wireless and Mobile Communications ICWMC'07.*

Kim, S., Ahn, T., & Oh, H. (2006). An efficient hierarchical group key management protocol for a ubiquitous computing environment. In *Computational Science and Its Applications—ICCSA 2006* (LNCS 3983).

Kwak, J., Oh, S., & Won, D. (2006). Efficient key distribution protocol for electronic commerce in mobile communications. In *Applied Parallel Computing* (LNCS 3732).

Lee, J.-H., & Chung, T.-M. (2006). Session key forwarding scheme based on AAA architecture in wireless networks. In *Parallel and Distributed Processing and Applications* (LNCS 4330).

Lim, J.-M., Lim, H.-J., & Chung, T.-M. (2006). Performance evaluation of public key based mechanisms for mobile IPv4 authentication in AAA environments. In *Information Networking. Advances in Data Communications and Wireless Networks* (LNCS 3961).

Nicholson, A. J., Corner, M. D., & Noble, B. D. (2006). Mobile device security using transient authentication. *IEEE Transactions on Mobile Computing, 5*(11), 1489-1502.

Noll, J., Ribeiro, V., & Thorsteinsson, S. E. (2005). Telecom perspective on scenarios and business in home services. In *Proceedings of the Eurescom Summit 2005* (pp 249-257).

Nützel, J., & Beyer, A. (2006). How to increase the security of digital rights management systems without affecting consumer's security, In *Emerging Trends in Information and Communication Security* (LNCS 3995).

Pfeifer, T., Savage, P., Brazil, J., & Downes, B. (2006). VidShare: A management platform for peer-to-peer multimedia asset distribution across heterogeneous access networks with intellectual property management. In *Autonomic Management of Mobile Multimedia Services* (LNCS 4267).

Phillips, T., Karygiannis, T., & Kuhn, R. (2005). Security standards for the RFID market. *IEEE Security & Privacy Magazine, 3*(6), 85-89.

Popescu, B. C., Crispo, B., Tanenbaum, A. S., & Kamperman, F. L. A. J. (2004). A DRM security architecture for home networks. In *Proceedings of the 4th ACM workshop on Digital rights management,* Washington, DC.

Ren, K., Lou, W., Kim, K., & Deng, R. (2006). A novel privacy preserving authentication and access control scheme for pervasive computing environments. *IEEE Transactions on Vehicular Technology, 55*(4), 1373-1384.

Rieback, M. R., Gaydadjiev, G. N., Crispo, B., Hofman, R. F. H., & Tanenbaum, A. S. (2006, December 3-8). *A platform for RFID security and privacy administration*. Paper presented at the 20th USENIX/SAGE Large Installation System Administration Conference—LISA 2006, Washington, DC.

Sufatrio, K. Y. L. (1999, June 23-25). *Registration protocol: A security attack and new secure mini-mal public-key based authentication*. Paper presented at the International Symposium on Parallel Architectures, Algorithms and Networks, ISPAN'99. Fremantle, Australia.

Sur, C., & Rhee, K. H. (2006). An efficient authentication and simplified certificate status management for personal area networks. In *Management of Convergence Networks and Services* (LNCS 4238).

Zou, X., Thukral, A., & Ramamurthy, B. (2006). An authenticated key agreement protocol for mobile ad hoc networks. In *Mobile Ad-hoc and Sensor Networks* (LNCS 4325).

KEY TERMS

Diffie-Hellman Key Exchange: Diffie-Hellman key exchange is a procedure, which allows negotiating a secure session key between parties, who do not have any former information about each other. The negotiation messages are in band, but because of the non-polynomial (NP) problem used in the procedure, adversaries are not able to compromise it.

Mutual Authentication: Mutual authentication occurs when the communicating parties can mutually check each others identity, thus reducing the possibility of a man-in-the-middle attack or other integrity attacks.

Out of Band Key Delivery: Out of band key delivery occurs when an encryption key is delivered with a mean, which is inaccessible from inside the network it will be used in. An example is to carry a key on an USB stick between parties, where the key will never be transmitted over the network.

Rootkit: Rootkit is a kind of software to hide other programs. Mainly used by Trojans, they enable hidden applications to access local resources without user knowledge.

Seamless Authentication: Seamless authentication is a method where the user is authenticated towards an entity without the burden of credential requests. For high security requirements, transparent methods are not applicable, but can provide additional security in traditional username/password or PIN-based sessions.

Session Key: Session key is a short life, randomly generated encryption key to protect one or a group of messages. The main purpose is to use expensive encryption operations only when starting a session and use a simpler to manage cipher in the later part.

This work was previously published in Handbook of Research on Wireless Security, edited by Y. Zhang; J. Zheng; M. Ma, pp. 145-157, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.12
Modeling Fault Tolerant and Secure Mobile Agent Execution in Distributed Systems

H. Hamidi
Iran University of Science & Technology, Iran-Tehran

K. Mohammadi
Iran University of Science & Technology, Iran-Tehran

INTRODUCTION

A mobile agent is a software program which migrates from a site to another site to perform tasks assigned by a user. For the mobile agent system to support agents in various application areas, the issues regarding reliable agent execution, as well as compatibility between two different agent systems or secure agent migration, have been considered. Some of the proposed schemes are either replicating the agents (Hamidi & Mohammadi, 2005) or check-pointing the agents (Park, Byun, Kim, & Yeom, 2002; Pleisch & Schiper, 2001;) For a single agent environment without considering inter-agent communication, the performance of the replication scheme and the check-pointing scheme is compared in Park et al. (2002) and Silva, Batista, and Silva (2000). In the area of mobile agents, only few works can be found relating to fault tolerance. Most of them refer to special agent systems or cover only some special aspects relating to mobile agents, such as the communication subsystem. Nevertheless, most people working with mobile agents consider fault tolerance to be an important issue (Izatt, Chan, & Brecht, 1999; Shiraishi, Enokido, & Takzawa, 2003). Mobile agents are becoming a major trend for designing distributed systems and applications in the last few years and foreseeable future. It can bring benefits such as reduced network load and overcoming of network latency (Chan, Won, & Lyu, 1993). Nevertheless, security is one of the limiting factors of the development of these systems. The main unsolved security problem lies in the possible existence of malicious hosts that can manipulate the execution and data of agents (Defago, Schiper, & Sergent, 1998). Most distributed applications we see today are deploy-

ing the *client/server paradigm*. There are certain problems with the client/server paradigm, such as the requirement of a high network bandwidth, and continuous user-computer interactivity.

In view of the deficiencies of the client/server paradigm, the *mobile code paradigm* has been developed as an alternative approach for distributed application design. In the client/server paradigm, programs cannot move across different machines and must run on the machines they reside on. The mobile code paradigm, on the other hand, allows programs to be transferred among and executed on different computers. By allowing code to move between hosts, programs can interact on the same computer instead of over the network. Therefore, communication cost can be reduced. Besides, *mobile agent* (Fischer, Lynch, & Paterson, 1983) programs can be designed to work on behalf of users autonomously. This autonomy allows users to delegate their tasks to the mobile agents, and not to stay continuously in front of the computer terminal. The promises of the mobile code paradigm bring about active research in its realization. Most researchers, however, agree that security concerns are a hurdle (Greenberg, Byington, & Harper, 1998).

In this article, we investigate these concerns. First, we review some of the foundation materials of the mobile code paradigm. We elaborate Ghezzi and Vigna's classification of mobile code paradigms (Ghezzi & Vigna, 1997), which is a collection of the *remote evaluation, code on demand,* and *mobile agent* approaches. In the next section, we address the current status of mobile code security. The following section presents the model for fault-tolerant mobile agent. In the next section, security issues of the mobile agent are discussed, and we discuss security modeling and evaluation for the mobile agent in the section after. In the following section, simulation results and influence of the size of agent are discussed. We then conclude the article.

THE MOBILE CODE PARADIGM

The mobile code paradigm is essentially a collective term, applicable wherever there is mobility of code. While different classes of code mobility have been identified, Ghezzi and Vigna proposed three of them, namely *remote evaluation, code on demand,* and *mobile agent* (1997). This classification, together with the client/server paradigm, is summarized in Table 1.

In particular, the "know-how" in Table 1 represents the code that is to be executed for the specific task. In the mobile code paradigms (*remote evaluation, code on demand*, and *mobile agent*), the *know-how* moves from one side to another side regarding where the computation takes place; while in the client/server paradigm, the *know-how* is stationary on the remote (server) side. *Resources* are the input and output for the code, whereas *processor* is the abstract machine that carries out and holds the state of the computation. The arrows represent the directions in which the specific item should move before the required task is carried out. Ghezzi and Vigna's classification is found to be comprehensive and representative of most existing mobile code paradigms (such as the rsh utility, Java applets and mobile agent systems), and we will base our discussion on this classification.

SECURITY CONCERNS OF MOBILE CODE PARADIGMS

In this section, we discuss some possible security attacks to different mobile code paradigms, and possible mechanisms against these attacks.

Security Attacks

A security attack is an action that compromises the security requirements of an application. Applications developed using different paradigms are

subject to different attacks. In the conventional client/server model, the local computer is usually assumed to be a secure premise ("*information fortress*") for code and data. This effectively limits the source of security attacks to outsiders of the local machine. Therefore, the main possible attacks are *masquerading* (pretending to be the server or the client), *eavesdropping* on the communication channel, and *forging messages* to the client or the server.

While the security fortress model is usually assumed in the client/server paradigm, it also applies to the *remote evaluation* and *code-on-demand* approaches, with the additional concern that the code receiving side must make sure the code is not harmful to run. In remote evaluation, the code receiving side is the remote side, while it is the local side in code-on-demand.

Mobile agent, on the other hand, is the most challenging area of mobile code security, due to the autonomy of agents. Mobile agent security is usually divided into two aspects: *host security* and *agent security.* Host security deals with the protection of hosts against malicious agents or other hosts, while agent security deals with the protection of agents against malicious hosts or other agents. For host security, the security fortress model can still apply. However, it hardly applies to agent security, due to the lack of trusted hardware with which to anchor security (Tschudin, 1999). There are two branches of new possible attacks to agents:

1. *Data tampering:* A host or another agent may modify the data or execution state being carried by an agent for malicious purpose.
2. *Execution tampering:* A host may change the code executed by an agent, or rearrange the code execution sequence for malicious purpose.

Security Mechanisms

Security mechanisms are mechanisms designed to prevent, detect or recover from security attacks. We see from the previous section that the main security challenges of the client/server paradigm are the mutual trust building between clients and servers, plus the protection of messages in transit. These problems can be satisfactorily solved by cryptographic techniques such as *security protocols* and *message encryption.* These mechanisms are already extensively employed in existing client/server applications. A lot of details can be found in Schneier (1996) and Stallings (1999).

As there are more possible attacks to mobile code paradigms, more mechanisms are required to secure mobile code applications. We see from a previous section that the main additional challenge to security of mobile code paradigms is the verification of the received code. One significant approach to this problem is the *sandbox model.* In the sandbox model, the code or agent received from a remote side can only access a dedicated portion of system resources. Therefore, even if the received code or agent is malicious, damage would be confined to the resources dedicated to that code or agent.

While the sandbox technique is well known and generally accepted for host security, there is yet no good mechanism for agent security. Some approaches have been proposed, and they can be classified into two categories. The first category is agent-tampering detection. These techniques aim at detecting whether an agent's execution or data have been tampered with along the journey. Some possible approaches are *range verification*, *timing information*, *addition of dummy data items and code*, and *cryptographic watermarks* (Tschudin, 1999). Another category is agent-tampering prevention. These techniques aim at preventing agent code or data being tampered with. Two representative approaches are the *execution of encrypted functions* (Sander & Tschudin, 1998) and *time-limited black-boxes* (Hohl, 1998). These approaches are enlightening in the way they open new areas in computer security. Yet they provide limited protection to agents for the time being.

Table 1. Ghezzi and Vigna's (1997) classification of mobile code paradigms

Paradigm		*Local side*	*Remote side*	*Computation takes place at*
Client/server		--	Know-how	Remote side
			Processor	
			Resources	
Mobile code Mobile code Mobile code …… …… …… …… ……	*Remote evaluation*	Know-how ----------	→	Remote side
			Processor	
			Resources	
	Code on demand	←	… Know-how	Local side
		Processor		
		Resources		
	Mobile agent	Know-how ……	→	Remote side
		Processor ----------	→	
			Resources	

Agent protection is still in its early stage, compared with the maturity of protection for hosts and client/servers, and efforts should be spent on improving the already-proposed mechanisms, or developing new protection mechanisms.

MODEL

We assume an asynchronous distributed system, that is, there are no bounds on transmission delays of messages or on relative process speeds. An example of an asynchronous system is the Internet. Processes communicate via message passing over a fully connected network.

Mobile Agent Model

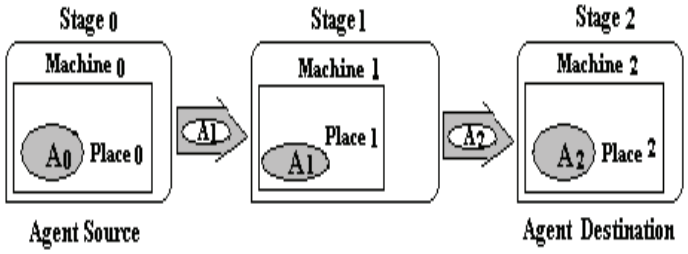
A mobile agent executes on a sequence of machines, where a place P_i ($0 \le i \le n$) provides the logical execution environment for the agent. Each place runs a set of services, which together compose the state of the place. For simplicity, we say that the agent "accesses the state of the place," although access occurs through a service running on the place. Executing the agent at a place P_i is called a stage S_i of the agent execution. We call the places where the first and last stages of an agent execute (i.e., P_i and P_n) the agent source and destination, respectively. The sequence of places between the agent source and destination (i.e., P_0, P_1, … P_n) is called the itinerary of a mobile agent. Whereas a static itinerary is entirely defined at the agent source and does not change during the agent execution, a dynamic itinerary is subject to modifications by the agent itself.

Logically, a mobile agent executes in a sequence of stage actions (Figure 1). Each stage action consists of potentially a multiple set of operations op_0, op_1,…op_n. Agent A_i ($0 \le i \le n$) at the corresponding stage S_i represents the agent that has executed the stage action on places P_j ($j < i$) and is about to execute on place P_i. The execution of A_i at place P_i results in a new internal state of the agent as well as potentially a new state of the place (if the operations of an agent have side effects, i.e., are non idempotent). We denote the resulting agent A_{i+1}. Place P_i forwards to P_{i+1} (for $i < n$).

Fault Model

Several types of faults can occur in agent environments. Here, we first describe a general fault model, and focus on those types, which are important in agent environments due to high

Figure 1. Model of mobile agent execution with three stages



occurrence probability, and those that have been addressed in related work insufficiently.

- Node failures: The complete failure of a compute node implies the failure of all agent places and agents located on it. Node failures can be temporary or permanent.
- Failures of components of the agent system: Failures of agent places, or components of agent places become faulty, for example, faulty communication units or incomplete agent directory. These faults can result in agent failures, or in reduced or wrong functionality of agents.
- Failures of mobile agents: Mobile agents can become faulty due to faulty computation, or other faults (e.g., node or network failures).
- Network failures: Failures of the entire communication network or of single links can lead to isolation of single nodes, or to network partitions.
- Falsification or loss of messages: These are usually caused by failures in the network or in the communication units of the agent systems, or the underlying operating systems. Also, faulty transmission of agents during migration belongs to this type.

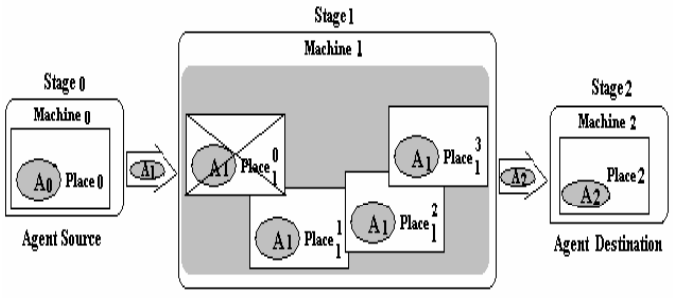
Especially in the intended scenario of parallel applications, node failures and their consequences are important. Such consequences are loss of agents, and loss of node specific resources. In general, each agent has to fulfill a specific task to contribute to the parallel application, and thus, agent failures must be treated with care. In contrast, in applications where a large number of agents are sent out to search and process information in a network, the loss of one or several mobile agents might be acceptable (Pleisch & Schiper, 2000, 2001).

Model Failures

Machines, places, or agents can fail and recover later. A component that has failed but not yet recovered is called down; otherwise, it is up. If it is eventually permanently up, it is called good (Aguilera, 2000). In this article, we focus on crash failures (i.e., processes prematurely halted). Benign and malicious failures (i.e., Byzantine failures) are not discussed. A failing place causes the failure of all agents running on it. Similarly, a failing machine causes all places and agents on this machine to fail as well. We do not consider deterministic, repetitive programming errors (i.e., programming errors that occur on all agent replicas or places) in the code or the place as relevant failures in this context. Finally a link failure causes the loss of messages or agents currently in transmission on this link and may lead to network partitioning. We assume that link failures (and network partitions) are not permanent. The failure of a component (i.e., agent, place, machine, or communication link) can lead to blocking in the mobile agent execution.

Assume, for instance that place P_1 fails while executing A_1 (Figure 2). While P_1 is down, the

Figure 2. The redundant places mask the place failure



execution of the mobile agent cannot proceed, that is, it is blocked. Blocking occurs if a single failure prevents the execution from proceeding. In contrast, an execution is non-blocking if it can proceed despite a single failure, the blocked mobile agent execution can only continue when the failed component recovers. This requires that recovery mechanism be in place, which allows the failed component to be recovered. If no recovery mechanism exists, then the agent's state and potentially its code may be lost. In the following, we assume that such a recovery mechanism exists (e.g., based on logging [Ghezzi & Vigna, 1997]). Replication prevents blocking. Instead of sending the agent to one place at the next stage, agent replicas are sent to a set of M_i places P_i^0, P_i^1,.., P_i^n (Figure 2). We denote by A_i^j the agent replica of A_i executing on place P_i^j, but will omit the superscripted index if the meaning is clear from the context. Although a place may crash (i.e., Stage 1 in Figure 2), the agent execution does not block. Indeed, P_2^1 can take over the execution of a1 and thus prevent blocking. Note that the execution at stages S_0 and S_2 is not replicated as the agent is under the control of the user. Moreover, the agent is only configured at the agent source and presents the results to the agent owner at the agent destination. Hence, replication is not needed at these stages.

Despite agent replication, network partitions can still prevent the progress of the agent. Indeed, if the network is partitioned such that all places currently executing the agent at stage S_i are in one partition and the places of stage S_{i+1} are in another partition, the agent cannot proceed with its execution. Generally (especially on the Internet), multiple routing paths are possible for a message to arrive at its destination. Therefore, a link failure may not always lead to network partitioning. In the following, we assume that a single link failure merely partitions one place from the rest of the network. Clearly, this is a simplification, but it allows us to define blocking concisely. Indeed, in the approach presented in this article, progress in the agent execution is possible in a network partition that contains a majority of places. If no such partition exists, the execution is temporally interrupted until a majority partition is established again. Moreover, catastrophic failures may still cause the loss of the entire agent. A failure of all places in M_1 (Figure 2), for instance, is such a catastrophic failure (assuming no recovery mechanism is in place). As no copy of A_1 is available any more, the agent A_1 is lost and, obviously, the agent execution can no longer proceed. In other words, replication does not solve all problems. The definition of non-blocking merely addresses single failures per stage as they cover most of the failures that occur in a realistic environment.

SECURITY ISSUES OF THE MOBILE AGENT

Any distributed system is subject to security threats, so is a mobile agent system. Issues such

as encryption, authorization, authentication, non-repudiation should be addressed in a mobile agent system. In addition, a secure mobile agent system must protect the hosts as well as the agents from being tampered with by malicious parties.

First, hosts must be protected because they continuously receive agents and execute them. They may not be sure where an agent comes from, and are at the risk of being damaged by malicious code or agents (Trojan horse attack). This problem can be effectively solved by strong authentication of the code sources, verification of code integrity, and limiting the access rights of incoming agents to local resources of hosts. This is mostly realized by the Java security model (Hohl, 1998). The main security challenge of mobile agent systems lies on the protection of agents. When an agent executes on a remote host, the host is likely to have access to all the data and code carried by the agent. If by chance a host is malicious and abuses the code or data of an agent, the privacy and secrecy of the agent and its owner would be at risk.

Seven types of attack by malicious hosts (Defago, Schiper, & Sergent, 1998) can be identified:

1. Spying out and manipulation of code;
2. Spying out and manipulation of data;
3. Spying out and manipulation of control flow;
4. Incorrect execution of code;
5. Masquerading of the host;
6. Spying out and manipulation of interaction with other agents; and
7. Returning wrong results of system calls to agents.

There are a number of solutions proposed to protect agents against malicious hosts (Chan et al., 1993), which can be divided into three streams:

- Establishing a closed network: Limiting the set of hosts among which agents travel such that agents travel only to hosts that are trusted.
- Agent tampering detection: Using specially designed state-appraisal functions to detect whether agent states have been changed maliciously during its travel.
- Agent tampering prevention: Hiding from hosts the data possessed by agents and the functions to be computed by agents, by messing up code and data of agents, or using cryptographic techniques.

None of the proposed solutions solve the problem completely. They either limit the capabilities of mobile agents, or are not restrictive enough. A better solution is being sought, and there is no general methodology suggested to protect agents. In the mean time, developers of mobile agent systems have to develop their own methodologies according to their own needs. Apart from attacks by malicious hosts, it is also possible that an agent attacks another agent. However, this problem, when compared with the problem of malicious hosts, is less important, because the actions of a (malicious) agent to another agent can be effectively monitored and controlled by the host on which the agent runs, if the host is not malicious.

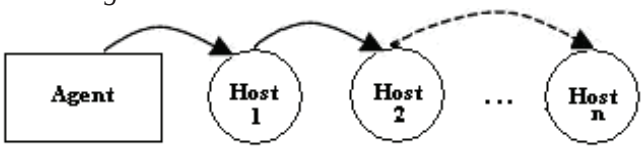
SECURITY MODELING AND EVALUATION FOR THE MOBILE AGENT

There is no well-established model for mobile agent security. One of the few attempts so far is given in Hohl (1998). Software reliability modeling is a successful attempt to give quantitative measures of software systems. In the broadest sense, security is one of the aspects of reliability. A system is likely to be more reliable if it is more secure. One of the pioneering efforts to integrate security and reliability is (Brocklehurst, Littlewood, Olovsoon, & Jonsson, 1994). In this article,

Table 2. Analogy between reliability and security

Security	Reliability
Vulnerabilities	Faults
Breach	Failure
Fail upon attack effort spent	Fail upon usage time elapsed

Figure 3. A mobile agent traveling on a network



the following similarities between security and reliability were observed.

Thus, we have *security function, effort to next breach distribution,* and *security hazard rate* similar to the *reliability function, time to next failure distribution,* and *reliability hazard rate* respectively as in reliability theory. One of the works to incorporate system security into a mathematical model is (Jonsson, 1997), which presents an experiment to model the attacker behavior. The results show that during the "standard attack phase," assuming breaches are independent and stochastically identical, the period of working time of a single attacker between successive breaches is found to be exponentially distributed.

Now, let us consider a mobile agent traveling through *n* hosts on the network, as illustrated in Figure 3. Each host, and the agent itself, is modeled as an abstract machine as in Hohl (1998). We consider only the standard attack phase described in Jonsson (1997)) by malicious hosts. On arrival at a malicious host, the mobile agent is subject to an attack effort from the host. Because the host is modeled as a machine, it is reasonable to estimate the attack effort by the number of instructions for the attack to carry out, which would be linearly increasing with time. On arrival at a non-malicious host, the effort would be constant zero. Let the agent arrive at host *i* at time T_i, for $i = 1, 2, ..., n$. Then the effort of host *i* at total time *t* would be described by the *time-to-effort function*:

$E_i(t) = k_i(t-T_i)$, where k is a constant

We may call the constant k_i the *coefficient of malice*. The larger the k_i, the more malicious host *i* is *(ki* = 0 if host *i* is non-malicious). Furthermore, let the agent stay on host *i* for an amount of time t_i, then there would be breach to the agent if and only if the following breach condition holds:

$E_i(t_i+T_i)$ > effort to next breach by host *i*

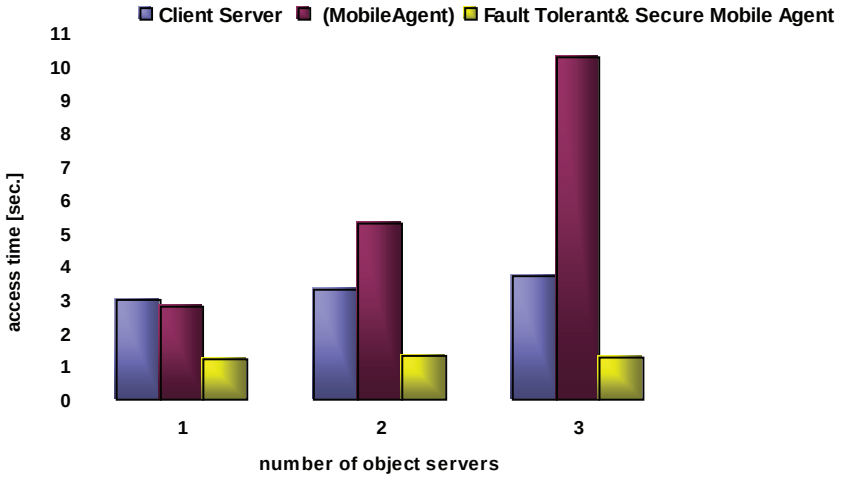
that is, $k_i t_i$ > effort to next breach by host *i*

As seen from Brocklehurst et al. (1994) and Jonsson (1997), it is reasonable to assume exponential distribution of the effort to next breach, so we have the *probability of breach at host i*,

P(breach at host *i*)
= P(breach at time $t_i + T_i$)
= P(breach at effort $k_i t_i$)
= $1 - exp(-vk_i t_i)$, *v* is a constant
= $1 - exp(-\lambda_i t_i)$, $\lambda_i = vk_i$

We may call v the *coefficient of vulnerability* of the agent. The higher the v, the higher is the

Figure 4. Access time for number of object servers



probability of breach to the agent. Therefore, the *agent security E* would be the probability of no breach at all hosts, that is,

Suppose that we can estimate the coefficients of malice k_i's for hosts based on trust records of hosts, and also estimate the coefficient of vulnerability v of the agent based on testing and experiments, then we can calculate the desired time limits T_i's to achieve a certain level of security *E*. Conversely, if users specify some task must be carried out on a particular host for a fixed period of time, we can calculate the agent security *E* for the users based on the coefficients of malice and vulnerability estimates.

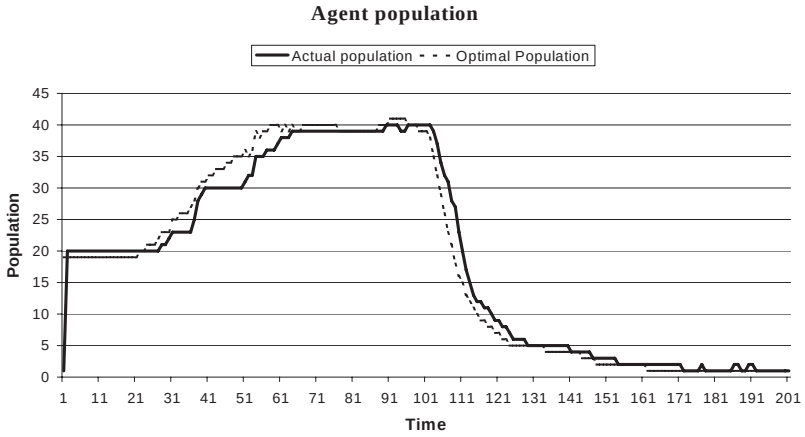
EVALUATION RESULTS AND INFLUENCE OF THE SIZE OF THE AGENT

We evaluate transactional agents in terms of access time compared with client- server model. The computation of mobile agents is composed of moving, class loading, manipulation of objects, creation of clone, and commitment steps. In the client-server model, there are computation steps of program initialization, class loading to client, manipulation of objects, and two-phase commitment.

Access time from the time when the application program starts to the time when the application program ends is measured for agents and the client-server model. Figure 4 shows the access time for a number of object servers. The non-fault tolerant and secure mobile agents show that mobile agent classes are not loaded when an agent A_i arrives at an object server. Here, the agent can be executed after Aglets classes are loaded. On the other hand, the fault tolerant and secure mobile agents mean that an agent manipulates objects in each object server where mobile agent classes are already loaded, that is, the agent comes to the object server after other agents have visited on the object server. As shown in Figure 4, the client-server model is faster than the transactional agent. However, the transactional agent is faster than the client-server model if object servers are frequently manipulated, that is, fault tolerant and secure mobile agent classes are a priori loaded.

A simulator was designed to evaluate the algorithm. The system was tested in several simulated network conditions and numerous parameters were introduced to control the behavior of the agents. We also investigated the dynamic functioning of the algorithm. Comparing to the previous case, the parameter configuration has a larger effect on the behavior of the system.

Figure 5. The size of the agent population under changing network conditions

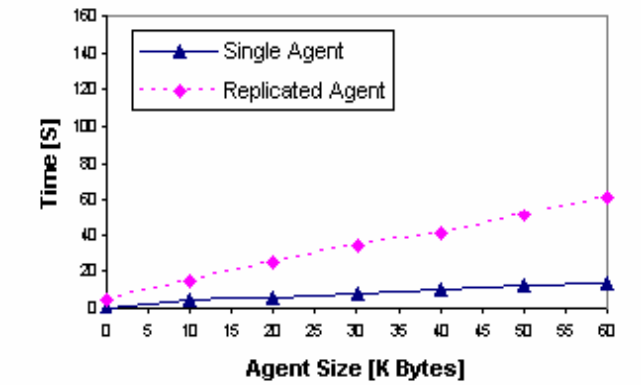


The most vital parameter was the frequency of the trading process and the pre-defined critical workload values.

Figure 5 shows the number of agents on the network in a dynamic network situation. The optimal agent population is calculated by dividing the workload on the whole network with the optimal workload of the agent. Simulation results show that by choosing the correct agent parameters the workload of agents is within ten percent of the predefined visiting frequency on a stable network. In a simulated network the population overload dynamically grows to meet the increased requirements and smoothly returns back to normal when the congestion is over.

To measure the performance of fault tolerant mobile agent system our test consists of sequentially sending a number of agents that increment the value of the counter at each stage of the execution. Each agent starts at the agent source and returns to the agent destination, which allows us to measure its round–trip time. Between two agents, the places are not restarted. Consequently, the first agent needs considerably longer for its execution, as all classes need to be loaded into the cache of the virtual machines. Consecutive agents benefit from already cached classes and thus execute much faster. We do not consider the first agent execution in our measurement results. For a fair comparison, we used the same

Figure 6. Costs of single and replicated agent execution increasing agent size



approach for the single agent case (no replication). Moreover, we assume that the Java class files are locally available in each place. Clearly, this is a simplification, as the class files do not need to be transported with the agent. Remote class loading adds additional costs because the classes have to be transported with the agent and then loaded into the virtual machine. However, once the classes are loaded into the class loader, other agents can take advantage of them and do not need to load these classes again.

The size of the agent has a considerable impact on the performance of the fault-tolerant mobile agent execution. To measure this impact, the agent carries a Byte array of variable length used to increase the size of the agent. As the results in Figure 6 show, the execution time of the agent increases linearly with increasing size of the agent. Compared to the single agent, the slope of the curve for the replicated agent is steeper.

CONCLUSION

In this article, we have presented the mobile code paradigm, which is a collection of remote evaluation, code on demand, and mobile agents, as an alternative to the conventional client/server paradigm. We examine security concerns of the mobile code paradigm, and survey existing security attacks and mechanisms to evaluate the current status of mobile code security. We conclude that the mobile code paradigm is still to be developed with respect to its security aspects and that mobile agent protection needs particular attention. To investigate the security threats to mobile agents, we implemented a simple Traveling Information Agent System, and discussed the possible attacks to the agents in this system, based on the attack model in [26].

We have identified two important properties for fault-tolerant mobile agent execution: non-blocking and exactly-once. Non-blocking ensures that the agent execution proceeds despite a single failure of either agent, place, or machine. Blocking is prevented by the use of replication. This article discussed a mobile agent model for processing transactions, which manipulate object servers. An agent first moves to an object server and then manipulates objects.

General possibilities for achieving fault tolerance in such cases were discussed and the respective advantages and disadvantages for mobile agent environments and the intended parallel and distributed application scenarios were shown. This leads to an approach based on warm standby and receiver side message logging. We have used dynamically changing agent domains to provide flexible, adaptive and robust operation. The performance measurement of Fault-Tolerant Mobile Agent System shows the overhead introduced by the replication mechanisms with respect to a non-replicated agent. Not surprisingly, it also shows that this overhead increases with the number of stages and the size of the agent.

REFERENCES

Aguilera, M. K., Chen, W. & Toueg, S. (2000). Failure detection and consensus in the crash-recovery model. *Distributed Computing, 13*(2), 99-125.

Brocklehurst, S., Littlewood, B., Olovsson, T., & Jonsson, E. (1994). On measurement of operational security. In *Proceedings of the Ninth Conference on Computer Assurance (COMPASS'94): Safety, Reliability, Fault Tolerance and Real Time, Security* (pp. 257-266).

Chan, H. W., Wong , K. M., & Lyu, R. (1993). Design ,implementation ,and experimentation on mobile agent security for electronic commerce application. In S. Mullender (Ed.), *Distributed systems* (2nd ed.) (pp. 199-216), Reading, MA: Addison-Wesley.

Chess, D., Harrison, C. G., & Kershenbaum, A. (1998). Mobile agents: Are they a good idea? In G. Vigna (Ed.), *Mobile agents and security* (pp. 25-47). Springer-Verlag.

Defago, X., Schiper, A. & Sergent, N. (1998, October). Semi-passive replication. In *Proceedings of the 17th IEEE Symposium on Reliable Distributed System (SRDS'98)* (pp. 43-50).

Fischer, M. J., Lynch, N. A. & Paterson, M. S. (1983, March). Impossibility of distributed consensus with one faulty process. In *Proceedings of the second ACM SIGACT-SIGMOD Symposium: Principles of Database System* (p. 17).

Ghezzi, C. & Vigna, G. (1997, April). Mobile code paradigms and technologies: A case study. In K. Rothermet, R. Popescu-Zeletin (Eds.), *Mobile Agents, First International Workshop, MA'97, Proceedings, LNCS 1219* (pp. 39-49), Berlin, Germany. Springer.

Greenberg, M. S., Byington, J. C., & Harper, D. G. (1998). Mobile agents and security. *IEEE Communications Magazine, 367.*

Hamidi, H. & Mohammadi, K. (2005, March). Modeling and evaluation of fault tolerant mobile agents in distributed systems. In *Proceedings of the 2nd IEEE Conference on Wireless & Optical Communications Networks (WOCN2005)* (pp. 91-95).

Hohl, F. (1998a) Time limited Blackbox security: Protecting mobile agents from malicious hosts. In G. Vigna (Ed.), *Mobile agents and security, LNCS 1419* (pp. 92-113). Springer.

Hohl, F. (1998b). A model of attacks of malicious hosts against mobile agents. In *Fourth Workshop on Mobile Object Systems (MOS'98): Secure Internet Mobile Computations.* Retrieved from http://cuiwww.unige.ch/~ecoopws/ws98/papers/hohl.ps

Hohl, F. (1998c). A model of attacks of malicious hosts against mobile agents. In *Proceedings of the ECOOP Workshop on Distributed Object Security and 4th Workshop on Object Systems*: *Secure Internet mobile computations* (pp. 105-120), Inria, France.

Izatt, M., Chan, P., & Brecht, T. (1999, June). Agents: Towards an environment for parallel, distributed and mobile Java applications. In *Proceedings of the 1999 ACM Conference on Java Grande* (pp. 15-24).

Jonsson, E. (1997). A quantitative model of the security intrusion process based on attacker behavior. *IEEE Transactions on Software Engineering, 23*(4).

Park, T., Byun, I., Kim, H. & Yeom, H. Y. (2002). The performance of checkpointing and replication schemes for fault tolerant mobile agent systems. In *Proceedings of the 21st IEEE Symposium on Reliable Distributed Systems.*

Pleisch, S. & Schiper, A. (2000). Modeling fault: Tolerant mobile agent execution as a sequence of agree problems. In *Proceedings of the 19th IEEE Symposium on Reliable Distributed Systems* (pp. 11-20).

Pleisch, S. & Schiper, A. (2001, July). FATOMAS — A Fault-Tolerant Mobile Agent System based on the agent-dependent approach. In *Proceedings of the 2001 International Conference on Dependable Systems and Networks* (pp. 215-224).

Pleisch, S. & Schiper, A. (2003). Fault-tolerant mobile agent execution. *IEEE Transactions on Computers, 52*(2).

Sander, T. & Tschudin, C. F. (1998). Protecting mobile agents against malicious hosts. In G. Vigna (Ed.), *Mobile agents and security, LNCS 1419* (pp. 44-60). Springer.

Schneier, B. (1996). *Applied cryptography.* Wiley.

Shiraishi, M., Enokido, T. & Takizawa, M. (2003). Fault-tolerant mobile agents in distributed objects

systems. In *Proceedings of the Ninth IEEE Workshop on Future Trends of Distributed Computer Systems (FTDCS, 03)* (pp. 11-20).

Silva, L. Batista, V., & Silva, L.G. (2000). Fault-tolerant execution of mobile agents. In *Proceedings of the International Conference on Dependable Systems and lIIenvorks.*

Stallings, W. (1999). *Cryptography and network security, principles and practice*. Prentice Hall.

Strasser, M. & Rothermel, K. (2000). System mechanism for partial rollback of mobile agent execution. In *Proceedings of the 20th International Conference on Distributed Computing Systems.*

Tschudin, C. F. (1999). Mobile agent security. In M. Klusch (Ed.), *Intelligent information agents* [Forthcoming *LNCS*]. Retrieved from http://www.docs.uu.se/~tschudin/pub/cft-1999-iia.ps.gz

This work was previously published in International Journal of Intelligent Information Technologies, Vol. 2, Issue 1, edited by V. Sugumaran , pp. 21-36, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.13
Security in 2.5G Mobile Systems

Christos Xenakis
University of Piraeus, Greece

ABSTRACT

The global system for mobile communications (GSM) is the most popular standard that implements second generation (2G) cellular systems. 2G systems combined with general packet radio services (GPRS) are often described as 2.5G, that is, a technology between the 2G and third generation (3G) of mobile systems. GPRS is a service that provides packet radio access for GSM users. This chapter presents the security architecture employed in 2.5G mobile systems focusing on GPRS. More specifically, the security measures applied to protect the mobile users, the radio access network, the fixed part of the network, and the related data of GPRS are presented and analyzed in detail. This analysis reveals the security weaknesses of the applied measures that may lead to the realization of security attacks by adversaries. These attacks threaten network operation and data transfer through it, compromising end users and network security. To defeat the identified risks, current research activities on the GPRS security propose a set of security improvements to the existing GPRS security architecture.

INTRODUCTION

The global system for mobile communications, (GSM) is the most popular standard that implements second generation (2G) cellular systems. 2G systems combined with general packet radio services (GPRS) (3GPP TS 03.6, 2002) are often described as 2.5G, that is, a technology between the 2G and third generation (3G) of mobile systems. GPRS is a service that provides packet radio access for GSM users. The GPRS network architecture, which constitutes a migration step toward 3G systems, consists of an overlay network onto the GSM network. In the wireless part, the GPRS technology reserves radio resources only when there is data to be sent, thus, ensuring the optimized utilization of radio resources. The fixed part of the network employs the IP technology and is connected to the public Internet. Taking advantage of these features, GPRS enables the provision of a variety of packet-oriented multimedia applications and services to mobile users, realizing the concept of the mobile Internet.

For the successful implementation of the new emerging applications and services over GPRS, security is considered as a vital factor. This is

because of the fact that wireless access is inherently less secure and the radio transmission is by nature more susceptible to eavesdropping and fraud in use than wire-line transmission. In addition, users' mobility and the universal access to the network imply higher security risks compared to those encountered in fixed networks. In order to meet security objectives, GPRS uses a specific security architecture, which aims at protecting the network against unauthorized access and the privacy of users. This architecture is mainly based on the security measures applied in GSM, since the GPRS system is built on the GSM infrastructure.

Based on the aforementioned consideration, the majority of the existing literature on security in 2.5G systems refers to GSM (Mitchell, 2001; Pagliusi, 2002). However, GPRS differs from GSM in certain operational and service points, which require a different security analysis. This is because GPRS is based on IP, which is an open and wide deployed technology that presents many vulnerable points. Similarly to IP networks, intruders to the GPRS system may attempt to breach the confidentiality, integrity, or availability, or otherwise attempt to abuse the system in order to compromise services, defraud users, or any part of it. Thus, the GPRS system is more exposed to intruders compared to GSM.

This chapter presents the security architecture employed in 2.5G mobile systems focusing on GPRS. More specifically, the security measures applied to protect the mobile users, the radio access network, the fixed part of the network, and the related data of GPRS are presented and analyzed in details. This analysis reveals the security weaknesses of the applied measures that may lead to the realization of security attacks by adversaries. These attacks threaten network operation and data transfer through it, compromising end users and network security. To defeat the identified risks, current research activities on the GPRS security propose a set of security improvements to the existing GPRS security architecture. The rest of this chapter is organized as follows. The next section describes briefly the GPRS network architecture. The third section presents the security architecture applied to GPRS and the fourth section analyzes its security weaknesses. The fifth section elaborates on the current research activities on the GPRS security and the sixth section presents the conclusions.

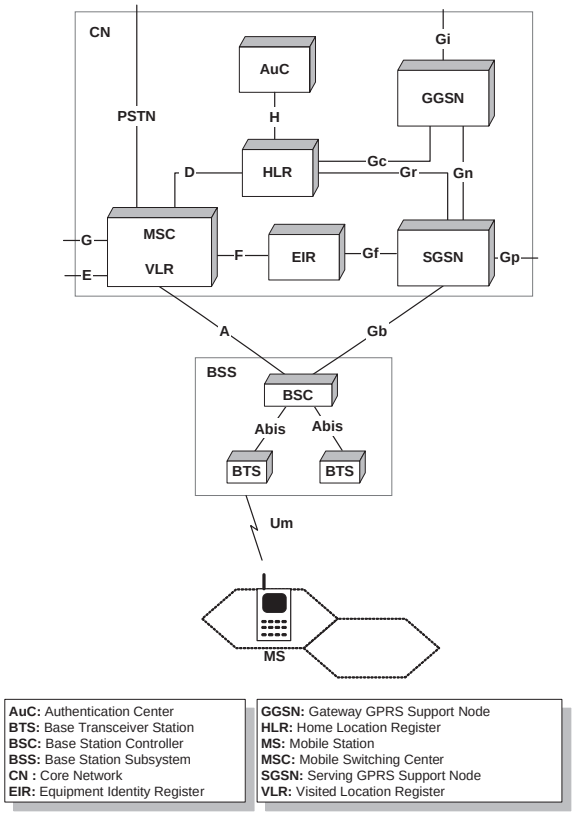
GPRS NETWORK ARCHITECTURE

The network architecture of GPRS (3GPP TS 03.6, 2002) is presented in Figure 1. A GPRS user owns a mobile station (MS) that provides access to the wireless network. From the network side, the base station subsystem (BSS) is a network part that is responsible for the control of the radio path. BSS consists of two types of nodes: the base station controller (BSC) and the base transceiver station (BTS). BTS is responsible for the radio coverage of a given geographical area, while BSC maintains radio connections towards MSs and terrestrial connections towards the fixed part of the network (core network).

The GPRS core network (CN) uses the network elements of GSM such as the home location register (HLR), the visitor location register (VLR), the authentication centre (AuC) and the equipment identity register (EIR). HLR is a database used for the management of permanent data of mobile users. VLR is a database of the service area visited by an MS and contains all the related information required for the MS service handling. AuC maintains security information related to subscribers identity, while EIR maintains information related to mobile equipments' identity. Finally, the mobile service switching centre (MSC) is a network element responsible for circuit-switched services (e.g., voice call) (3GPP TS 03.6, 2002).

As presented previously, GPRS reuses the majority of the GSM network infrastructure. However, in order to build a packet-oriented mobile network some new network elements (nodes) are

Figure 1. GPRS network architecture



required, which handle packet-based traffic. The new class of nodes, called GPRS support nodes (GSN), is responsible for the delivery and routing of data packets between an MS and an external packet data network (PDN). More specifically, a serving GSN (SGSN) is responsible for the delivery of data packets from, and to, an MS within its service area. Its tasks include packet routing and transfer, mobility management, logical link management, and authentication and charging functions. A gateway GSN (GGSN) acts as an interface between the GPRS backbone and an external PDN. It converts the GPRS packets coming from the SGSN into the appropriate packet data protocol (PDP) format (e.g., IP), and forwards them to the corresponding PDN. Similar is the functionality of GGSN in the opposite direction. The communication between GSNs (i.e., SGSN and GGSN) is based on IP tunnels through the use of the GPRS tunneling protocol (GTP) (3GPP TS 09.60, 2002).

GPRS SECURITY ARCHITECTURE

In order to meet security objectives, GPRS employs a set of security mechanisms that constitutes the GPRS security architecture. Most of these mechanisms have been originally designed for GSM, but they have been modified to adapt to the packet-oriented traffic nature and the GPRS network components. The GPRS security architecture, mainly, aims at two goals: (1) to protect the network against unauthorized access, and (2) to protect the privacy of users. It includes the following components (GSM 03.20, 1999):

- Subscriber identity module (SIM)
- Subscriber identity confidentiality
- Subscriber identity authentication
- User data and signaling confidentiality between the MS and the SGSN
- GPRS backbone security

Subscriber Identity Module (SIM)

The subscription of a mobile user to a network is personalized through the use of a smart card named SIM (ETSI TS 100 922, 1999). Each SIM card is unique and related to a user. It has a microcomputer with a processor, ROM, persistent EPROM memory, volatile RAM, and an I/O interface. Its software consists of an operating system, file system, and application programs (e.g., SIM application toolkit). The SIM card is responsible for the authentication of the user by prompting for a code (PIN), the identification of the user to a network through keys, and the protection of user data through cryptography. To achieve these functions it contains a set of security objects including:

- A (4-digit) PIN code, which is used to lock the card preventing misuse;
- A unique permanent identity of the mobile user, named international mobile subscriber identity (IMSI) (3GPP TS 03.03, 2003);
- A secret key, Ki, (128 bit) that is used for authentication; and
- An authentication algorithm (A3) and an algorithm that generates encryption keys (A8) (GSM 03.20, 1999).

Since the SIM card of a GSM/GPRS subscriber contains security critical information, it should be manufactured, provisioned, distributed, and managed in trusted environments.

Subscriber Identity Confidentiality

The subscriber identity confidentiality deals with the privacy of the IMSI and the location of a mobile user. It includes mechanisms for the protection of the permanent identity (IMSI) when it is transferred in signaling messages, as well as measures that preclude the possibility to derive it indirectly from listening to specific information, such as addresses, at the radio path.

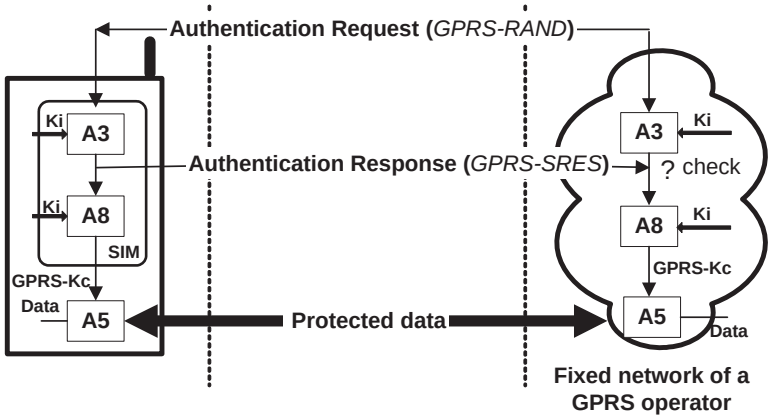
The subscriber identity confidentiality is mainly achieved by using a temporary mobile subscriber identity (TMSI) (3GPP TS 03.03, 2003; GSM 03.20, 1999), which identifies the mobile user in both the wireless and wired network segments. The TMSI has a local significance and thus it must be accompanied by the routing area identity (RAI) in order to avoid confusions. The MS and the serving VLR and SGSN only know the relation between the active TMSI and the IMSI. The allocation of a new TMSI corresponds implicitly for the MS to the de-allocation of the previous one. When a new TMSI is allocated to the MS, it is transmitted to it in a ciphered mode. The MS stores the current TMSI and the associated RAI in a non-volatile memory, so that these data are not lost when the MS is switched off.

Further to the TMSI, a temporary logical link identity (TLLI) (3GPP TS 03.03, 2003) identifies also a GPRS user on the radio interface of a routing area. Since the TLLI has a local significance, when it is exchanged between the MS and the SGSN, it should be accompanied by the RAI. The TLLI is either derived from the TMSI allocated by the SGSN or built by the MS randomly and thus, provides identity confidentiality. The relationship between the TLLI and the IMSI is only known in the MS and in the SGSN.

Subscriber Identity Authentication

A mobile user that attempts to access the network must first prove his/her identity to it. User authentication (3GPP TS 03.6, 2002) protects against

Figure 2. GPRS authentication



fraudulent use and ensures correct billing. GPRS uses the authentication procedure already defined in GSM with the same algorithms for authentication and generation of encryption key, and the same secret key, Ki, (see Figure 2). However, from the network side, the whole procedure is executed by the SGSN (instead of the BS) and employs a different random number (GPRS-RAND) and thus, it produces a different signed response (GPRS-SRES) and encryption key (GPRS-Kc) than the GSM voice counterpart.

To achieve authentication of a mobile user, the serving SGSN must possess security-related information for the specific user. This information is obtained by requesting the HLR/AuC of the home network that the mobile user is subscribed. It includes a set of authentication vectors, each of which includes a random challenge (GPRS-RAND), the related signed response (GPRS-SRES), and the encryption key (GPRS-Kc) for the specific subscriber. The authentication vectors are produced by the home HLR/AuC using the secret key Ki of the mobile subscriber.

During authentication the SGSN of the serving network sends the random challenge (GPRS-RAND) of a chosen authentication vector to the MS. The latter encrypts the GPRS-RAND by using the A3 hash algorithm, which is implemented in the SIM card, and the secret key, Ki. The first 32 bits of the A3 output are used as a signed response (GPRS-SRES) to the challenge (GPRS-RAND) and are sent back to the network. The SGSN checks if the MS has the correct key, Ki, and, then, the mobile subscriber is recognized as an authorized user. Otherwise, the serving network (SN) rejects the subscriber's access to the system. The remaining 64 bits of the A3 output together with the secret key, Ki, are used as input to the A8 algorithm that produces the GPRS encryption key (GPRS-Kc).

Data and Signalling Protection

User data and signaling protection over the GPRS radio access network is based on the GPRS ciphering algorithm (GPRS-A5) (3GPP TS 01.61, 2001), which is also referred to as GPRS encryption algorithm (GEA) and is similar to the GSM A5. Currently, there are three versions of this algorithm: GEA1, GEA2, and GEA3 (that is actually A5/3), which are not publicly known and thus, it is difficult to perform attacks on them. The MS device (not the SIM-card) performs GEA using the encryption key (GPRS-Kc), since it is a strong algorithm that requires relatively high processing capabilities. From the network side, the serving

SGSN performs the ciphering/deciphering functionality protecting signaling and user data over the Um, Abis, and Gb interfaces.

During authentication the MS indicates which version(s) of the GEA supports and the network (SGSN) decides on a mutually acceptable version that will be used. If there is not a commonly accepted algorithm, the network (SGSN) may decide to release the connection. Both the MS and the SGSN must cooperate in order to initiate the ciphering over the radio access network. More specifically, the SGSN indicates whether ciphering should be used or not (which is also a possible option) in the *Authentication Request* message, and the MS starts ciphering after sending the *Authentication Response* message (see Figure 2).

GEA is a symmetric stream cipher algorithm (see Figure 3) that uses three input parameters (GPRS-Kc, INPUT, and DIRECTION) and produces an OUTPUT string, which varies between 5 and 1,600 bytes. GPRS-Kc (64 bits) is the encryption key generated by the GPRS authentication procedure and is never transmitted over the radio interface. The input (INPUT) parameter (32 bits) is used as an additional input so that each frame is ciphered with a different output string. This parameter is calculated from the logical link control (LLC) frame number, a frame counter, and a value supplied by the SGSN called the input offset value (IOV). The IOV is set up during the negotiation of LLC and layer 3 parameters. Finally, the direction bit (DIRECTION) specifies whether the output string is used for upstream or downstream communication.

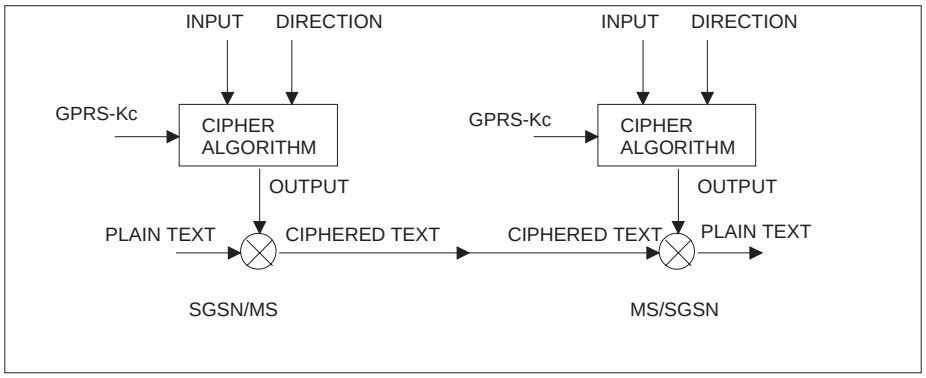
After the initiation of ciphering, the sender (MS or SGSN) processes (bit-wise XOR) the OUTPUT string with the payload (PLAIN TEXT) to produce the CIPHERED TEXT, which is sent over the radio interface. In the receiving entity (SGSN or MS), the original PLAIN TEXT is obtained by bit-wise XORed the OUTPUT string with the CIPHERED TEXT. When the MS changes SGSN, the encryption parameters (e.g., GPRS-Kc, INPUT) are transferred from the old SGSN to the new SGSN, through the (inter) routing area update procedure in order to guarantee service continuity.

GPRS Backbone Security

The GPRS backbone network includes the fixed network elements and their physical connections that convey user data and signaling information. signaling exchange in GPRS is mainly based on the signaling system 7 (SS7) technology (3GPP TS 09.02, 2004), which does not support any security measure for the GPRS deployment. Similarly, the GTP protocol that is employed for communication between GSNs does not support security. Thus, user data and signaling information in the GPRS backbone network are conveyed in cleartext exposing them to various security threats. In addition, inter-network communications (between different operators) are based on the public Internet, which enables IP spoofing to any malicious third party who gets access to it. In the sequel, the security measures applied to the GPRS backbone network are presented.

The responsibility for security protection of the GPRS backbone as well as inter-network communications belongs to mobile operators. They utilize private IP addressing and network address translation (NAT) (Srisuresh & Holdrege, 1999) to restrict unauthorized access to the GPRS backbone. They may also apply firewalls at the borders of the GPRS backbone network in order to protect it from unauthorized penetrations. Firewalls protect the network by enforcing security policies (e.g., user traffic addressed to a network element is discarded). Using security policies the GPRS operator may ensure that only traffic initiated from the MS and not from the Internet should pass through a firewall. This is done for two reasons: (1) to restrict traffic in order to protect the MS and the network elements from external attacks; and (2) to protect the MS from receiving unrequested traffic. Unrequested traffic may be

Figure 3. GPRS ciphering



unwanted for the mobile subscribers since they pay for the traffic received as well. The GPRS operator may also want to disallow some bandwidth-demanding protocols preventing a group of subscribers to consume so much bandwidth that other subscribers are noticeably affected. In addition, application-level firewalls prevent direct access through the use of proxies for services, which analyze application commands, perform authentication, and keep logs.

Since firewalls do not provide privacy and confidentiality, the virtual private network (VPN) technology (Gleeson, Lin, Heinanen, Armitage, & Malis, 2000) has to complement them to protect data in transit. A VPN is used for the authentication and the authorization of user access to corporate resources, the establishment of secure tunnels between the communicating parties, and the encapsulation and protection of the data transmitted by the network. In current GPRS implementations, pre-configured, static VPNs can be employed to protect data transfer between GPRS network elements (e.g., an SGSN and a GGSN that belong to the same backbone), between different GPRS backbone networks that belong to different mobile operators, or between a GPRS backbone and a remote corporate private network. The border gateway, which resides at the border of the GPRS backbone, is a network element that provides firewall capabilities and also maintains static, pre-configured VPNs to specific peers.

GPRS SECURITY WEAKNESSES

Although GPRS have been designed with security in mind, it presents some essential security weaknesses, which may lead to the realization of security attacks that threaten network operation and data transfer through it. In the following, the most prominent security weaknesses of the GPRS security architecture are briefly presented and analyzed.

Subscriber Identity Confidentiality

A serious weakness of the GPRS security architecture is related to the compromise of the confidentiality of subscriber identity. Specifically, whenever the serving network (VLR or SGSN) cannot associate the TMSI with the IMSI, because of TMSI corruption or database failure, the SGSN should request the MS to identify itself by means of IMSI on the radio path. Furthermore, when the user roams and the new serving network cannot contact the previous (the old serving network) or cannot retrieve the user identity, then, the new serving network should also request the MS to identify itself by means of IMSI on the radio path.

This fact may lead an active attacker to pretend to be a new serving network, to which the user has to reveal his/her permanent identity. In addition, in both cases the IMSI that represents the permanent user identity is conveyed in cleartext over the radio interface violating user identity confidentiality.

Subscriber Authentication

The authentication mechanism used in GPRS also exhibits some weak points regarding security. More specifically, the authentication procedure is one way and thus, it does not assure that a mobile user is connected to an authentic serving network. This fact enables active attacks using a false BS identity. An adversary, who has the required equipment, may masquerade as a legitimate network element mediating in the communication between the MS and the authentic BS. This is also facilitated by the absence of a data integrity mechanism on the radio access network of GPRS, which defeats certain network impersonation attacks. The results of this mediation may be the alternation or the interception of signaling information and communication data exchanged.

Another weakness of the GPRS authentication procedure is related to the implementation of the A3 and A8 algorithms, which are often realized in practise using COMP128. COMP128 is a keyed hash function, which uses two 16-byte (128 bits) inputs and produces a hash output of 12 bytes (96 bits). While the actual specification of COMP128 was never made public, the algorithm has been reverse engineered and cryptanalyzed (Barkan, Biham, & Neller, 2003). Thus, knowing the secret key, Ki, it is feasible for a third party to clone a GSM/GPRS SIM-card, since its specifications are widely available (ETSI TS 100 922, 1999).

The last weakness of the GPRS authentication procedure is related to the network ability of reusing authentication triplets. Each authentication triplet should be used only in one authentication procedure in order to avoid man-in-the-middle and replay attacks. However, this depends on the mobile network operator (home and serving) and cannot be checked by mobile users. When the VLR of a serving network has used an authentication triplet to authenticate an MS, it shall delete the triplet or mark it as used. Thus, each time that the VLR needs to use an authentication triplet, it shall use an unmarked one, in preference to a marked. If there is no unmarked triplet, then the VLR shall request fresh triplets from the home HLR. If fresh triplets cannot be obtained, because of a system failure, the VLR may reuse a marked triplet. Thus, if a single triplet is compromised, a false BS can impersonate a genuine GPRS network to the MS. Moreover, as the false BS has the encryption key, Kc, it will not be necessary for the false BS to suppress encryption on the air interface. As long as the genuine SGSN is using the compromised authentication triplet, an attacker could also impersonate the MS and obtain session calls that are paid by the legitimate subscriber.

Data and Signalling Protection

An important weakness of the GPRS security architecture is related to the fact that the encryption of signalling and user data over the highly exposed radio interface is not mandatory. Some GPRS operators, in certain countries, never switch on encryption in their networks, since the legal framework in these countries do not permit that. Hence, in these cases signaling and data traffic are conveyed in cleartext over the radio path. This situation is becoming even more risky from the fact that the involved end users (humans) are not informed whether their sessions are encrypted or not.

As encryption over the radio interface is optional, the network indicates to the MS whether and which type(s) of encryption it supports in the *authentication request* message, during the GPRS authentication procedure. If encryption is activated, the MS start ciphering after sending the

authentication response message and the SGSN starts ciphering/deciphering when it receives a valid *authentication response* message from the MS. However, since these two messages are not protected by confidentiality and integrity mechanisms (data integrity is not provided in the GPRS radio interface except for traditional non-cryptographic link layer checksums), an adversary may mediate in the exchange of authentication messages. The results of this mediation might be either the modification of the network and the MS capabilities regarding encryption, or the suppression of encryption over the radio interface.

GPRS Backbone

Based on the analysis of the GPRS security architecture (see the *GPRS security architecture* section) it can be perceived that the GPRS security does not aim at the GPRS backbone and the wireline connections, but merely at the radio access network and the wireless path. Thus, user data and signaling information conveyed over the GPRS backbone may experience security threats, which degrade the level of security supported by GPRS. In the following, the security weaknesses of the GPRS security architecture that are related to the GPRS backbone network for both signaling and data plane are presented and analyzed.

Signaling Plane

As mentioned previously, the SS7 technology used for signaling exchange in GPRS does not support security protection. Until recently, this was not perceived to be a problem since SS7 networks belonged to a small number of large institutions (telecom operator). However, the rapid deployment of mobile systems and the liberalization of the telecommunication market have dramatically increased the number of operators (for both fixed and mobile networks) that are interconnected through the SS7 technology. This fact provokes a significant threat to the GPRS network security, since it increases the probability of an adversary to get access to the network or a legitimate operator to act maliciously.

The lack of security measures in the SS7 technology used in GPRS results also in the unprotected exchange of signaling messages between a VLR and a VLR/HLR, or a VLR and other fixed network nodes. Although these messages may include critical information for the mobile subscribers and the networks operation like ciphering keys, authentication data (e.g., authentication triplets), user subscription data (e.g., IMSI), user billing data, network billing data, and so forth, they are conveyed in a cleartext within the serving network as well as between the home network and the serving network. For example, the VLR of a serving network may use the IMSI to request authentication data for a single user from its home network, and the latter forwards them to the requesting VLR without any security measure. Thus, the exchanges of signaling messages, which are based on SS7, may disclose sensitive data of mobile subscribers and networks, since they are conveyed over insecure network connections without security precautions.

Data Plane

Similarly to the signaling plane, the data plane of the GPRS backbone presents significant security weaknesses, since the introduction of IP technology in the GPRS core shifts towards open and easily accessible network architectures. In addition, the data encryption mechanism employed in GPRS does not extend far enough towards the core network, also resulting in a cleartext transmission of user data in it. Thus, a malicious user, which gains access to the network, may either obtain access to sensitive data traffic or provide unauthorized/incorrect information to mobile users and network components. As presented previously, the security protection of users' data in the fixed segment of the GPRS network mainly relies on two independent and complementary

technologies, which are not undertaken by GPRS but from the network operators. These technologies include: (1) firewalls that enforce security policies to a GPRS core network that belongs to an operator; and (2) pre-configured VPNs that protect specific network connections.

However, firewalls were originally conceived to address security issues for fixed networks and thus are not seamlessly applicable in mobile networks. They attempt to protect the cleartext transmitted data in the GPRS backbone from external attacks, but they are inadequate against attacks that originate from malicious mobile subscribers as well as from network operator personnel or any other third party that gets access to the GPRS core network. Another vital issue regarding the deployment of firewalls in GPRS has to do with the consequences of mobility. The mobility of a user may imply roaming between networks and operators, which possibly results in the changing of the user address. This fact in conjunction with the static configuration of firewalls may potentially lead to discontinuity of service connectivity for the mobile user. Moreover, in some cases the security value of firewalls is considered limited as they allow direct connection to ports without distinguishing services.

Similarly to firewalls, the VPN technology fails to provide the necessary flexibility required by typical mobile users. Currently, VPNs for GPRS subscribers are established in a static manner between the border gateway of a GPRS network and a remote security gateway of a corporate private network. This fact allows the realization of VPNs only between a security gateway of a large organization and a mobile operator, when a considerable amount of traffic requires protection. Thus, this scheme can provide VPN services neither to individual mobile users that may require on demand VPN establishment, nor to enterprise users that may roam internationally. In addition, static VPNs have to be reconfigured every time the VPN topology or VPN parameters change.

CURRENT RESEARCH ON GPRS SECURITY

The analyzed security weaknesses of the GPRS security architecture increase the risks associated with the usage of GPRS networks influencing their deployment, which realizes the mobile Internet. In order to defeat some of these risks, a set of security improvements to the existing GPRS security architecture may be incorporated. Additionally, some complementary security measures, which have been originally designed for fixed network and aim at enhancing the level of security that GPRS supports, may be applied (Xenakis, 2006). In the following, the specific security improvements and the application of the complementary security measures are briefly presented and analyzed.

SIM Card

The majority of the security weaknesses that are related to a MS and the SIM card of a mobile user have to do with the vulnerabilities of COMP128. To address these, the old version of COMP128 (currently named as COMP128-1) is replaced by two newer versions COMP128-2 and COMP128-3, which defeat the known weaknesses. There is an even newer version COMP128-4, which is based on the 3GPP algorithm MILENAGE that uses advanced encryption standard (AES). In addition, it is mentioned to the GPRS operators that the COMP128 algorithm is only an example algorithm and that every operator should use its own algorithm in order to support an acceptable level of security (Xenakis, 2006).

User Data

User data conveyed over the GPRS backbone and the public Internet most likely remain unprotected (except for the cases that the operator supports pre-established VPNs over the public Internet)

and thus are exposed to various threats. The level of protection that GPRS provides to the data exchanged can be improved by employing two security technologies: (1) the application of end-user security, and (2) the establishment of mobile IPsec-based VPN, dynamically. End-user security is applied by using application layer solutions such as the secure sockets layer (SSL) protocol (Gupta & Gupta, 2001). SSL is the default Internet security protocol that provides point-to-point security by establishing a secure channel on top of TCP. It supports server authentication using certificates, data confidentiality, and message integrity. On the other hand, IPsec protects traffic on a per connection basis and thus is independent from the applications that run above it. An IPsec-based VPN is used for the authentication and the authorization of user access to corporate resources, the establishment of secure tunnels between the communicating parties, and the encapsulation and protection of the data transmitted by the network. On-demand VPNs that are tailored to specific security needs are especially useful for GPRS users, which require any-to-any connectivity in an ad hoc fashion. Regarding the deployment of mobile VPNs over the GPRS infrastructure, three alternative security schemes have been proposed: (1) the end-to-end (Xenakis, Gazis, Merakos, 2002), (2) the network-wide (Xenakis, Merakos: IEEE Network, 2002), and (3) the border-based (Xenakis, Merakos: IEEE PIMRC, 2002). These schemes mainly differ in the position where the security functionality is placed within the GPRS network architecture (MS, SGSN, and GGSN), and whether data in transit are ever in cleartext or available to be tapped by outsiders.

Signaling Plane of the GPRS Backbone

The lack of security measures in the signaling plane of the GPRS backbone gives the opportunity to an adversary to retrieve critical information such as the permanent identities of mobile users (IMSI), temporary identities (TMSI, TLLI), location information, authentication triplets (RAND, SRES, Kc), charging and billing data, and so forth. The possession of this information enables an attacker to identify a mobile user, to track his/her location, to decipher the user data transferred over the radio interface, to over bill him/her, and so forth. To address this inability of GPRS, it has been proposed the incorporation of the network domain security (NDS) features (Xenakis, 2006; Xenakis & Merakos, 2004) into the GPRS security architecture. NDS features, which have been designed for the latter version of UMTS, ensure that signaling exchanges in the backbone network as well as in the whole wire-line network are protected. For signaling transmission in GPRS the SS7 and IP protocol architectures are employed, which incorporate the mobile application part (MAP) (3GPP TS 09.02, 2004) and the GTP protocol (3GPP TS 09.60, 2002), respectively. In NDS both architectures are designed to be protected by standard procedures based on existing cryptographic techniques. Specifically, the IP-based signaling communications will be protected at the network level by means of the well-known IPsec suite (Kent & Atkinson, 1998). On the other hand, the realization of protection for the SS7-based communications will be accomplished at the application layer by employing specific security protocols (Xenakis & Merakos, 2004). However, until now only the MAP protocol from the SS7 architecture is designed to be protected by a new security protocol named MAPsec (3GPP TS 33.200 2002).

CONCLUSION

This chapter has presented the security architecture employed in 2.5G mobile systems focusing on GPRS. This architecture comprises a set of measures that protect the mobile users, the radio access network, the fixed part of the network, and the related data of GPRS. Most of these measures

have been originally designed for GSM, but they have been modified to adapt to the packet-oriented traffic nature and the GPRS network components. The operational differences between the application of these measures in GSM and GPRS have been outlined and commented. In addition, the security measures that can be applied by GPRS operators to protect the GPRS backbone network and inter-network communications, which are based on IP, have been explored. Although GPRS has been designed with security in mind, it presents some essential security weaknesses, which may lead to the realization of security attacks that threaten network operations and data transfer through it. These weaknesses are related to: (1) the compromise of the confidentiality of subscriber's identity, since it may be conveyed unprotected over the radio interface; (2) the inability of the authentication mechanism to perform network authentication; (3) the possibility of using COMP128 algorithm (which has been cryptoanalyzed) for A3 and A8 implementations; (4) the ability of reusing authentication triplets; (5) the possibility of suppressing encryption over the radio access network or modifying encryption parameters; and (5) the lack of effective security measures that are able to protect signaling and user data transferred over the GPRS backbone network. To defeat some of these risks, a set of security improvements to the existing GPRS security architecture may be incorporated. Additionally, some complementary security measures, which have been originally designed for fixed network and aim at enhancing the level of security that GPRS supports, may be applied.

ACKNOWLEDGMENT

Work supported by the project CASCADAS (IST-027807) funded by the FET Program of the European Commission.

REFERENCES

3rd Generation Partnership Project (3GPP) TS 03.6 (V7.9.0). (2002). *GPRS service description, Stage 2*. Sophia Antipolis Cedex, France: Author. Retreived from ftp://ftp.3gpp.org/specs/2006-12/R1998/03_series

3rd Generation Partnership Project (3GPP) TS 09.60 (V7.10.0). (2002*). GPRS tunneling protocol (GTP) across the Gn and Gp interface*. Sophia Antipolis Cedex, France: Author.Retrived from ftp://ftp.3gpp.org/specs/2006-12/R1998/09_series

3rd Generation Partnership Project (3GPP) TS 03.03 (v7.8.0). (2003). *Numbering, addressing and identification.* Sophia Antipolis Cedex, France: Author. Retrieved from ftp://ftp.3gpp.org/specs/2006-12/R1998/03_series

3rd Generation Partnership Project (3GPP) TS 01.61 (v7.0.0). (2001). *GPRS ciphering algorithm requirements.* Sophia Antipolis Cedex, France: Author. Retrieved from ftp://ftp.3gpp.org/specs/2006-12/R1999/01_series

3rd Generation Partnership Project (3GPP) TS 09.02 (v7.15.0). (2004). *Mobile application part (MAP) specification.* Sophia Antipolis Cedex, France: Author. Retrieved from ftp://ftp.3gpp.org/specs/2006-12/R1998/09_series

3rd Generation Partnership Project (3GPP) TS 33.200 (v4.3.0), (2002*). 3G security; network domain security; MAP application layer security.* Sophia Antipolis Cedex, France: Author. Retrieved from ftp://ftp.3gpp.org/specs/2006-12/Rel-4/33_series

Barkan, E., Biham, E., & Neller, N. (2003). Instant ciphertext-only cryptanalysis of GSM encrypted communication. In *Proceedings of Advances in Cryptology (CRYPTO 2003)* (LNCS 2729, 600-616).

ETSI TS 100 922 (v7.1.1). (1999). Subscriber identity modules (SIM) functional characteristics. Retrieved from http://pda.etsi.org/pda/queryform.asp

Gleeson, B., Lin, A., Heinanen, J., Armitage, G., & Malis, A. (2000). *A framework for IP based virtual private networks* (RFC 2764). Retrieved from http://www.faqs.org/rfcs/rfc2764.html

GSM 03.20. (1999). Security related network functions. Retrieved from ftp://ftp.3gpp.org/specs/2006-12/R1999/03_series

Gupta, V., & Gupta, S. (2001). Securing the wireless Internet. *IEEE Communications Magazine, 39*(12), 68-74.

Kent, S., & Atkinson, R. (1998). *Security architecture for the Internet protocol* (RFC 2401). Retrieved from http://www.javvin.com/protocol/rfc2401.pdf

Mitchell, C. (2001). *The security of the GSM air interface protocol*. Retrieved August, 2001, from http://www.ma.rhul.ac.uk/techreports/

Pagliusi, P. (2002). A contemporary foreword on GSM security. In *Proceedings of the Infrastructure Security International Conference (InfraSec)* (LNCS 2437, pp. 129-144). Springer-Verlag.

Srisuresh, P., & Holdrege, M. (1999). *IP network address translator (NAT) terminology and considerations* (RFC 2663). Retrieved from http://www.faqs.org/rfcs/rfc2663.html

Xenakis, C. (2006). Malicious actions against the GPRS technology. *Journal in Computer Virology, 2*(2), 121-133.

Xenakis, C., Gazis, E., & Merakos, L. (2002). Secure VPN deployment in GPRS mobile network. In Proceedings of European Wireless, Florence, Italy (pp. 293-300).

Xenakis, C., & Merakos, L. (2002). On demand network-wide VPN deployment in GPRS. *IEEE Network, 16*(6), 28-37.

Xenakis, C., & Merakos, L. (2002). Dynamic network-based secure VPN deployment in GPRS. In *Proceedings of IEEE PIMRC,* Lisboa, Portugal, (pp. 1260-1266).

Xenakis, C., & Merakos, L. (2004). Security in third generation mobile networks. *Computer Communications, 27*(7), 638-650.

KEY TERMS

General Packet Radio Service (GPRS): GPRS is a mobile data service available to users of GSM.

Global System for Mobile Communications (GSM): GSM is the most popular standard for mobile phones in the world.

GPRS Tunneling Protocol (GTP): GTP is an IP-based protocol that carries signaling and user data with the GPRS core network.

International Mobile Subscriber Identity (IMSI): IMSI is a unique number associated with all GSM network mobile phone users.

Second Generation (2G): 2G is a short for second-generation wireless telephone technology.

Second and a Half Generation (2.5G): 2.5G is used to describe 2G systems that have implemented a packet-switched domain in addition to the circuit-switched domain.

Signaling System 7 (SS7): SS7 is a set of telephony signaling protocols which are used to set up the vast majority of the world's public switched telephone network telephone calls.

Subscriber Identity Module (SIM): SIM is a removable smart card for mobile phones that stores network specific information used to authenticate and identify subscribers on the network.

Temporary Mobile Subscriber Identity (TMSI): TMSI is a randomly allocated number that is given to the mobile the moment it is switched on and serves as a temporary identity between the mobile and the network.

This work was previously published in Handbook of Research on Wireless Security, edited by Y. Zhang, J. Zheng, and M. Ma, pp. 351-363, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.14
Evaluation of Security Architectures for Mobile Broadband Access

Symeon Chatzinotas
University of Surrey, UK

Jonny Karlsson
Arcada University of Applied Sciences, Finland

Göran Pulkkis
Arcada University of Applied Sciences, Finland

Kaj Grahn
Arcada University of Applied Sciences, Finland

ABSTRACT

During the last few years, mobile broadband access has been a popular concept in the context of fourth generation (4G) cellular systems. After the wide acceptance and deployment of the wired broadband connections, such as DSL, the research community in conjunction with the industry have tried to develop and deploy viable mobile architectures for broadband connectivity. The dominant architectures which have already been proposed are Wi-Fi, universal mobile telecommunications system (UMTS), WiMax, and flash-orthogonal frequency division modulation (OFDM). In this chapter, we analyze these protocols with respect to their security mechanisms. First, a detailed description of the authentication, confidentiality, and integrity mechanisms is provided in order to highlight the major security gaps and threats. Subsequently, each threat is evaluated based on three factors: likelihood, impact, and risk. The technologies are then compared taking their security evaluation into account. Flash-OFDM is not included in this comparison since its security specifications have not been released in public. Finally, future trends of mobile broadband ac-

cess, such as the evolution of WiMax, mobile broadband wireless access (MBWA), and 4G are discussed.

INTRODUCTION

During the last decade, wireless network technologies have greatly evolved and have been able to provide cost-efficient solutions for voice and data services. Their main advantages over wired networks are that they avoid expensive cabling infrastructure and they support user mobility and effective broadcasting. As a result, mobile wireless networks have managed to take over a large percentage of the "voice" market, since the global system for mobile communications (GSM) cellular technology has promoted the worldwide expansion of mobile telephony. Furthermore, nowadays broadband Internet has become a necessity for many home and business users. Moreover, in the context of all-IP network convergence, an increasing share of telephony subscribers is migrating towards VoIP solutions mainly due to the decreased cost compared to fixed telephony. Therefore, the main challenge is to find spectrum- and cost-efficient solutions for the provision of mobile broadband services. In this direction, a large research community of academic and industrial origin has dedicated considerable effort on designing, implementing, and deploying systems for mobile broadband access, such as Wi-Fi, universal mobile telecommunications system (UMTS), WiMax, and flash-orthogonal frequency division modulation (OFDM). According to the predictions, in the years to come, more and more of our voice samples and data packets will be carried over wireless broadband links through the Internet. Therefore it becomes imperative that these messages are secured from malicious eavesdroppers and attackers. Especially in applications such as e-banking, e-commerce, and e-government the revelation of sensitive data to unauthorized persons, unauthorized data submission, and/or the interruption of system availability can cause financial damage, user preferences' surveillance, industry espionage, and/or administrative overhead.

The purpose of this chapter is to analyze and compare the security architectures of the dominant mobile broadband technologies. More specifically, the objectives are to:

- Describe and analyze the security architectures of mobile broadband technologies.
- Identify the strong and weak points of each technology in terms of access control based on authentication, confidentiality, integrity, and physical layer resilience.
- Compare the investigated security architectures based on a risk evaluation of the identified security vulnerabilities.

MOBILE BROADBAND TECHNOLOGIES

This section discusses the mobile technologies Wi-Fi, UMTS, WiMax, and flash-OFDM. Authentication performance, confidentiality, and integrity mechanisms for each technology are analyzed.

Wi-Fi

Wi-Fi was the first widely-deployed technology for wireless computer networks. It was originally designed to provide portability support in local area networks (LANs). However, Wi-Fi has also been utilized in other scenarios, such as wireless metropolitan area networks (WMANs), since it was the first wireless technology with support for mobile communication and for a wide range of portable and mobile devices.

The Wi-Fi radio interface is based on the IEEE 802.11 standard and is available in three versions:

- **802.11a**
  - **Frequency:** 5.5 GHz,
  - **Modulation:** OFDM
  - **Bandwidth:** 54 Mbps
- **802.11b**
  - **Frequency:** 2.4 GHz
  - **Modulation:** Direct sequence spread spectrum (DSSS)
  - **Bandwidth:** 11 Mbps
- **802.11g**
  - **Frequency:** 2.4 GHz
  - **Modulation:** OFDM
  - **Bandwidth:** 54 Mbps

In this context, Wi-Fi alliance is an organization testing products in order to evaluate that they correctly implement the set of standards defined in the IEEE 802.11 specification. After the products have successfully passed these tests, they are allowed to use the Wi-Fi logo.

Security Architecture

Wi-Fi security standards include wired equivalent privacy (WEP), Wi-Fi protected access (WPA), and WPA2. WEP was the first introduced security standard. WPA was designed to be a security protocol that corrects the security deficiencies of WEP and to be backward compatible with existing hardware. The last development in Wi-Fi security is the WPA2 standard which was published in June 2004 by the IEEE 802.11i group. WPA2 was designed to offer a further improved security scheme (Edney & Arbaugh, 2003). The aforementioned security specifications are analyzed and compared in the following paragraphs.

Authentication

Authentication services are utilized to allow a client to communicate with the serving access point. After successful authentication, a session is initiated and it can be terminated by either the client or the access point. Wi-Fi provides the following link-layer authentication schemes:

- Closed system authentication
- Media access control (MAC) filtering
- WEP suthentication—Shared RC4 key
- WPA and WPA2 authentication—802.1X/extensible authentication protocol (EAP)

Closed system authentication, MAC filtering, and WEP authentication are not recommended due to their well-known serious security flaws (Borisov, Goldberg, & Wagner, 2001; Lynn & Baird, 2002; Welch & Lathrop, 2003).

WPA and WPA2 security schemes have some major design differences from WEP, since the authentication and the confidentiality processes operate totally independently from each other (Baek, Smith, & Kotz, 2004). The authentication process of WPA and WPA2 adopts the three-entity model of IEEE 802.1x which was originally designed for the point-to-point protocol (IEEE, 2001). The three entities involved in this protocol are the client, the access point (AP), and the authentication server (AS). First, the client request to obtain access to the network. The AP acts as a network guard, allowing access only to the clients that the AS has authenticated. Finally, the AS is responsible for deciding whether the client is allowed to access the network. These three entities utilize EAP to exchange communication messages in order to coordinate the authentication process (Stanley, Walker, & Aboba, 2005).

In addition, there is a lighter version of WPA, called WPA-preshared key (WPA-PSK). This version is based on a shared secret key or passphrase in order to authenticate the wireless clients. As a result, an attacker can use a wireless sniffer to capture the 4-way WPA handshake, log the packets, and then try a brute force attack using a dictionary file (Van de Wiele, 2005). Thus, if WPA-PSK is deployed, the robustness of the

network security totally depends on the length and the complexity of the secret key.

Encryption

Encryption services are utilized to provide confidentiality over wireless communication links. In Wi-Fi networks the following encryption schemes are available:

- WEP based on the RC4 (Ron's Code 4) stream cipher
- WPA encryption based on the temporal key integrity protocol (TKIP)
- WPA2 encryption based on the advanced encryption standard (AES)

WEP is a weak implementation of the RC4 stream cipher and WEP encryption is thus not recommended (Borisov et al., 2001; Stubblefield, Ioannidis, & Rubin, 2002; Welch & Lathrop, 2003).

WPA encryption is based on TKIP. It incorporates the basic functionalities of WEP, but improvements have been made to address the security flaws. The length of the initialization vector (IV) has been increased from 24 bits to 48 bits and therefore the possibility of reused keys has been significantly decreased. Furthermore, WPA does not directly utilize the master keys. Instead it constructs a hierarchy of derived keys to be utilized in the encryption process. Finally, WPA dynamically cycles keys while transferring data. Since keys are regularly changed, a malicious user has a very short time window to attempt an attack.

WPA2 was designed from scratch taking the vulnerabilities of the previous security architectures into account. WPA2 allows various network implementations, but the default configuration utilizes the advanced encryption standard (AES) and the counter mode CBC MAC protocol (CCMP). AES is a block cipher, operating on blocks of 128 bit data, and is a replacement of the RC4 algorithm used by WPA. AES is much more robust since it has already been tested in various security architectures without revealing serious vulnerabilities. CCMP comprises of two main parts. The first is the counter mode (CM) which is responsible for the privacy of the data in combination with AES. The second is the cipher block chaining message authentication code (CBC-MAC) providing data integrity checking and authentication.

Integrity

Integrity services are responsible for making sure that transmitted information is not replayed or modified during transmission. The following techniques are applicable in Wi-Fi networks:

- WEP cyclic redundancy heck 4 (CRC-32) Checksum
- WPA Integrity
- WPA2 Integrity

WEP checksum is a noncryptographic linear function of the plaintext. This means that multiple messages may correspond to a single 32-bit number. Hence, an experienced intruder could modify the plaintext in such a way that the checksum remains unchanged. Furthermore, due to the linearity of both the RC4 stream cipher and the CRC-32 checksum, the attacker is able to change the message even when he does not know the plaintext (Welch & Lathrop, 2003).

WPA has incorporated mechanisms for the prevention of replay attacks. More specifically, the TKIP sequence counter (TSC) based on the IVs is utilized, so that the receiver can identify and reject "replayed" messages. Furthermore, WPA uses an improved integrity mechanism in order to generate the message integrity check (MIC). This mechanism, called Michael, is able to detect possible attacks and deploy countermeasures to prevent new attacks.

WPA2 utilizes CCMP for providing integrity services. CCMP generates a MIC using the CBC-

MAC method. In this method, even the slightest change in the plaintext will produce a totally different checksum.

Security Vulnerabilities

Although the Wi-Fi security architecture has been greatly improved since WEP, there are still vulnerabilities which cannot be addressed by WPA2. These vulnerabilities can lead to a number of link layer denial-of-service (DoS) attacks (Van de Wiele, 2005). All the DoS techniques described here are fairly easy to use with freely available tools found on the Internet. In most of the cases, the attacker will use different forged MAC addresses to mount DoS attacks. These attacks can be detected by specialized hardware (e.g., air monitor, security aware access point) which can detect the misuse of the infrastructure. Furthermore, this specialized hardware can notify the people responsible for the follow-up of a DoS incident and give an estimate on where the attacker is located by considering the signal and noise levels.

Disassociation Storm

Before any wireless communication can occur, a client has to send an association frame to the access point asking to join the network. Similarly, after the end of the wireless session, the access point or client has to send a disassociation frame to terminate the connection. The frames of these messages are broadcasted and can be sniffed by an attacker. The attacker can then flood the network with spoofed disassociation frames every time the client tries to join the network, thus disrupting the association process and the network access.

Authenticated / Deauthenticated Storm

The aforementioned principle can be exploited in order to disconnect a client and try to keep the client disconnected. This technique starts by sending a spoofed deauthentication frame followed by a disassociation frame in order to make sure that the client has disconnected from the legitimate access point. In a more advanced version of this attack, a fake probe request and some beacon frames are transmitted in order to force the client to connect to a rogue access point which ignores or monitors the client's traffic.

UMTS

Universal mobile telecommunications system (UMTS) is one of the third generation (3G) wireless cellular technologies for mobile communication. Mobile devices like smartphones, laptops, and handheld computers can be used. UMTS is standardized by the 3G partnership project (3GPP) and it is mainly deployed in Europe and Japan. Theoretically UMTS supports up to 1920 Kbps data transfer rates, but currently the real world performance can reach 384 Kbps. It uses the W-code division multiple access (CDMA) technology over two 5 MHz channels, one for uplink and one for downlink. The specific frequency bands originally defined by the UMTS standard are 1885-2025 MHz for uplink and 2110-2200 MHz for downlink.

In UMTS network topology, a mobile station is connected to a visited network by means of a radio link to a particular base station (Node B). Multiple base stations of the network are connected to a radio network controller (RNC) and multiple RNCs are controlled by a general packet radio service (GPRS) support node (GSN) in the packet-switched case. The visitor location register (VLR) and the serving GSN keep track of all mobile stations that are currently connected to the network. Every subscriber can be identified by its international mobile subscriber identity (IMSI). In order to protect against profiling attacks, this permanent identifier is sent over the air interface as infrequently as possible. What is more, locally valid temporary mobile subscriber identities (TMSI) are used to identify subscrib-

ers whenever possible. Every UMTS subscriber has a dedicated home network with which the subscriber shares a long term secret key Ki. The home location register (HLR) keeps track of the current location of all subscribers of the home network. Mutual authentication between a mobile station and a visited network is carried out with the support of the current serving GSN (SGSN) or the mobile switching center (MSC)/VLR respectively.

The new series of 3.5G mobile telephony technologies, known as high speed packet access (HSPA), will provide more bandwidth to the end-user, improved network capacity to the operator, and enhanced interactivity for data applications. HSPA refers to the improvements made in the UMTS downlink, known as high speed downlink packet access (HSDPA), and the UMTS uplink, usually referred to as high speed uplink packet access (HSUPA) but also referred to as enhanced dedicated channel (E-DCH).

HSDPA provides a bandwidth of 14.4 Mbps/user. For multiple-input-multiple-output (MIMO) systems up to 20 Mbps can be achieved. Both HSDPA and HSUPA can be implemented in the standard 5 MHz carrier of UMTS networks and can coexist with original UMTS networks. As HSPA specifications refer only to the access network, there is no change required in the core network (CN) except from the high data-rate links required to handle the increase in clients' traffic generated by HSPA.

Security Architecture

The 3G security architecture is based on GSM, but certain improvements are added in order to correct the described security vulnerabilities.

Authentication

Authentication and key agreement (AKA) is the main security protocol of UMTS in the 3GPP specification. According to AKA, a mobile device and a base station have to authenticate each other. Figure 1 provides an overview of the AKA process. The authentication vector includes the following components:

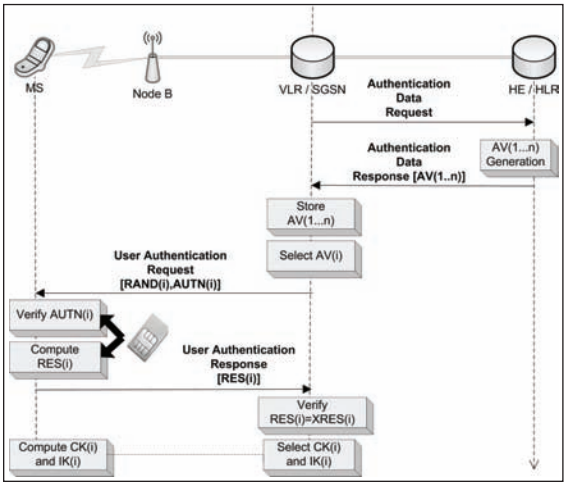
a. A random number (RAND)
b. An expected response (XRES)
c. A cipher key (CK)
d. An integrity key (IK)
e. An authentication token (AUTN)

RAND and XRES are utilized by the network to authenticate the mobile station (MS), whereas AUTN is utilized by the MS to authenticate the network. After the mutual authentication, the two communicating parties can agree on the CK and the IK which will be used throughout the rest of the session.

Confidentiality and Integrity

UMTS employs the UMTS encryption algorithm (UEA) in order to provide information confidentiality. The encryption process of UEA is based on the f8 algorithm. One of the main improvements

Figure 1. 3GPP authentication and key agreement (AKA)



of UMTS is that the link layer encrypted channel is established between the MS and the GSN instead of the BS, as in GSM. Furthermore, UEA is utilized to protect not only the data channels but also certain signalling channels.

For user confidentiality UMTS utilizes the same mechanism as GSM. Instead of the IMSI, a temporary identity (TMSI) assigned by VLR is used to identify the subscriber in the communication messages exchanged with the BS. However, the IMSI is still transmitted in clear-text over the air while establishing the TMSI. This has been proved to be a starting point for security attacks against UMTS.

Data integrity in 3GPP is assured explicitly through the UMTS integrity algorithm (UIA). The UIA operation is based on the f9 algorithm. UIA is utilized to protect both communication and signalling. UEA and UIA are presented in Figure 2.

GSM Compatibility

UMTS has been designed to be backwards compatible with GSM. It includes standardized security features in order to ensure world-wide interoperability and roaming. More specifically, GSM user parameters are derived from UMTS parameters using a set of predefined conversion functions. However, GSM subscribers roaming in 3GPP networks are supported by the GSM security context, which is vulnerable to the aforementioned GSM vulnerabilities.

Security Vulnerabilities

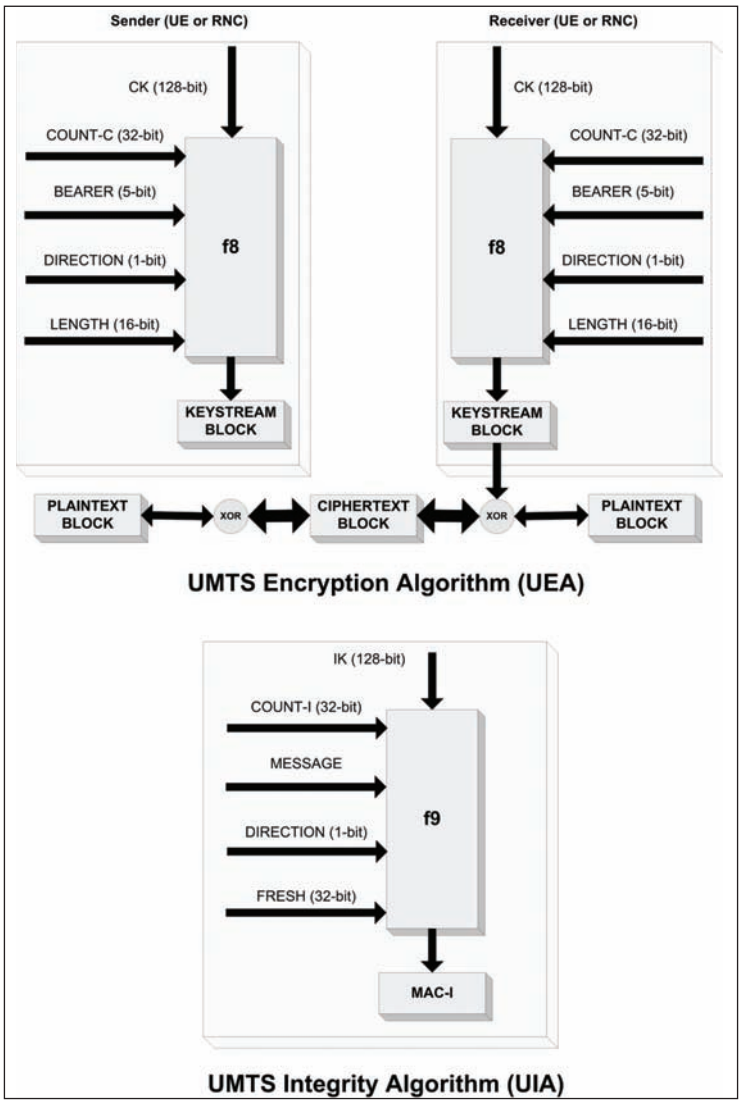
3G security has been significantly improved compared to GSM. However, there are still vulnerabilities related to the backwards compatibility with GSM. Meyer and Wetzel (2004a, 2004b) present a man-in-the-middle attack which can be mounted even if the subscriber utilizes a 3G enabled device within a 3G base station coverage. The described attack goes far beyond the anticipations of the 3GPP group. UMTS subscribers are vulnerable to what 3GPP calls a "false base station attack" even if subscribers are roaming in a pure UMTS network and even though UMTS authentication is applied.

This attack can be categorized as a "roll-back attack." This category of attacks exploits weaknesses of old versions of algorithms and protocols by means of the mechanisms defined to ensure backward compatibility of newer and stronger versions. According to this technique, the attacker acts on behalf of the victim's mobile station in order to obtain a valid authentication token AUTN from any real network. It is assumed that the attacker has already retrieved the IMSI of the targeted subscriber, since the latter is sent in clear-text when establishing a TMSI. The attacker can capture the AUTN by initiating the AKA procedure with any legitimate network. The next step is to impersonate a valid GSM base station to the victim mobile station. The mobile station connects and verifies the rogue BS, since it possesses a valid AUTN. Subsequently, the rogue BS is configured by the attacker to utilize "no encryption" or weak encryption. Finally, the attacker can send to the mobile station the GSM cipher mode command including the chosen encryption algorithm. The man-in-the-middle attack is mounted and the attacker can use passive or active eavesdropping without being detected.

WIMAX

The IEEE 802.16 or broadband wireless access (BWA) Working Group was established in 1999 to prepare specifications for broadband wireless metropolitan area networks. The first 802.16 standard was approved in December 2001 and was followed by three amendments: 802.16a, 802.16b and 802.16c. In 2004 the 802.16-2004 standard (IEEE-SA, 2006) was released and the earlier 802.16 documents including the a/b/c amendments were withdrawn. An amendment

Figure 2. UMTS encryption and integrity algorithm



to the standard 802.16e (IEEE-SA, 2006) addressing mobility was introduced in 2005. The main additions of the 802.16e were low density parity check (LDPC) codes at the physical layer, enhanced MIMO setup functions, new states for MS operation, parameter-defined power saving classes of mobiles, and enhanced FFT sizes for scalable OFDMA.

WiMax aims at providing high data rate triple-play wireless services to fixed users, to nomadic users, and to users of mobile devices. It is based on a low latency quality of service (QoS) architecture in order to provide real-time multimedia services. It operates on the 2-6 GHz (IEEE802.16e) and 10-66 GHz (IEEE802.16-2004) frequency bands and it uses the OFDMA technology for modulation and medium access.

Security Architecture

WiMax has been designed with security in mind, especially after the serious vulnerabilities dis-

covered in the original Wi-Fi security protocol. The IEEE 802.16 specifications include a security sublayer within the MAC layer. The IEEE 802.16 security architecture is based on the following issues:

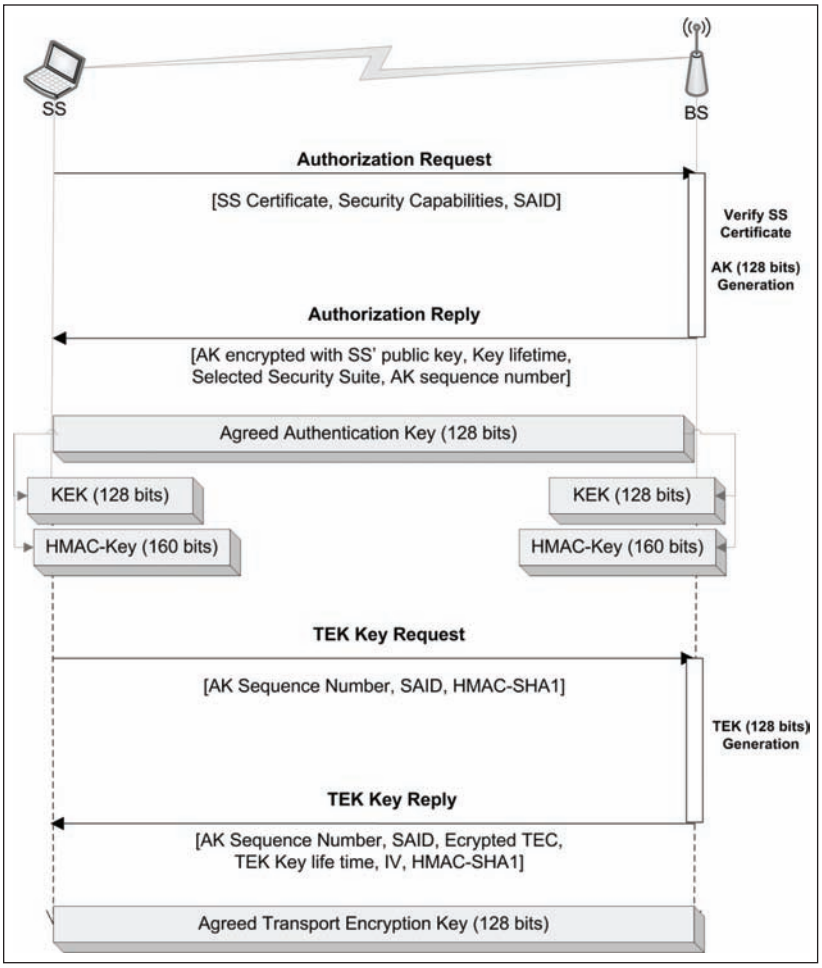
- **Authentication:** The baseline authentication architecture, by default, employs a public key infrastructure (PKI) based on X.509 certificates. The base station (BS) validates the client's certificate before permitting access to the physical layer (see Figure 3). First, the subscriber station (SS) sends to the BS an authorization request containing the certificate, the available security capabilities, and the security association identifier (SAID). The BS verifies the certificate and generates a 128 bit authentication key (AK). Then, the BS sends to the SS an authorization reply, which contains the AK encrypted with SS's public key, the AK's lifetime, the selected security suite, and an AK sequence number. The SS uses its private key to recover the AK, which can now be utilized as an authentication token in further communication.
- **Key exchange:** The SS and the BS can agree on a transport encryption key (TEK), which will be utilized for data encryption (see Figure 3).
  TEK is randomly generated by the BS. The AK established during authentication is used to derive two additional keys:
  - Message authentication key (HMAC key), which is utilized to provide message integrity and AK confirmation during the key exchange process.
  - Key encryption key (KEK), which is utilized for encrypting the TEK before sending it back to the SS. The modes for encrypting TEK are:
    a. 3DES with a 112 bit KEK
    b. AES with a 128 bit KEK
    c. RSA using SS's public key
- **Data encryption and integrity:** The modes for implementing data privacy are:
  - Data encryption standard (DES) with a 56 bit key and cipher block chaining (CBC), which utilizes the Initialization Vectors obtained during Key Exchange,
  - AES with a 128 bit key and counter mode with cipher block chaining message authentication code protocol, which provides message integrity and replay protection.

Security Vulnerabilities

WiMax supports unilateral device level authentication (Barbeau, 2005), which can be implemented in a similar way as Wi-Fi MAC filtering based on the hardware device address. Therefore, address sniffing and spoofing make a MS masquerade attack possible. In addition, the lack of mutual authentication makes a man-in-the-middle attack from a rogue BS possible. However, a successful man-in-the-middle attack is difficult because of the time division multiple access (TDMA) model in WiMax. The attacker must transmit at the same time as the legitimate BS using a much higher power level in order to "hide" the legitimate signal. Furthermore, WiMax supports mutual authentication at user network level based on the generic extensible authentication protocol (EAP) (Aboba, Blunk, Vollbrecht, Carlson, & Levkowetz, 2004). EAP variants, EAP- transport layer security (TLS) (X.509 certificate based) (Aboba & Simon, 1999) and EAP-subscriber identity module (SIM) (Haverinen & Salowey, 2004), are supported.

In the data privacy domain, the main security threat is the transmission of unencrypted management messages over the wireless link. Eavesdropping of management messages is a critical threat for users and a major threat to a system. For example, an attacker could use this vulnerability to verify the presence of a victim at its location before perpetrating a crime. Addi-

Figure 3. WiMax authentication and key exchange process



tionally, it might be used by a competitor to map the network. Another major vulnerability is the encryption mode based on DES. The 56 bit DES key is easily broken by brute force with modern computers. Furthermore, the DES encryption mode includes no message integrity or replay protection functionality and is thus vulnerable to active or replay attacks. The secure AES encryption mode should be preferred over DES.

Finally, there is a potential for DoS attacks because authentication operations trigger the execution of long procedures. For example, a DoS attack could flood a MS with a high number of messages to authenticate. Due to low computational resources, the MS will not be able to handle a large amount of invalid messages, rendering the DoS attack successful.

FLASH-OFDM

Fast low-latency access with seamless handoff orthogonal frequency division multiplexing (flash-OFDM) is an OFDM-based proprietary system which specifies the physical layer, as well as higher protocol stack layers. It is an all IP technology and it aims to compete with GSM/3G networks. Already implemented flash-OFDM technology

operating in the 450 MHz frequency band can offer a maximum download speed of 5.3 Mbps and an upload speed of 1.8 Mbps.

Design objectives have included design of a high capacity physical layer, a packet-switched air interface, a contention-free and QoS-aware MAC layer, and efficient operations using existing Internet protocols. The air interface is designed and optimized across all protocol stack layers. Fast hopping across all tones in a pseudorandom predetermined pattern is employed. Channel coding and modulation are carried out on a per-segment basis and can be individually optimized for each channel. The ability to send segments of arbitrary size enables the MAC layer to perform efficient packet switching over the air interface. Given segments can be dedicated for use with predefined functionality. Thus there is no need to send overheads, such as message headers. Therefore, network layer traffic experiences small delays and no significant delay jitter.

Security Architecture

The security relies on "defence in depth," that is, virtual private network (VPN) tunnelling and end-to-end encryption are used. Security specifications for flash-OFDM have not been presented in public (Lehtonen, Ahonen, Savola, Uusitalo, Karjalainen, Kuusela et al., 2006).

Security Analysis

A security analysis of the mobile broadband technologies Wi-Fi, UMTS, and WiMax is presented. Inclusion of flash-OFDM in this comparison is not possible because of the unavailability of public security specifications. Threats are analyzed with respect to the likelihood of occurrence, the impact on the network operation, and the global risk they represent. In the following paragraphs, we first describe in detail the evaluation and comparison methodology, and then a group of tables is presented in which the security threats of the investigated technologies are evaluated. Security threats are classified based on four main axes: authentication, confidentiality, integrity, and physical layer resilience. Finally, the security evaluations of the studied technologies are compared and presented in a concise overview table.

Methodology

The evaluation and comparison methodology was based on the method described by Barbeau, (2005) and ETSI (2003). More specifically, three main criteria are considered: likelihood, impact, and risk. "Likelihood" refers to the probability that an attack associated with a specific threat is successfully launched. In this context, two variables are considered:

a. The technical difficulties of mounting the attack in terms of the required software, hardware, and estimated time duration.
b. The attacker's motivation in terms of the level of network access or the severity of the system malfunction that the attack achieves.

Three levels of likelihood are available as described in Table 1. "Impact" refers to the consequences of an attack in terms of user and network security. The two variables of impact are:

a. User impact in terms of the severity of network access degradation.
b. System impact in terms of the severity of network degradation or outage.

Three levels of impact are available as described in Table 1. According to the level of likelihood and impact, numerical values from a predefined range are assigned to each criterion (see Table 1). For a specific threat, the "risk" refers to an overall threat level which is determined by the product of the likelihood value and impact value.

Table 1. Evaluation and comparison methodology

		Variables		
Criteria	**Cases**	**Difficulty**	**Motivation**	**Rank**
Likelihood	**Unlikely**	Strong	Low	**1**
	Possible	Solvable	Reasonable	**2**
	Likely	None	High	**3**
		User	**System**	
Impact	**Low**	Annoyance	Very limited outages	**1**
	Medium	Loss of service	Limited outages	**2**
	High	Long time loss of service	Long time outages	**3**

Risk = Likelihood x Impact			
Risk	**Minor**	No need for countermeasures	**1-3**
	Major	Threat need to be handled	**3-6**
	Critical	High priority	**6-9**

Security threats which result in a high evaluated risk value are critical and additional measures should be taken to protect the network perimeter, whereas threats which have a low risk can be tolerated without employing countermeasures.

In this point, it is worth noting that this quantitative ranking is subjective. However, this is a useful evaluation and comparison methodology which can stimulate a structured discussion based on the evaluation criteria, that is, likelihood, impact, and risk. The comparison axes are authentication, confidentiality, integrity, and physical layer resilience.

Objective-Based Comparison

This section applies the aforementioned methodology on four main objectives of wireless security architectures: authentication, confidentiality, integrity, and physical layer resilience. For each objective, a thorough discussion describes the rationale behind the ranking of the security threats.

Authentication Evaluation

Wi-Fi includes four security threats which are all ranked to have a high impact on the system, since the attacker can exploit them to override the authentication checks or launch a combination of attacks which will grant him full network access. However, the likelihood ranking greatly varies. Closed system authentication and MAC filtering are very likely to be attacked by sniffing software which is readily available on the Internet. WEP attacks are more complicated, because a combination of software is required to induce and capture network traffic and then exploit the weak IVs in order to crack the key. WPA-PSK is even more difficult to break since it requires a brute force attack. The resilience of WPA-PSK is greatly dependent on the length and the complexity of the preshared key.

UMTS is far more resilient to authentication attacks, since most of the security gaps have been identified during the deployment of GSM and tackled in the specification design of UMTS. However, UMTS includes two main authentication vulnerabilities which can be exploited to launch a man-in-the-middle attack (high impact). The IMSI hijack threat refers to the deployment of a rogue BS in order to initiate an authentication procedure and steal the IMSI of a mobile user. The motivation for this attack is high, but the equipment is expensive and complicated to configure. AUTN capture is the second step of the attack and it refers to capturing an authentication token by masquerading a MS. It assumes that the IMSI Hijack attack has been already successfully launched. However, this attack does not require the deployment of a rogue BS and therefore it is more possible to happen.

In the WiMax architecture, the main security threat is the device-level authentication mode. When this mode is utilized without certificate support, it is as vulnerable as MAC filtering and it can be exploited to launch MS or BS masquerading attacks. A less critical vulnerability is the DoS attack which can be launched by flooding authentication requests. This attack mostly affects the MS due to its limited processing resources, but it is not a major threat since it has a medium

impact and a low motivation.

Confidentiality Evaluation

Wi-Fi includes some major vulnerabilities. It supports a null mode encryption which is configured as default in the majority of the commercial access points. WEP encryption can provide an elementary level of protection, but it is still too weak to keep the intruders out. WPA-PSK offers a satisfactory level of confidentiality, if long and complex keys are utilized. The ranking of the Wi-Fi confidentiality vulnerabilities is similar to authentication ranking, since both objectives are based on the same mechanisms.

UMTS incorporates strong encryption algorithms which have eliminated the deficiencies of its predecessor GSM. Nevertheless, the backwards compatibility with GSM can be exploited to compromise dual-band mobile devices by launching a man-in-the-middle attack. In this attack, the rogue BS can mandate the MS to use null mode encryption or one of the GSM encryption modes which can be easily broken (Biham & Dunkelman, 2000; Biryukov, Shamir, & Wagner, 2000). However, this is an unlikely attack since it requires the deployment of a BS and a prior successful launch of the IMSI hijack and AUTN capture attacks.

WiMax security architecture includes two main shortcomings. First of all, the DES encryption mode provides an inadequate level of confidentiality, since it can be easily broken. In addition, the eavesdropping of unencrypted management frames can be easily established, but it cannot greatly affect the system if robust authentication and integrity mechanisms have been deployed.

Integrity Evaluation

Wi-Fi supports null mode which leaves the messages totally unprotected against modification and replay attacks. WEP CRC-32 integrity mechanism provides a moderate level of protection, but there is no replay protection and the integrity protection can be overridden by an experienced attacker.

The UMTS architecture includes a major shortcoming, namely the inadequate replay protection of authentication tokens. This vulnerability can have a high impact since it allows the reuse of the token retrieved by an AUTH capture attack and the completion of the UMTS man-in-the-middle attack. However, it requires a prior successful launch of IMSI hijack and AUTN capture. Therefore it results in a high technical difficulty.

WiMax supports two modes that can greatly compromise information integrity. The first is the DES mode which does not support integrity and replay protection of data frames. The second is the null MAC mode for management frames, which can allow the intruder to inject modified management frames and affect the network operation.

Physical Layer Resilience Evaluation

The resilience of the physical layer of each technology is evaluated with respect to jamming and scrambling. Jamming is achieved by introducing a source of noise strong enough to significantly reduce the capacity of the channel. Scrambling is similar to jamming, but it takes place for short intervals of time and it is targeted to specific frames or parts of frames.

Wi-Fi comprises of the three different specifications IEEE 802.11a/b/g which all utilize random medium access techniques but operate on different physical channels. IEEE 802.11a/g operate on a 5 MHz OFDM channel, whereas IEEE 802.11b operates on a 5 MHz DSSS channel. The DSSS is more resilient to narrowband jamming than OFDM and therefore jamming has a higher impact on IEEE802.11a/g. However, if the attacker wants to jam all the channels, the attacker has to jam a bandwidth of 40 MHz, which is quite difficult. Scrambling is easier to launch because of the random medium access layer.

UMTS operates on two 5 MHz DSSS chan-

nels, one for the uplink and one for the downlink. It is resilient to narrowband jamming because of the DSSS modulation, but it is still vulnerable to scrambling because of the random access.

WiMax operates on a 1.25-20 MHz OFDM channel and it employs TDMA techniques. Thus, it can be vulnerable to jamming especially if it operates on a narrow channel, but it is resilient to scrambling due to the TDMA.

OVERALL COMPARISON

The results from authentication, confidentiality, integrity, and physical layer resilience evaluation are presented in Table 2.

As follows, the overall comparison results:

- **Wi-Fi:**
  - **Authentication:** 6.75
  - **Confidentiality:** 6
  - **Integrity:** 6
  - **PHY Resilience:** 5
  - **AVERAGE RISK:** 5.94
- **UMTS**
  - **Authentication:** 4.5
  - **Confidentiality:** 3
  - **Integrity:** 3
  - **PHY Resilience:** 3.5
  - **AVERAGE RISK:** 5.94
- **WiMax**
  - **Authentication:** 6.5
  - **Confidentiality:** 6
  - **Integrity:** 7.5
  - **PHY Resilience:** 3
  - **AVERAGE RISK:** 5.75

Wi-Fi has the highest average risk, which is quite reasonable because of the initial lack of security mechanisms in the Wi-Fi specification and the subsequent failure of WEP. WPA and WPA2 modes are much more secure, but the poor usability and the limited security awareness have constrained their wide deployment. UMTS proved to be quite robust by eliminating the security inefficiencies of its predecessor GSM. However, an attacker can still exploit some backward-compatibility issues to launch a man-in-the-middle attack. WiMax's performance was not satisfactory enough mainly due to the provision of weak security modes. Nevertheless, the practical performance is greatly dependent on the actual security decisions of the network operators. These decisions vary according to the provided service requirements.

Table 2. Security evaluation

AUTHENTICATION EVALUATION				
Technology	**Threat**	**Likelihood**	**Impact**	**Risk**
Wi-Fi	Closed System	3	3	9
	MAC Filtering	3	3	9
	WEP	2	3	6
	WPA-PSK	1	3	3
Average Risk				**6,75**
UMTS	IMSI Hijack	2	3	6
	AUTN Capture	1	3	3
Average Risk				**4,5**
WiMAX	Device-level Authentication	3	3	9
	DoS on MS	2	2	4
Average Risk				**6,5**
CONFIDENTIALITY EVALUATION				
Technology	**Threat**	**Likelihood**	**Impact**	**Risk**
Wi-Fi	Null	3	3	9
	WEP	2	3	6
	WPA-PSK	1	3	3
Average Risk				**6**
UMTS	Rogue BS – Null / Weak	1	3	3
Average Risk				**3**
WiMAX	DES mode	3	3	9
	Management Frames	3	1	3
Average Risk				**6**
INTEGRITY EVALUATION				
Technology	**Threat**	**Likelihood**	**Impact**	**Risk**
Wi-Fi	Null	3	3	9
	WEP	1	3	3
Average Risk				**6**
UMTS	AUTN Replay	1	3	3
Average Risk				**3**
WiMAX	DES mode – Null integrity	3	2	6
	Management Frame-Null MAC	3	3	9
Average Risk				**7,5**
PHYSICAL LAYER RESILIENCE EVALUATION				
Technology	**Threat**	**Likelihood**	**Impact**	**Risk**
Wi-Fi	Jamming (IEEE 802.11a/g)	2	3	6
	Scrambling (IEEE 802.11a/g)	3	3	9
	Jamming (IEEE 802.11b)	2	2	4
	Scrambling (IEEE 802.11b)	3	2	6
Average Risk				**5**
UMTS	Jamming	1	2	2
	Scrambling	2	2	4
Average Risk				**2,5**
WiMAX	Jamming	1	3	3
	Scrambling	1	3	3
Average Risk				**3**

FUTURE TRENDS

Broadband wireless access networking is presently a rapidly evolving ICT area. Three important development trends can be identified:

- WiMax evolution for long range broadband wireless access.
- Development of a broadband wireless access technology supporting high speed mobility.
- Emerging 4G wireless cellular technology.

WiMax Evolution

The WiMax standard was finalized in June 2004. WiMax has the potential to change telecommunications as it is known today. "It eradicates the resource scarcity that has sustained incumbent service providers for the last century. As this technology enables a lower barrier to entry, it will allow true market-based competition in major telecommunications services like voice, video and data" (Ohrtman, 2005).

WiMax can offer a point-to-point range of 50 km with a throughput of 72 Mbps. The WiMax technology will make personal broadband services profitable to service providers and will be available to business and consumer subscribers at affordable prices. The first mobile WiMax products are expected to be introduced into the market in the first quarter of 2007. New technologies such as MIMO and beam forming for higher throughput and capacity will be introduced in 2007 (WiMax Forum, 2006).

Mobile Broadband Wireless Access (MBWA)

The IEEE 802.20 (or MBWA) Working Group was established in December 11, 2002, with the aim to develop a specification for an efficient packet-based air interface that is optimized for the transport of IP based services. The goal is to enable worldwide deployment of affordable, always-on, and interoperable BWA networks. The group will specify the lower layers of the air interface, operating in licensed bands below 3.5 GHz and enabling peak user data rates exceeding 1 Mbps at speeds of up to 250 km/h. A draft version of the specification was approved in January 18, 2006.

4G – Future Wireless Cellular Technology

Frameworks for future 4G networks, which seamlessly integrate heterogeneous mobile technologies in order to provide enhanced service integration, QoS, flexibility, scalability, mobility, and security, are currently being developed. However, these frameworks raise security vulnerabilities. An international consortium presents requirements and recommendations for the evolving 4G mobile networking technology (Akhavan, Vivek Badrinath, & Geitner, 2006). The 4G technology, which is at its infancy, is supposed to allow data transfer up to 100 Mbps outdoor and 1 Gbps indoor. The International Telecommunications Union (ITU) defines 4G as downlink throughput of 100 Mbps or more, and corresponding uplink speeds of at least 50 Mbps.

The 4G technology will support roaming for interactive services such as video conferencing. The cost of the data transfer will be comparatively low and global mobility will be possible. The networks will be all IPv6 networks. WLAN, 2.5G, 3G, and other networks such as SATCOM, WiMAX, and Bluetooth will be integrated in 4G networks. The antennas will be much smarter and improved access technologies like OFDM and MC-CDMA will be used. More efficient algorithms at the physical layer will reduce the inter-channel interference and cochannel interference.

Security Issues

Seamless convergence of heterogeneous wireless networks provides new security challenges for the research community. Global authentication architectures are needed which can operate independently of the wireless physical protocol. In addition, specifications are needed for maintaining the confidentiality and the integrity of the communication data while the user terminal is in a hand-off state. In this direction, a forum of mobile operators called fixed mobile convergence alliance (FMCA) is working on defining specifications for the convergence of heterogeneous networks in the context of all IP 4G wireless systems.

Security policy issues are:

- The use of lightweight and flexible authentication, authorization, account, and audit (AAAA) schemes,
- The use of Trusted Computing (Reid, Nieto, & Dawson, 2003), and
- Different security polices for different services are recommended for 4G systems (Zheng, He, Xu, & Tang, 2005a).

Several security architecture proposals for 4G wireless systems have been made:

- Zheng, He, Yu, and Tang (2005b) propose a security architecture with:
  - Network access security features.
  - Network area security features for secure data exchange between network nodes.
  - User area security features for secure access to ME/USIM.
  - Application security for secure end-to-end data exchange.
- Integration of the SSL security protocol and a public key infrastructure is outlined and evaluated by Kambourakis, Rouskas, and Gritzalis (2004).
- A hierarchical trust model for 4G wireless networks is proposed by Zheng et al. (2005a).

CONCLUSION

In this chapter, the dominant mobile broadband technologies have been evaluated and compared based on their security performance. Three technologies were taken into consideration: Wi-Fi, UTMS, and WiMax. Their security architectures have been presented and analyzed in order to highlight the main security deficiencies. The evaluation and comparison methodology was based on assigning qualitative rankings to security threats with respect to the following criteria: likelihood, impact, and risk. The methodology was applied on four evaluation axes: authentication, confidentiality, integrity, and physical layer resilience. According to the comparison results, Wi-Fi is more liable to security attacks, followed by WiMax and UMTS. However, WiMax has not been widely tested under real-world systems due to its recent release. More security vulnerabilities may therefore be discovered in the future. Finally, the security architecture of UMTS is quite robust because of the lessons learned from GSM, but it is still not invincible against an experienced attacker with the right equipment.

REFERENCES

Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., & Levkowetz, H. (2004). *Extensible authentication protocol (EAP)* (IETF RFC 3748).

Aboba, B., & Simon, D. (1999). *PPP EAP TLS authentication protocol* (IETF RFC 2716).

Akhavan, H., Vivek Badrinath, V., & Geitner, T. (2006). *Next generation mobile networks beyond HSPA & EVDO* (White Paper.NGMN—Next

generation mobile networks Ltd.) Retrieved April 24, 2007, from http://www.ngmn.org/

Baek, K., Smith, W., & Kotz, D. (2004). *A survey of WPA and 802.11i RSN authentication protocols* (Tech. Rep. TR2004-524). Dartmouth College, Computer Science.

Barbeau, M. (2005). WiMax/802.16 threat analysis. In *Proceedings of the 1st ACM Workshop on QoS and Security for Wireless and Mobile Networks (Q2SWinet)*, Montreal, (pp. 8-15).

Biham, E., & Dunkelman, O. (2000). Cryptanalysis of the A5/1 GSM stream cipher. In *Proceedings of the First International Conference on Progress in Cryptology* (pp. 43-51).

Biryukov, A., Shamir, A., & Wagner, D. (2000). *Real time cryptanalysis of A5/1 on a PC*. Paper presented at the Fast Software Encryption Workshop 2000, New York.

Borisov, N., Goldberg, I., & Wagner, D. (2001). Intercepting mobile communications: The insecurity of 802.11. In *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking*, Rome, (pp. 180-189).

Edney, J., & Arbaugh, W. A. (2003). *Real 802.11 security: Wi-Fi protected access and 802.11i* (1st ed.). Addison-Wesley Professional.

ETSI. (2003). *Technical specification ETSI TS 102 165-1 V4.1.1.*

Haverinen, H., & Salowey, J. (2004). *Extensible authentication protocol method for GSM subscriber identity modules (EAP-SIM)* (Internet draft [work in progress]). Internet Engineering Task Force.

IEEE. (2001). IEEE standards for local and metropolitan area networks: Standard for port based network access control. *IEEE Std 802.1x-2001.* Retrieved April 24, 2007, from http://standards.ieee.org/getieee802/download/802.1X-2001.pdf

IEEE-SA. (2006). IEEE 802.16 LAN/MAN broadband wireless LANS. *IEEE 802.16 standards.* Retrieved April 24, 2007, from http://standards.ieee.org/getieee802/802.16.html

Kambourakis, G., Rouskas, A., & Gritzalis, S. (2004). Performance evaluation of public key-based authentication in future mobile communication systems. *EURASIP Journal on Wireless Communications and Networking, 1,* 184-197

Lehtonen, S., Ahonen, P., Savola, R., Uusitalo, I., Karjalainen, K., Kuusela, E., et al. (2006, September). *Information security in wireless networks.* Ministry of Transport and Communication. Finland: LUOTI Publications. ISBN 952-201-783-3. Retrieved April 24, 2007, from http://www.luoti.fi/material/InfoSec_in_WNetworks_final.pdf

Lynn, M., & Baird, R. (2002). *Advanced 802.11 attack*. Paper presented at the Black Hat 2002 Conference, Las Vegas. Retrieved April 24, 2007, from http://www.blackhat.com/presentations/bh-usa-02/baird-lynn/bh-us-02-lynn-802.11attack.ppt

Meyer, U., & Wetzel, S. (2004a). On the impact of GSM encryption and man-in-the-middle attacks on the security of interoperating GSM/UMTS networks. In *Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC2004).*

Meyer, U., & Wetzel, S. (2004b). A man-in-the-middle attack on UMTS. In *Proceedings of ACM Workshop on Wireless Security (WiSe 2004).*

Ohrtman, F. (2005). *WiMax handbook. Building 802.16 wireless networks.* McGraw-Hill Communications.

Reid, J., Nieto, J., & Dawson, E. (2003). Privacy and trusted computing. In *Proceedings of the 14th International Workshop on Database and Expert Systems Applications* (pp. 383-388).

Stanley, D., Walker, J., & Aboba, B. (2005). *Extensible authentication protocol (EAP) method requirements for wireless LANs* (IETF RFC 4017).

Stubblefield, A., Ioannidis, J., & Rubin, A. (2002). *Using the Fluhrer, Mantin, and Shamir attack to break WEP.* Paper presented at the NDSS.

Van de Wiele, T. (2005). *Wireless security: Risks and countermeasures* (UNISKILL Whitepaper).

Welch, D. J., & Lathrop, S. D. (2003). *A survey of 802.11a wireless security threats and security mechanisms* (Tech. Rep. ITOC-TR-2003-101). United States Military Academy.

WiMax Forum. (2006). *Mobile WiMax—Part I: A technical overview and performance evalation*. Retrieved April 24, 2007, from http://www.wimaxforum.org/home/

Zheng, Y., He, D., Xu, L., & Tang, X. (2005a). Security scheme for 4G wireless systems. In *Proceedings of 2005 International Conference on Communications, Circuits and Systems* (Vol. 1, pp. 397-401).

Zheng, Y., He, D., Yu, W., & Tang, X. (2005b). *Trusted computing-based security architecture for 4G mobile networks***.** Paper presented at the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies PDCAT 2005 (pp. 251-255).

KEY TERMS

Authentication: Verification of the identity of a user or network node who claims to be legitimate.

Broadband: A network connection with a bandwidth of about 2 Mbps or higher.

Confidentiality: A cryptographic security service which allows only authorized users or network nodes to access information content.

EAP: Extensible authentication protocol (EAP) is an authentication protocol used with 802.1X to pass authentication information messages between a suppliant and an authentication server.

Integrity: A security service which verifies that stored or transferred information has remained unchanged.

UMTS: Universal mobile telecommunication system (UMTS) is a global third generation wireless cellular network for mobile telephony and data communication with a bandwidth up to 2 Mbps which can be upgraded up to 20 Mbps with high speed packet access (HSPA).

Wi-Fi: Wireless local area networking based on IEEE 802.11 standards.

WiMax: Wireless metropolitan area networking based on IEEE 802.16 standards.

WPA, WPA2: Wi-Fi protected access (WPA) is a protocol to secure wireless networks created to patch the previous security protocol WEP. WPA implements part of and WPA2 implements the entire IEEE 802.11i standard. In addition to authentication and encryption, WPA also provides improved payload integrity.

This work was previously published in Handbook of Research on Wireless Security, edited by Y. Zhang, J. Zheng, and M. Ma, pp. 759-775, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.15
Developing a Theory of Portable Public Key Infrastructure (PORTABLEPKI) for Mobile Business Security

Sashi Nand
Rushmore University, Grand Cayman, BWI

INTRODUCTION

This chapter looks at how a public key infrastructure (PKI) can increase the wireless network's security by requiring certificate-based authentication for access. It also develops a theory of PORTABLEPKI. Finally, a framework for testing PORTABLEPKI and future research opportunities are discussed.

MOBILE BUSINESS

Mobile Business (m-business) can simplistically be understood as follows:

M-Business = Internet + E-Business + Wireless

M-business is the application infrastructure required to maintain business relationships by means of mobile devices. M-business is also the logical extension of electronic business (e-business) to address new customer channels and integration challenges. There is an interconnection of business processes within an organization and between external parties. For the notion of "business without boundaries" to prevail, back-end applications and data must be re-engineered to take complete advantage of the features offered by m-business (Kalakota & Robinson, 2002).

The most challenging and complex aspects of the m-business revolution are the design implementation, security, and integrity of mobile-enhanced business processes because they transcend traditional and regulatory boundaries (Stanley, 2004).

WIRELESS NETWORK

Wireless technologies are based on communication without land-based physical connections. For example, traditional telephone handsets use continuous cabling for connectivity, hence it is wired. Wireless telephony, on the other hand, uses radio waves rather than cables to broadcast network traffic and data transmission.

The two primary areas of wireless technology are mobile phones and mobile computers. Mobile implies portability—a device such as a mobile phone, PalmPilot, or laptop that travels with the user and can be used either off-line or online:

- *Mobile and off-line* means that the device can be used to run self-contained applications while not connected to the Internet or other telephony devices.
- *Mobile and online* is commonly called wireless. This means that the experience is based on a live connection supplied via satellite, cellular, or radio transmission. An online device will always be 'on' in the presence of any wireless network—seamlessly connecting to the Internet or some other system (Kalakota & Robinson, 2002).

What is a Wireless Network?

In a wireless network, radio waves carry the signal at least part of the way. The greater the proportion of the wireless to wired, the more wireless we consider the network. Three basic wireless networking technologies include:

- **Wireless Private Area Networks (WPANs):** Refer to confined short-range networks, for example computers connected while traveling such as mobile phones, laptops, and personal digital assistants (PDAs).
- **Wireless Local Area Networks (WLANs):** Refer to same local-range networks, for example computers connected within the same area such as an office building or home.
- **Wireless Wide Area Networks (WWANs):** Refer to long-range networks, for example computers connected over long distances such as a university campus, city, or town (Shaw, 2003).

SECURITY

With any new technology—especially wireless networking—concerns and questions arise about security of data transmission (Shaw, 2003). Security is a process of minimizing risk, threat, or the likelihood of harm (Pipkin, 2000).

Wireless communications are inherently more open to attack than wired data transfer because the physical layer is the uncontained cyber-space (Campbell, Calvert, & Boswell, 2003).

An insecure wireless connection exposes users to intrusion, which can lead to a loss of protection for confidential information, interception of messages, or abused connections. Some examples are:

- E-mail can be intercepted, read, or changed.
- A hacker who hijacks a session can replace a user's credentials with false information gaining access to the system.
- An unauthorized person can log on to a wireless network that is not secure and use the resources, or obtain financial gain through deception including free connectivity to the Internet (Chan, 2004).

Security dominates discussions about wireless communication. The reason is simple: removing the wires simultaneously removes the access restrictions. In fact, many wireless networks begin life completely unsecured because vendors design wireless access points (WAPs) and WLAN cards with ease of installation and usage in mind. Configuration of security settings does not equate

with ease of use. For this reason, a secure network needs to be set up intentionally and consciously (Randall & Sosinsky, 2005).

Even the most technically efficient and well-managed wireless network will be of little use if the network is not secured (Shaw, 2003). When implementing a wireless network, a plan must be developed for securing the network to reduce the likelihood of risks and threats.

METHODS OF SECURING WIRELESS NETWORKS

Security of wireless networks is specifically covered by the Institute of Electrical and Electronic Engineers (IEEE) 802.11i security specification. Some of the common methods of protecting a wireless network are as follows.

Media Access Control (MAC) Filtering

One of the most basic ways of protecting a wireless network is to implement MAC filtering. At the WAP, configure those MAC addresses (the low-level firmware address of a wireless card) that are allowed to connect to the WAP. Although this sounds like an ideal and easy way to secure a wireless network, consider the following weaknesses:

- It is easy to spoof an approved MAC address.
- MAC filtering is hard to manage.
- MAC filtering authenticates only the computer, not the user.
- The size of the approved MAC list is limited.

Wired Equivalent Privacy (WEP)

WEP provides encryption services to wireless networking. When a wireless connection enables WEP, the wireless network interface card (NIC) encrypts each data packet transmitted on the network using the Rivest Cipher version 4 (RC4) stream cipher algorithm. The WEP then decrypts the data packets on receipt. The weakness in WEP's implementation is two-fold:

- The symmetric encryption key is rarely changed.
- The initialization vector (IV) is only 24 bits and is re-used over time, thereby giving rise to a pattern of usage which can be easily identified and exploited (Komar & Microsoft PKI Team, 2004).

Wi-Fi Protected Access (WPA)

This is an encryption standard produced by the Wireless Fidelity (Wi-Fi) Alliance to address the security issues found in WEP. The following enhancements are included in WPA:

- **Increased Data Encryption:** WPA implements Temporal Key Integrity Protocol (TKIP), which uses a per-packet key mixing function, a message integrity check (MIC) known as *Michael*, and an extended IV with rules on sequencing. In addition, WPA implements a re-keying mechanism so that the same key is not used for long periods of time.
- **Dependency on 802.1x Authentication:** The use of 802.1x authentication is optional for WEP encryption only. WPA requires 802.1x authentication to ensure that only authorized users or computers are allowed to connect to the wireless network. 802.1x authentication also ensures mutual authentication so that a wireless client does not connect to a rogue network, rather than an authorized network.

Weaknesses in the current WEP algorithm implemented in current WLANs have been ex-

posed. The new security supplement to the 802.11 MAC standard is 802.11i, which will address security holes in the 802.11a, b, and g protocols, and improve encryption, key management, distribution, and user authentication. This standard is worth remembering, because these improvements to security may be available as firmware and later hardware upgrades for existing Wi-Fi networks (McCullough, 2004).

The current WPA definition includes forward compatibility with the new 802.11i security specification. 802.11i adds secure fast handoffs, secure de-authentication, and secure disassociation with WAPs. 802.11i also implements strong forms of authentication from the Advanced Encryption Standard (AES) (Komar & Microsoft PKI Team, 2004).

Public Key Infrastructure

PKI significantly increases the security of wireless networks because it requires encryption as well as a certificate-based authentication for access. PKI uses pairs of cryptographic keys (public key and private key) provided by a trusted third party, known as a certification authority (CA), which is verified by a registration authority (RA). Central to the workings of PKI, a CA issues a digital certificate, which positively identifies a holder of keys. The CA maintains accessible directories of valid certificates and also a list of certificates it has revoked.

PKI brings to the electronic world the security and confidentiality normally provided by physical documents such as handwritten signatures, sealed envelopes, and established trust relationships that are part of traditional paper-based transactions. These security and confidentiality features are as follows:

- **Confidentiality:** Ensures that only intended recipients can read files, or changes can only be implemented with a valid key.
- **Data Integrity:** Ensures files cannot be changed.
- **Authentication:** Ensures that participants in an electronic transaction are who they claim to be.
- **Non-Repudiation:** Prevents participants from denying involvement in an electronic transaction (Austin, 2001).

Vendors can provide security solutions that install digital certificates on the end devices itself, optimizing the PKI implementation especially for the wireless environment. For example, Microsoft Windows Server 2003 PKI provides the necessary certificates for 802.x authentication for wireless as well as wired networks. When a user or computer performs 802.1x authentication for wireless or wired network, the following two authentication types are available:

- **Extensible Authentication Protocol with Transport Layer Security (EAP/TLS):** A certificate-based authentication method that provides mutual authentication between the user or computer and the Remote Authentication Dial-In User Service (RADIUS) server when implemented for a wireless networking solution.
- **Protected Extensible Authentication Protocol (PEAP):** Allows the transmission of other EAP types within a TLS secure channel (Komar & Microsoft PKI Team, 2004).

Thus, in an open, untrusted, and insecure wireless network environment, cryptography provides the security and PKI provides the trust to enhance m-business (Deloitte & Touche Research Team, 2001).

The lack of security in mobile business is the fundamental problem. In order to better understand and examine ways in which this problem may be solved, a research theory that addresses

this critical issue of security using PKI is developed in the following section.

DEVELOPING A THEORY OF PORTABLE PKI (PORTABLEPKI)

There can be situations where it is absolutely critical to use PKI, but it comes at a cost. Nevertheless, despite its costs and complexity, laying down a sound theoretical foundation, combined with best business practices and robust technological infrastructure, will enhance the usage of PKI and the security of mobile business. Hence Portable Theory of PKI (PORTABLEPKI) has been developed.

There is no definite meaning given to the term "theory," and there are many views on what constitutes a theory. The standard or the orthodox view has been used to construct PORTABLEPKI theory. According to this view there exists a phenomena in the real world (P-Field). Observation of phenomena leads to abstractions by an individual's reason (C-Field) (Staunton, 1976). A theory begins in the 'unreal' world of abstraction, that is, in the human mind (C-Field). In order for it to be useful, theory must eventually relate to the 'real' world, the world of experience (P-Field). Three types of relationships in the theoretical structure are:

1. **Syntactics:** Rules of language. If expressed in English, then the relationship refers to the rules of grammar. If the theory is mathematical, then the relationship refers to the rules of mathematics.
2. **Semantics:** Rules of correspondence or operational definitions which link the concepts to objects in the real world. Semantics concern the relationship of a word, sign, or symbol to a real-world object or event. It is the semantic relationship that makes a theory realistic and meaningful.
3. **Pragmatics:** The effect of words or symbols on people. We are interested in how concepts and their measured correlations in the real world affect people's behavior (Nand & Unhelkar, 2003).

The first step of PORTABLEPKI theory is to identify the research problem in the P-Field by observing the use of PKI in the real world of mobile business. The next step is to develop the conceptual and theoretical structure, including the causal links and chains, and state the hypothesis (H). Then the hypothesis can be written in simple English stating the relationship of each clause whether they are directly or indirectly related. The real-world effect on stakeholders also has to be shown. The overall theory of PORTABLEPKI can be viewed as a set of principles for the purpose of enhancing growth and acceptance of PKI, as well as to enhance the security of m-business. A framework for testing this theory is provided in the next section.

A FRAMEWORK FOR TESTING PORTABLEPKI

An empirical research program based on the inductive-deductive approach developed by Abdel-Khalik and Ajinkya can be modified to test this theory of PORTABLEPKI (Godfrey, Hodgson, & Holmes, 1997). This involves the following eight stages:

Stage 1: Identify a Research Problem by Observation

PKI is one of the remedies to m-business security problems. An examination can be made of the following six factors influencing use of PKI technology in m-business in Australia (Nand & Unhelkar, 2003):

- **Environment:** Includes security, globalization, market competition, regulating

forces, telecommunications, and political influence.

- **Organization:** Includes corporate governance, management, organizational structure, and resources.
- **Business Strategy:** Includes strategic planning, business process re-engineering, total cost of ownership, and return on investment.
- **IT Strategy:** Includes strategic planning, system development, system maintenance, technological risk (including wireless), and complexity of PKI.
- **PKI Technology:** Includes necessity of trust, PKI initiatives, PKI availability, and PKI success stories.
- **People:** Includes PKI skills, PKI training and dissemination of information, and employee culture.

Stage 2: Develop the Conceptual and Theoretical Structure, Including Causal Links and Chains

To develop a rationale as to why firms do or do not use PKI, the study would test the effect of some selected independent variables for the use of PKI. Two independent variables which influence use of PKI are industry type (service or non-service) and the number of years of IT experience. The dependent variable is the level of usage of PKI.

Stage 3: Operationalize the Theoretical Constructs and Relationships, and State the Specific Hypothesis to be Tested

Two hypotheses that have been developed to test this theory are:

- **Hypothesis 1:** Higher usage of PKI technology is expected in the service industry compared with the non-service industry.
- **Hypothesis 2:** Greater usage of PKI technology is expected in organizations that have a greater number of years of IT experience.

Stage 4: Construct the Research Design

The survey research method can be adopted to obtain data from organizations Australia-wide.

Stage 5: Implement this Design by Sampling and Gathering Data

A sample of Australian companies from at least one service industry and one non-service industry can be selected and company details recorded using a database.

Stage 6: Analyze Observations in Order to Test Each Hypothesis

Descriptive statistics and Chi-Square Test can be used to process and analyze the collected data using Microsoft Excel together with PHStat, Prentice-Hall's statistical add-in for Excel. With descriptive statistics frequency, distributions of all responses to the national survey can be recorded using simple tabulations and cross-tabulations on the Microsoft Excel spreadsheet. By using Chi-Square Test, hypotheses 1 and 2 can be tested together with PHStat. This test involves comparison of actual frequency with expected frequency.

Stage 7: Evaluate the Results

Determine whether or not the results support the theory of PORTABLEPKI.

Stage 8: Consider the Specific Limitations and Constraints

Refer to the procedures undertaken in Stages 1-7, and ask: Are there any limitations to the

way the theory was developed or tested? Do any refinements to the theory appear warranted? If the answer is 'yes' to either question, then return to the appropriate stage and attempt to remedy the limitation.

CONCLUSION AND FUTURE DIRECTIONS

Theories play an important role in understanding and changing the world. This chapter has developed a theory of PORTABLEPKI, which breaks new ground. The next step is to test the hypotheses stated in this chapter using the framework provided here and to validate the reality of PKI usages. This will lead to the refinement of PORTABLEPKI theory. People dealing with information security systems have to be ever vigilant because security is an unending mission. While creativity and innovation are what drives new technology, it also gives rise to its associated security problems. Hence the implementation of PORTABLEPKI theory will lead to increased usage of PKI which consequently will enhance the security of both wired and wireless networks and mobile businesses.

This is an initial step for the development of the theory of PORTABLEPKI for the purpose of increasing the security of m-business. The future direction is for the PORTABLEPKI theory to be tested by selecting one specific service industry (e.g., automobile telematics (wireless telemetry) industry) and one non-service industry (e.g., mining oil and gas). This testing will either confirm or negate the PORTABLEPKI theory. Further work could then be undertaken to refine this theory for other stages of development in m-business.

REFERENCES

Austin, T. (2001). *PKI: A Wiley tech brief.* New York: John Wiley & Sons.

Campbell, P., Calvert, B., & Boswell, S. (2003). *Security + guide to network security fundamentals.* Boston: Cisco Learning Institute, Thomson Course Technology.

Chan, D. (2004). What auditors should know about encryption. *Information Systems Control Journal, 3,* 32.

Deloitte & Touche Research Team. (2001). *E-commerce security: Public key infrastructure: Good practices for secure communications.* Rolling Meadows, IL: Information Systems Audit and Control Foundation.

Godfrey, G., Hodgson, A., & Holmes, S. (1997). *Accounting theory* (3rd ed.). Sydney: John Wiley & Sons.

Kalakota, R., & Robinson, M. (2002). *M-business: The race to mobility* (pp. 8-10, 19). New York: McGraw-Hill.

Komar, B., & Microsoft PKI Team. (2004). *Microsoft Windows Server 2003 PKI and certificate security* (pp. 467-471). Redmond, WA: Microsoft Press.

McCullough, J. (2004). *185 wireless secrets: Unleash the power of PDAs, cell phones, and wireless networks.* Indianapolis: Wiley Publishing.

Nand, S., & Unhelkar, B. (2003, November 24). Progress report on development of "Investigations Theory of PKI" and its application to Australian information systems. *Proceedings of the 1st Australian Information Security Management Conference*, Perth, Australia (p. 3).

Nand, S., & Unhelkar, B. (2003, December 16-18). Development of an Australian trust scheme of PKI to enhance confidence in security for e-transforming organisations: A study of a cluster of SMEs in Australia. *Proceedings of the 2003 International Business Information Management Conference*, Cairo, Egypt (p. 5).

Pipkin, D. L. (2000). *Information security: Protecting the global enterprise*. Upper Saddle River, NJ: Prentice-Hall.

Randall, N., & Sosinsky, B. (2005). *PC Magazine: Wireless solutions.* Indianapolis: Wiley Publishing.

Shaw, R. (2003). *Wireless networking made easy.* New York: AMACOM.

Stanley, R. A. (2004). Security, audit and control issues for managing risk in the wireless LAN environment. *Information Systems Control Journal, 3,* 23.

Staunton, J. J. (Ed.). (1976). *Theory construction and verification in accounting.* Armidale, Australia: University of New England.

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 393-400, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.16
Authentication, Authorisation, and Access Control in Mobile Systems

Josef Noll
University Graduate Center – UniK, Norway

György Kálmán
University Graduate Center – UniK, Norway

ABSTRACT

Converging networks and mobility raise new challenges towards the existing authentication, authorisation, and accounting (AAA) systems. Focus of the research is towards integrated solutions for seamless service access of mobile users. Interworking issues between mobile and wireless networks are the basis for detailed research on handover delay, multi-device roaming, mobile networks, security, ease-of-use, and anonymity of the user. This chapter provides an overview over the state of the art in authentication for mobile systems and suggests extending AAA mechanisms to home and community networks, taking into account security and privacy of the users.

INTRODUCTION

Today's pervasive computing environments raise new challenges against mobile services. In future visions, a converged user access network is projected. This means, that one network will be used to deliver different services, for example, broadcast TV, telephony, and Internet. Composed from mobile (e.g., Universal Mobile Telecommunications System [UMTS]), wireless (IEEE 802.11, IEEE 802.16, IEEE 802.20), and wired (cable, Asymmetric Digital Subscriber Line [ADSL]), these networks hide the border between the telecom, broadcast, and computer networks. The common service enables roaming terminals, which can access services independently of the currently used networking technology. Market players in both areas transform into wireless service providers across access networks. Telecom provide packet

switched data and mobile services over the fixed network, while Internet service providers run voice over IP (VoIP) and video on demand (VoD) over mobile networks.

The changing environment also changes the management plane of the underlying networks. Providers on converged networks have to change their accounting and billing methods and need to redefine their business models. While commercial players demonstrate early examples, research in the AAA area focuses on providing a backplane for the upcoming ubiquitous services run over converged networks.

BACKGROUND

The AAA methods employed in current networks were developed for a single type of network, resulting in two different systems, one for telecommunication services and one for computer networks. This chapter addresses AAA in global system for mobil communications (GSM) and UMTS and computer network solutions based on Internet Engineering Task Force (IETF) standards.

The computer networks provide a unified AAA access, and research focuses on extending the existing methods to be suitable for telecommunication services. Extensions for Remote Authentication Dial In User Service (RADIUS) and Diameter are proposed. RADIUS is the current de facto standard for remote user authentication. It uses Universal Datagram Protocol (UDP) as transport. Authentication requests are protected by a shared secret between the server and the client, and the client uses hash values calculated from this secret. The requests are sent in plaintext except for the user password attribute. The Diameter protocol provides an upgrade possibility as compared to RADIUS. While enhancing the security through supervised packet transmission using the transmission control protocol (TCP) and transport layer encryption for reducing man-in-the-middle attacks, it lacks backward compatibility.

Both methods have a different background. The computer networks targeted the person using a computer in a fixed network environment, while mobile systems addressed a personal device in a mobile network. Thus a challenge for telcos is to enhance seamless network authentication towards user authentication for service access. Most companies are also Internet service providers (ISPs), this would be a natural unification of their AAA systems.

A generic approach is taken by extension of the Extensible Authentication Protocol (EAP) family. Development efforts of the Internet and telecommunication world were united on EAP. This protocol family has the potential for becoming the future common platform for user authentication over converged networks. EAP is a universal authentication framework standardised by IETF, which includes the authentication and key agreement (AKA) and Subscriber Identity Module (SIM) methods. EAP-AKA is the standard authentication method of UMTS networks.

Beside the fundamental differences of communication and computer networks, mobility is the key issue for both. Network services should not only be accessible from mobile terminals, but they should be adapted to the quality of service (QoS) requirements of a mobile/wireless link. Improvements of AAA methods are of fundamental importance for mobility, providing fast handover, reliable and secure communications on a user-friendly and privacy protecting basis.

Subscriber Authentication in Current Networks

In GSM networks, the integrated AAA is used for any type of user traffic. The authentication is just one way the user has to authenticate himself/herself towards the network.

To be more precise, the user is authenticated with a PIN code towards the SIM in the mobile phone, then the device authenticates itself towards the network. Device authentication instead of user

authentication can hinder the upcoming personalised services because it is hiding the user behind the device. In UMTS, the authentication of the device is two-way. A device can also check the authenticity of the network with the help of keys stored on the SIM.

Integration of the mobile authentication with different external services is not widespread. The telecom providers have some internal services, which can authenticate the subscriber based on the data coming from the network. Credentials could be basically the CallerID, the Temporary International Mobile Subscriber Identity (TIMSI) or other data transformed with a hash function. Access control and authorisation is more an internal network task. Without considerable extension, the current mobile networks are more islands than connecting networks in the area of AAA. Equipment manufacturers are now recommending various IP multimedia subsystem (IMS) solutions for mobile providers in order to enable integrated and third party service convergence and to enable multimedia content over today's networks.

AAA protocols employed in computer networks are meant to provide services for authenticated users. Current single sign-on (SSO) protocols, like RADIUS, Diameter, or Kerberos provide the identity of the user to third parties. SSOs can use digital certificates, public key infrastructure (PKI) and other strong encryption methods. But, none of them is able to provide such a complete solution like the integrated AAA of the mobile network. Computer network protocols lack the support for fast mobility of moving clients and optimise resource usage for low bandwidth connections.

With incorporating seamless authentication used in network internal services in telecom world and SSO solutions provided by various protocols from computer networks, a unified AAA system will achieve a enhanced user acceptance and service security. In such a system, secure key storage and tamper resistant handling is crucial. Smart cards for key storage and generation will fulfil the security requirements, but usage and distribution of the smart cards is cumbersome. As most users have a mobile phone, the SIM card is a candidate to be a primary smart card used for AAA in a ubiquitous environment (Kálmán & Noll, 2006).

AAA IN CONVERGED NETWORKS

A converged network carries several types of traffic and enables seamless information exchange between different terminals, regardless of transport medium. To enable converged AAA, research work is going on in different areas: enabling wireless LAN (WLAN)-mobile network interworking, enhancing network mobility in wireless computer networks, and reducing resource requirements in cryptography.

Interworking Between Mobile and Wireless Networks

Network convergence is most significant in the wireless environment, having to face varying QoS measures on the radio interface, for example, propagation delay, variation of delay, bit error rate, error free seconds, distortion, signal to noise ratio, duration of interruption, interruption probability, time between interruption, bit rate, and throughput. These parameters will depend on the user and terminal environment and underline that an optimum access will have to use all available wireless and mobile connections. Leu, Lai, Lin, and Shih (2006) have provided the fundamental differences of these networks, summarised in Table 1.

Increased demand for security has improved the security on wireless links, resulting in Wi-Fi protected access (WPA) and WPA2 as draft implementations of the IEEE 802.11i standard. This standard aims at incorporating protocols of the EAP family, especially transport layer security (TLS) and SIM.

Table 1. Comparison of cellular and WLAN networks

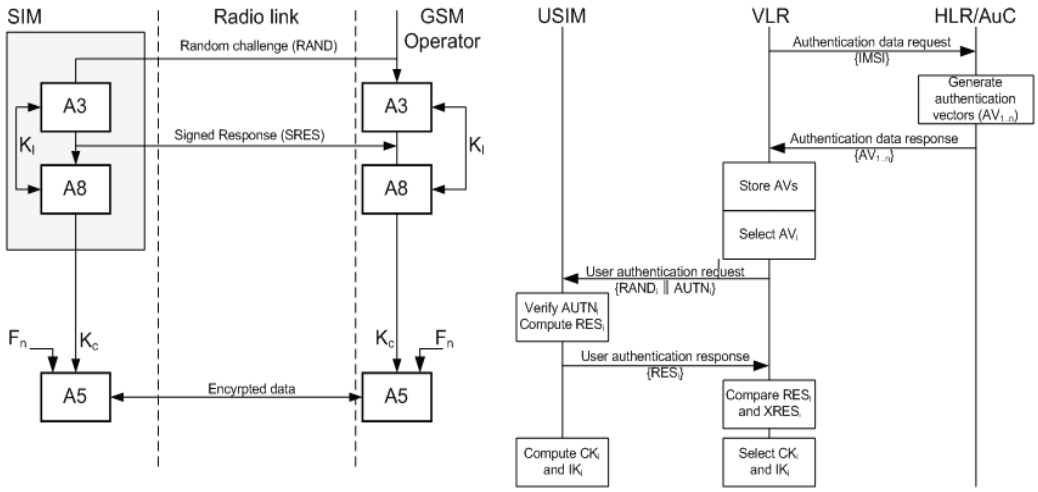
	Cellular	**WLAN**
Coverage	Country-wide	Local
Security	Strong	Depends on setup
Transmission rate	Low	High
Deployment cost	High	Low
License fee	Very high	No need
Construction	Difficult	Easy
Mobility support	High	Poor

Most cellular operators are now providing WLAN services using the Universal Access Method (UAM) for authentication. UAM uses a layer 3 authentication method, typically a Web browser to identify the client for access to the WLAN. This raises the problem of mutual authentication, which has been a problem also in GSM networks. By extending to EAP-SIM it would be possible to enable SIM-based authentication in these environments for SIM-enabled devices.

Roaming between access providers is a second issue. Since data between access points are carried over an IP backbone, it is natural to use a network-based protocol such as Radius, suggested by Leu et al. (2006). Transport encryption inside the backbone is indifferent from normal wired practice, hence out of scope for this chapter. In a converged network, where users can switch between mobile networks and WLAN services, a common AAA system has to be operational to ensure correct operation. A unified billing scheme is proposed by Janevski et al. (2006), suggesting to use 802.1x on the WLAN side as shown on Figure 2. The mobile networks WLAN connection is suggested through the RADIUS server used also for access control in 802.1x.

The use of the IEEE 802.1x standard allows seamless authentication, since preshared certificates and key negotiation are provided to the cellular network, where the user is already

Figure 1. Authentication in GSM and UMTS

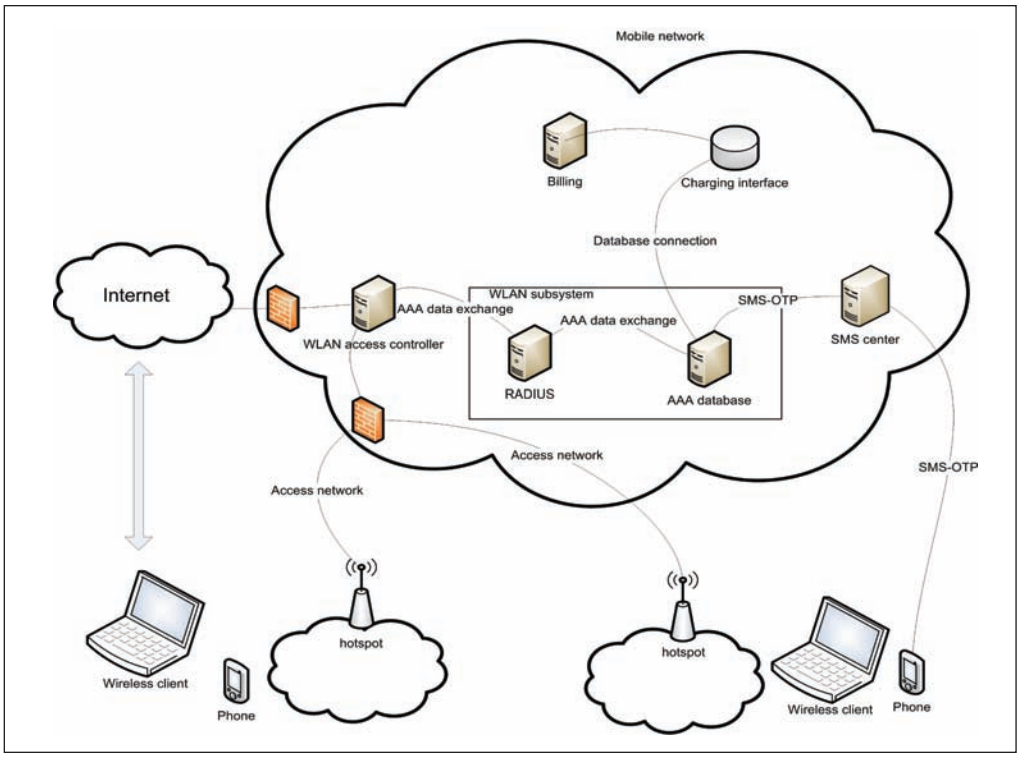


authenticated. With the use of digital certificates, the system is getting closer to the preferred view of pervasive systems, where the user and the service providers are mutually identified. Since these systems authenticate the user towards several services, privacy is a primary concern. A possible solution, recommended by Ren, Lou, Kim, and Deng (2006) has a secure authentication scheme while preserving user privacy.

In pervasive environments a user connected will experience seamless authentication to all services when connected through a SSO service. Malicious tracking of his/her behaviour or eavesdropping of authentication messages can compromise the user credentials. The SSO service has to be extremely prudent when sending user-related information. Keeping a reasonable level of privacy, the system should deal with questions in location privacy, connection anonymity, and confidentiality (Ren et al., 2006). The recommendations are based on blind signatures and hash chains. Using hash is highly recommended, since a good hash function can provide good foundation for anonymous access and its resource needs are not too high for the current mobile devices, as sometimes blind signatures based on Rivest-Shamir-Adleman (RSA) scheme may be. In certain environments, the GSM integrated functions may also be used.

The user retains full control over authentication credentials when composing and generating authentication tokens like the identities suggested by Chowdhury and Noll (2007). Initial service access can be achieved showing one of these tokens after mutual identification between the service and the user. Based on these tokens, no user data can be retrieved nor traced back. If all of the initial identification steps succeed, the exchange of the required credentials can proceed using a freshly negotiated session key.

Figure 2. Integration of radius and mobile network authentication



The base of most authentication techniques is a preshared key, delivered to the user device out-of-band. Authentication can be done for example in mobile phones by inserting a master private key on the SIM at the activation of the card (Kálmán & Noll, 2006).

A different approach is to extend the current mobile network with additional elements to enable network integrated AAA also in an Internet environment. Khara, Mistra, and Saha (2006) suggest including a new node, called Serving GPRS Access Router. This entity acts as a gateway for the WLAN traffic to enter the general packet radio services (GPRS) backbone and enable GPRS signalling to control WLAN. The new protocol set eliminates the need of Signalling System 7 (SS7) in addition to the IP backbone. Khara et al. claim that this solution is superior in terms of speed and overhead compared to the RADIUS-based methods suggested previously. The main drawback is the need of special dual mode devices with a split IP layer, a solution which might not be practical having in mind the basis of 2.5 billion mobile phones available in the market.

For mobile devices limited computational resources and battery power require an effective AAA mechanism. Extension of the GPRS/UMTS network could be potentially more expensive than deploying RADIUS authentication. Handover delay caused by terminal mobility is an issue which might favour GPRS/UMTS protocols.

Authentication in Converged Networks

From the data traffic's point of view, the speed of the network's internal routines does not play a primary role, in VoIP and other sensitive services, QoS is a key parameter. Delay reduction is currently the topic having intensive focus. Interconnecting mobile and IP networks for data traffic is not a challenge, since GPRS has an IP backbone, and UMTS is practically an IP network. Most of the problems begin when the network has to provide a certain QoS in order to support service with time-critical transmission, that is, voice or video calls. Delay in the wired network can be reduced by additional bandwidth to reduce collisions, alternate routing paths, or other methods. But in wireless environments, where terminals move around and connect to different networks, which may be "far" away in terms of network topology, switching the data transfer path is a challenging task.

In the IP world, Mobile IPv6 (MIPv6) was introduced to deal with mobility problems. This protocol works flawlessly for clients that are changing networks with quite low frequency and are connected to a wired network, where additional signalling and other overheads are not causing bandwidth problems. The convergence time of the routing in MIPv6 is quite slow. In a wireless environment every additional message exchange or signalling overhead has a direct influence on usability. When the terminal is moving fast between these distant networks, it may reach a speed, where the routing of MIPv6 can not keep the connection in a correct state. This means that while data traffic could be able to transmit with low average speed, QoS cannot be kept on an adequate level to support VoIP or VoD services, for example. To fight this problem, several micro-mobility (local area) protocols were developed to support fast moving nodes. Different approaches are used, for example in hierarchical MIPv6 with fast handover adds a *local home agent* into the network. Seamless handoff for MIPv6 tries to lower the handover time with instructing the nodes to change networks based on precalculated patterns.

Handoff between neighbouring IP networks could be done in reasonable time if they are cooperating, but with introducing converged network access, it is likely that the terminal moves between WLAN and UMTS networks and back in less than a minute. Session mobility, for example a VoIP call without interruption, cannot be achieved using current protocols. The key is to reduce

the handoff delay in interworking networks. To reduce delay inside the UMTS network, Zhang and Fujise (2006) show a possible improvement for the integrated authentication protocol. One cause of the long delay is getting an authentication vector (AV) if the Serving GPRS Support Node (SGSN) and Home Location Register (HLR) are far away. While roaming, the AV consumption is higher, if the terminal is moving frequently or it is producing significant traffic. The specifications allow a high blocking ratio of 20% for the UTMS network in case of requesting new AVs. The proposal claims to lower this rate to 2%. For each authentication instance, the SGSN consumes one AV from a first in first out (FIFO) storage.

A fundamental question is to allow the size of the AV vector to be customised based on the terminal's behaviour. In the default way, the SGSN executes a distribution of authentication vector (DAV) procedure if all AVs are consumed. Communication between the terminal and the SGSN cannot proceed until the reply is received from the HLR, inserting a potentially high delay into the system. This can lead to call failure, errors in location update, or unacceptable delays in services running on GPRS. The proposed protocol from Zhang and Fujise (2006) implies no change in case of the first authentication to the SGSN, but keeps track of the number of available AVs and sends out a new request when hitting a predefined level. This level can be customised for a network, to reduce or even remove the possible delay of waiting for an AV. The proposal also changes the basic behaviour, asking for new AVs when they are consumed. The original 3rd Generation Partnership Project (3GPP) system asks for them when a new event comes in and no AVs are available.

While reducing delay inside the GPRS network can reduce block probability in reaching network services, also handover functions in IP have to be revised in order to achieve reasonably fast mobility support. The basic challenge is that currently AAA and MIPv6 are operated independently. This means that the terminal has to negotiate with two different entities in order to get access to the new network.

In MobileIPv6, the terminal is allowed to keep connections to a home agent (HA) and a correspondent node (CN), even when the terminal changes point of attachment to that network. The terminal has two addresses, the home address (HoA) and the care-of address (CoA). The HoA is fixed, but the CoA is generated by the visited network. The mobile IP protocol binds these two addresses together. To ensure an optimal routing in the network, the terminals switch to *route optimalisation* mode after joining a new network. Then it executes a *return routability* procedure and a *binding update* (BU) to communicate to the correspondent node directly. The return routability procedure consists of several messages, which together induce a long delay.

The handover between networks implies even more steps and consumes more time: movement detection, address configuration, home BU, return routability procedure, and a BU to the correspondent node. The terminal cannot communicate with the CN before the end of the procedure.

Fast handover capability is a major research item in IETF for MIPv6, including the standards FMIPv6 and HMIPv6. In addition to these schemes, Ryu and Mun (2006) introduce an optimisation in order to lower the amount of signalling required and thus lower the handover delay between domains. In an IPv6 system, the IP mobility and AAA are handled by different entities. This architecture implies unnecessary delays. Several solutions are proposed to enable the mobile terminal to build a security association between the mobile node and the HA. This enables home BU during the AAA procedure. Route optimisation is a key topic in efficient mobility service provision. MIPv6 optimises the route with the use of the return routability procedure. In wireless environments, the generated signalling messages represent a considerable part of the whole overhead. Moving route optimisation into the AAA procedure can reduce the delay

by nearly 50% (Ryu & Mun, 2006). This was enabled by embedding the BU message into the AAA request message and so optimising the route while authenticating. This solution can solve MIPv6's basic problem of supporting different administrative domains and enable scalable large scale deployment.

Lee, Huh, Kim, and Lee (2006) define a novel communication approach to enable communication between the visited AAA servers for a faster and more efficient authentication mechanism. If a terminal visits a remote network, the AAA must be done by the remote system. IETF recommends integrating Diameter-based authentication into the MIPv6 system. But, when the user is using services on the remote network, the remote AAA has to keep a connection with the home AAA. The proposed new approach of Lee, Huh, et al. suggests enabling faster authentication when the terminal moves between subnets inside a domain by exchanging authentication data between visited AAA servers without the need of renegotiation with the HA. Connection to the HA is needed only after the authentication when the terminal executes a BU.

One other aspect is shown by Li, Ye, and Tian (2006) suggesting a topology-aware AAA overlay network. This additional network could help MIPv6 to make more effective decisions and to prepare for handovers and other changes in network configuration. Based on the AAA servers and connections between, a logical AAA backbone can be created, which can serve as administration backbone for the whole network. Signals delivered over this network are topologically aware, so the optimal route can easily be selected and signalling messages can be transmitted over the best route. In exchange to the build cost of this backbone network and some additional bandwidth consumed, MIPv6's security and performance can be enhanced.

As the route of the service access is secured, optimised and delay reduced, one basic problem still remains: how to ensure that the user is the one, the network thinks he/she is. Lee, Park, and Jun (2006) suggest using smart cards to support interdomain roaming. The use of the SIM might be preferable because of its widespread use and cryptographic capabilities (Kálmán & Noll, 2006). The problem of having multiple devices is also raised here, since a system based on the SIM as smart card will require SIM readers in every device—if a secure key exchange method between the devices is not in place.

Lee, Park, et al. (2006) suggest an entity called *roaming coordinator* ensuring seamless roaming services in the converged network. This additional node provides context management services and enables seamless movement between the third generation (3G) network and WLAN to enforce security in converged networks. In order to provide good user experience in a pervasive environment, additional intelligence needs to be added to the traditional AAA systems to ensure that the terminal selects the most appropriate connection method. This method has to be based on the context and has to be supported in all networks. A smart-card-based secure roaming management framework enables the transfer of the terminals context without renegotiating the whole security protocol set. When the terminal moves into a new network, the roaming coordinator, AAA servers, and proxies take charge of the authentication process. The coordinator, having received a roaming request, evaluates the available networks and chooses the best available one, and then triggers the context transfer between the corresponding AAA servers. When transferring whole user contexts, the system has to consider privacy requirements of the user's identity and his/her profile.

Anonymity and Identity

In pervasive environments, privacy is of key importance. With computers all around, gathering information about traffic, movements, service access, or physical environment, customer privacy

must be protected. Køien (in press) suggests a protocol, which is able to provide better protection for the user's privacy than the normal 3G network. Changes in the EAP-AKA protocol are suggested to use only random generated user authentication values. He defines three user contexts implying different key management and authentication schemes, like existing keys for short-term and fresh keys for medium-term access. Identity-based encryption is recommended to enable a flexible binding of the security context to protect the permanent subscriber identity and location data, which will only be discoverable by the home register. The main drawback of the suggested protocol is its higher computing requirements as compared to EAP-AKA, potentially limiting the applicability.

Security and Computing Power

A security protocol in a wireless environment should be fast and secure, and it has to be effective in terms of computing power and low data transfer need. In low power environments an authentication scheme with high security and low computing power is advised. One solution is based on hash functions and smart cards, allowing minimised network traffic and short message rounds used for authentication. Anonymity can be ensured through one-time passwords. While accepting the advantages of a system with smart cards, the use of extra hardware like a card reader is not advisable, due to compatibility issues and power requirements.

Software-based solutions have an advantage, as they only require computing power. Showing the importance of power consumption, a comparison of cryptographic protocols is presented by Lee, Hwang, and Liao (2006) and Potlapally, Ravi, Raghunathan, and Jha (2006) showing, that twice of the transmit energy of one bit is needed to run asymmetric encryption on that piece of information. Symmetric encryption needs, in contrast, around one half of the transmit energy. Most overhead is generated by session initialisation, meaning longer sessions induce lower overhead. There is a trade-off between security and session length. While negotiation overhead is getting lower with long sessions, security risks are getting higher.

This overhead can be lowered by special hardware or software solutions. Hardware needs some power and bigger silicon, while software requires a faster CPU. Hash functions have an energy requirement of around half a percent compared to PKI in generating session keys (Potlapally et al., 2006). Key exchange protocols using elliptic curve Diffie-Hellman (DH) come out much more energy efficient as compared to the same traditional strength DH. The DH calculations demonstrate the trade-off between power consumption and security. In order to have an efficient operation, the security protocol needs to have the possibility to adapt encryption to the needs of the current application. Authentication token generation can be problematic for devices with limited computing capabilities. Personal area networks (PAN) with multiple devices raise this problem by their very nature.

Security in Personal Area and Home Networks

Efficient authentication and certificate management ensures better usability of PAN devices. By using efficient security protocols, content-adaptive encryption, efficient key and certificate management, considerably longer battery operation is achievable. To enable key management in a PAN a personal certificate authority (CA) entity is suggested (Sur & Rhee, 2006; Sur, Yang, and Rhee, 2006), which will be responsible for generating certificates for all mobile devices within the PAN or home device domain (Popescu, Crispo, Tanenbaum, & Kamperman, 2004). Because of the context of use, the authentication protocol is focused on efficiency by reducing computational overheads for generating and verifying

signatures.

Main focus is on reducing PKI operations, which have been proven to be energy consuming. Instead, it proposes to use hash chains to lower communication and computational costs for checking certificates. Former research suggested hash trees in order to authenticate a large number of one-time signatures. By extending these with fractal-based traversal, it has been proven that these trees provide fast signature times with low signature sizes and storage requirements. The personal CA has to be a unique trusted third party in the PAN. It needs to have a screen, a simple input device, and has to always be available for the members of the network. A cell phone with the SIM is a perfect candidate to be a personal CA (Kálmán & Noll, 2006).

In home environments, basically two types of authentication are distinguished: (1) user authentication, and (2) device authentication (Jeong, 2006). Mutual authentication has to be used in order to prevent impersonation attacks (*identity theft).* This requires an SSO infrastructure, which can be for example Kerberos or RADIUS. A special aspect of resource access over the home LAN is that specific privileges are given to selected programs. The AAA server maintains an access control list to ensure correct privilege distribution.

To build the initial trust relationships some kind of user interaction is needed. The key should initially be distributed out-of-band, for example on an USB stick, or by using short range wireless technology, Near Field Communication (NFC), for example (Noll, Lopez Calvet, & Myksvoll, 2006). On home networks, where power consumption is not a problem, PKI may be used for negotiating session keys between devices, since key management in a PKI is simpler than in symmetric encryption and the delay caused by checking certificates and so forth will not be noticeable in this environment. Users authenticated towards the AAA infrastructure can access the resources seamlessly. Initial authentication is done with PKI. In case of mobile devices, also the home AAA can use previously calculated hash values in chain to lower computational cost. These AAA infrastructures can be connected to a providers AAA, for example to use in digital rights management (DRM) or home service access from a remote network (Popescu et al., 2004).

A user moving with his/her devices to the home raises another AAA challenge, the mobile nodes.

Mobile Nodes (Network Mobility)

Movement of whole networks like PANs or networks deployed on a vehicle, introduce a new level of AAA issues. In a conventional network a standard mobility support does not describe route optimisation. Several procedures are suggested to provide this functionality for mobile nodes, like Recursive Binding Update Plus (RBU+), where route optimisation is operated by MIPv6 instead of the network mobility (NEMO) architecture. This means, that every node has to execute its own BU with the corresponding HAs. To solve problems with pinball routing, it uses the binding cache in the CN. When a new BU message arrives, the RBU+ has to execute a recursive search, which leads to serious delays with a growing cache size. One potential route optimisation is presented by Jeong (2006).

A designated member of the network, called a mobile router is elected to deal with mobility tasks to reduce network overhead. The AAA protocol for this environment defines a handover scheme and tree-based accounting to enable efficient optimisation. They recommend using dual BU (DBU) procedure instead of the existing procedures like RBU+ as a solution for the reverse routing problem raised by mobility. DBU operates with additional information placed into the messages sent in a BU process. This is the CoA of the top level mobile router (TLMR). By monitoring the

messages, the CNs in the subnet can keep optimal route towards the TLMR.

Moving subnets are the subject of eavesdropping and possible leakage of the stored secrets. A secure AAA is proposed for network mobility over wireless links, which deals with these problems (Fathi et al., 2006). Secret leakage can be caused by malicious eavesdroppers, viruses, or Trojans. A possibility is to store the keys in tamper resistant modules, like smart cards, the SIM, or trusted hardware modules. Deploying additional modules can be problematic and expensive. Fathi et al. propose a protocol based on a short secret, which can be remembered by humans and used in a secure protocol called Leakage-resilient authenticated key exchange protocol (LR-AKE). This protocol is used for AAA to reduce NEMO latency under 300 ms in order to provide session continuity, for example in VoIP applications, which is important in keeping a good user experience. However, short passwords as proposed with LR-AKE are not advisable. If complex, they will be noted down by the user, and if weak, they are easy to guess.

As network mobility has considerable security issues, it may be not the way to go. Functionality of a mobile network might be achieved by using a dedicated device as a gateway of the PAN. Only this device will show up in the wireless network, and all traffic originating and arriving to the PAN will go through this device and its HA.

After these technical issues of authentication the next chapter will deal with authentication from the user viewpoint.

Customer Ergonomics

There is always a trade-off between user security and ease of use. If the system is prompting for a password for every transaction, it can assume with quite high probability, that the access is enabled just for the correct user. But, that is unacceptable for most of the users in private environments, where convenience is more valued than security. In corporate networks, policies are just enforced and users have to accept it. It would however be problematic if the credentials were only asked once at start-up or connecting to the network, since mobile devices are threatened by theft, loss, and other dangers by their nature of use.

Smart cards could be a solution to have a good trade-off between the usability and security. Since the user will have a token, which he/she has to care of, and exchange keys generated by it, at least it could be secured that the user who is accessing a specified service holds the authentication token. The mobile phone with the integrated smart card, the SIM, is a potential tool for this purpose. As indicated by Leu et al. (2006) the requirement of carrying a SIM reader or equipping all the equipment with SIM cards is neither convenient nor cost effective. The possibility of secure key exchange between user equipment shall be provided.

The cell phone can act as a key negotiator, with its tamper resistant cryptographic functions integrated into the SIM and then exchange the session keys with other terminals with the use of a short range wireless solution. Currently, most of the security problems, besides the user behaviour, are coming from security holes in the software. Having the capability to download new software over the air to the phone ensures the use of recent updates and eliminates this type of security threat (Kálmán & Noll, 2006). Compared to a security token, it may be better to use the phone, since the SIM card can be locked by the provider, so if the device gets lost, the authentication credentials can be withdrawn within short time.

OUTLOOK

Current research is focused on merging basic network functions to enable pervasive computing and network access. The result of these efforts is a converged infrastructure, which is able to handle

most of user needs in high quality. The problem of QoS control in wireless systems remains an open one, but experiences of VoIP and VoD services in wireless networks show the adaptability of the user to the current environment.

Mobility of packet data is still to be enhanced, with the challenge of reducing the handover delay. Remote access to home content is just beginning to be spread between early adopters. MIPv6 will address most of the issues sometime in the future, and with the promising extensions, the protocol will be able to handle sessions together with the AAA infrastructure without service interruption. Mobile networks will use WLAN as a high capacity data service, although upcoming solutions and MIPv6 extensions may be able to threaten their use inside dense populated areas, assuming global Wi-Fi roaming mechanisms are in place.

Efforts are being made towards an easy deployable home AAA infrastructure, which can later bear the tasks associated with inner (user management, remote access, user content DRM, purchased media DRM) and outer (authentication towards corporate, provider- or public-based AAA) authentication and access control.

Educating the user might be the biggest challenge, as mobile phone users represent the whole population, and not just the *educated* computer community. The enforcement of the use of smart cards is advisable, where the possible use of the mobile phone shall be investigated.

Now, we can experience the dawn of new social and community services over the Internet. This raises the problem of privacy protection as never before. AAA services must take care of user credentials, and even must ensure that data collected from different AAA providers cannot be merged. So, research in the area of one-way functions, blind signatures, and different PKI methods is recommended.

Finally, current market players also have to change their business plans. Research in the economical area has to point out new objectives to ensure a good working, open, and secure AAA infrastructure which can be used by every service provider while keeping information exchange on the required minimal level.

CONCLUSION

The biggest effort in AAA systems is on extending the capabilities of the existing solutions in telecommunication and in computer networks to an integrated network approach enabling seamless service access of mobile users.

While telecom solutions are usually more secure, user privacy is not a primary concern here. In computer networks AAA solutions are more open and flexible, while the widespread model of "web of trust" methods is not acceptable for commercial service exchange. Ongoing research indicates the potential for a common mobile/Internet authentication suite, potentially based on the EAP.

Interworking issues between mobile and wireless networks are the basis for detailed research on handover delay, multi-device roaming, mobile networks, security, ease-of-use, and anonymity of the user. This chapter provided an overview of the state of the art in authentication for mobile systems.

Extended AAA mechanisms are suggested for home and community networks, taking into account security and privacy of the users. These networks will keep a high amount of personal data, and thus need stronger privacy protection mechanisms. By using link layer encryption, smart cards, and secure key transfer methods the security and privacy protection can be greatly enhanced.

REFERENCES

Chowdhury, M. M. R., & Noll, J. (2007). Service interaction through role based identity. In *Proceedings of the The International Confer-*

ence on Wireless and Mobile Communications (ICWMC2007).

Fathi, H., Shin, S., Kobara, K., Chakraborty, S. S., Imai, H., & Prasad, R. (2006). LR-AKE-based AAA for network mobility (NEMO) over wireless links. *IEEE Journal on Selected Areas in Communications, 24*(9), 1725-1737.

Janevski, T., Tudzarov, A., Janevska, M., Stojanovski, P., Temkov, D., Kantardziev, D., et al. (2006). Unified billing system solution for interworking of mobile networks and wireless LANs. In *Proceedings of the IEEE Electrotechnical Conference MELECON 2006* (pp. 717-720).

Jeong, J., Chung, M. Y., & Choo, H. (2006). Secure user authentication mechanism in digital home network environments. In *Embedded and Ubiquitous Computing* (LNCS 4096).

Jeong, K. C., Lee, T.-J., Lee, S., & Choo, H. (2006). Route optimization with AAA in network mobility. In *Computational Science and Its Applications—ICCSA 2006* (LNCS 3981).

Kálmán, Gy., Chowdhury, M. M. R., & Noll, J. (2007). Security for ambient wireless services. In *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC2007).*

Kálmán, Gy., & Noll, J. (2006). SIM as a key of user identification: Enabling seamless user identity management in communication networks. In *Proceedings of the WWRF meeting #17.*

Khara, S., Mistra, I. S., & Saha, D. (2006). An alternative architecture for WLAN/GPRS integration. In *Proceedings of the IEEE Vehicular Technology Conference, 2006, VTC 2006* (pp. 37-41).

Køien, G. M. (in press). Privacy enhanced mobile authentication. *Wireless Personal Communications.*

Lee, C.-C., Hwang, M.-S., & Liao, I.-E. (2006). Security enhancement on a new authentication scheme with anonymity for wireless environments. *IEEE Transactions on Industrial Electronics, 53*(5), 1683-1687.

Lee, M., Park, S., & Jun, S. (2006). A security management framework with roaming coordinator for pervasive services. In *Autonomic and Trusted Computing* (LNCS 4158).

Lee, S.-Y., Huh, E.-N., Kim, Y.-W., & Lee, K. (2006). An efficient authentication mechanism for fast mobility service in MIPv6. In *Computational Science and Its Applications—ICCSA 2006* (LNCS 3981).

Leu, J.-S., Lai, R.-H., Lin, H.-I., & Shih, W.-K. (2006). Running cellular/PWLAN services: Practical considerations for cellular/PWLAN architecture supporting interoperator roaming. *IEEE Communications Magazine, 44*(2), 73-84.

Li, J., Ye, X.-M., & Tian, Y. (2006). Topologically-aware AAA overlay network in mobile IPv6 environment. In *Networking 2006* (LNCS 3976).

Long, M., & Wu, C.-H. (2006). Energy-efficient and intrusion-resilient authentication for ubiquitous access to factory floor information. *IEEE Transactions on Industrial Informatics, 2*(1), 40-47.

Noll, J., Lopez Calvet, J. C., & Myksvoll, K. (2006). Admittance services through mobile phone short messages. In *Proceedings of the International Conference on Wireless and Mobile Communications ICWMC'06.*

Popescu, B. C., Crispo, B., Tanenbaum, A. S., & Kamperman, F. L. A. J. (2004). A DRM security architecture for home networks. In *Proceedings of the 4th ACM Workshop on Digital Rights Management.*

Potlapally, N. R., Ravi, S., Raghunathan, A., & Jha, N. K. (2006). A study of the energy consumption characteristics of cryptographic algorithms and security protocols. *IEEE Transactions on Mobile Computing, 5*(2), 128-143.

Ren, K., Lou, W., Kim, K., & Deng, R. (2006). A novel privacy preserving authentication and access control scheme for pervasive computing environments. *IEEE Transactions on Vehicular Technology, 55*(4), 1373-1384.

Ryu, S., & Mun, Y. (2006). An optimized scheme for mobile IPv6 handover between domains based on AAA. In *Embedded and Ubiquitous Computing* (LNCS 4096).

Sur, C., & Rhee, K.-H. (2006). An efficient authentication and simplified certificate status management for personal area networks. In *Management of Convergence Networks and Services* (LNCS 4238).

Sur, C., Yang, J.-P., & Rhee, K.-H. (2006). A new efficient protocol for authentication and certificate status management in personal area networks. In *Computer and Information Sciences—ISCIS 2006* (LNCS 4263).

Zhang, Y., & Fujise, M. (2006). An improvement for authentication protocol in third-generation wireless networks. *IEEE Transactions on Wireless Communications, 5*(9), 2348-2352.

KEY TERMS

Authentication, Authorisation, and Accounting (AAA): AAA is a system that handles all users of the system to ensure appropriate right management and billing.

Converged Network: Converged network is a network carrying various types of traffic. Such a network is providing services to different terminals, which can access and exchange content regardless of the current networking technology they are using.

Diameter: Diameter is a proposed successor of RADIUS. It uses TCP as a transport method and provides the possibility to secure transmissions with TLS. It is not backward compatible with RADIUS.

Digital Rights Management (DRM): DRM is a software solution that gives the power for the content creator to keep control over use and redistribution of the material. Used mostly in connection with digital media provider companies, but in pervasive environments, users may also require a way to have a fine-grained security infrastructure in order to control access to own content.

Extensible Authentication Protocol (EAP): EAP, a flexible protocol family, which includes TLS, IKE protocols, and also the default authentication method of UMTS, EAP-AKA.

International Mobile Subscriber Identity (IMSI), Temporary-IMSI (TMSI): IMSI and TIMSI is the unique identity number used in UMTS to indentify a subscriber. The temporary one is renewed from time to time, and that is the only one that is used over the air interface.

Public Key Infrastructure (PKI): PKI is a service that acts as a trusted third party, manages public keys, and binds users to a public key.

Remote Authentication Dial in User Service (RADIUS): RADIUS is the de facto remote authentication standard over the Internet. It uses UDP as a transport method and is supported by software and hardware manufacturers. Privacy problems may arise when used on wireless links, since only the user password is protected by an MD5 hash.

Rivest-Shamir-Adleman (RSA): RSA is the de facto standard of public key encryption.

Smart Card: Smart card is a tamper resistant pocket sized card, which contains tamper resistant non-volatile storage and security logic.

Subscriber Identity Module (SIM): SIM is the smart card used in GSM and UMTS (as USIM) networks to identify the subscribers. It has integrated secure storage and cryptographic functions.

This work was previously published in Handbook of Research on Wireless Security, edited by Y. Zhang, J. Zheng, and M. Ma, pp. 176-188, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.17
Antecedents of Consumer Trust in B2C Electronic Commerce and Mobile Commerce

Dan J. Kim
University of Houston Clear Lake, USA

ABSTRACT

Despite the importance of trust in electronic commerce including mobile commerce, there is insufficient theory and model concerning the determinants of consumer trust in business-to-consumer electronic commerce. Thus, the purpose of this chapter is to (1) identify the major antecedents of a consumer's trust in electronic commerce and mobile commerce contexts through a large-scale literature review, (2) develop an integrative trust antecedent reference model summarizing the antecedents of consumer trust, and (3) discuss six categories of mobile applications as future trends of technologies and key issues related to consumer trust area in electronic commerce. In addition, to provide the validity of the proposed reference model, this chapter also proposes a research model derived from the reference model and discusses the constructs of the proposed model in detail. The chapter concludes that building trust is not simply an issue related to consumer-technology-buyer, but it is a complex issue that involves the interactions of key elements (buyer, seller, third-party, technology, and market environment) at least.

INTRODUCTION

Trust is important in exchange relations because it is a key element of social capital (Mayer, Davis, & Schoorman, 1995), and is related to firm performance, satisfaction, competitive advantage, and other favorable economic outcomes. Trust is identified as an important factor in several literatures, including marketing, behavioral science, and electronic commerce (Beatty, Mayer, Coleman, Reynolds, & Lee, 1996; Czepiel, 1990; Dirks & Ferrin, 2001, 2002; Hoffman, Novak, & Peralta, 1999; Jarvenpaa, Knoll, & Leidner, 1998; Kramer, 1999). According to the study conducted by Urban, Sultan, and Qualls (2000), consumers make electronic commerce (e-commerce) transac-

tion decisions based on trust. Therefore, lack of trust is one of the most frequently cited reasons for online consumers not engaging in exchange relationships with Internet vendors in e-commerce (Lee & Turban, 2001).

Mobile commerce (m-commerce) extends current e-commerce channels into more convenient "anytime, anyplace, and personalized" environment. As an emerging subset of e-commerce, m-commerce faces the same problems troubling e-commerce plus a few of its own due to the limitations of mobile technology (Siau & Shen, 2003). The limitations include restricted computation powers, memory, small screens, low-resolution displays, tiny multifunction keypads, battery life, unfriendly user interface for mobile devices, low bandwidth, unstable network connection, relatively high usage cost, and vulnerability of wireless data transmission. Therefore, building consumer trust in m-commerce is a particularly intimidating task due to the unique limitations of mobile technology.

Since consumer trust plays an essential role in online transactions, it is important to identify antecedents that affect a consumer's trust in e-commerce and m-commerce areas. Several researchers and professionals (Ba, Whinston, & Zhang, 1999; Beatty et al., 1996; Brynjolfsson & Smith, 2000; Czepiel, 1990; Hoffman et al., 1999; Jarvenpaa et al., 1998; Ratnasingham, 1998; Urban et al., 2000) have focused on various issues of trust in e-commerce. Even so, some scholars (Ratnasingham, 1998) have argued that the study of trust has been problematic for several reasons. These include problems with the definition of trust, confusion between trust and its antecedents, difficulties of observing and measuring trust, the tendency of particular disciplines to provide only partial descriptions of trust antecedents, and a lack of specificity about who the parties are (e.g., trustor and trustee) in research contexts in which trust is relevant (Mayer et al., 1995).

This chapter attempts to consider some of the above issues. First, we identify the major antecedents of a consumer's trust in electronic commerce and mobile commerce contexts through a large-scale literature review, second, develop an integrative trust antecedent reference model summarizing the antecedents of consumer trust, and finally discuss six categories of mobile applications as future trends of technologies and key issues related to consumer trust area in electronic commerce. In addition, this study also proposes a theoretical research model derived from the integrative trust antecedent reference model and discusses the constructs of the proposed model in detail to provide the validity of the reference model.

BACKGROUND: ANTECEDENTS OF TRUST

Trust Antecedents in E-Commerce Studies

Several researchers have tried to categorize antecedents or factors of a consumer trust (Barney & Hansen, 1994; Doney & Cannon, 1997; McKnight, Choudhury, & Kacmar, 2002b; Walczuch, Seelen, & Lundgren, 2001; Zucker, 1986). Zucker (1986) proposed three major ways to build trust: (1) process-based (e.g., reputation, experience), (2) characteristic-based (e.g., disposition), and (3) institutional-based (e.g., third-party certification). Mayer et al. (1995) defined trust as a behavioral intention based upon the expectations of another person. Based on this definition, they proposed a model of dyadic trust in organizational relationships that includes the characteristics of both the trustor and trustee that influence the formation of trust. The three characteristics included in the model, representing the perceived trustworthiness of the trustee, are benevolence, integrity, and ability. Doney and Cannon (1997) developed five distinct trust building processes in business relationships: (1) calculative process (trustor calculates the costs and/or rewards of

a target acting), (2) prediction process (trustor develops confidence that target's behavior can be predicted), (3) capability process (trustor assesses the target's ability to fulfill its promises), (4) intentionality process (trustor evaluates the target's motivations), and (5) transference process (trustor draws on proof sources from which trust is transferred to the target). They also categorized characteristics of supplier firm, salesperson, and the relationship into four types. Barney and Hansen (1994) and Lewis and Weigett (1985) defined the three levels of customer trust: (1) strong trust, (2) semistrong trust, (3) weak trust. Bhattacherjee (2002) proposed three key dimensions of trust: (1) trustee's ability, (2) benevolence, and (3) integrity, based on cross-disciplinary literature review on dimensions of trust. Recently, Kim, et al. (2005) identified four different entities of e-commerce market structure: consumer, seller, third party, and technology. Based on the four entities, they investigated the determinants of online trust and divide the determinants into six dimensions: consumer-behavioral, institutional, information content, product, transaction, and technology dimension.

Trust and National Culture

National culture also influences individual and organizational trust development processes (Doney, Cannon, & Mullen, 1998). Hofstede (1991, 1994) revealed the five cultural dimensions: individualism/collectivism, uncertainty avoidance, power distance, masculinity/femininity, and long/short term orientation on life. *Individualism* refers to the degree the society reinforces individual or collective achievement and interpersonal relationships; *uncertainty avoidance* refers to the degree of tolerance for uncertainty and ambiguity within the society—that is, unstructured situations; *power distance* refers to the degree of equality, or inequality, between people in the country's society; *masculinity* refers to the degree the society reinforces, or does not reinforce, the traditional masculine work role model of male achievement, control, and power; and *long/short term orientation of life* refers to the degree the society embraces, or does not embrace, long-term devotion to traditional, and forward thinking values (Hofstede, 1980, 1991, 1994).

Based on Hofetede's framework and using individualism/collectivism and power distance as independent variables, Strong and Weber (1998) examined the theory that trust is culturally determined in organization's contexts. They concluded that differences in trust exist globally between cultures. Griffith, Hu, and Ryans (2000) designated the United States and Canada as *Type I culture* with an "individualistic-small power distance-weak uncertainty avoidance" type of culture to contrast with *Type II culture* countries (Chile and Mexico) with "collectivistic-large power distance-strong uncertainty avoidance" characteristics. Although no significant difference in the strength of the trust-commitment relationship was found between Type I and Type II cultures, the study discovered that Type I cultures have a higher possibility of forming a trusting relationship with other Type I cultures, rather than with Type II cultures.

Several cultural studies (Mayer & Tan, 2002; Park & Jun, 2003; Png, Tan, & Wee, 2001; Soh, Kien, & Tay-Yap, 2000; Tan, Wei, Watson, Clapper, & McLean, 1998; Tan, Wei, Watson, & Walczuch, 1998) have shown that the dimensions of national culture affect the development, adoption, and impact of information communication technology (ICT) infrastructure and its applications in the field of information systems. However, only a handful of studies (Gefen & Heart, 2006; Jarvenpaa, Tractinsky, Saarinen, & Vitale, 1999; Lim, Leung, Sia, & Lee, 2004; Pavlou & Chai, 2002) to date have aimed at the effect of national culture on trust in computer-mediated electronic commerce transactions.

Jarvenpaa et al. (1999) used Hofstede's dimensions to compare Internet trust in individualistic and collectivistic cultures to conduct a study on a

cross-cultural validation of an Internet consumer trust model. They found that consumers in different cultures may have differing expectations of what makes a Web merchant trustworthy. Although no strong cultural effects were found regarding the antecedents of trust, their study ignited examinations of cultural differences in the antecedents of trust and the levels of trust in the context of e-commerce. Incorporating Hofstede's three cultural dimensions (i.e., individualism/collectivism, power distance, and long-term orientation) along with the theory of planned behavior, Pavlou and Chai (2002) conducted an empirical study to explain e-commerce adoption across cultures using data from consumers in the United States and China. The results of the study support the theory that cultural differences play a significant role in consumers' e-commerce adoption. Lim et al. (2004) identified two national culture dimensions (i.e., individualism-collectivism and uncertainty avoidance) and their interaction that influences Internet shopping rates across countries. They also found that trust mediates the relationship between cultural differences and Internet shopping adoption decisions. Cross-validating the scale of trust and its antecedents in both the U.S. and Israel, a cross cultural study by Gefen and Heart (2006) found that trust beliefs may be a relatively unvarying aspect of e-commerce but the effects of predictability and familiarity on trust beliefs may differ across national cultures.

Trust Antecedent in M-Commerce Studies

Mobile commerce is defined as business activities and processes related to an e-commerce transaction conducted through wireless communications networks that interface with mobile devices (Tarasewich, Nickerson, & Warkentin, 2002). Several studies (Anckar & D'Incau, 2002; Booz, 2000; Kannan, Chang, & Whinston, 2001; Malhotra & Segars, 2005; Siau, Lim, & Shen, 2001) identified the following distinctive mobile capabilities or values which drive one of the most promising innovative application services in near future: ubiquity, time-criticality, spontaneity/immediacy, constancy, convenience, personalization, location discovery, and so forth.

Ubiquity is the ability to allow mobile users to obtain information and conduct mobile transactions any place through Internet-enabled mobile devices. *Time-criticality* refers to the ability to access time-sensitive information immediately (Malhotra & Segars, 2005; Sadeh, 2002). A similar value to time-criticality, *spontaneity/immediacy* refers to the mobile capability for mobile users to get information and complete transactions in real-time. *Constancy* refers to the accessibility to network applications anytime and anywhere (Baldi & Thaung, 2002; Clarke, 2001; Malhotra & Segars, 2005). The constancy feature of mobile service provides the mobile value related to *convenience*. Since mobile devices are personal devices, they contain individual information as well as personal preferences. Thus, *personalization* refers to the ability to customize content and uses of mobile devices (Sadeh, 2002). Another mobile value is *location discovery* which allows mobile service providers to do location-based marketing and to deliver promotional offerings based on a user's current geographic position (Clarke, 2001). Since mobile devices are always on and carry user identity, the location of the mobile user can be tracked (Baldi & Thaung, 2002; Kannan et al., 2001; Malhotra & Segars, 2005).

Studies on trust in m-commerce are scarce due to the novelty of mobile commerce area. Siau and Shen (2003) developed a framework for building customer trust in mobile commerce. They identified two components of customer trust in mobile commerce: (1) mobile technology and (2) mobile vendor. Another study of trust in m-commerce conducted by Siau, Sheng, and Nah (2003) proposed a framework for trust in mobile commerce which outlines the variables influencing trust

Table 1. Selected studies of antecedents/processes of trust in e-commerce and m-commerce

Study topic and author(s)	Category of Antecedents	Sub categories or Set of Antecedents
Three levels of customer trust (Barney & Hansen, 1994; Lewis & Weigert, 1985)	Strong trust	Interactions, cognitive trust (e.g. the similarity), emotional trust
	Semi-strong trust	Rational-calculation-based trust (e.g. a company's reputation, the threat of punishment)
	Weak trust	Transferred trust (e.g. a well developed market, or word-of-mouth)
Three central modes of trust production (Zucker, 1986)	Process-based	Reputation, brands, gift-giving
	Characteristic-based	Family background, ethnicity, sex
	Institutional-based	Professional, firm associations, bureaucracy, banks, regulation
Three dimensional generic typology of trust (Mayer et al., 1995)	Ability	Competency, experience, institutional endorsements, knowledgeability
	Integrity	Fairness, fulfillment, loyalty, honestly, dependability, reliability,
	Benevolence	Concern, empathy, faith, receptivity
Five distinct trust building processes (Doney & Cannon, 1997)	Calculative process	Firm's reputation, size, willingness to customize, confidential information sharing, length of relationship with firm, length of relationship with salesperson
	Prediction process	Length of relationship with firm, salesperson likeability, salesperson similarity, frequent social contact with salesperson, frequent business contact with salesperson, length of relationship with salesperson
	Capability	Salesperson expertise, salesperson power
	Intentionality	Firm's willingness to customize, firm's confidential information sharing, salesperson likeability, salesperson similarity, frequent social contact with salesperson
	Transference	Firm's reputation, supplier firm size, trust of supplier firm, trust of salesperson
Trust of a supplier firm and salesperson (Doney & Cannon, 1997)	Characteristics of the supplier firm and firm relationship	Reputation, size, willingness to customize, confidential information sharing, length of relationship
	Characteristics of the salesperson and salesperson relationship	Expertise, power, likeability, similarity, frequent business contact, frequent social contact, length of relationship
A trust model for consumer internet shopping (Lee & Turban, 2001)	Trustworthiness of Internet merchant	Ability, integrity, benevolence
	Trustworthiness of Internet shopping medium	Technical competence, reliability, medium understanding
	Context factors	Effectiveness of third party certification, effectiveness of security infrastructure
	Other factors	Individual trust propensity, etc

Table 1. continued

Study topic and author(s)	Category of Antecedents	Sub categories or Set of Antecedents
An integrative typology of trust (McKnight, Choudhury, & Kacmar, 2002a)	Disposition to trust	Faith in humanity, trusting stance
	Institution-based trust	Situational normality, general competence, integrity, benevolence, structural assurance
	Trusting beliefs	Competence beliefs, benevolence beliefs, and integrity beliefs
	Trusting intentions	Willingness to depend, subjective probability of depending
Online trust: a stakeholder perspective (Shankar, Urban, & Sultan, 2002)	Website characteristics	Navigation, user friendliness, advice, error free
	User characteristics	Internet savvy, past Internet shopping behavior, feeling or control
	Other characteristics	Online medium, trustworthiness of firm, perceived size of firm
Psychological antecedents of consumer trust (Walczuch & Lundgren, 2004)	Personality-based	Extraversion, neuroticism, agreeableness, conscientiousness, openness to experience, propensity to trust
	Perception-based factors	Perceived reputation (e.g., word-of-mouth), perceived investment, perceived similarity, perceived normality, perceived control, perceived familiarity
	Experience-based	Experience over time, satisfaction, communication
	Knowledge-based factors	Information practices, security technology
	Attitude	Computers & the internet, Shopping
Process-oriented Multi-dimensional Trust Formation (Kim et al., 2005)	Consumer-Behavioral Dimension	Demographic factors, experience, familiarity, individual culture, traditions, privacy, etc.
	Institutional Dimension	Reputation, accreditation, authentication, approvals (e.g., advisors and guarantors), customer communities (e.g. eBay's feedback forum), legal requirements and authorities, etc.
	Information Content Dimension	Accuracy, currency, completeness, non-bias, credibility, website brand royalty, entertainment, usefulness, etc.
	Product Dimension	Durability, reliability, brand equity, quality, variety, customization, competitiveness and availability, etc.
	Transaction Dimension	Transparency, pricing and payment options, financial planning (complexity), sales-related service (refund policy, after-sales, etc.), promotions, delivery fulfillment, etc.
	Technology Dimension	Quality of media transmission, interface design and contents, security, reversibility, digital certificate, public-key cryptography (infrastructure), authenticity, integrity, confidentiality, non-repudiation, attributes of the system (benevolence, competency, predictability), etc.

Table 1. continued

Study topic and author(s)	Category of Antecedents	Sub categories or Set of Antecedents
A Trust-based Consumer Decision Making (Kim, Ferrin, & Rao, (Forthcomming))	Cognition (observation)-based	Privacy protection, security protection, system reliability, information quality, etc.
	Affect-based	Reputation, presence of third-party seals, referral, recommendation, buyers' feedback, word-of-mouth, etc.
	Experience-based	Familiarity, Internet experience, e-commerce experience, etc.
	Personality-oriented:	Disposition to trust, shopping style, etc.
Framework of trust-inducing features (Wang & Emurian, 2005)	Graphic design	Use of three-dimensional, dynamic, and half-screen size clipart, symmetric use of moderate pastel color of low brightness and cool tone, use of well-chosen, good-shot photographs
	Structure design	Easy-to-use navigation, accessible information, navigation reinforcement, application of page design techniques
	Content design	Brand-promoting information, disclosure of all aspects of the customer relationship, seals of approval or third-party certificate, use of comprehensive, correct, and current product information, use of a relevant domain name
	Social-cue design	Inclusion of representative photograph or video clip, use of synchronous communication media
Customer Trust in Mobile Commerce (Siau & Shen, 2003)	Mobile Technology	Initiate trust formation (feasibility) Continuous trust development (reliability, consistency)
	Mobile Vendor	Initiate trust formation (familiarity, reputation, information quality, third-party recognition, attractive reward, Continuous trust development (site quality, competence, integrity, privacy policy, security controls, open communication, community building, external auditing)
Trust in mobile commerce (Siau et al., 2003)	Vendor Characteristics	Reputation, brand reputation, availability, privacy policy
	Website Characteristics	Website design, ease of input and navigation, readability, accuracy, richness
	Technology of wireless services	Connection speed, coverage area, transaction data, authentication
	Technology of mobile services	User interface, ease of input and navigation, readability
	Other factors	Third-party regulation, word-of-mouth

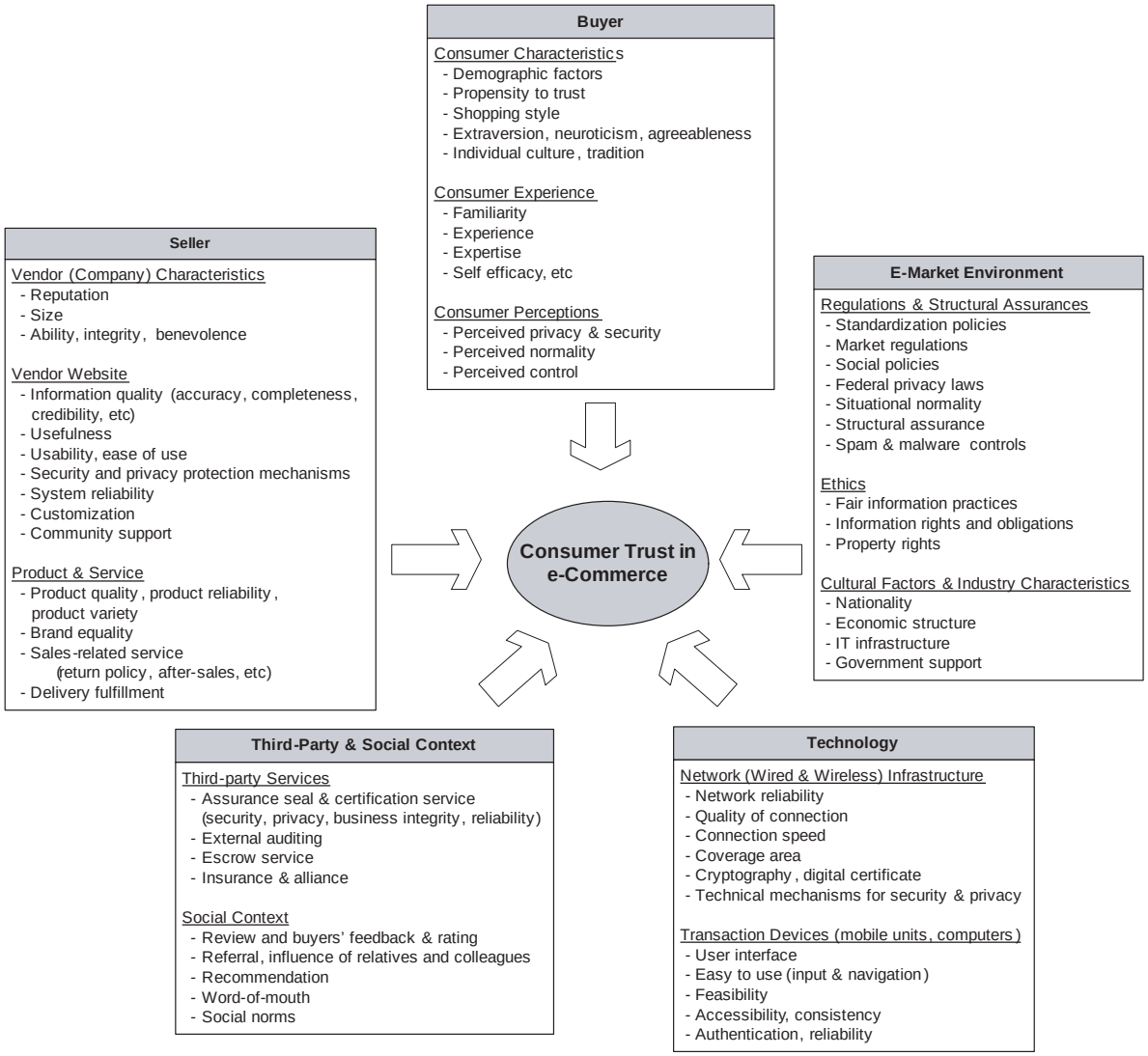
building in mobile commerce. Table 1 provides a summary of selected studies of antecedents/processes of trust in e-commerce.

AN INTEGRATIVE TRUST ANTECEDENT REFERENCE MODEL

The literature review depicts that various factors and entities influence the complex process

of engendering customer trust in e-commerce. A process- oriented, multi-dimensional trust formation model recently proposed by Kim et al. (2005) is well reflected in the actual online exchange process. The model consists of six dimensions of trust formation process and four different entities representing three ingredients of e-commerce transactions: trustor (buyer), trustee (seller), and environment (third party and technology). Although the model describes a holistic, multi-dimensional trust formation processes in a succinct manner the phenomena of trust formation in e-commerce transaction, it does not capture some environmental factors which influence trust

Figure 1. An integrative trust antecedent reference model



formation process such as cultural factors, national industry characteristics, market regulations, ethics, social context, and so forth. Therefore, along with the four entities of e-commerce markets suggested by Kim et al. (2005), I suggest five entities of e-commerce markets, to include buyer, seller, third-party and social context, technology, and market environment factors. Finally, after reclassifying and reorganizing determinants of trust in e-commerce and m-commerce areas, an integrative trust antecedent reference model (see Figure 1) is proposed in an effort to synthesize existing literature on enhancing consumer trust in e-commerce and m-commerce.

The integrative trust antecedent reference model shows that cultivating consumer trust involves the interactions of five entities at least. A *buyer* (i.e., trustor) has several subdimensional factors influencing his or her trust belief such as personal characteristics (e.g., propensity to trust, individual culture, demographic elements, and so on), individual experiences (e.g., familiarity, Web experience, self-efficacy, and so on), and individual perceptions (e.g., perceived privacy, perceived security, perceived normality, and so on). As a trustee, a seller also possesses several sub-dimensional factors.

Plank, Reid, and Pullins (1999) suggested a definition of trust toward multiple objects: salesperson, product, and company. According to their definition of trust, trust is a global belief on the part of the buyer that the salesperson, product, and company will fulfill their obligations as understood by the buyer. In e-commerce context, a seller could be multiple objects: Web site, product, and company. Thus, three subdimensional factors of an e-commerce *seller* (i.e., trustee) are vendor (company) characteristics (e.g., size, reputation, ability, integrity, and benevolence), Web site elements (e.g., information quality, usefulness, usability, system reliability, and so on), and product service factors (e.g., product quality, product reliability, product variety, after-sales service, delivery fulfillment, and so on). *Third-party and social context* are important entities in e-commerce transactions. Third parties are impartial organizations which include individual mechanisms delivering business confidence through an electronic transaction (Kim et al., 2005). Social contexts are about how the trustee is viewed by the people around. Third-party services include assurance seals and business certification services, escrow service, and so on. Examples of social context are buyers' reviews and feedbacks, referral, word-of-mouth, and so forth.

Technology is the major entity which makes a difference between e-commerce and traditional brick-and-mortar transactions because all e-commerce transactions take place primarily through wired and/or wireless network infrastructure. Network infrastructure and end-unit devices for electronic transactions are identified as subdimensional factors. Network reliability, connection quality, speed, and coverage area for wireless networks and user interface, easy to use, and reliability for mobile units are specifically important. Although Web site characteristics could be classified as technology subdimensions, they are arranged as a seller side component because a Web site is a seller's storefront. Finally, electronic market (e-market) environmental factors are another important entities influencing consumer trust in e-commerce. *E-market environment* has several subdimensional factors that include regulations and structural assurances (e.g., standardization policies, market regulations, structural assurances, and so forth), ethics (e.g., fair information practices, information and property rights, and so forth), and national culture and industry characteristics (e.g., nationality, economical structure, government support, and so forth).

FUTURE TRENDS AND KEY ISSUES RELATED TO CONSUMER TRUST

The exponential growth of wired broadband and wireless mobile networks will be expected

to drive the future development of e-commerce and provide new opportunities in m-commerce beyond e-commerce (Maamar, 2003). Enhancing the current e-commerce applications and business models in the market, there are six categories of mobile applications which utilizing the major unique features of mobile technology (i.e., anytime, anywhere, and personalized service).

The six categories of mobile applications are: (1) commerce transaction applications (e.g., mobile-shopping, micro-payments, bill payment, mobile banking, mobile trading, hotel reservation, and so forth), (2) communication applications (e.g., e-mail, char/SMS, multi-media SMS, mobile conferencing, broadcast, news flash, and so forth), (3) content delivery applications (e.g., information browsing, and directory service, interactive online gaming, music/video/game downloading, off-line games, flight schedules, weather information, and so forth),(4) community applications (finding buddies, mobile blog, dating, mobile community for referral and recommendation, and so forth), (5) customization (e.g., scheduling, location based services, personal dieting, information filtering, and so forth), (6) connection (e.g., mobile tracking, mobile inventory management, geographic positioning systems, and so forth).

While there are many potential advantages of the new "niche" technology, there are many problems and issues as well. Using the five entities of the integrative trust antecedent reference model, some key challenges are identified in e-commerce and m-commerce areas (Cavoukian & Gurski, 2002; Maamar, 2003; Yeo & Huang, 2003).

1. Issues related to trustors (buyers)

- User comfort level of e-commerce transaction
- Privacy and security issues because of tracking and location based service
- Restricted data collection and control of personal information
- Individual culture
- Experiences
- Self-efficacy
- Different perceptions

2. Issues related to trustees (sellers)

- Pricing issue
- Marketing issue
- Consumer retention issue
- Fulfillment issue
- Customization and advertising issues
- Web interface development issue for mobile devices
- Information quality
- Application development issue

3. Issues related to third-party and social context

- Effectiveness of third-party assurance services
- Fair feedback and rating systems
- Open community in e-commerce and m-commerce areas
- Social influence

4. Issues related to technology

- Wired and wireless technology infrastructure
  - Global standardizations of new technologies
  - The lack of network security
  - Slow bandwidth and efficient use of limited bandwidth
  - Strong encryption technology
  - Open source technology
  - Mobile payment issues
  - Virus and malware (spyware, adware, phishing, and hacking) control issues
- Transaction device technology
  - Small display screen

- Comfortable user interface
- Open platform for wireless devices
- Computational power—hardware and software

5. Issues related e-market environment

- Cultural issues
- Market regulations and social polices
- International and inter-states taxation
- Information and property rights
- Digital dividend
- Government regulation and support issues

Supplemental Study

In order to provide the validity of the proposed integrative trust antecedent reference model, a research model titled "Antecedents of Consumer Trust in B2C E-Commerce" is developed. The research constructs of the model are discussed in detail below.

A RESEARCH MODEL: ANTECEDENTS OF CONSUMER TRUST IN B2C E-COMMERCE

In traditional commerce, trust is affected by the characteristics of customers and the selling party (salespersons and company) and interactions between the two parties involved (Burt & Knez, 1996; Doney & Cannon, 1997; Shapiro, Sheppard, & Cheraskin, 1992; Swan, Bowers, & Richardson, 1999). It is also true in electronic commerce. Therefore, drawing from a part of the integrative trust antecedent model, three categories of antecedents influencing a consumer's trust toward an electronic commerce vendor are selected. The three categories and some trust antecedents from previous studies are summarized as follow:

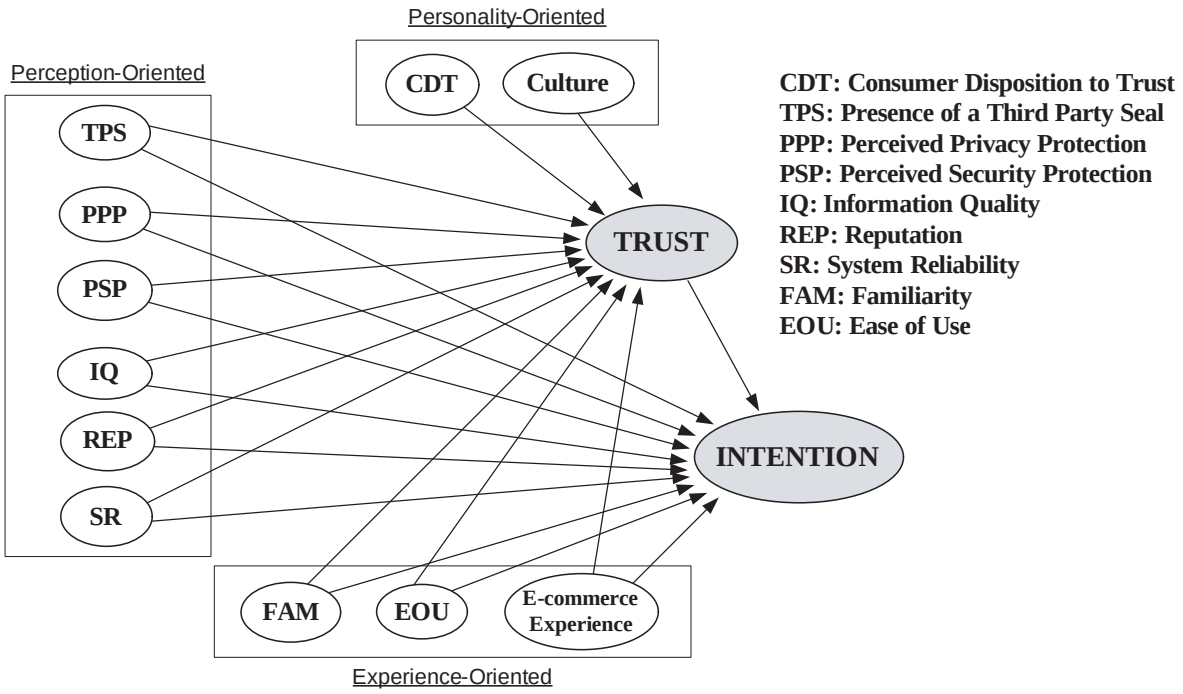
Consumer Side Antecedents

1. **Consumer personality-oriented:** Disposition to trust, shopping style, culture, and so forth
2. **Consumer experience-oriented:** Familiarity, ease of use, Internet experience, e-commerce experience, satisfaction, and so forth
3. **Consumer perception (observation) toward e-commerce vendor Web site:** Presence of third party assurance services, privacy protection, security protection, information quality, system reliability, and so forth.

The personality-oriented and experience-oriented antecedents are related to the characteristics of consumers, which are not easy to improve and manage by selling party perspectives. The perception-oriented antecedents are associated with salespersons (Web sites), company (brand image), and interactions (interface) between the two parties. In light of the difficulty of controlling all antecedents at the same time, this study proposes a research model mainly focusing on the perception-oriented antecedents with some personality and experience-oriented antecedents. Consumer disposition to trust, culture, familiarity with a selling party, ease-of-use, and Internet experience are included in the research model because some studies have shown evidence that they are strong antecedents of consumer trust (Gefen, 2000; Luhmann, 1979; Mayer et al., 1995; Rotter, 1971).

Even though we are interested in the antecedents of trust, there is concern that some antecedents of trust may have a direct effect on purchase intention (McKnight & Chervany, 2002; McKnight, Cummings, & Chervany, 1998). Therefore it is necessary at least to propose the direct effects from antecedents to a consumer's

Figure 2. Research model: Antecedents of consumer trust in B2C e-commerce



purchase intention. Figure 2 shows the research model including direct paths from antecedents to trust and intention, and the description of each construct and their relationships with trust are following.

An online *consumer trust (TRUST)* is defined as a consumer's subjective belief that the selling party or entity will fulfill its transactional obligations as the consumer understands them and as such transactions are enabled by electronic processes. Trust plays a vital role in almost any commerce involving monetary transactions (Gefen, 2002; Jarvenpaa et al., 1999; Urban et al., 2000). Internet business is much more based on the consumer's trust in the processes, in contrast to that of traditional business involving brick and mortar stores, where trust is based on face-to-face personal relationships. Peter Grabosky, in *The Nature of Trust Online*, supports the idea that the key to success in Internet business is the establishment of trusted processes (Grabosky, 2001). This fact mandates that Internet sellers create an environment in which a prospective consumer can be relaxed and confident about any prospective transactions. Thus we propose that a consumer trust positively influences a consumer's purchase intention of electronic transaction.

Intention to purchase (INTENTION) refers to the degree to which a consumer intends to purchase from a certain vendor through the Web. The theory of reasoned action (TRA) presumes that volitional behavior is determined by inten-

tions to act. Ajzen and Fishbein (1980) point out that behavior intention (intention to purchase, in this study) is a predictor of actual behavior (purchase), and there is a strong correlation between behavioral intentions and actual behavior (Sheppard, Hartwick, & Warshaw, 1988; Venkatesh & Davis, 2000). Consumer's purchase intention is one of the interesting variables for most e-shopping vendors.

Consumer disposition to trust (CDT) refers to a customer's personality traits that lead to generalized expectations about trustworthiness, which is a consumer-specific antecedent of trust. Since consumers have different developmental experiences, personality types, and cultural backgrounds, they differ in their inherent propensity to trust (Gefen, 2000). This tendency is not based upon experience with or knowledge of a specific trusted party, but it is the result of ongoing lifelong experience and socialization (Kahneman, 2003; McKnight et al., 1998; Rotter, 1971). If a consumer has a high tendency to trust others in general, this disposition is especially influential when customers have not had an extensive personal interaction with the selling parties (McKnight et al., 1998; Rotter, 1971). Consumer disposition to trust is an antecedent of trust, but it is not directly related to a consumer behavior intention.

Culture is defined by Hofstede (1994) as "the collective programming of mind which distinguishes one national group or category of people from another" (p. 5). Several studies (Mayer & Tan, 2002; Png et al., 2001; Soh et al., 2000; Tan, Wei, Watson, Clapper, et al., 1998; Tan, Wei, Watson, & Walczuch, 1998) have shown that the dimensions of national culture affect development, adoption, and impact of information communication technology (ICT) infrastructure and its applications in the field of information systems. Even though culture is a crucial aspect of trust, it has been overlooked by previous e-commerce studies. Only a handful of studies (Gefen & Heart, 2006; Jarvenpaa et al., 1999; Lim et al., 2004; Pavlou & Chai, 2002) to date have aimed at the effect of culture on trust in computer-mediated electronic commerce transactions. Since e-commerce transactions are sometimes required international interactions, understanding the cross-national aspects (i.e., culture) of trust building is essential (Gefen & Heart, 2006).

Familiarity with the online selling party (FAM) is a consumer experience-oriented antecedent of trust, which refers to the degree of consumer's acquaintance with the selling party. Familiarity would include enough knowledge to search for products and information and to order through the Web site's purchasing interface. Familiarity is a "precondition or prerequisite of trust" (Luhmann, 1979), which is an antecedent of trust because familiarity leads to an understanding of the current actions while trust deals with beliefs about the future actions of other entities (Gefen, 2000). For example, a consumer's familiarity based on previous good experience with salesperson (i.e., Web site), their services (i.e., searching products and information, and so forth) let the consumer create concrete ideas of what to expect for the future. As in electronic commerce in general, the more customers are familiar with such a selling party, the more their favorable expectations (trust) are likely to have been confirmed. It is thus hypothesized that more familiarity with a selling party should affect customer's trust on the selling entity.

Ease of use (EOU) of a Web site primarily deals with ease of navigation, ease of searching for products and information, and ease of understanding content. These trappings, along with the user's movement throughout the site, are as integral to the overall user experience as the transaction the user wants to execute. Like the importance of user interface design for software development, the Internet Web site interface design has received enormous research attention, since poorly designed sites have an adverse influence on consumer's shopping behavior (Lohse & Spiller, 1998). We posit that ease of use increases a consumer's trust toward the selling party.

The relationship between *e-commerce experience* and trust is found to be strongly associated (Gefen, 2000). In the traditional "brick-and-mortar" business environment, trust is mainly build through repeated successful transaction experiences (Lunn & Suman, 2002). It could be true at the "brick-and-click" or "pure-click" business environments. Thus, a positive e-commerce transaction experience is an antecedent of consumer trust, which is also directly related to a consumer purchase intention.

The presence of a third party seal (TPS) refers to the assurance of Internet vendors by third party certifying bodies (e.g. banks, accountants, consumer unions, and computer companies). Recently, a wide variety of third party seals were introduced to help create trust in electronic commerce. The purpose of seals is to provide assurance to consumers that a Web site discloses and follows its operating practices, that it handles payments in a secure and reliable way, that it has certain return policies, or that it complies with a privacy policy that says what it can and cannot do with the collected personal data (Castelfranchi & Tan, 2001; Koreto, 1997; Shapiro, 1987). An example of the third party involved in the trust of online transactions is TRUSTe, a non-profit, privacy seal program. The TRUSTe trust mark on Web sites informs buyers that the owners have openly agreed to disclose their information gathering and dissemination practices, and that their disclosure is backed by credible third-party assurance (Benassi, 1999). The basic argument of the presence of a seal and consumer trust is that the seals on a vendor's site issued from certificate authorities may assure consumers that the site is a reliable and credible place to do business. Therefore, when Internet customers see the seal on a given site, it creates extra trust in that selling site.

Perceived privacy protection (PPP) refers to a consumer's perception of the likelihood or intention of Internet vendors to protect consumers' personal information, which is collected during electronic transactions, from unauthorized use or the disclosure of confidential information. At the time of a transaction, the online seller collects the names, e-mail addresses, phone numbers, and home addresses of buyers. Some sellers pass the information on to telemarketers. For many online consumers, loss of privacy is a main concern. In a recent survey, 92% of survey respondents indicated that they do not have confidence that companies will keep their information private, even when the companies promise to do so (Light, 2001). These increasing consumer concerns are forcing sellers to take privacy protection measures to increase their trustworthiness and thereby to encourage online transactions. Consumers often perceive that one of the obligations of a seller is that the seller should not share or distribute the buyer's private information. Since this is a perceived obligation of the seller under the contract, buyers will be more likely to trust a seller who they believe will protect personal privacy.

Perceived security protection (PSP) refers to a consumer's perception that the Internet vendor will fulfill security requirements, such as authentication, integrity, encryption, and nonrepudiation. How a consumer perceives security protection when making online transactions depends on how clearly she or he understands the level of security measures implemented by the seller (Friedman, 2000). When an ordinary consumer finds security features (e.g., a security policy, a security disclaim, encryption, a safe shopping guarantee, SSL technology, and so forth) in the seller's Web site, he or she can recognize the seller's intention to fulfill the security requirements during the online transactions. This positively affects the trustworthiness of the seller as far as security is concerned, and, thus the consumer feels comfortable completing the transaction.

Even the definition of information is a complex concept and quality of information may be interpreted in multiple ways (e.g., accuracy, relevance, timelines, reliability, sufficiency, and so forth), *information quality (IQ)* refers to a consumer's general perception of the accuracy and

completeness of Web site information as it relates to products and transactions. It is well recognized that information on the Internet varies a great deal in quality, ranging from highly accurate and reliable, to inaccurate and unreliable, to intentionally misleading. As well, it is often very difficult to tell how frequently the information in Web sites is updated and whether the facts have been checked or not (Pack, 1999). Thus, potential purchasers on the Internet are likely to be particularly attentive to the quality of information on a Web site because the quality of information should help them make good purchasing decisions. To the extent that consumers perceive that a Web site presents quality information, they are more likely to have confidence that the vendor is reliable, and therefore will perceive the vendor as trustworthy. As buyers perceive that the Web site presents quality information, they will perceive that the seller is interested in maintaining the accuracy and currency of information, and, therefore, will be more inclined to fulfill its obligations and be in a better position to fulfill its obligations.

Reputation of selling party (REP) refers to the degree of esteem in which public consumers hold a selling party. Positive reputation has been considered a key factor for creating trust in organizations by marketing (Doney & Cannon, 1997; Ganesan, 1994) and electronic commerce (Jarvenpaa et al., 1999). Reputation building is a social process dependent on the past interactions (e.g., whether that business partner was honest before) between consumers and selling party (Zacharia & Maes, 2000).

A positive reputation provides information that the selling party has honored or met its obligations toward consumers in the past, or, in the case of a negative reputation, that it has failed to honor or meet its obligations. Based on this reputation information, a consumer may infer that the selling party is likely to continue in its behavior. In the case of a positive reputation, one is likely to infer that the company will honor its specific obligations to oneself, and therefore conclude that the selling party is trustworthy. By the same reasoning, an individual may conclude that the selling party will not honor its specific obligations, and hence conclude that it is untrustworthy. A positive reputation generates a feeling of trust and willingness to engage in the transaction.

System reliability (SR) refers to the consumer's perception that a Web vendor system is always available and fast and makes few errors at all levels, that the transaction record is correct, and that services will not fail during a transaction. As a technical dimension to support electronic commerce, system reliability considers key factors such as the following: access is always fast and available, very few errors are allowed at all levels, the transaction record is correct and remains correct, and services do not fail during a transaction. For example, a site may not totally fail but site access may become so slow that sales may be lost. This is not a hard failure, but may be classified as a soft failure. Even under soft failure, consumer's trust regarding that site may be negatively affected.

CONCLUSION

Wired and wireless technologies bring together a broad range of evolution or revolution influencing today's business life. Many studies have indicated that trust is critical for the growth and success of e-commerce. Since we already have observed the negative consequences of a lack of confidence and trust on the growth of e-commerce, trust issues including security and privacy concerns must be addressed in the early stage of mobile commerce development. In the electronic business world, building trust is not simply an issue related to consumer-technology-buyer, but it is a complex issue that involves the key interactions of five elements (i.e., buyer, seller, third-party, technology, and market environment) at least.

ACKNOWLEDGMENT

This study is supported in part by the Faculty Research and Support Fund (FRSF) (Award #908) of the University of Houston Clear Lake.

REFERENCES

Ajzen, I., & Fishbein, M. (1980). *Understanding attitude and predicting social behavior.* Englewood Cliffs, NJ: Prentice-Hall.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. *The Journal of Information Technology Theory and Application (JITTA), 4*(1), 43-64.

Ba, S., Whinston, A. B., & Zhang, H. (1999). *Building trust in the electronic market through an economic incentive mechanism.* Paper presented at the 1999 International Conference on Information Systems.

Baldi, S., & Thaung, H. (2002). The Entertaining Way to M-Commerce: Japan's Approach to the Mobile Internet—A Model for Europe? *Electronic Markets, 12*(1).

Barney, J. B., & Hansen, M. H. (1994). Trustworthiness as a source of competitive advantage. *Strategic Management Journal, 15*, 175-190.

Beatty, S. E., Mayer, M., Coleman, J. E., Reynolds, K. E., & Lee, J. (1996). Customer-sales associate retail relationships. *Journal of Retailing, 72*(3), 223-247.

Benassi, P. (1999). TRUSTe: An online privacy seal program. *Communications of the ACM, 42*(2), 56-59.

Bhattacherjee, A. (2002). Individual trust in online firms: Scale development and initial test. *Journal of Management Information Systems, 19*(1), 213-243.

Booz, A. H. (2000). The wireless internet revolution [Electronic Version]. *Insights: Communications, Media & Technology Group, 6*(1). Retrieved from http://www.boozallen.com/media/file/34103.pdf

Brynjolfsson, E., & Smith, M. (2000). Frictionless commerce? A comparison of Internet and conventional retailers. *Management Science, 46*(4), 563-585.

Burt, R., & Knez, M. (1996). Trust and third-party gossip. In R. M. Kramer & T. R. T. (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 68-89). Thousand Oaks, CA: Sage Publications.

Castelfranchi, C., & Tan, Y.-H. (2001). *Trust and deception in virtual societies.* Norwell, MA: Kluwer Academic Publishers.

Cavoukian, A., & Gurski, M. (2002). Privacy in a wireless world. *Business Briefing: Wireless Technology.* Retrieved from http://www.ipc.on.ca

Clarke, I. (2001). Emerging value propositions for m-commerce. *Journal of Business Strategies, 18*(2), 133-148.

Czepiel, J. A. (1990). Service encounters and service relationships: Implications for research. *Journal of Business Research, 20*(1), 13-21.

Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *Journal of Applied Psychology, 87*(4), 611-628.

Dirks, K. T., & Ferrin, D. L. (2001). The role of trust in organizational settings. *Organization Science, 12*(4), 450-467.

Doney, P. M., & Cannon, J. P. (1997). An examination of the nature of trust in buyer-seller relationships. *Journal of Marketing, 61*(2), 35-51.

Doney, P. M., Cannon, J. P., & Mullen, M. R. (1998). Understanding the influence of national

culture on the development of trust. *Academy of Management Journal, 23*(3), 601-620.

Friedman, B. (2000). Trust online. *Communications of the ACM, 43*(12), 34-40.

Ganesan, S. (1994). Determinants of long-term orientation in buyer-seller relationships. *Journal of Marketing, 58*, 1-19.

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega: The International Journal of Management Science, 28*(5), 725-737.

Gefen, D. (2002). Reflections on the dimensions of trust and trustworthiness among online consumers. *ACM SIGMIS Database, 33*(3), 38-53.

Gefen, D., & Heart, T. (2006). On the need to include national culture as a central issue in e-commerce trust beliefs. *Journal of Global Information Management, 14*(4), 1-30.

Grabosky, P. (2001, April 23). The nature of trust online [Electronic Version]. *The Age*, pp. 1-12. Retrieved from http://www.aic.gov.au/publications/other/online_trust.html

Griffith, D. A., Hu, M. Y., & Ryans, J. K. (2000). Process standardization across intra- and intercultural relationships. *Journal of International Business Studies, 31*(2), 303-325.

Hoffman, D. L., Novak, T. P., & Peralta, M. (1999). Building consumer trust online. *Communications of the ACM, 42*(4), 80-85. Association for Computing Machinery.

Hofstede, G. (1980). Motivation, leadership, and organization: Do American theories apply abroad? *Organizational Dynamics, 9*(1), 42-63.

Hofstede, G. (1991). *Cultures and organizations: Software of the mind*. London: McGraw-Hill.

Hofstede, G. (1994). *Cultures and organizations: Software of the mind: Intercultural*. London: HarperCollins.

Jarvenpaa, S. L., Knoll, K., & Leidner, D. E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. *Journal of Management Information Systems, 14*(4), 29-64.

Jarvenpaa, S. L., Tractinsky, N., Saarinen, L., & Vitale, M. (1999). Consumer trust in an Internet store: A cross-cultural validation [Electronic Version]. *Journal of Computer Mediated Communications, 5*(2). Retrieved from http://jcmc.indiana.edu/vol5/issue2/jarvenpaa.html

Kahneman, D. (2003). Maps of Bounded Rationality, Psychology for Behavioral Economics. *American Economic Review, 93*(5), 1449-1475.

Kannan, P., Chang, A., & Whinston, A. (2001, January 3-6). *Wireless commerce: Marketing issues and possibilities*. Paper presented at the the 34th Annual Hawaii International Conference on System Sciences (HICSS-34).

Kim, D. J., Ferrin, D. L., & Rao, H. R. (Forthcoming). A trust-based consumer decision making model in electronic commerce: The role of trust, risk, and their antecedents. *Decision Support Systems*.

Kim, D. J., Song, Y. I., Braynov, S. B., & Rao, H. R. (2005). A multi-dimensional trust formation model in B-to-C e-commerce: A conceptual framework and content analyses of academia/practitioner perspective. *Decision Support Systems, 40*(2), 143-165.

Koreto, R. (1997). In CPAs we trust. *Journal of Accountancy, 184*(6), 62-64.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50*, 569-598.

Lee, M. K. O., & Turban, E. (2001). A trust model for consumer Internet shopping. *International Journal of Electronic Commerce, 6*(1), 75-91.

Lewis, J. D., & Weigert, A. (1985). Trust as social reality. *Social Forces, 63,* 967-985.

Light, D. A. (2001). Sure, you can trust us. *MIT Sloan Management Review, 43*(1), 17.

Lim, K. H., Leung, K., Sia, C. L., & Lee, M. K. (2004). Is eCommerce boundary-less? Effects of individualism-collectivism and uncertainty avoidance on internet shopping. *Journal of International Business Studies, 35,* 545-559.

Lohse, G. L., & Spiller, P. (1998). Electronic shopping. *Communications of the ACM, 41*(7), 81-87.

Luhmann, N. (1979). *Trust and power.* Chichester, UK: Wiley.

Lunn, R. J., & Suman, M. W. (2002). Experience and trust in online shopping In B. Wellman & C. A. Haythornthwaite (Eds.), *The Internet in everyday life* (pp. 549-577). Blackwell Publishing.

Maamar, Z. (2003). Commerce, e-commerce, and m-commerce: What comes next? *Communications of the ACM, 46*(12), 251-257.

Malhotra, A., & Segars, A. H. (2005). Investigating wireless Web adoption patterns in the U.S. *Communications of the ACM, 48*(10), 105-110.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734.

Mayer, M. D., & Tan, F. B. (2002). Beyond models of national culture in information systems research. *Journal of Global Information Management, 10*(1), 24-32.

McKnight, D. H., & Chervany, N. L. (2002). What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International Journal of Electronic Commerce, 6*(2), 35-60.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002a). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(4), 334-359.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002b). The impact of initial consumer trust on intentions to transact with a Web site: A trust building model. *Journal of Strategic Information Systems, 11*(3-4), 297-323.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review, 23*(3), 473-490.

Pack, T. (1999). Can you trust Internet information? *Link - up, 16*(6), 24.

Park, C., & Jun, J.-K. (2003). A cross-cultural comparison of Internet buying behavior. *International Marketing Review, 20*(5), 534-553.

Pavlou, P. A., & Chai, L. (2002). What drives electronic commerce across cultures? A cross-cultural investigation of the theory of planned behavior. *Journal of Electronic Commerce Research, 3*(4), 240-253.

Plank, R. E., Reid, D. A., & Pullins, E. B. (1999). Perceived trust in business-to-business sales: A new measure. *The Journal of Personal Selling & Sales Management, 19*(3), 61-71.

Png, I. P. L., Tan, B. C. Y., & Wee, K.-L. (2001). Dimensions of national culture and corporate adoption of IT infrastructure. *IEEE Transactions on Engineering Management, 48*(1), 36-45.

Ratnasingham, P. (1998). The importance of trust in electronic commerce. *Internet Research: Electronic Networking Applications and Policy, 8*(4), 313-321.

Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist, 26,* 443-450.

Sadeh, N. (2002). *M-Commerce: Technologies, services, and business models.* Boston: Wiley.

Shankar, V., Urban, G. L., & Sultan, F. (2002). Online trust: A stakeholder perspective, concepts, impleications, and future directions. *Journal of Strategic Information Systems, 11*, 325-344.

Shapiro, S. P. (1987). The social control of impersonal trust. *American Journal of Sociology, 93*(3), 623-658.

Shapiro, D., Sheppard, B., & Cheraskin, L. (1992). Business on a handshake. *The Negotiations Journal, 8*, 365-377.

Sheppard, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A meta analysis of past research with recommendations for modifications in future research. *Journal of Consumer Research, 15*(3), 325-343.

Siau, K., Lim, E., & Shen, Z. (2001). Mobil commerce: Promises, challenges, and research agenda. *Journal of Database Management, 12*(3), 4-13.

Siau, S., & Shen, Z. (2003). Building customer trust in mobile commerce. *Communication of ACM, 46*(4), 91-94.

Siau, K., Sheng, H., & Nah, F. (2003). *Development of a framework for trust in mobile commerce.* Paper presented at the Workshop on HCI Research in MIS.

Soh, C., Kien, S. S., & Tay-Yap, J. (2000). Cultural fits and misfits: Is ERP a universal solution? *Communication of ACM, 43*(4), 47-51.

Strong, K., & Weber, J. (1998). The myth of the trusting culture. *Business & Society, 37*(2), 157-183.

Swan, J. E., Bowers, M. R., & Richardson, L. D. (1999). Customer trust in the salesperson: An integrative review and meta-analysis of the empirical literature. *Journal of Business Research, 44*(2), 93-107.

Tan, B. C. Y., Wei, K.-K., Watson, R. T., Clapper, D. L., & McLean, E. R. (1998). Computer-mediated communication and majority influence: Assessing the impact in an individualistic and a collectivistic culture. *Management Science, 44*(9), 1263-1278.

Tan, B. C. Y., Wei, K.-K., Watson, R. T., & Walczuch, R. M. (1998). Reducing status effects with computer-mediated communication: Evidence from two distinct national cultures. *Journal of Management Information Systems, 15*(1), 119-141.

Tarasewich, P., Nickerson, R. C., & Warkentin, M. (2002). Issues in mobile e-commerce. *Communications of the Association for Information Systems, 8*, 41-64.

Urban, G. L., Sultan, F., & Qualls, W. J. (2000). Placing trust at the center of your Internet strategy. *Sloan Management Review, 42*(1), 39-48.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186-204.

Walczuch, R., & Lundgren, H. (2004). Psychological antecedents of institution-based consumer trust in e-retailing. *Information & Management, 42*, 159-177.

Walczuch, R., Seelen, J., & Lundgren, H. (2001). Psychological determinants for consumer trust in e-retailing. *Eighth Research Symposium on Emerging Electronic Markets* (*RSEEM 01*). Retrieved March 8, 2007, from http://www-i5.informatik.rwth-aachen.de/conf/rseem2001/

Wang, Y. D., & Emurian, H. H. (2005). An overview of online trust: Concepts, elements, and implications. *Computer in Human Behavior, 21*, 105-125.

Yeo, J., & Huang, W. (2003). Mobile E-commerce outlook. *International Journal of Information Technology & Decision Making, 2*(2), 313-332.

Zacharia, G., & Maes, P. (2000). Trust management through reputation mechanisms. *Applied Artificial Intelligence, 14*(9), 881-907.

Zucker, L. (1986). Production of trust: Institutional sources of economic structure (1840-1920). *Research in Organizational Behavior, 8*, 53-111.

This work was previously published in Computer-Mediated Relationships and Trust: Managerial and Organizational Effects, edited by L. Brennan and V. Johnson, pp. 158-176, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.18
Trust Models for Ubiquitous Mobile Systems

Mike Burmester
Florida State University, USA

ABSTRACT

This chapter introduces the notion of trust as a means to establish security in ubiquitous mobile network systems. It argues that trust is an essential requirement to enable security in any open network environments, and in particular, in wireless ad hoc environments where there is no network topology. In such environments, communication can only be achieved via routes that have to be trusted. In general it may be hard, or even impossible, to establish, recall, and maintain trust relationships. It is therefore important to understand the limitations of such environments and to find mechanisms that may support trust either explicitly or implicitly. We consider several models that can be used to enable trust in such environments, based on economic, insurance, information flow, and evolutionary paradigms.

INTRODUCTION

Wireless mobile networks are a paradigm for mobile communication in which wireless nodes do not rely on any underlying static network infrastructure for services such as packet routing, name resolution, node authentication, or distribution of computational resources. The communication medium is broadcast. Nodes in range communicate in a direct peer-to-peer manner, while nodes out of range establish routing paths dynamically through other nodes where possible. The recent rise in popularity of mobile wireless devices and technological developments have made possible the deployment of wireless mobile networks for several applications. Examples include emergency deployments, disaster recovery, search-and-rescue missions, sensor networks, military (battlefield) operations, and more recently e-commerce. Since the network nodes are mobile, the network topology frequently changes: Communication links are

established or broken as nodes move in and out of range, and the network may get partitioned with the connectivity restricted to the partitions. As a result it may be much harder (or even impossible) to establish trust associations.

The trend in trust management is to view trust implicitly through delegation of privilege via certificates. Certificates can be chain-linked (linking à priori trust relationships) and used to propagate and distribute trust over insecure media, without the danger of being manipulated.

In this chapter, we give an overview of several models that can be used to support trust in mobile networks, based on economic, insurance, information flow, and evolutionary paradigms.

TRUST IN WIRELESS MOBILE NETWORKS

We consider environments in which there may be no fixed underlying network infrastructure, such as static base stations, for services such as packet routing, name resolution, node authentication, or the distribution of computational resources. In such environments, recalling and maintaining trust relationships is particularly challenging. Mobile systems share many of the complexities of fixed infrastructure systems. For example, nodes may have (Burmester & Yasinsac, 2004):

1. No prior relationship or common peers
2. No shared proprietary software
3. Different transmission, memory and processing capabilities
4. Different mobility characteristics
5. Different lifetime properties

Defining Trust

Trust is a highly abstract concept and it is unlikely that any simple definition can comprehensively capture all the subtleties of its essence. Informally we may define trust as a behavioral expectation of one party toward another. There are two perspectives in this definition, one in which a party *awards* trust to another (Alice trusts that Bob's public key is PK(Bob)), the other in which a party *gains* trust from another (Alice has convinced Bob that her public key is PK(Alice)).

Representing Trust: Certificates vs. Tokens

In any stateful trust model, trust must be represented by some type of persistent structure. Certificates are the de facto standard for representing trust relationships that are protected by cryptography. Certificates are portable and bind a cryptographic key (a digital string) to an entity, thus guaranteeing the authenticity of actions performed by that entity. Trust tokens are another structure that can be used to represent trust in a more direct way, analogous to the relation between checks and cash. Checks guarantee payment by tying the purchaser to some identifying information (like a certificate), while the value of cash is self-contained.

Trusted Third Parties

A trusted third party (TTP) can facilitate significantly the establishment of trust in mobile environments. For example, if two parties *A* and *B* who do not know each other have a trust relationship with a third party *T*, then *T* can be an effective intermediary for transactions between *A* and *B*. However in general, wireless mobile networks may not have any infrastructure components that are typically used as TTPs. In such cases, TTPs have to be elected or assigned by using an appropriate election or assignment protocol.

MODELS FOR TRUST IN WIRELESS MOBILE ENVIRONMENTS

Trust is *context* driven (e.g., A may trust *B* for event *x*, but not for event *y*). Trust may also be qualitative rather than Boolean (e.g., A may trust *B* more than *C*). Finally, trust relationships may be fixed or dynamic. Dynamic trust relationships are most appropriate for the requirements of mobile environments.

Models for dynamic trust must support establishing, changing, and permanently revoking trust between parties, and must also consider network environment issues. In the following sections we shall consider several models that can be used to support trust in wireless mobile networks (Burmester & Yasinsac, 2004).

A Mathematical Model for Trust: The Trust Graph

We may represent the trust in a network by a directed graph, the *trust graph*, whose links (A, B) correspond to the explicit trust that node *A* has in node *B*. Such links are indicated by $A \Rightarrow B$. The implicit trust that a node *X* has in another node *Y* is then represented by a trust path from *X* to *Y* :

$$X = X_0 \Rightarrow X_1 \Rightarrow X_2 \ldots \Rightarrow X_{n-1} \Rightarrow X_n = Y,$$

in which node *X* awards trust to node *Y* via a chain of intermediary nodes X_i, where X_i awards trust explicitly to the next node X_{i+1} in the chain. Such trust may be supported by certificates. For example, node X_i may certify (digitally sign) that key $PK(X_{i+1})$ is the public key of node X_{i+1}. A chain of certificates can then be used for implicit certification. This is essentially the trust model for the X509 PKI authentication infrastructure (ISO/IEC 9594-8, 1995). This particular trust infrastructure is hierarchical, with trust centrally managed (by a Root Certifying Authority, which is also a single-point-of-failure). PGP (Zimmermann, 1995) uses a web of trust in which trust is distributed "horizontally." See Burmester and Desmedt (2004) for a discussion on security issues of hierarchical vs. horizontal infrastructures.

In the basic trust graph model, trust is transitive but not necessarily reflexive. That is, even though *A* may award trust to *B*, *B* may not award trust to *A*. However, trust is binary: $A \Rightarrow B$ is either true or false. Therefore, there is a natural trust metric which is one unit for explicit trust. This is also the trust of a trust path that links A to *B*. In this model the trust that *A* awards to *B* is represented by the trust flow of *A*, *B*, which is also the connectivity of *A*, *B*. This model is appropriate for Byzantine faults environments in which the adversary can corrupt a bounded number of nodes, and trust has to be based on à priori beliefs, and not statistical profiles.

A Model Based on a Weighted Trust Graph

There are several other ways to define trust. For a stochastic model based on statistical profiling, we can define the explicit trust that *A* awards to (or has in) *B* as the probability with which *A* trusts *B*, based on, say, a history of good behavior by *B*. See the next section for a discussion on trust based on observed behavior. In this model we have a weighted trust graph in which each link $A \Rightarrow B$ is assigned a weight $t \in [0,1]$, which corresponds to the (explicit) trust that *A* has in *B*. If $\pi 1, \pi 2, \ldots, \pi n$ are (all) the trust paths that link *X* to *Y*, then the implicit trust that *X* has in *Y* can be computed as follows (Burmester, Douligeris, & Kotzanikolaou, 2006):

$$\Sigma_{\pi_i} \prod_{t \in \pi_i} t \; - \; \Sigma_{\pi_i \neq \pi_j} \prod_{t \in \pi_i \cup \pi_j} t + \ldots + (-1)^{n+1} \prod_{t \in \pi_1 \cup \ldots \cup \pi_n} t$$

For example, if there are three disjoint paths from *X* to *Y* with trust weights (t_1,t_2), (t_3,t_4), (t_5,t_6)

respectively, then the implicit trust that *X* has in *Y* is:

$$t_1t_2 + t_3t_4 + t_5t_6 - t_1t_2t_3t_4 - t_3t_4t_5t_6 + t_1t_2t_3t_4t_5t_6 .$$

One can extend this model to allow for a dynamic model in which trust is regularly updated, by using a trust-ranking algorithm similar to that used by Web search engines (e.g., PageRank of Google [PageRank, 1997]).

A Model Based on Observed Behavior

A natural way to acquire trust is through direct observation. At its most fundamental level, trust is a decision, subject to emotions and intuition. In this scenario, personal observation is preferred to second-hand methods because of hints, nuances, and feelings that can be garnered. Though feelings are not considered in computer trust systems, there are advantages in doing so. Not all actions give insight into trustworthiness. The challenge is to translate such observations into trust decisions.

A challenge to trust management systems is that trust relationships need to be constructed *before* they are exercised. There are four basic categories of activity that affect trust (Burmester & Yasinsac, 2004):

1. Trust earning actions over time
2. Trust earning actions by count
3. Trust earning actions by magnitude
4. Trust defeating actions

Combinations of the first three allow cautious parties to grant trust frugally. Untrustworthy parties will be challenged to conduct a sufficient quality and quantity of trustworthy actions to gain trust. On the other hand, observation of malicious, reckless, or otherwise unpredictable actions allows reduction or revocation of awarded trust.

A Model Based on the Internet Paradigm

The economic opportunity provided by the Internet has driven rapid establishment of many new trust models. Companies like eBay, Amazon, and Priceline conduct all of their business with customers with whom they have no personal relationship or interaction with. Early work on supporting trust models was from a business perspective (Pardue, 2000). Some work has been done more recently to identify models that support cryptographic protection of trust relationships. In Zhong, Chen, and Yang (2003), a token-based trust model is proposed in which parties accumulate trust, transaction-by-transaction. For trust-earning actions, parties are awarded tokens that can be retained and later presented to reflect the earned trust. If no additional trust information is gathered, tokens may be revoked or restricted. This novel approach to trust acquisition has many properties that are well-suited to mobile networks. Tokens can be created, awarded, and verified via distributed algorithms, allowing a global aspect to trust decisions. Conversely, if the trust algorithm is well understood, parties that desire to perform malicious acts can become sleepers, behaving perfectly until they acquire sufficient trust to allow successful mischief.

Transitive Trust

Transitivity is in many respects a natural attribute of trust and is encountered in some of the most used security systems (Steiner, Neuman, & Schiller, 1988; Zhong et al., 2003). With transitive trust models, trust must be explicit (i.e., parties must know that if they place their trust in one party, then they are automatically placing their trust in other potentially unknown parties as well). For example, if Alice trusts Bob and Bob trusts Carol, then Alice must trust Carol. Such models make

strong trust requirements on intermediaries or third parties. Unfortunately, there are inherent dangers in models with transitive trust (Christianson & Harbison, 1997).

A Model Based on Trust Classes

Trust may be considered as a two party relationship or there may be environments where nodes take on *class* trust properties, as in the Bell-LaPadula model (Bell & LaPadula, 1973). One way to form trust management functionality is to establish a trust promotion system. For example, consider a simple trust environment in which nodes can be categorized into the following five trust classes (from most to least trusted): *Highly trusted, Trusted, Unknown, Untrusted, Highly untrusted.* We can then establish a set of rules for promoting and demoting members between groups. These rules will be identified by the desired promotion rule. If promotion is not allowed for highly untrusted parties, then no rule is established for this class. The model may be further extended by designating a subset of the class of most trusted nodes as *promoters.* Promoters are responsible for determining if requestors meet the promotion requirements as designated in the promotion rules and in taking action to effect the justified group movement. While promotion is requested directly, demotion must be requested second hand.

A Financial Model

Trust can also be *contractually* secured. In this case, a Trusted Third Party guarantees the trust. As with secured loans, if the guaranteed trust is violated, the guarantor will deliver the promised security to the offended party. Secured trust is a pure form of transitive trust. It is unique in that its trust graph tree has height one and trust is secured by a contractually agreed value. As with secured financial interactions, the secured value may take many forms, including the following: a *co-signed trust certificate, a trust insurance policy, a trust bond and a trust collateral.*

These correspond to security mechanisms of the financial world. For a co-signed certificate, the co-signing party would have credentials that exceed those of the target and would assume liability for any adverse events that occur as a result of a trust breech. The insurance policy model is similar, except that the security is provided by a well recognized organization that promises benefits to the executor of the policy. The last two models are similar in that the trust target provides the value that secures the trust. The value can be monetary, property, or other items or issues of suitable value to the source.

CONCLUSION

We have considered several models that can be used to manage the trust in mobile wireless environments. These models are highly distributed and address many of the trust management properties that are needed to secure mobile environments.

ACKNOWLEDGMENTS

This material is based on work supported in part by the National Science Foundation under grant number NSF 0209092 and in part by the U.S. Army Research Laboratory and the Army Research Office under grant DAAD19-02-1-0235.

REFERENCES

Bell, D. E., & LaPadula, L. (1973). Secure computer systems: Mathematical foundations and model, *MITRE Corp.* M74-244, Bedford, MA.

Burmester, M., & Desmedt, Y. (2004). Is hierarchical public-key certification the next target for hackers? *Communications of the ACM, 47*(8), 68-74.

Burmester, M., & Yasinsac, A. (2004). Trust infrastructures for wireless mobile networks. *WSAES Transactions on Telecommunications* (pp. 377-381).

Burmester, M., Douligeris, C., & Kotzanikolaou, P. (2006). Security in mobile ad hoc networks. In C. Douligeris & D. Serpanos (Eds.), *Network security: Current status and future directions.* Piscataway, NJ: IEEE Press.

Christianson, B., & Harbison, W. S. (1997). Why isn't trust transitive? In *Proceedings of the 4th International Workshop on Security Protocols* (LNCS 1189, pp. 171-176).

ISO/IEC 9594-8. (1995). Information technology, Open Systems Interconnection. *The Directory: Overview of concepts, models, and services. International Organization for Standardization.* Geneva, Switzerland.

PageRank. (1997). Google. Retrieved from http://www.google.com/technology/

Pardue, H. (2000). A trust-based model of consumer-to-consumer online auctions. *The Arrowhead Journal of Business, 1*(1), 69-77.

Steiner, J., Neuman, C., & Schiller, J. I. (1988). Kerberos and authentication service for open network systems. In *Proceedings of USENIX*, Dallas, TX.

Zhong, S., Chen, J., & Yang, R. (2003). Sprite: A simple, cheat-proof, credit-based system for mobile ad hoc networks. In *Proceedings of INFOCOM 2003.*

Zimmermann, P. (1995). *The official PGP user's guide.* Cambridge, MA: MIT Press.

This work was previously published in Information Security and Ethics: Concepts, Methodologies, Tools, and Applications, edited by H. Nemati, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.19
Quality of Service in Mobile Ad Hoc Networks

Winston K. G. Seah
Institute for Infocomm Research, Singapore

Hwee-Xian Tan
National University of Singapore, Singapore

INTRODUCTION

Mobile ad hoc networks (MANETs) form a class of multi-hop wireless networks that can easily be deployed on-the-fly. These are autonomous systems that do not require existing infrastructure; each participating node in the network acts as a host as well as a packet-forwarding router. In addition to the difficulties experienced by conventional wireless networks, such as wireless interference, noise and obstructions from the environment, hidden/exposed terminal problems, and limited physical security, MANETs are also characterized by dynamically changing network topology and energy constraints.

While MANETs were originally designed for use in disaster emergencies and defense-related applications, there are a number of potential applications of ad hoc networking that are commercially viable. Some of these applications include multimedia teleconferencing, home networking, embedded computing, electronic classrooms, sensor networks, and even underwater surveillance.

The increased interest in MANETs in recent years has led to intensive research efforts which aim to provide quality of service (QoS) support over such infrastructure-less networks with unpredictable behaviour. Generally, the QoS of any particular network can be defined as its ability to deliver a guaranteed level of service to its users and/or applications. These service requirements often include performance metrics such as throughput, delay, jitter (delay variance), bandwidth, reliability, etc., and different applications may have varying service requirements. The performance metrics can be computed in three different ways: (i) concave (e.g., minimum bandwidth along each link); (ii) additive (e.g., total delay along a path); and (iii) multiplicative (e.g., packet delivery ratio along the entire route).

While much effort has been invested in providing QoS in the Internet during the last decade, leading to the development of Internet QoS models such as integrated services (IntServ) (Braden, 1994) and differentiated services (DiffServ) (Blake, 1998), the Internet is currently able to provide only best effort (BE) QoS to its applications. In such networks with predictable resource availability, providing QoS beyond best effort is already a challenge. It is therefore even more difficult to achieve a BE-QoS similar to the Internet in networks like MANETs, which experience a vast spectrum of network dynamics (such as node mobility and link instability). In addition, QoS is only plausible in a MANET if it is combinatorially stable, i.e., topological changes occur slow enough to allow the successful propagation of updates throughout the network. As such, it is often debatable as to whether QoS in MANETs is just a myth or can become a reality.

BACKGROUND

The successful deployment of QoS in MANETs is a challenging task because it depends on the inherent properties of the network: node mobility, variable (and limited) capacity links, as well as rapid deployment and configuration. These factors lead to a network with dynamic topology, complex route maintenance, and unpredictable resource availability. It is difficult to implement existing Internet QoS models on MANETs because these mechanisms cannot be efficiently deployed in a network with unpredictable and varying resource availability.

A very critical assumption is made by most, if not all, MANET protocols, which is the willingness of all nodes to participate in the forwarding of packets for other nodes in the network. QoS provisioning in MANETs is therefore a multi-faceted problem which requires the cooperation and integration of the various network layers, which will be discussed in the following subsections.

1. **Physical layer:** The physical layer of any network is used to provide the means to transmit sequences of bits between any pair of nodes joined by a communication channel. In MANETs, the radio channel is used to provide wireless communication between the nodes in the network. In contrast with wired networks, which offer predictability and stability, radio channels are affected by the effects of reflection, diffraction, and scattering from environmental interferences. As such, the wireless medium is often unreliable and subject to drastic variations in signal strength, leading to higher bit rate errors (BER) at the physical layer. Due to node mobility and the erratic behaviour of the wireless channel, the link characteristics of the network experience rapid changes. The effects of large-scale/small-scale fading, shadowing, and path loss may also cause these communication links to be asymmetric. Hence, the physical mechanisms must be able to adapt to the changes and deterioration in link quality during data transmission and change their modulation scheme accordingly to suit the current channel state.
2. **Medium access control (MAC) layer:** The wireless channel in MANETs is a broadcast and shared medium where nodes are often subject to interference from neighbouring nodes within the transmission and interference ranges, and often suffer from hidden/exposed terminal problems. Although many solutions have been proposed to alleviate the exposed/hidden terminal problems, these problems are more pronounced in autonomous, mobile environments; wireless channels are also subjected to errors which are bursty, location-, and mobility-dependent. The MAC layer for MANETs has to cope with these problems, as well as the challenges of minimizing collisions, allowing fair ac-

cess, and providing reliable data transport under rapidly changing conditions.

3. **Network layer:** The main challenge of the network layer in a MANET is to determine and distribute routing information efficiently under changing link conditions and scarce bandwidth. In addition, it must be able to interoperate with traditional non-ad-hoc networks, such as the Internet and other wireless networks. Existing MANET routing protocols can be broadly grouped under reactive, proactive, or hybrid routing protocols. If the network topology changes too rapidly due to high node mobility, topology updates in these routing protocols may not propagate fast enough to form stable routes. Most of these protocols are also based on shortest-path algorithms, which may not result in routes that have the required resources to meet the requirements of the applications they support. An ideal QoS routing protocol should be able to adaptively select its paths based on the currently available resources to provide the service desired by a particular application.
4. **Transport layer:** In the wired Internet, there are two transport-layer protocols available to the application layer: (i) UDP (user datagram protocol), which provides unreliable and connectionless service; and (ii) TCP (transmission control protocol), which provides reliable, connection-oriented service to the invoking applications. Besides having to provide logical communication between applications running on mobile hosts, the transport layer in a MANET also needs to handle delay and packet loss arising from conditions unlike wired networks. In TCP (which is used for most applications in the Internet), packet losses are due to congestion, and a back-off mechanism will then be invoked to reduce the sending rate of data packets from the source nodes. However, in wireless media, packet loss is mainly due to transmission errors and current flow control or congestion control techniques that might lead to lowered throughput. There are also large variations in delay when the route changes, which is not addressed by the design of the existing transport layer protocols.
5. **Application layer:** According to Kurose (2003), the service requirements of an application can be broadly classified into data loss, bandwidth, and delay (which can be average end-to-end delay or delay variance). Loss-tolerant applications include multimedia applications such as real-time audio and video, which are not adversely affected by occasional loss of data but are highly sensitive to bandwidth and delay. Other applications involving sensitive data integrity, such as electronic mail and banking transactions, require fully reliable data transfer, but may not be time-sensitive and can work with elastic bandwidth. To cater for QoS in MANETs, the application layer must be designed to handle frequent disconnections and reconnections caused by the dynamic network topology and varying signal quality of the wireless channel. It must also be able to adapt to widely varying delay and packet losses.

DEVELOPMENTS IN MANET QoS

From the discussion above, we can see that hard QoS in MANET is unlikely to be plausible because of the inherent dynamic nature of a mobile ad hoc environment. It may be more feasible to implement soft QoS, whereby there may exist transient periods of time when the network is allowed to fall short of QoS requirements, up to a permitted threshold. The level of QoS satisfaction is thus quantified by the fraction of total disruption. We can then make QoS a function of the available network resources, and applications should ideally adapt to the quality of the network.

There have been many research efforts to provide QoS support in MANETs. Wu (2001) categorizes these efforts into QoS models, QoS resource reservation signaling, QoS routing, and QoS MAC, and provides an overview of how these different components can work together to deliver QoS in MANETs. In the following subsections, we describe some of the recent developments by the networking community and evaluate their effectiveness in providing QoS in MANET scenarios.

MANET QoS Models

A QoS model defines the methodology and architecture for providing certain types of service in the network, but it does not define the specific protocols, algorithms, or implementations to realize QoS provisioning. An ideal MANET QoS model should take into account the various network dynamics and constraints experienced by the nodes, such as mobility and varying link characteristics, and ensure that the architecture is able to provide some form of QoS guarantees.

Xiao (2000) proposes a flexible QoS model for MANETs (FQMM), which adopts a hybrid provisioning policy by dividing traffic into different classes and applying different QoS handling mechanisms to these classes. IntServ-like per flow provisioning is used for the class with highest priority, while DiffServ-like per aggregate is used for the remaining classes. FQMM is suited for relatively small-sized networks of up to 50 nodes; as in DiffServ, a node can be an ingress node (source node), egress node (destination node), or an interior node which forwards packets for other nodes. Depending on the topology and traffic pattern, the role that each node undertakes changes dynamically.

An integrated MANET QoS (iMAQ) model proposed by Chen (2002) defines a cross-layer architecture for multimedia traffic. The framework is comprised of: (i) an application layer that generates multimedia data, (ii) a middleware layer that uses location information from the network layer to predict network partitioning, and (iii) a network layer that uses predictive location-based routing to select a path to the future location of a node. Information is exchanged among the data advertising, lookup, replication services, and the QoS routing protocol in the form of system profiles, and this helps to achieve a higher quality in data access as well as enhances communications between the respective layers.

A two-layer QoS (2LQoS) model has also been proposed (Nikaein, 2002) which provides differentiated services and soft guarantees to network resources for admitted applications by using class-based weighted fair queuing (CB-WFQ) at the intermediate nodes. The model is comprised of two main phases: (i) path generation, in which the quality of the route is computed based on the network layer quality of intermediate nodes; and (ii) path selection based on the desired QoS class (which are mapped onto various application level metrics). In this architecture, network layer metrics (NLMs), which are used to determine the quality of individual nodes, are separated from the Application Layer Metrics (ALMs). NLMs refer to the hopcount, buffer level, and stability level, whereas ALMs comprise delay, throughput, and enhanced best-effort. The work is extended to include MAC Layer Metrics, such as the signal-to-noise-ratio (SNR) of the link and the coding scheme used.

QoS MAC for MANETs

Medium access control (MAC) protocols for MANETs are non-deterministic and distributed, with no base station of a centralized controller to coordinate the channel access. Therefore, mobile nodes have to contend for access to the shared medium in a random access manner. The industry standard MAC access scheme used in wireless networks is IEEE 802.11, which includes both the point coordination function (PCF) and distributed coordination function (DCF). However,

this base standard is not directly applicable to MANETs because of its lack of traffic prioritization mechanism.

A number of contention-based QoS MAC protocols have been proposed, some of which include: (i) priority queuing schemes such as IEEE 802.11e, (ii) multi-channel schemes with separate channels for data and control packets, and (iii) black-burst contention schemes with varying delay for traffic of different priorities.

The IEEE 802.11e MAC standard (and its enhanced distributed coordination function extension) is an enhancement to the original standard which aims to support QoS in wireless networks. Differentiated service is provided to nodes by having four queues with different access categories, each of which corresponds to a different set of channel access parameters. Traffic with higher priority can then contend for channel access more successfully than low priority traffic by modifying parameters such as inter-frame spacing (IFS) and contention window size (He, 2003).

Conventional MAC protocols, such as IEEE 802.11, are single-channel models which experience higher collisions and contention for channel access as the number of nodes in the network increases. Consequently, the network performance degrades significantly because the overall throughput is limited by the bandwidth of the channel. In contrast, multi-channel MAC schemes, such as those proposed by Tian (2003) and Wu (2002), have better throughput performance, decreased propagation delay per channel, and QoS provisioning in MANETs.

The black-burst (BB) contention scheme (Sobrinho, 1999) is a distributed MAC scheme that provides real-time access to ad hoc CSMA wireless networks. The real-time data traffic contend for access to the wireless channel by jamming the media with pulses of energy known as BBs, which have lengths that are functions of the delay being experienced by the nodes. Hence, the BB contention scheme gives priority to real-time traffic, enforces a round-robin discipline among real-time nodes, and results in bounded access delays to real-time packets.

Another MAC protocol that provides multiple priority levels is proposed by Sheu (2004). It adopts the black-burst mechanism, as described earlier, to differentiate between high and low priority stations. This helps to guarantee that the frames with higher priority, such as multimedia real-time traffic, will always be transmitted earlier than frames with lower priority. Furthermore, stations with the same priority will access the shared channel in a round-robin manner.

MANET QoS Routing

QoS routing is considered by Chakrabarti (2001) to be the most important element in the network because it specifies the process of selecting routes to be used by the packets of a logical connection in attaining the associated QoS guarantee. Crawley (1998) presents a framework for QoS-based routing in the Internet whereby paths for flows are determined based on some knowledge of resource availability in the network as well as the QoS requirements of the connection. There have been significant research efforts on QoS routing in MANETs, some of which include: (i) QoS extensions to existing routing protocols, (ii) AQOR (ad hoc QoS on-demand routing), (iii) CEDAR (core extraction distributed ad hoc routing), (iv) multi-path QoS routing, and (v) QoS-GRID, which uses topology management. Other QoS routing schemes for MANETs are discussed in Jawhar (2004).

As conventional MANET routing protocols implicitly select the shortest paths during route establishment and/or route maintenance, they are unable to offer QoS support to the nodes in the network. As such, many extensions for existing MANET protocols have been proposed to take into account the type of resources desired by the requesting application. Perkins (2003) proposes changes to AODV to provide QoS support by adding extensions to the messages used by the

route-discovery process. These extensions specify the service requirements, such as maximum delay and minimum bandwidth, in the route request (RREQ) and route reply (RREP) messages. Badis (2004) has also proposed QOLSR, an extension to the original OLSR. Instead of using the number of hops for route selection, metrics such as the available bandwidth, delay, jitter, loss probability, etc., are also added to the OLSR functionality and control messages format, to be used for multi-point relay (MPR) selection and routing table calculation.

The ad-hoc QoS on-demand routing (AQOR) protocol (Xue, 2003) is another QoS routing protocol that provides end-to-end QoS support in terms of bandwidth and delay. Besides performing accurate admission control and resource reservation via detailed computations, AQOR is also equipped with signaling capabilities to handle temporary reservation and destination-initiated recovery processes.

CEDAR (Sivakumar, 1999) performs QoS routing for small to medium-sized MANETs. It is comprised of three main components: (i) establishment and maintenance of a self-organizing routing infrastructure called the core, by approximating a minimum dominating set, (ii) propagation of the link state of high bandwidth and stable links in the core to all core nodes, and (iii) a QoS-route computation algorithm that is executed at the core nodes using only locally available states. The route is then selected from the dominator of the source node to the dominator of the destination that satisfies the required bandwidth.

In multi-path QoS routing protocols such as Liao (2002), Chen (2004), Leung (2001), and Chen (2004), the route discovery process selects multiple paths (ideally disjoint) from the source to destination. The multiple paths can be used collectively to satisfy the required QoS requirements (such as bandwidth), and in the event of link breakages along the main paths, the backup paths will then take over the routing immediately, thus reducing the time needed for another route-computation process.

QoS-GRID (Liu, 2003) is a location-based routing protocol with QoS provisioning. It uses a two-tier grid system to reduce the transmission power of each node so as to enhance the bandwidth utilization and provide stable bandwidth guarantees.

ALTERNATIVE QoS MECHANISMS

In the previous sections, we have offered a multi-layered overview of the problems and issues that surround QoS provisioning in MANETs. Despite numerous efforts to overcome these challenges and add guarantees to data delivery in autonomous, distributed, and ad hoc environments, it is inherently difficult to provide QoS support in MANETs due to the following factors: (i) unreliable and unpredictable wireless transmission media, (ii) node mobility induced topology and route changes, which lead to inaccurate locality information, and (iii) power control and energy constraints.

In addition, existing algorithms and mechanisms do not provide any form of assurance that routes will be found nor that broken routes will be recovered within a given time. To overcome these uncertainties, techniques like topology control and mobility prediction have been exploited.

In topology control, certain system parameters, such as the transmission radii of the nodes, can be varied using power control. However, this is not straightforward and may increase contention in the nodes. A QoS routing mechanism with mobility prediction has also been proposed by Wang (2001), which uses node movement patterns to determine the future location of nodes. It then selects the most stable path based on mobility prediction and QoS requirements on bandwidth and delay, but this does not eliminate the possibility that link breakages can still occur along the selected paths.

In the following, we propose some key alternatives to overcome the transient and unpredictable characteristics of MANETs and provision for QoS in MANETs.

- **Controlled node movement:** The system/protocols can be empowered with the ability to control the movement of a subset of nodes in the network. This can be done by making use of swarms of mobile robots with sensors and actuators (Seah, 2006), unmanned autonomous vehicles (UAVs), and public transportation (such as buses and trains), which have more predictable mobility patterns.
- **Clustering:** The ad-hoc nature of MANETs, along with their decentralized architecture, poses much difficulty in the coordination and functioning of the network. Clustering techniques enable dynamic hierarchical architectures to be formed and improve the network performance.
- **Vertical coupling (cross layer interactions):** Although traditional networking paradigms promote the usage of a multi-layered protocol stack in which the different layers have minimal impact on each other, this does not lead to optimal performance. Cross layered designs, such as that proposed by Chen (2002) can help to improve network performance by sharing information across the different layers, at the cost of interdependency between adjacent layers.
- **QoS adaptation:** Conventional network protocols have static behavior, i.e., they perform a fixed set of actions at all times, irrespective of the current network conditions. Since a MANET is generally dynamic in nature, QoS adaptation, whereby the protocols adapt to the network conditions at all times, may be able to produce better performance.

FUTURE TRENDS

Although there have been vast amounts of studies in the different aspects of QoS support in MANETs—QoS models, QoS MAC protocols, QoS routing protocols, and QoS signaling techniques, there are currently very few practical deployments of such networks. This is a consequence of the fact that the current research in MANETs is still unable to support the QoS requirements of the envisioned applications.

Future trends in QoS provisioning in MANETs appear to follow a cross-layered approach, with the different protocol layers working together to enhance the reliability, robustness, and overall performance of the network. The dynamic nature of MANETs also necessitates the need for the protocols to adapt their behaviors according to the prevailing network conditions–a mechanism that can generally be defined as QoS adaptation. In addition, the unpredictability and constraints of MANETs push the need for soft QoS to be considered as a compromising principle in MANETs. To provide QoS guarantees in the network, many other issues and assumptions have to be further studied. These include security, node reliability, node misbehaviors, node mobility, and the possible further interoperation of MANETs with the wired Internet.

CONCLUSION

The distributed architecture and autonomous nature of the nodes in MANETs contribute to its attractiveness as a communication network that can be easily deployed on-the-fly. However, the inherent characteristics of MANETs—node mobility, decentralized architecture, multi-hop communications, limited resources, and unstable link quality—contribute to the impediment of network performance. As such, it is difficult to provide QoS support in the network using

traditional network techniques (such as resource reservation and traffic differentiation) that are used in the wired Internet. As QoS provisioning in MANETs typically involve the collaboration of various layers of the networking protocol stack, researchers are increasingly considering the use of cross-layered designs, adaptivity, and mobility predictions to achieve QoS guarantees in the network. Nevertheless, there are still several outstanding QoS issues that must be addressed, and alternative forms of mechanisms must be studied in greater depth to facilitate the development of QoS in MANETs.

REFERENCES

Badis, H., Agha, K. A., & Munaretto, A. (2004). *Quality of service for ad hoc optimized link state routing protocol (QOLSR).* IETF Internet draft, draft-badis-manet-qolsr-00.txt, Work in Progress.

Blake, S., Black, D., Carlson, N., Davies, E., Wang, Z., & Weiss W. (1998). *An architecture for differentiated services.* IETF RFC 2475.

Braden, B., Clark, D., & Shenker, S. (1994). Integrated services in the Internet architecture: An overview. IETF RFC1633, June.

Chakrabarti, S., & Mishra, A. (2001). QoS issues in ad hoc wireless networks. *IEEE Communications Magazine, 39*(2), 142-148.

Chen, K., Shah, S. H., & Nahrstedt, K. (2002). Cross-layer design for data accessibility in mobile ad hoc networks. *Journal of Wireless Personal Communications,* Special Issue on Multimedia Network Protocols and Enabling Radio Technologies, Kluwer Academic Publishers, 21, 49-75, 104-116.

Chen, Y. S., Tseng, Y. C., Sheu, J. P., & Kuo, P. H. (2004). An on-demand, link-state, multi-path QoS routing in a wireless mobile ad-hoc network. *Computer Communications, 27*(1), 27-40.

Chen, Y. S., & Yu, Y. T. (2004). Spiral-multi-math QoS routing in a wireless mobile ad hoc network, *IEICE Transactions on Communications,* E87-B, No. 1.

Crawly, E. S., Nair, R., Rajagopalan, B., & Sandick, H. (1998). *A framework for QoS-based routing in the Internet.* IETF RFC 2386.

He, D., & Shen, C. Q. (2003). Simulation study of IEEE 802.11e EDCF. In *Proceedings of IEEE Vehicular Technology Conference* (VTC 2003, Spring), Seoul, Korea, Vol. 1, 685-689.

Jawhar, I., & Wu, J. (2004). *Quality of service routing in mobile ad hoc networks.* Kluwer Academic Publishers.

Kurose, J. R., & Ross, K. W. (2003). *Computer networking: A top-down approach featuring the Internet* (2nd ed.). Addison Wesley.

Leung, R., Liu, J., Poon, E., Chan, C., & Li, B. (2001). MP-DSR: A QoS-aware multi-path dynamic source routing protocol for wireless ad-hoc networks. In *Proceedings of 26th IEEE Annual Conference on Local Computer Networks (LCN 2001),* Tampa, Florida, USA, 132-141.

Liao, W. H., Wang, S. L., Sheu, J. P., & Tseng, Y. C. (2002). A multi-path QoS routing protocol in a wireless mobile ad hoc network. *Telecommunications Systems,* 19, 329-347.

Liu, H.,& Li, Y. (2003). A location based QoS routing protocol for ad hoc networks. In *Proceedings of 17th International Conference on Advanced Information Networking and Applications (AINA '03),* Xi'an, China, 830-833.

Nikaein, N., Bonnet, C., Moret, Y., & Rai, I. A. (2002). 2LQoS–Two-layered quality of service model for reactive routing protocols for mobile ad hoc networks. In *Proceedings of 6th World*

Multiconference on Systemics, Cybernetics and Informatics (SCI 2002), Orlando, Florida, USA.

Perkins, C. E., & Belding-Royer, E. M. (2003). *Quality of service for ad hoc on-demand distance vector routing.* IETF Internet draft, draft-perkins-manet-aodvqos-02.txt, Work in Progress.

Seah, W. K. G., Liu, Z., Lim, J. G., Rao, S. V., & Ang, M. H. Jr. (2006). TARANTULAS: Mobility-enhanced wireless sensor-actuator networks. In *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC2006)*, Taichung, Taiwan, 548-551.

Sheu, J. P., & Liu, C. H. (2004). A priority MAC protocol to support real-time traffic in ad hoc networks. *Wireless Networks*, 10, 61-69.

Sivakumar, R., Sinha, P., & Bharghavan, V. (1999). CEDAR: A core-extraction distributed ad hoc routing algorithm. *IEEE Journal on Selected Areas in Communications, 17*(8), 1454-1465,

Sobrinho, J. L., & Krishnakumar, A. S. (1999). Quality-of-service in ad hoc carrier sense multiple access wireless networks, *IEEE Journal on Selected Areas in Communications, 17*(8), 1353-1368.

Tian, H., Li, Y. Y., Hu, J., & Zhang, P. (2003). A MAC protocol supporting multiple traffic over mobile ad hoc networks. In *Proceedings of 57th IEEE Semiannual Vehicular Technology Conference* (VTC 2003, Spring), Seoul, Korea, Vol. 1, 665-669.

Wang, J., Tang, Y., Deng, S., & Chen, J. (2001). QoS routing with mobility prediction in MANET. In *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, Canada*, Vol. 2, 357-360.

Wu, K., & Harms, J. (2001). QoS support in mobile ad hoc networks. *Crossing Boundaries: The GSA Journal of University of Alberta, 1*(1), 92-106

Wu, S. L., Tseng, Y. C., Lin, C. Y., & Sheu, J. P. (2002). A multi-channel MAC protocol with power control for multi-hop mobile ad hoc networks. *The Computer Journal, 45*(1), 101-110.

Xiao, H., Seah, W. K. G., Lo, A., & Chua, K. C. (2000). A flexible quality of service model for mobile ad-hoc networks In *Proceedings of IEEE 51st Vehicular Technology Conference*, Tokyo, Japan, Vol. 1, 445-449.

Xue, Q., & Ganz, A. (2003). Ad hoc QoS on-demand routing (AQOR) in mobile ad hoc networks. *Journal of Parallel and Distributed Computing, 63*(2), 154-165.

KEY TERMS

Clustering: A networking technique in which nodes in the network group themselves according to some network attributes to form hierarchical architectures.

Cross-Layer Design: A protocol design that leverages on the interactions and dependencies between different layers of the networking protocol stack to achieve better performance. MANET (mobile ad hoc network)—self-configuring and self-maintaining network in which nodes are autonomous and distributed in nature.

QoS (quality of service): The ability of a network to deliver a guaranteed level of service to its users and/or applications.

QoS Adaptation: The adaptation of the behavior of one or more network protocols according to the prevailing network conditions, so as to achieve QoS in the network.

Service Requirements: Performance metrics such as throughput, delay, jitter (delay variance), bandwidth, and reliability which are usually application-specific.

Soft QoS: A compromising principle of QoS support whereby there may exist transient periods of time when the network is allowed to fall short of QoS requirements, up to a permitted threshold. The level of QoS satisfaction is thus quantified by the fraction of total disruption.

UAV (Unmanned Autonomous Vehicle): A machine that can move through the terrain intelligently and autonomously without the need for any human intervention.

This work was previously published in Encyclopedia of Internet Technologies and Applications, edited by M. Freire and M. Pereira, pp. 441-448, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.20
Quality of Service Issues in Mobile Multimedia Transmission

Nalin Sharda
Victoria University, Australia

ABSTRACT

The focus of this chapter is on the quality of service (QoS) aspects involved in transmitting multimedia information via mobile systems. Multimedia content and applications require sophisticated QoS protocols. These protocols need to manage throughput, delay, delay variance, error rate, and cost. How errors are handled in a multimedia session can have significant impact on the delay and delay variance. Coding and compression techniques also influence how the final presentation is transformed by the impediments encountered on a mobile network. Providing the user with the ability to negotiate between cost, quality, and temporal aspects is important, as this allows the user to strike a balance between these factors. In moving from 2G to 3G, and, over the next decade to 4G mobile networks, the ability to transmit multimedia information is going to improve constantly. Nonetheless, providers must develop viable economic models and user interfaces for providing differentiated QoS to the users.

INTRODUCTION

Transmission of multimedia information over mobile networks to portable devices, such as laptops, mobile phones, and PDAs (personal digital assistants), is leading to the development of new applications. However, successful transmission of multimedia information over mobile networks cannot be taken for granted. Understating the impediments to successful transmission of multimedia information is of paramount importance. This chapter focuses on multimedia applications that use mobile networks, and issues involved in the delivery of multimedia content with the desired quality of service (QoS). Current and future challenges in achieving successful mobile multimedia information transmission are also discussed.

Multimedia applications require more sophisticated QoS protocols than those for simple data transmission. The main parameters that underpin QoS are throughput, delay, delay variance, error rate, human perception of quality, and cost (Sharda, 1999). The interplay between these factors is rather complex, therefore, some simplifying assumptions must be made in developing methodologies for delivering multimedia content with the desired QoS.

For the delivery of desired QoS, one of the most promising concepts developed over the last few years is that of resource reservation. This entails reserving resources such as bandwidth on interconnects, and buffer space and processing power on switching nodes.

Packet switching networks embody the idea of statistical time division multiplexing (STDM); that is, resources are allocated to a communication session based on the demands of the traffic. This leads to more efficient, and therefore, more economical usage of the resources. However, the need to allocate resources dynamically adds complexity to the communication system's operation and management. Mobile multimedia communications are further complicated due to their variable transmission quality, the need to keep track of end system location, restrictions placed due to limited battery life, reduced screen size, and the cost of the connection.

Over the last decade, some progress has been made in establishing mobile multimedia transmission systems. However, much research and development is still required before we can take it for granted that a multimedia application, such as videoconferencing, would run with the desired QoS over a mobile communication infrastructure on a hand-held device as we zoom down a freeway at high speed, and, all this at a reasonable cost.

The next section of this chapter presents the challenges introduced by the mobile multimedia content, applications, and communication systems. It begins with an overview of mobile multimedia systems, and then presents the implications of coding and compression techniques for transmitting multimedia. Requirements of various multimedia applications and their relationship to mobile communication systems are also presented.

The third section presents QoS issues in transmitting multimedia content over mobile systems. Fundamentals of QoS concepts and different QoS models are introduced, and a novel model for managing QoS in real time is presented.

The fourth section presents directions for future research, and the final section gives the conclusions.

MOBILE MULTIMEDIA SYSTEMS

Overview

This section presents an overview of coding methods used for various media types, multimedia applications, and current mobile communication systems. QoS issues related to each of these are also discussed.

Multimedia communication systems combine different types of media contents, such as text, audio, still images, and moving images, to achieve the overall objective of a communication session. Therefore, the network needs to provide a service which works well for all media types.

The requirements for successfully transmitting a particular media type depend upon its coding and compression techniques, and the application in which it is being used. Media content that must be transmitted live, or processed in real time, poses more stringent requirements. Consequently, live video conferencing is one of the most challenging multimedia applications.

The network infrastructure and the communications protocols used for transmission play a vital role in satisfying the demands of a given application. In general, multimedia transmission requires high bandwidth, low error rate, low delay, and very low delay variance. To date, we have

not solved all of these challenges for even wired media. Fulfilling these requirements for achieving high-quality multimedia transmission over wireless connections is even more challenging.

The transition from the 2nd generation (2G) mobile systems to the 3rd generation (3G) mobile communication infrastructure presents new opportunities; however, still there are many challenging problems that need to be overcome. One of the key features missing in the current systems is the facility for the user to negotiate with the system and strike a compromise between the three key service aspects—quality, cost, and time—just as any market-oriented goods or services have to strike a balance between the quality, cost, and its delivery time.

Errors encountered in any transmission system can be either ignored, or detected and corrected. Errors can be ignored only if the received message is usable even with some errors. If errors in the received message are not acceptable, then these errors must be detected and corrected. Reverse error correction protocol requests retransmission of packets received with errors. This not only adds delay to the final reception of packets, it also adds delay jitter, as different packets encounter different delays. Forward error correction protocols include additional error correction bits, so that some of the errors can be corrected from the received data; this adds to the total data traffic. The choice of error handling method depends upon the type of data, its coding methodology, and the application.

Multimedia Content

By definition, a multimedia system combines different media types: text, audio, still and moving images. Each of these content types can be further categorised into sub-types. For example, still images can be bi-tonal, greyscale, or full-colour; furthermore, these can have continuous variation in tone—as in a photograph, or have sudden variation in the intensity—as in a printed page. A variety of techniques are used for digitally coding still images, depending upon the image type and application. Similarly, many text representation techniques and associated digital coding techniques are used. Audio and video are even more complex, as these are time varying quantities and involve continuous sampling over time. Errors and delays introduced at any stage of sampling, encoding, transmitting, and decoding of audio and video can lead to reduction in the quality of the final presentation.

Most multimedia content needs to be compressed to reduce the storage space and transmission bandwidth. Uncompressed multimedia content has in-built redundancy, and a few corrupted bits do not change the contents dramatically. Conversely, compressed media is compact, and has much less redundancy. Consequently, any errors during transmission affect compressed content more severely.

Mobile transmission systems are inherently more error-prone than wired transmission systems. The requirements for successfully transmitting a particular media type over a network depend not only on its coding and compression techniques, but on its application as well. However, all multimedia content is for human consumption, therefore, the criteria for acceptable quality of presentation ultimately depends upon human perception. For example, streamed video can accept a few seconds of delay, but live video conferencing becomes rather ineffective if the round-trip delay exceeds even a tenth of a second.

Text Coding

Despite the move towards graphical information, text remains a vital part of any multimedia presentation. One of the most enduring text codes is the American Standard Code for Information Interchange (ASCII). ASCII began its life as a 7-bit code designed for use with teletypes. Today, if

someone talks of an ASCII document, they essentially refer to a text document with no formatting. Applications such as Notepad create ASCII text, and word processors can save a file as "text only". Extended ASCII codes were designed for computers to be able to handle additional characters from other languages. It took some time to get a single standard for these additional characters, and there are a few Extended ASCII sets.

Unicode provides a text code that is independent of platform, program, or language. In Unicode, a unique 16-bit number is reserved for every character. The Unicode standard aims to provide a universal repertoire with logical ordering that is efficient. The latest version of Unicode Standard is Unicode 4.0.1, and supports around a hundred international scripts.

ASCII and Unicode have been used extensively over wire-line communication systems, and can be used over wireless media as well. In general, transmission of text codes does not require high bandwidth or stringent limits on delay and delay variance. Hence maintaining QoS in transmitting text is often not much of a problem. Nonetheless, a new code set was designed for sending short text messages over mobile systems.

Short message service (SMS) uses a 7-bit code set that enables one to send and receive text messages of up to 160 characters on mobile phones. Some 8-bit messages are used for sending smart messages (such as images and ring tones) and for changing protocol settings. For Unicode-based text messages, 16-bit codes of maximum 70 characters can be used. These are viewable by most phones, and some appear as a flash SMS, that is, appear on the screen immediately upon arrival, without pressing any button. The SMS code was originally developed for the 2G technology, and therefore works well with 2G as well as 3G systems. The only possible issue with respect to QoS can be errors; bandwidth, delay, and delay jitter do not impede the transmission of SMS messages.

Non-Textual Information

The standard developed to transmit multimedia information over the Internet is the multipurpose Internet mail extensions (MIME). This standard was developed by the Internet Engineering Task Force (IETF) to support the transmission of mixed-media messages across TCP/IP networks. This also became the standard for transmitting foreign language text which the ASCII code could not represent.

Multimedia messaging service (MMS) provides the ability to send messages that combine text, sounds, images, and video over wireless networks. This requires handsets that are MMS capable. MMS is an open wireless standard specified by the WAP (wireless application protocol) forum—which has now been consolidated into the Open Mobile Alliance (OMA). In the WAP protocol, a notification message triggers the receiving terminal to start retrieving the message automatically using the WAP GET command. This retrieval may be modified by applying filters defined by the user. The content that can be transmitted with the WAP protocol can use a variety of media types and encoding standards.

Audio Coding

The basic technique for digitising analog audio signals is called pulse code modulation (PCM). In this technique, an analog audio signal is sampled at a rate double that of the maximum frequency that needs to be captured, and each sample is stored using 8-bit or 16-bit words.

Phone quality audio signals are sampled at 8,000 samples per second, and stored with 8-bit resolution; this generates 64 Kbps data rate. CD quality audio has two channels; it is sampled at 44,000 samples per second and saved with 16-bit resolution, giving a 1.4 Mbps data rate. A variety of compression techniques are used to reduce the bandwidth required to transmit audio signals.

Compression becomes particularly important for CD quality stereo music, as the required 1.4 Mbps bandwidth is not economically available even in wire-line networks, much less so in wireless networks.

The MP3 (MPEG audio Layer 3) compression format has become one of the most widely used standards for transmitting high quality stereo audio. MP3 is one of three audio coding schemes associated with the MPEG video compression standard. The MP3 standard provides the highest level of compression and uses perceptual audio coding and psychoacoustic compression to remove all redundant and irrelevant parts of a sound signal that the human ear does not hear. MP3 uses modified discrete cosine transform (MDCT) and improves the frequency resolution 18 times with respect to that of the MPEG audio Layer 2 coding scheme. It manages to reduce the CD bit rate of 1.4 Mbps down to 112-128 Kbps (a factor of 12) without sacrificing sound quality. Since MP3 files are small, they are easily transferred across the Internet, and are also suitable for transmission over wireless networks.

The next generation of MP3 standard is called mp3PRO. It is fully compatible with MP3, while halving the storage and bandwidth requirements. With this standard CD quality stereo can be transmitted at 64 Kbps. Furthermore, it can be used with digital rights management software, and can be ported transparently to any MP3-friendly application.

Advanced audio coding (AAC) is a wideband audio coding algorithm that exploits two main coding strategies to reduce the amount of data needed to encode high-quality digital audio. First, it removes signal components that are not important from a human perception point of view, and second, it eliminates redundancies in the coded audio signal. The MPEG-4 AAC standard incorporates MPEG-2 AAC, for data rates above 32 Kbps per channel. Additional techniques increase the effectiveness of the AAC technique at lower bit rates, and are able to add scalability and/or error resilience. (These techniques extend AAC into its MPEG-4 version: ISO/IEC 14496-3, Subpart 4.) The MPEG-4 aacPlus standard combines advanced audio coding techniques such as spectral band replication (SBR), and parametric stereo (PS). The SBR techniques deliver the same audio quality at half the bit rate, while the PS techniques (optimised for the 16-40 Kbps range) provide high audio quality at bit rates as low as 24 Kbps (Dietz & Meltzer, 2002).

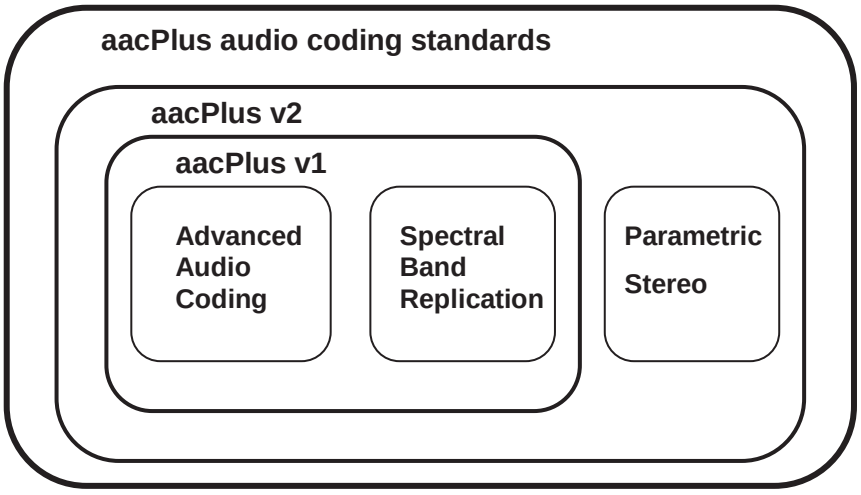
The aacPlus codec family includes two versions. Version 2 of aacPlus is the high quality audio codec targeted for use in the 3GPP (3rd Generation Partnership Project). The aacPlus version 1 standard is adopted by 3GPP2 and ISMA (Internet Streaming Media Alliance) for digital video broadcasting (DVB).

The relationship between the aacPlus codec family members is shown in Figure 1 (Dietz & Meltzer, 2002). To compress the incoming stereo audio, the encoder extracts parametric representation of the stereo aspect of the audio. The stereo parametric information takes 2-3 Kbps and is transmitted along with the mono signal. Based on the parametric representation of the stereo information, the decoder regenerates the stereo signal from the received mono audio signal.

To be able to transmit high quality stereo audio, it is necessary to compress it to reduce the bandwidth, otherwise it may not be possible to obtain the desired QoS, especially over wireless networks. However, high level of compression makes the transmitted signal highly susceptible to errors, especially if the audio is being transmitted in real time. Any loss in the parametric information will severely degrade the quality of the reproduced stereo signal. Stereo is often used for music, and the slightest imperfection in music gets noticed by even non-experts.

Human ears are more sensitive to errors than human eyes. Human hearing faculties behave like differentiators, accentuating any variations, while

Figure 1. Relationship between aacPlus audio codecs v1 and v2 (Dietz & Meltzer, 2002)



human eyes behave like integrators, smoothening out variations (Sharda, 1999). Therefore, an audio stream should be given higher priority as compared to text or an image data stream.

Still Image Coding

Still image coding depends upon the type of image and its compression algorithm. Standards such as JPEG (Joint Photographic Experts Group), GIF (Graphics Interchange Format), and PNG (Portable Network Graphics) have dominated the field so far. JPEG is generally used for lossy compression of continuous tone images, such as photographs. GIF is a bitmap image format for pictures with 256 colours. PNG is a lossless bitmap image format. PNG improves upon the GIF format and is freely available.

The newer JPEG 2000 image compression standard uses a wavelet transform instead of the discrete cosine transform used in JPEG (Taubman & Marcellin, 2002). Therefore, JPEG 2000 can give higher compression ratio without generating the blocky and blurry artefacts introduced by the original JPEG standard. It also allows progressive downloads to extract various image resolutions, qualities, components, or spatial regions, without having to decompress the entire image. Distortion performance is also improved over the original JPEG standard, especially at low bit rates and at extremely high quality settings. JPEG 2000 is more error resilient as compared to the original JPEG standard (Secker & Taubman, 2004). This makes JPEG 2000 much better suited for applications requiring image transmission over wireless networks, as errors and delays introduce fewer observable artefacts in the displayed image.

Wireless networks experience higher error rates, and have lower bandwidth. Therefore, they are more severely challenged when transmitting digital images. Since the JPEG 2000 standard provides higher compression ratios, it is more

suitable for the low bandwidth wireless networks; however, some additional issues need to be addressed (Santa-Cruz, Grosbois, & Ebrahimi, 2002). Issues such as error resilience over wireless networks are being addressed by the JPEG 2000 Wireless (JPWL) team. Their aim is to standardise tools and methods for efficient transmission of JPEG 2000 images over error-prone wireless networks. One of the techniques being developed by JPWL make the code stream more error resilient by adding redundancy, or by interleaving data (Dufaux & Nicholson, 2004). The decoder not only detects errors, but also corrects some, where possible. Another technique changes the sensitivity of different parts of the code stream to errors. More sensitive sections of the code stream are more heavily protected than the less sensitive sections. The third technique describes the locations of the remaining errors in the code stream; the decoder then uses this information to exclude the corrupted parts of the code stream from the decoding process.

By standardising these techniques in JPWL, JPEG 2000 is being made more resilient to transmission errors, making it an ideal choice for the transmission of digital images and video over wireless media.

Moving Image Coding

Moving image coding can entail storing up to 20-30 image frames in every second. This demands very high bandwidth for high quality uncompressed video. Development of video coding standards that provide low resolution and low frame rate video suitable for transmission over networks began with the H.261 standard published by the ITU (International Telecom Union) in 1990, with data rates in multiples of 64 Kbps. The H.263 version provided a replacement for H.261 (in 1995) to work at all bit rates. It was further enhanced as H.263v2 (in 1998) and H.263v3 (in 2000). H.263 is similar to H.261, with improved performance and error recovery, and supports CIF,[1] QCIF, SQCIF, 4CIF, and 16CIF images. As these standards are designed for multiples of 64Kbps rates these are sometimes called px64 (where p can be 1-30). Originally these data rates were expected to suit ISDN (integrated services digital network) lines, nonetheless, these standards are useful in transmitting video over other wire-line and wireless networks also. H.263 is the baseline standard for the new 3G-324M standard, which targets the 3G wireless networks (Smith & Jabri, 2004).

Another option within the 3G-324M specification is the next generation video coding standard MPEG-4 AVC. It was approved in 2003 and called MPEG-4 AVC or ITU-T H.264, or simply advanced video coding (AVC).

MPEG-4 AVC doubles the compression efficiency of earlier standards for the same picture quality, which leads to 50% lower bandwidth (Navakitkanok & Aramvith, 2004). Therefore, it is far better than the earlier standards for wireless transmission. It offers improved resilience to transport errors, improved bit rate scalability, and stream switching for transmission over less reliable network infrastructure, such as wireless networks.

Motion JPEG 2000 (like Motion JPEG) can perform video compression applying only intra-frame compression. This makes Motion JPEG 2000 well suited for video transmission over wireless networks. It has been demonstrated that Motion JPEG 2000 outperforms MPEG-4 in terms of coding efficiency, error resilience, complexity, scalability, and coding delay (Tabesh, Bilgin, Krishnan, & Marcellin, 2005).

The JPWL work has taken into consideration the general principle underpinning networking protocols, with particular attention given to 3G networks (3GPP/3GPP2), wireless LANs (WLAN based on the IEEE 802.11 standards family), and Digital Radio Mondiale (DRM), making motion JPEG 2000 particularly suitable for wireless networks.

Multimedia Applications

Applications which have so far been bound to wire-line networks and desktop computers, now want to be let loose. The only option is to use wireless networks and portable devices. Areas for such mobile multimedia applications include both personal and business communications. E-learning, marketing, travel, and tourism are just a few of the burgeoning application areas that can make good use of mobile multimedia systems. Some of the potential killer applications based on the JPEG 2000 Wireless (JPWL) methods include video streaming and video conferencing (Liu & Choudary, 2004).

Mobile systems offer new opportunities and challenges as they become capable of transmitting multimedia information. Such applications need to transmit not only the core information, but also some associated meta-information. Most electronic systems use multi-tier information transmission processes, which include: intimation of arrival (bell, ring, beep, and vibrate); abbreviated information (subject, caller ID); textual information (text message, SMS); multimedia information, and meta-information for layered retrieval of the information.

How a particular information type and associated meta-information is used depends upon the application, the user preference, user device, and the required QoS (Cheng & Shang, 2005).

Text Applications

Text is very useful for communication. It is often said that a picture is worth a thousand words; nonetheless, we should not forget that a few well chosen words can be worth scores of pictures. Additionally, text requires much lower bandwidth, and has greater certainty of meaning. It is more reliable in the face of transmission errors, especially if we use either reverse or forward error correction protocols.

Text is easy to transmit asynchronously or synchronously. One can send an SMS to a friend during work, without the fear of disturbing her in an import meeting. The receiver can reply in her own time, or the two can engage in a brief chat session to fix their evening rendezvous. The runaway success of SMS follows the predicate that "brevity is the soul of wit," as SMS allows succinct messages that convey the meaning quickly. Coded messages based on SMS have also become prevalent, further reducing the time taken to enter and read the message.

Some commonly used SMS codes include: *ATB—All the best; BRB—Be right back; GR8—Great; LUV—Love; PCM—Please call me; TTYL—Talk to you later; 2DAY—Today; and WER R U—Where are you?* SMS codes are also being used to download information to mobile phones, such as snow photos to check the condition on ski slopes.

In Japan, codes called Emoji have been developed. These are colourful, and often animated inline graphics used for mobile messaging. However, these are not standardised or interoperable between carriers. Emonji's are treated as characters, and each carrier has its own set.

In conclusion, text or text-like messages are, and will remain, an important aspect of mobile communications, especially because these are inexpensive, highly expressive, and are least problematic with respect to delivery with the desired QoS.

Audio Applications

Transmission of voice was the original motivation for developing the mobile communications technology. However, digital radio is also coming online, and integrating digital radio in mobile phones is in the offing. An Austrian company Livetunes has developed UMTS-enabled handset with digital radio. SIRIUS Satellite Radio can transmit commercial-free music and other audio

entertainment to cars and homes. Mobile audio commercials over such digital radio channels allow advertisers to send audio commercials to their customers' mobile phone. The customer receives a phone call; upon answering the call, the audio commercial is played. It can include new offers, promotions, and announcements. To avoid spamming, companies have to provide their own subscriber database and the audio clip.

The QoS requirements for audio are different for bi-directional conversation than those for uni-directional digital radio transmission. For digital radio, buffering can be used to remove any delay jitter; however, excessive buffering can add unacceptable delay to conversational applications (Sharda, 1999).

Human hearing is very sensitive to any distortion in audio. For conversational audio, we can tolerate some errors, as long as the meaning of the spoken words is clear. If there is a problem in understanding the meaning, then the listener can always ask the speaker to repeat what was said. This is like reverse error correction working at the highest communication layer, that is, the user layer. However, this cannot work for stereo music; as human hearing works like a differentiator, and any distortion gets accentuated. Furthermore, our hearing is capable of picking slightest variation between the two channels of stereo music. Therefore, QoS issues are very important when high quality stereo music is transmitted, but not so important for conversational audio.

Still Image Applications

Applications needing still image transmission can use multimedia messaging service (MMS). Examples of MMS based applications include: weather reports giving images, stock prices displayed as graphs, football goals displayed as a slide show, and many more. An extension of still image transmission is animated text messages. The main QoS factor that affects still images is delay. As delay jitter does not effect still image transmission, any errors can be overcome by using reverse error correction protocols. If such additional delays are not acceptable, then images can be displayed with errors. Uncompressed images can tolerate a high level of errors; however, the ability to tolerate errors reduces for compressed images. The original JPEG type compression techniques lead to blocky images when errors occur, as they use discrete cosine transform. However, JPEG 2000 compression standard overcomes this problem by using wavelet transform. Images compressed with JPEG 2000 degrade "gracefully" in face of errors. As wireless communication systems are inherently more error prone, image-based applications will benefit from the use of JPEG 2000 standard for their compression (Dufaux & Nicholson, 2004).

Content repurposing is also becoming important, so that the content creator can compile content only once, and the system can vary image size and resolution depending upon the display screen size and the communications channel bandwidth (Rokou & Rokos, 2004)

The aim of content repurposing is to push the content with the most appropriate resolution, so that it can be transmitted over the available network to meet the QoS goals. In general, this would imply pushing lower resolution images over wireless networks. However, with JPEG 2000 and JPWL, the system can push a rough image to begin with, which keeps improving as more data bits are transmitted.

Video Applications

Video phones are a natural extension of the current audio telephony. Wire-line based video phones were demonstrated decades ago, however, these never became popular. Mobile video telephony is likely to become popular once the cost of transmitting acceptable quality video becomes affordable. In the meanwhile, look-at-this (LAT) applications

will generate demand for mobile Internet and 3G wireless networks, as these will create large amount of real-time mobile video. Some possible LAT application areas include:

a. **Retail:** Before purchasing an item, the consumer sends an image of the item to their partner for comment or approval.
b. **Real Estate:** An agent sends images of the building and its surrounding areas to the prospective customer.
c. **General Business:** A worker sends live video to colleague(s) at other location(s) while holding a voice conversation. This could be applied to developing new ideas; designing new products; repairing faulty equipment; maintaining, installing, or inspecting a system.

QoS is of great importance in video transmission. Video conferencing is the most challenging multimedia application for transmission over mobile systems. Much effort has gone in to migrating from 2G networks to 3G networks to provide the desired QoS for video transmission. However, the cost of transmission is still high enough for it to be an impediment in its large-scale adoption.

Mobile Communication Systems

The desire to communicate over long distances has been an innate need for human beings since time immemorial. We can reflect that the earliest telecommunications systems devised by human beings were wireless systems, namely, smoke signals, semaphore flags, drums, and yodelling and so forth. Therefore, it is not surprising that electronic communications are also moving towards wireless systems.

Evolution of Telecommunications

Electric telecommunications began with the telegraph demonstrated by Morse in 1837 and the telephone developed by Bell in 1876. Marconi began his experiment with radio transmission in 1895. Automation of circuit switching systems began in 1919 with the Strowger exchange. The era of satellite communications dawned with the Telstar satellite in 1950. Saber became the first major data network in 1962.

Evolution of Mobile Systems

An early landmark in the development of wireless communications was the patent for the spread spectrum concept, proposed in 1941 by Hedy Lamarr. The first mobile telephone service was setup in St. Louis by AT&T as far back as 1946. Some theoretical breakthroughs also occurred around this time. In 1948, Claude Shannon published the Shannon-Hartley equation, and in 1949 Claude Shannon and Robert Pierce develop the underlying concepts for CDMA (code-division multiple access). In 1950, Sture Lauhrén made the world's first cellphone call, and by 1956, Swedish PTT Televerket operated a mobile telephone service. In 1969, the Nordic Mobile Telephone Group started a mobile service. CDMA was deployed for military systems in the 1970s. In 1973, Motorola vice presidents Marty Cooper and John Mitchell demonstrated the first public call from a handheld wireless phone.

Evolution of Digital Mobile Systems

First Global System for Mobile Communications (GSM) technology based networks were implemented by Radiolinja in Finland in 1991. In 1992, the Japanese Digital Cellular (JDC) system was introduced. By 1993, the IS-95 CDMA standard got finalised. First meetings of the 3GPP (3rd Generation Partnership Project) Technical Specification Group was held in December 1998. In 2000, Siemens demonstrated the world's first 3G/UMTS (3rd Generation Universal Mobile Telecommunications System) call over a TD-CDMA (time division-CDMA) network.

In 2000, commercial GPRS (general packet radio service) networks were launched. These networks supported data rates up to 115 Kbps, as compared to GSM systems with 9.6 Kbps data rates. In 2001 NTT (Nippon Telegraph & Telephone Corp.) produced commercial WCDMA (wide-band CDMA) 3G mobile network. In 2003, Ericsson demonstrated the transmission of IPv6 traffic over 3G UMTS using WCDMA technology.

Fixed Wireless vs. Mobile Communications

We need to distinguish between fixed wireless communication systems and mobile communication systems. A mobile communication system frees the end systems from the tyranny of being connected to a wall socket, and provides the ability to communicate anytime and anywhere. It allows the freedom to roam outside the home or the office.

Fixed wireless communication systems are local alternatives to wired communication systems. These do not provide mobility outside the home or the office, nonetheless, they provide a cost effective telecommunications connection for a given location with the ability to move around within a specified boundary.

For remote locations, satellite-based communication systems may be the only means of establishing a connection; however, these can be expensive. Satellite connections add about half a second round trip delay, making full-duplex audio or video connections rather difficult. While there is appreciable delay in a satellite connection, the delay variance is not very high, as the number of hops is fixed at two—transmitter to the satellite, and satellite to the transmitter. Error rates can be high on a satellite connection, especially burst errors—in case of atmospheric disturbances.

Universal Mobile Telecommunications System (UMTS)

Today, there are more than 60 3G/UMTS networks using WCDMA technology. Over 25 countries have adopted this technology, and there is a choice of over 100 terminal designs in Asia, Europe, and the U.S. The 3G mobile technologies identified by ITU for 3G/UMTS offer broadband capabilities to support a large number of voice and data customers, and offer much higher data rates at a lower incremental cost than the 2G technologies (Myers, 2004).

One of the issues driving the development and proliferation of 3G technologies is the recognition that there is a need for guaranteeing the QoS for multimedia traffic. Without guaranteed QoS, many applications fail to perform as per the users' expectations. Until the users are confident of getting the quality they need for running mobile multimedia applications effectively, they will not shift from their current mode of operation and adopt the new wireless networking technologies for multimedia information transmission.

QUALITY OF SERVICE IN MOBILE SYSTEMS

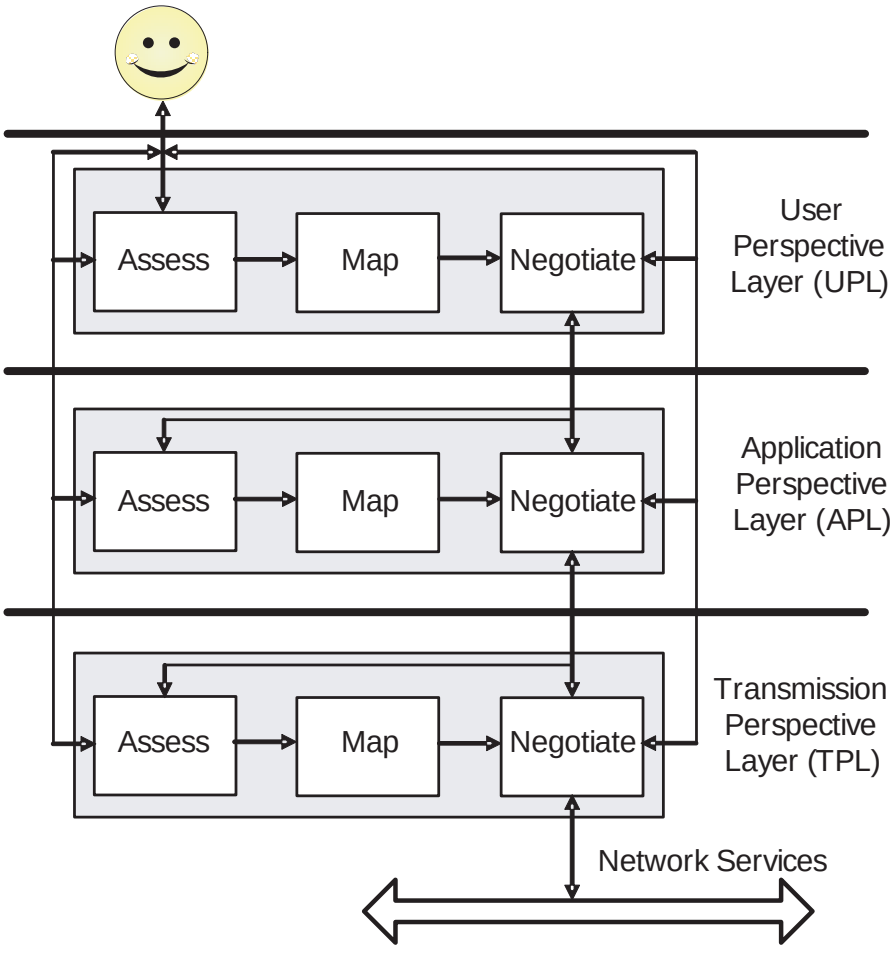
This section gives an overview of the various approaches being trailed for the provision of QoS in mobile networks. While much work has been done in providing QoS guarantees at the network infrastructure level, a holistic approach to providing end-to-end QoS has been missing to some extent. We begin by presenting a QoS model that focuses on the user, and develops a methodology for allowing the user to negotiate with the system to find a compromise between cost, quality, and temporal aspects.

QoS Concepts and Models

The three layer quality of service (TRAQS) model shown in Figure 2 comprises three layers for QoS management in multimedia communications (Sharda & Georgievski, 2002). These three layers are: user perspective layer, application perspective layer, and transmission perspective layer. Each layer performs QoS processing for a set of QoS parameters that are related to the specific perspective. The main functions of the three perspective layers are:

- **User Perspective Layer (UPL)** interacts and performs QoS negotiations with the user and then transfer the QoS request to the APL.
- **Application Perspective Layer (APL)** first assesses the QoS request received from the UPL, and aims to satisfy the needs of the

Figure 2. Three layer QoS (TRAQS) model (Sharda, 1999)



multimedia application by requesting the required services from the TPL.

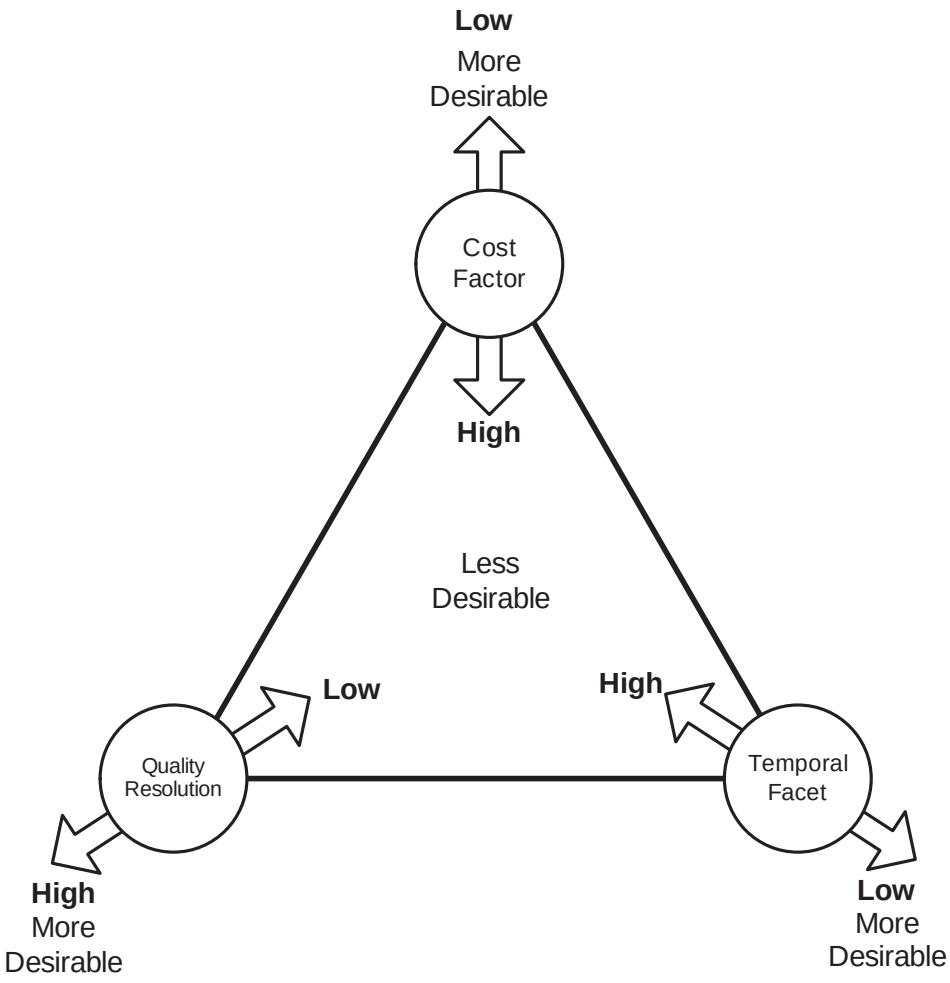
- **Transmission Perspective Layer (TPL)** is responsible for negotiating with the network infrastructure to obtain appropriate communication services that can guarantee QoS.

Similarly, the various QoS protocols developed at the network infrastructure level need to be able to communicate with the TPL to allow the user to specify the desired compromise between cost, quality, and temporal issues such as delay and jitter.

Quality, Cost, Temporal Triangle (QCTT)

In purchasing any goods or services one needs to find a compromise between three important factors: cost, quality, and time. While one would

Figure 3. Quality, cost, temporal triangle (QCTT) model (Georgievski & Sharda, 2005a)



like to get the best quality at the least cost and in the shortest time, in practice, this is not possible. One must strike a compromise between these three factors. So far mobile communications systems have not come to grips with this reality. Future communication systems must provide users the ability to specify what quality and temporal aspects (such as delay and jitter) they want, and then systems should respond with the cost it would charge to provide that quality.

Only with a differentiated cost can the telecommunications service providers afford to deliver the required QoS. If the cost is too low, the network may be overwhelmed with traffic, and none of the users can then obtain the desired QoS. Over and above this, the network services provider may not be able to make profit. On the other hand, if the cost is too high, there will not be enough consumers using the service, once again making it difficult for the service provider to get return on investment. Whereas, by providing the user the ability to negotiate, the consumer and the service provider can both have a win-win situation; some consumers pay high cost as they need higher quality, while other consumers can pay lower cost, as they have lower QoS requirements.

In the following sections, we first explain the concepts involved in the quality, cost, temporal triangle (QCTT), and then present an implementation of the same (Georgievski & Sharda, 2005a).

The three performance aspects—quality, cost, and time—are bound by a tri-partite dependency and thus can be modelled as a triangular relationship, as shown in Figure 3. The QCTT model embodies an inherent restriction on the delivery of QoS, that is, it is possible to achieve the more desirable parameter values only for two of the three performance aspects, while the third aspect must be forced to the less desirable value (Georgievski & Sharda, 2005b).

For example, if a user chooses to have high quality resolution (e.g., large image size, high frame rate), and, the more desirable, low temporal facet (e.g., low delay and jitter), then the cost factor has got to be high. By embedding the quality, cost temporal (QCTT) model in a user interface, we can provide the ability to dynamically manage QoS even while a multimedia session is in progress.

A multimedia communication session first needs to enter static QoS specifications, and then carry out dynamic QoS management as the session proceeds. An interface based on the QCTT model provides the ability to dynamically manage QoS. Such interfaces are described in the following sections.

Static QoS Specification

Figure 4 shows the user interface developed for negotiating static QoS prior to initiating a multimedia communication session. Using this interface, the user is able to specify the desired QoS, and then interactively negotiate with the system. It uses intuitive GUI elements such as a four colour system, a user status response, and a system status signalling system. These GUI elements allow the user to request the desired QoS, and get feedback if the network can deliver the same (Georgievski & Sharda, 2005a).

Dynamic QoS Management with QCTT

A dynamic QoS management interface is shown in Figure 5. This interface uses the QCTT model for re-negotiating QoS while a communication session is taking place. This is achieved by using three GUI elements: three sliders, buttons, and pivot point displacement. The system feedback GUI elements include: system QoS provision ring and values, and QCT threshold line (Georgievski & Sharda, 2005a).

To specify the desired QoS, the user moves the pivot point in the QCT triangle to a location which indicates the desired values for quality, cost, and temporal parameters.

The system provides visual feedback as follows:

Figure 4. Static QoS negotiation user interface (Georgievski & Sharda, 2005a)

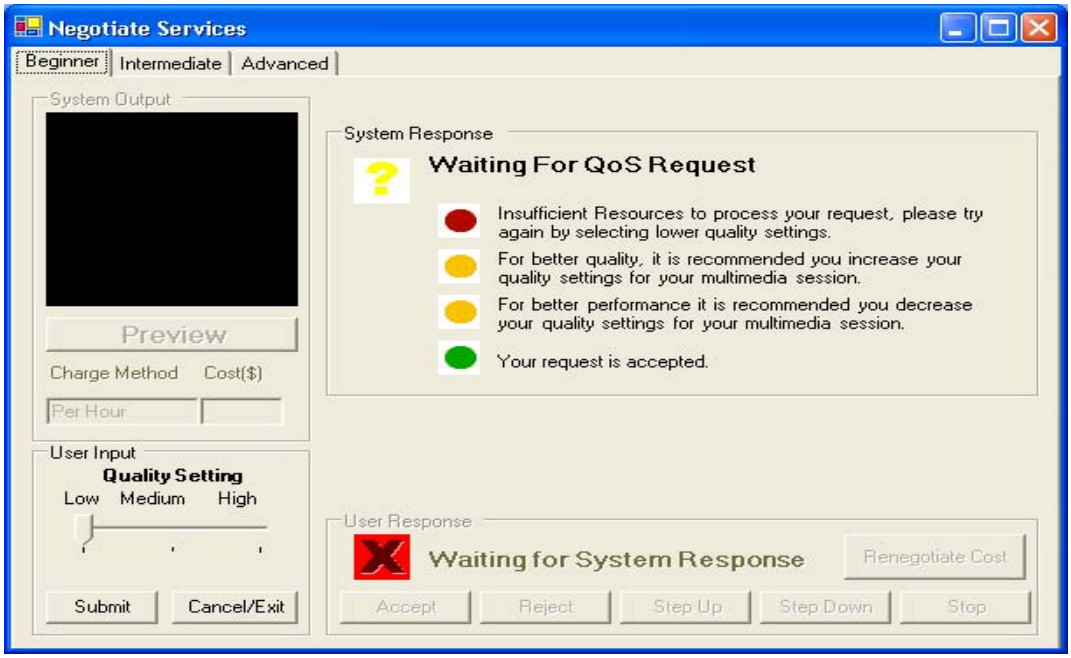
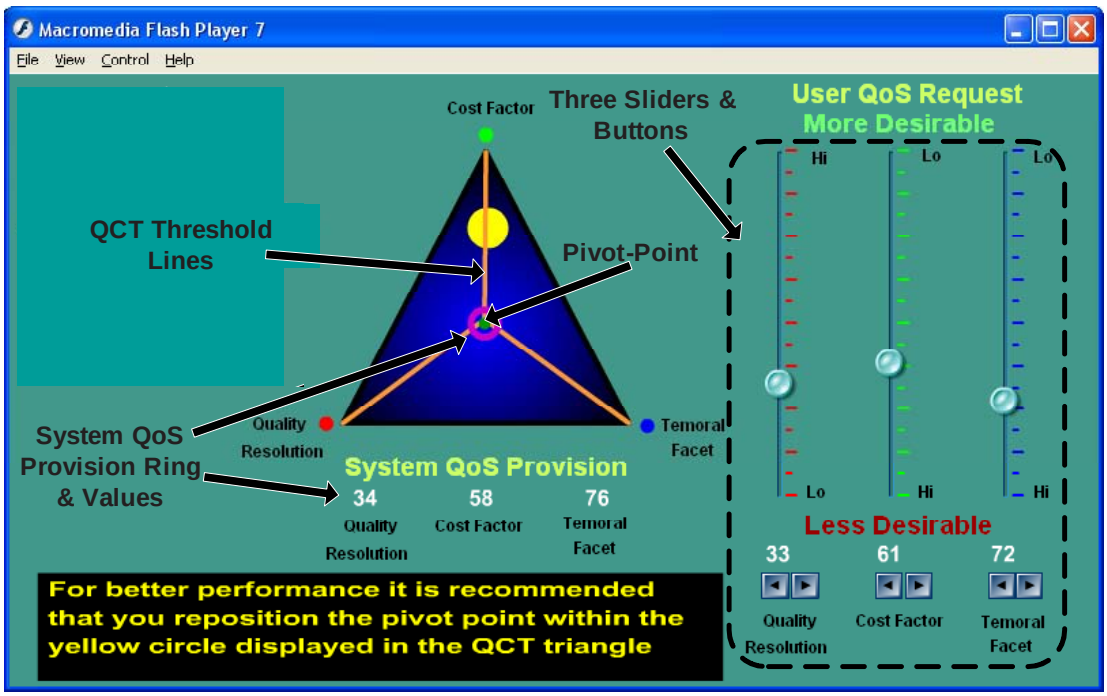


Figure 5. Dynamic QoS management user interface (Georgievski & Sharda, 2005a)



1. **QoS Provision Ring** displays the current QoS parameter values that the system is able to provide.
2. **QoS Provision** values display the current numerical values set for QoS parameters.
3. **QCT Threshold Line** uses a three-colour scheme to provide feedback for displaying desirable and non-desirable values for each aspect.

This system has been tested and a usability analysis has been carried out on the same. While some improvements have been stipulated in the current implementation, overall, it received good assessment from the users (Georgievski & Sharda, 2005a).

Quality of Service on the Move

The ability to provide the requested QoS while roaming will be an important aspect in differentiating various mobile operators. This will determine their ability to hold on to their customers and therefore their revenue stream.

Providing QoS to a customer on the move is highly complex. Factors such as continuous handover, variable quality, dropout, and environmental factors make delivery of consistent QoS highly problematic. QoS provisioning has three main aspects: (1) resource reservation, (2) QoS routing protocol, and (3) Call admission control policy.

The integrated services (IntServ) framework developed under the RFC 1633 aims to provide customised QoS to individual applications (Aggélou, 2003). This is based on two aspects:

1. **Resource Reservation:** Each router needs to know the amount of buffer space and link bandwidth it needs to reserve for a session.
2. **Call Admission:** Each router determines the resources already committed to current sessions it is serving, before accepting the request from a new session.

QoS routing is the most important protocol for mobile networks, the main objective specified for this protocol in the RFC 2386 are:

1. **Dynamic Determination of Feasible Paths:** This is based on policy and cost constraints.
2. **Optimisation of Resource Usage:** This requires state-dependent routing schemes.
3. **Graceful Performance Degradation:** This aspect compensates for transient inadequacies using the state-dependent routing scheme.

In summary, a mobile network needs the ability to reserve resources, ensure that a new call is admitted only if enough resources are available, choose the most suitable path to optimise the utilisation of resources, and provide graceful degradation in performance as resources become overloaded.

Quality of Services in Mobile Ad-Hoc Networks

Mobile ad-hoc networks are becoming an important area of investigation. As routing paths are not fixed in an ad-hoc network, QoS routing becomes an even more dynamic problem (Aggélou, 2004). In any ad-hoc network, a variety of routes with differing node capacity and power may be available to transmit data to the destination.

In general, not all routes are capable of providing the required QoS to satisfy the needs of the mobile users. Even when a route is selected that initially meets the user requirements, its error characteristics will not remain constant with time, due to the dynamic nature of routing and node placement in mobile ad-hoc networks. Therefore, ongoing re-routing will be required in an ad-hoc mobile network.

MOSQUITO: Mobile Quality of Service Provision in the Multi-Service Network

The MOSQUITO project, at the University College London, explored a microeconomic approach to resource allocation for providing QoS over multi-service network.

In this protocol, a base station sells bandwidth and QoS guarantees in small auctions to mobile terminals. A simple price setting / bidding function is used to determine the outcome of the auction. This research project aims to explore if:

- Microeconomics can be used for resource allocation.
- The performance of such a system can be measured.
- The algorithm creates a stable system or a chaotic one.
- Chaos can be characterised and controlled.

To use microeconomics for QoS provisioning, such questions need to be answered. Additionally, pricing functions need to be established using some simplifying assumptions; because, without simplifying heuristics, the juxtaposition of a myriad of factors such as pricing, routing, and quality selection will make real-time negotiations impossible.

FUTURE DIRECTIONS

There is no doubt that the future is heading towards mobile communications. And multimedia information will increasingly become the main traffic being transmitted, or blocked, on these networks. One solution to this problem is the so called "brute force" method: that is, "throwing" more bandwidth at the multimedia applications. However, experience shows that as more resources are made available without a viable economic model, the system ultimately gets overloaded. Therefore, developing QoS systems that provide the user with the ability to negotiate with the network infrastructure are going to be of paramount importance.

Some of the developments in this area point to the following:

1. In the near future, mobile computing and communication systems will suffer from low bandwidth and low performance due to battery limitations.
2. Increasingly, mobile systems will provide higher bandwidth and combine different wireless technologies, such as high performance local wireless networks and wide area networks.
3. Third generation mobile systems will combine IP-traffic with traditional voice traffic.

The next generation of mobile networking technology is called 4G, or "3G and beyond" by IEEE (Aggélou & Tafazolli, 2001). In Japan, NTT DoCoMo is conducting tests under the 4G banner for 100 Mbps speeds with moving terminals, and 1 Gbps for stationary terminals. The first commercial release by NTT DoCoMo is expected in 2010. This technology aims to provide on demand high quality video and audio. 4G will use OFDM (orthogonal frequency division multiplexing), and also OFDMA (orthogonal frequency division multiple access) to better allocate network resources, and service multiple users simultaneously. Unlike the 3G networks, which use both circuit switching and packet switching, 4G will use packet switching only. Additionally, many QoS issues will be handled by developing new protocols. Nonetheless, the author contends that providing the ability to negotiate a compromise between cost, quality, and temporal aspects will remain an important issue.

CONCLUSION

Transmission of multimedia information over mobile networks is becoming increasingly important. New applications are in the offing if such multimedia information can be transmitted with the desired QoS. Text and still images do not pose much problem when transmitting these over mobile networks, as delay and delay variance do not adversely effect the operation of applications using text or still images. JPEG 2000 standard provides a marked improvement over the current standards such as JPEG, GIF, and PNG for still image transmission. Audio and video transmission, especially for full-duplex applications requiring real-time operation, poses the most demanding requirements for providing the desired QoS. While 3G networks, and 4G networks of the future, are capable of providing the required infrastructure for delivering multimedia content with the desired QoS, their user interfaces need to provide the ability to strike the desired balance between quality, cost, and temporal aspects.

ACKNOWLEDGMENTS

The author would like to thank Dr. Mladen Georgievski for his useful suggestions and other contributions towards the preparation of this chapter.

REFERENCES

Aggélou, G. (2004). *Mobile ad hoc networks.* New York: McGraw-Hill Professional.

Aggélou, G., & Tafazolli, R. (2001). QoS support in 4th generation mobile multimedia ad hoc networks. *Proceedings of the Second International Conference on 3G Mobile Communication Technologies,* London, March 26-28 (pp. 412-416). London: Institute of Electrical Engineers.

Aggélou, G. N. (2003). An integrated platform for quality-of-service support in mobile multimedia clustered ad hoc networks. In M. Ilyas (Ed.), *The handbook of ad hoc wireless networks* (pp. 443-465). Boca Raton, FL: CRC Press, Inc.

Cheng, A., & Shang, F. (2005). Priority-driven coding of progressive JPEG images for transmission in real-time applications. *11th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA'05),* Hong Kong, August 17-19 (pp. 129-134). Washington, DC: IEEE Computer Society.

Dietz, M., & Meltzer, S. (2002, July). CT-aacPlus: A state-of-the-art audio coding scheme. *EBU Technical Review,* (291), 1-7. Retrieved from http://www.ebu.ch/en/technical/trev/trev_291-dietz.pdf and http://www.ebu.ch/en/technical/trev/trev_index-digital.html

Dufaux, F., & Nicholson, D. (2004). JPWL: JPEG 2000 for wireless applications. Photonic devices and algorithms for Computing VI. In K. M. Iftekharuddin, & A. A. S. Awwal (Eds.), *Proceedings of the SPIE, 5558,* 309-318.

Georgievski, M., & Sharda, N. (2005a). Enhancing user experience for networked multimedia systems. *Proceedings of the 4th International Conference on Information Systems Technology and its Applications (ISTA2005*), Massey University, Palmerston North, New Zealand, May 23-25 (pp. 73-84). Bonn: Lecture Notes in Informatics (LNI), Gesellschaft für Informatik (GI).

Georgievski, M., & Sharda, N. (2005b). Implementation and usability of user interfaces for quality of service management. *Tencon'05: Proceedings of the Annual technical Conference of IEEE Region 10,* Australia, November 21-24. New Jersey: IEEE.

Liu, T., & Choudary, C. (2004). Content-aware streaming of lecture videos over wireless networks. *IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04),*

Miami, FL, December 13-15 (pp. 458-465). Washington, DC: IEEE Computer Society.

Myers, D. (2004). *Mobile video telephony.* New York: McGraw-Hill Professional.

Navakitkanok, P., & Aramvith, S. (2004). Improved rate control for advanced video coding (AVC) standard under low delay constraint. *International Conference on Information Technology: Coding and Computing (ITCC'04), 2*, Las Vegas, NV, April 5-7 (p. 664). Washington, DC: IEEE Computer Society.

Rokou, F. P., & Rokos, Y. (2004). Integral laboratory for creating and delivery lessons on the Web based on a pedagogical content repurposing approach. *Fourth IEEE International Conference on Advanced Learning Technologies (ICALT'04),* Joensuu, Finland, August 30-September 1 (pp. 732-734). Washington, DC: IEEE Computer Society.

Santa-Cruz, D., Grosbois, R., & Ebrahimi, T. (2002). JPEG 2000 performance evaluation and assessment. *Signal Processing: Image Communication, 17*(1), 113-130.

Secker, A., & Taubman, D. S. (2004). Highly scalable video compression with scalable motion coding. *IEEE Transactions on Image Processing, 13*(8), 1029-1041.

Sharda, N. (1999). *Multimedia information networking.* New Jersey: Prentice Hall.

Sharda, N., & Georgievski, M. (2002). A holistic quality of service model for multimedia communications. *International Conference on Internet and Multimedia Systems and Applications (IMSA2002),* Kaua'i, Hawaii, August 12-14 (pp. 282-287). Calgary, Alberta, Canada: ACTA Press.

Smith, J. R., & Jabri, M. A. (2004). The 3G-324M protocol for conversational video telephony. *IEEE MultiMedia, 11*(3), 102-105.

Tabesh, A., Bilgin, A., Krishnan, K., & Marcellin, M. W. (2005). JPEG2000 and motion JPEG2000 content analysis using codestream length information. *Proceedings of The Data Compression Conference (DCC'05),* Snowbird, UT, March 29-31 (pp. 329-337). Washington, DC: IEEE Computer Society.

Taubman, D., & Marcellin, M. (2002). *JPEG2000: Image compression fundamentals, standards and practice*. Netherlands: Kluwer Academic Publishers.

ENDNOTE

[1] CIF: Common Intermediate Format. A video format used in videoconferencing systems. It is part of the ITU H.261 videoconferencing standard, and specifies a data rate of 30 frames per second (fps), with each frame containing 288 lines and 352 pixels per line. Other CIF based standards include: QCIF - Quarter CIF (176x144), SQCIF - Sub quarter CIF (128x96), 4CIF - 4 x CIF (704x576), and 16CIF - 16 x CIF (1408x1152).

This work was previously published in Mobile Multimedia Communications: Concepts, Applications, and Challenges, edited by G. Karmakar and L. Dooley, pp. 45-63, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.21
Classification of 3G Mobile Phone Customers

Ankur Jain
Inductis India Pvt. Ltd., India

Lalit Wangikar
Inductis India Pvt. Ltd., India

Martin Ahrens
Inductis India Pvt. Ltd., India

Ranjan Rao
Inductis India Pvt. Ltd., India

Suddha Sattwa Kundu
Inductis India Pvt. Ltd., India

Sutirtha Ghosh
Inductis India Pvt. Ltd., India

ABSTRACT

In this article we discuss how we have predicted the third generation (3G) customers using logistic regression analysis and statistical tools like Classification and Regression Tree (CART), Multivariate Adaptive Regression Splines (MARS), and other variables derived from the raw variables. The basic idea reflected in this paper is that the performance of logistic regression using raw variables standalone can be improved upon, by the use for various functions of the raw variables and dummies representing potential segments of the population.

INTRODUCTION

An Asian telecommunication operator which has successfully launched a 3G mobile telecommunications network would like to make use of existing customer usage and demographic data to identify which customers are likely to switch to using their 3G network.

The objective of this competition was to develop a prioritization mechanism that will accurately predict as many current 3G customers as possible from the "holdout" sample provided. It also involved identifying the profiles of 3G customers that can be used in identifying potential 3G customers among the existing second generation (2G) base.

The competition organizers were provided with a sample of 24,000 mobile phone subscribers, out of which customer type was provided for 18,000 subscribers, 15,000 being 2G and the rest 3G. Around 250 variables describing call and usage-related information was provided for all of the 18,000 subscribers. A holdout sample of another 6,000 subscribers was provided with the same set of variables, but without the 2G/3G flag. The task was to accurately predict as many 3G customers as possible from the holdout sample.

The organization of the article is as follows: We discuss the methodology approach taken and the modeling techniques used to develop the logistic model. Then we discuss the model results and the cutoff we have selected to generate the predictions. Finally, we discuss an alternative approach that we have tried.

METHODOLOGY APPROACH MODELING METHODOLOGY

The modeling approach used for determining the 3G customers is a combination of logistic regression, CART, MARS, and other derived variables. The CART and MARS are modeling tools of Salford Systems. This combination is an improvement over the logistic regression model with raw variables only. The potential segments of the population are identified by CART, and potential splines for various important variables obtained by MARS are used along with the other variables. Logistic regression is used as the dependent variable is dichotomous (reference Hosmer W. David, Stanley Lemeshow: Applied Logistic Regression, Wiley, New York (1989) Chapter 1 Pages 8-10, Chapter 2 Pages 25-29). In addition, we have selected specific segments of some of the raw variables, which have very high or low event rates.

The variables obtained from CART are indicators of potential segments of the population. By potential segments, we mean segments of population with very high or low event rates. These indicators are used in the logistic model as independent variables. MARS, on the other hand, generates splines from variables, thereby capturing important segments of a variable. These splines, termed as basis functions, are then used in the logistic model. Some variables have a very high or low event rate in a particular range. We have analyzed these ranges and created segments to be used as independent variables in the model. The CART, MARS, and other derived variables, when included in the model, show a higher predictive power than what is obtained from the raw variables standalone.

Once the CART and MARS variables have been included, a stepwise logistic regression is used to reach an optimum model. The stepwise regression is used for the sake of parsimony as the number of variables (raw, CART, MARS, and derived variables combined) is large, thereby creating a scope of overfitting. Moreover, by using a stepwise procedure it is ensured that the variables in the model are all significant at the desired level. The variance inflation factors of each of the variables entering the model are scrutinized in order to prevent multi-collinearity.

MODELING TECHNIQUES USED: MISSING IMPUTATION AND OUTLIER TREATMENT

In order to prepare the population for building the model, missing values had to be imputed and outliers had to be smoothed out. Missing imputation is done on variables which have less than 70% missing values. Variables with more than 70% missing values are omitted. The respective medians of the variables are used for the missing imputation.

Outlier treatment is done for variables with high and/or low extreme values. Variables with maximum/99th percentile ratio greater than 5, and those with 1st percentile/minimum ratio greater than 5 are considered for outlier treatment. The outlier treatment is done using exponential smoothing. The higher outliers are treated as follows:

If observation > 99th percentile then observation = (99th percentile)$^{(4/5)}$* (observation)$^{(1/5)}$

The lower outlier is treated as follows:

If observation < 1st percentile then observation = (1st percentile)$^{(4/5)}$* (observation)$^{(1/5)}$

For variables with 1st percentile as 0 and minimum less than 0, all observations less than 0 are set to 0.

CART

The CART is used to build a classification tree on the population to find segments with very high and low event rates. The modeling population selected for building the CART tree is a 66.67% simple random sample without replacement of the overall population, on which missing imputation and outlier treatment have been done. The remaining 33.33% is used as a validation sample. The target variable is 0 if the customer is 2G and 1 if 3G. CART trees are built on this modeling sample (see Figure 1).

Figure 1. A typical snapshot of a CART run

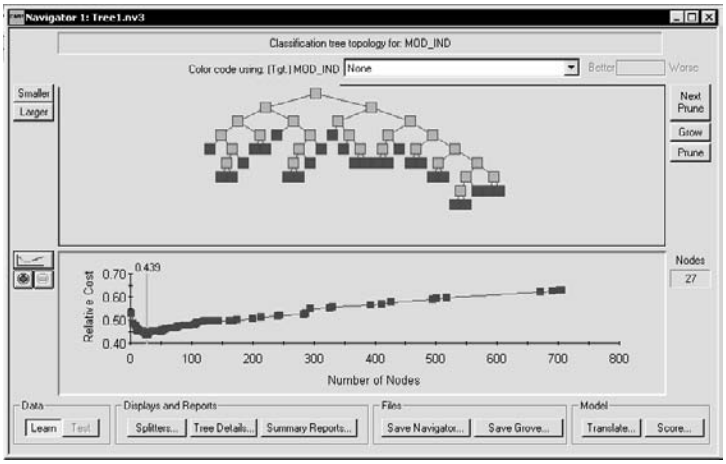
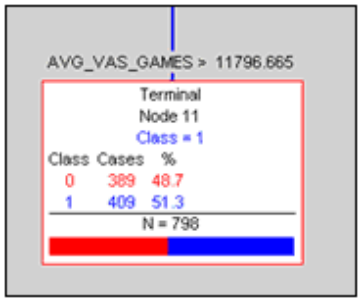


Figure 2. An example of a terminal node



Each node of the CART tree would give a segment of the population. The terminal nodes together would divide the population into a set of mutually exclusive and exhaustive segments (see Figure 2).

The modeling population has an event rate of 16.7%, but the population segment captured by this terminal node has an event rate of 51.3%. Terminal nodes with very high (as in the previous case) or very low event rates would be used in the logistic model as independent variables. For this purpose, indicator variables are created using the logic that CART has used to arrive at the terminal node. The translation generated by CART is a sas code that creates nodes and terminal nodes and is used to make these indicator variables.

MARS

MARS is used to create spline transformations of variables for use in the model, hence improving their accuracy. The core building block of a MARS model is the basis function transformation of a predictor variable X:

Basis function = Max(0, X-C) *where c is a constant discovered by the algorithm*

The value of this basis function is equal to 0 for all values of x up to a threshold c and equal to X-C for all values of x greater than c. The basis function defines a knot (c) where a regression changes slope.

The same modeling population that is used to build CART was used for MARS. The target variable was 0 if the customer is 2G and 1 if 3G.

A typical basis function would be

BF1 = max(0, HS_AGE - 4.000);

The basis functions are used in the logistic model as independent variables.

OTHER DERIVED VARIABLES

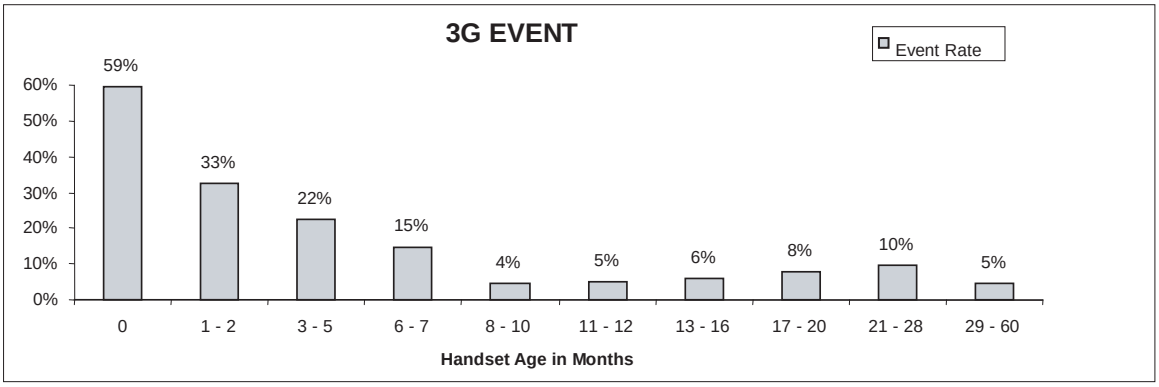
Some variables are derived by analyzing bivariate plots of various variables with the dependent variable. The plots are created by dividing the range of a variable into 10 to 20 bins and computing the event rates at each of those bins. Indicator variables are created for bins with very high or very low event rates. A typical example of such a derived variable is shown in Figure 3.

From the plot in Figure 3 we see that the event rate is very high when the handset age is 0 months. Hence, we can create an indicator variable for the 0 bin and use it in the model.

MODEL RESULTS MODEL PERFORMANCE

A logistic model is built on the modeling population to predict whether a given customer is 3G or 2G. The model consists of nine variables. All the variables are accepted at 99.95% level of significance. The concordance is becoming as high as 90.6%. The variables showed negligible multi-collinearity among themselves, which is reflected by the fact that the highest variance inflation factor is 1.74. The Ks statistic is becom-

Figure 3. Bivariate plot for handset age in months



ing 68% at the 25th percentile for the modeling population and 65% at the 20th percentile for the validation population. The Hosmer-Lemeshow statistic is 6.7852 with 0.56 p-value.

MODEL VARIABLES AND SIGNIFICANCE

The most important variable appearing in the model is the *handset model.* Three subsegments of the population with a high 3G rate appeared in the model, of which, two were becoming the most important variables, in terms of percentage variation explained. Subsegments with a low average billing amount over the last 6 months, low variation in usage of games, and low handset age are also becoming important. The ongoing hypotheses and results for the variables entering the model are stated as follows:

Handset model. Customers who have bought 3G-enabled handsets have a higher chance of making a 3G connection. Indicator variables for three segments based on handset age are entering the model. All of these variables show a high positive effect on 3G enrollment and together they capture 53% of the variance in the model.

Low average bill amount over the last 6 months. Customers with lower usage have a lower chance of making a 3G connection. This variable shows a small negative effect on 3G enrollment and captures 11% of the variance in the model.

Low handset age. Customers who have bought a handset very recently have a higher chance of making a 3G connection. As discussed in section 2.2.4, low handset age shows a positive effect on 3G enrollment and captures 10% of the variance in the model.

Low variation in usage of games. Customers who are not used to accessing games on mobiles have a lower chance of making a 3G connection. This variable shows a very small negative effect on 3G enrollment and captures 8% of the variance in the model.

High number of retention campaigns in the last 6 months. Customers who have received a higher number of retention campaigns have

a higher chance of making a 3G connection. This variable shows a negative effect on 3G enrollment and captured 7% of the variance in the model.

Segments with a high usage of games and subscription plans with a high 3G rate. Customers accessing games on mobiles and those that have certain subscription plans have a higher chance of switching to the 3G connection. The indicator variable created for this segment shows a considerably higher positive effect on 3G enrollment and captures 6% of the variance in the model.

Low variation in usage of GPRS. Customers who do not use GPRS are less likely to use the 3G connection. This variable shows a very small negative effect on 3G enrollment and captures 5% of the variance in the model.

MODEL CUTOFF

The model accuracy is maximized at a 90th percentile cutoff, but it captures only around 53% of the 3G customers. There is a steep increase in sensitivity up to the 75% percentile of the predicted probabilities, after which it plateaus. Specificity falls steeply after the 80th percentile of the predicted probabilities. The sensitivity, specificity, and accuracy are shown in Figures 4 to 6.

From Figures 4, 5, and 6 we find that the best results for specificity, sensitivity, and accuracy are obtained in the 75th - 80th percentile of the scores. Given the fact that the event rate is 16.7% we choose to go with the 80th percentile. Hence, customers who are falling in the top 20% of the model score have been assigned as potential 3G customers. The lift at the top of the 20% cutoff is becoming 76% for the modeling population and 74% for the validation population.

CONTINGENCY TABLE OF THE CUTOFF

The modeling population consists of 11,955 records. Based on the top 20% score cutoff we get 481 3G customers misclassified as 2G, 867 2G

Figure 4.

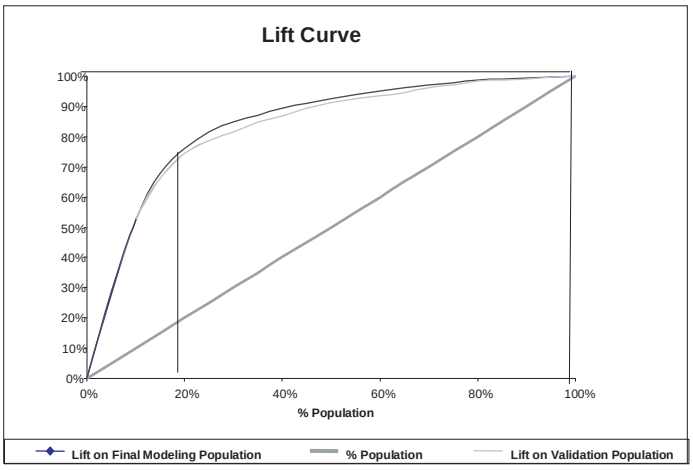


Figure 5.

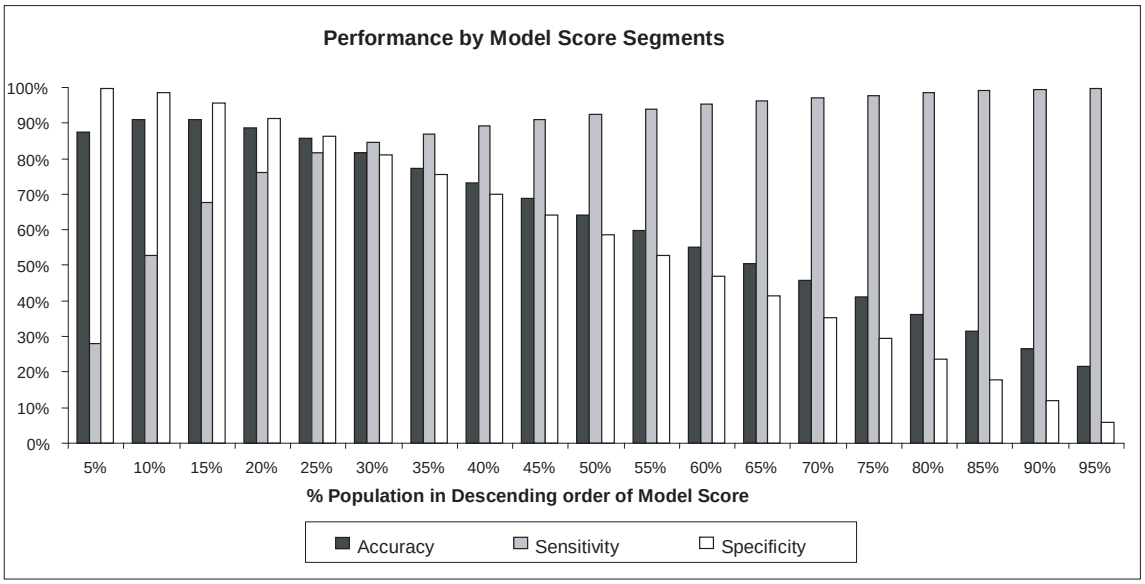
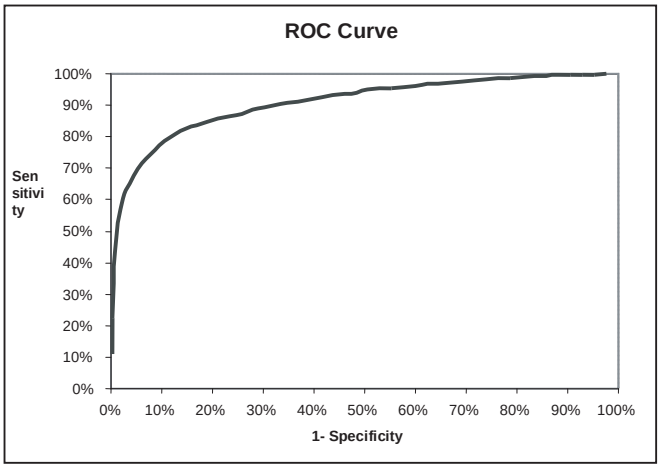


Figure 6.



customers misclassified as 3G, 9,088 customers correctly classified as 2G and 1,519 customers correctly classified as 3G. The predictions on the modeling population have an accuracy of 89%, a sensitivity of 76%, and a specificity of 91%. The definitions of accuracy, sensitivity, and specificity are shown in Table 1.

Table 1.

	Actual Event (1)	Actual Non-Event (0)
Predicted Event (1)	True Positive (TP)	False Positive (FP)
Predicted Non-Event (0)	False Negative (FN)	True Negative (TN)

Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)
Accuracy = (TP + TN) / (TP + TN + FP + FN)

The validation population consists of 6,045 records. Based on the same score cutoff as the modeling population, we get 266 3G customers misclassified as 2G, 434 2G customers misclassified as 3G, 4,611 customers correctly classified as 2G, and 734 customers correctly classified as 3G. The predictions on the validation population have an accuracy of 88%, a sensitivity of 73%, and a specificity of 91%.

COMPARISON WITH MODEL USING RAW VARIABLES

The model using CART dummies, MARS basis functions, and other derived variables along with the raw variables (MODEL 1) is performing better than the best model obtained using raw variables standalone (MODEL 2).

MODEL 2 is a 10-variable model with a concordance of 84.3% where as MODEL 1 is a 9 variable model with a concordance of 90.6%. Moreover, MODEL 1 gives a lift of 76% for the modeling population, as compared to a 63% lift achieved from MODEL 2.

ALTERNATIVE APPROACH

An alternative approach is tried while building the model after creating all the derived variables that are discussed in earlier sections. Instead of building a single logistic model on the entire modeling population, various subsegments of the population are taken and a separate logistic model is built on each of them. The ongoing hypothesis behind this approach is that the population may show more homogeneity within proper subsegments and hence a model based on a particular subsegment may have more predictive power as compared to the overall model.

The subsegments are derived using CART. The final segmentation chosen for this approach is based on average billing amount in the last 6 months. Based on this variable the modeling population is segmented into two parts. One part (PART A) contains 44.5% of the population and has a 3G event rate of 28.46%, and the other part (PART B) contains 55.5% of the population and has a 3G event rate of 7.32%.

The model built on PART A consists of 9 variables and had a concordance of 88.30%, whereas the model built on PART B consists of

10 variables and had a concordance of 87.20%. The combination of these two models gives an overall lift of 75.05% at a top 20% cutoff on the modeling population. The overall summary of classification is as follows:

Accuracy = 88%
Sensitivity = 75%
Specificity = 91%

The combined model using this approach shows a slightly inferior predictive power as compared to the overall model. Hence, we have decided to use the overall model.

CONCLUSION

In this article we have discussed the various techniques used to improve upon the predictive power of a logistic regression model by implementing various types of variables.

We have discussed how the raw data is cleaned by imputing missing values and treating for lower and upper outliers. We have then discussed how to create indicator variables for potential segments of the population that are captured by CART and use these variables in the logistic model. We have also discussed how basis functions derived from MARS are used in the model. Furthermore, we have other derived variables that are obtained by analyzing the bivariate plots of various important variables with the dependent variable.

After this, we have discussed the ongoing hypotheses and the effect on 3G enrollment of various independent variables that appear in the final logistic model. We have discussed the cutoff used for this model and the various analyses like lift curve and receiver operating characteristic (ROC) curve that have guided us to the cutoff point. We have looked at the various results like sensitivity, specificity, and accuracy, from the contingency table of the cutoff, for both the modeling and the validation population.

We have concluded our discussion by showing that the model we have built is indeed better as compared to the model using raw variables.

REFERENCES

Hosmer, W. D., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

CART user manual: San Diego, CA: Salford Systems.

Steinberg, D., Colla, P.L., & Martin, K. (1999). *MARS user guide*. San Diego, CA: Salford Systems.

This work was previously published in the International Journal of Data Warehousing and Mining, edited by D. Taniar, Volume 3, Issue 2, pp. 22-31, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.22
Mobile Phone Customer Type Discrimination via Stochastic Gradient Boosting

Dan Steinberg
Salford Systems, USA

Mikhaylo Golovnya
Salford Systems, USA

Nicholas Scott Cardell
Salford Systems, USA

ABSTRACT

Mobile phone customers face many choices regarding handset hardware, add-on services, and features to subscribe to from their service providers. Mobile phone companies are now increasingly interested in the drivers of migration to third generation (3G) hardware and services. Using real world data provided to the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2006 Data Mining Competition we explore the effectiveness of Friedman's stochastic gradient boosting (Multiple Additive Regression Trees [MART]) for the rapid development of a high performance predictive model.

INTRODUCTION

The PAKDD 2006 Data Mining Competition required the analysis of real world data from an industry that includes people of all ages and walks of life. In economically developed regions it is increasingly common for elementary school children to have their own mobile phones, and mobile communication is now preferred over fixed lines for undeveloped regions. Evolving 3G technologies offer a considerable expansion of the communication services routinely supported by mobile phone networks to include multi-player games, video conferencing, and enhanced Web browsing. Excitement over 3G technology has

waxed and waned since 2000 as the early promises were not fulfilled, but 3G is now becoming a fixture of the global mobile marketplace. Thus, a competition focused on analysis of 3G mobile phone customers is both topical and readily understood by data analysts, modelers, and business decision makers from all industries.

This Salford Systems report is organized as follows. In the first sections we offer our understanding of the competitive challenge, the data available, and how we framed the modeling objectives. The competition organizers have provided their own description of the nature of the modeling challenge and the data, but we believe that our perspective on these topics is somewhat different and is thus needed to explain our strategy. In the second section we provide a summary of the key descriptive statistics that gave us our initial picture of the nature of the data and its adequacy for modeling purposes. The third section describes our modeling methods and reports our results and performance based on the labeled data. The fourth section delves further into the results to examine specific findings at the predictor level. Finally, the last section summarizes our results and offers conclusions.

THE MODELING CONTEXT

The data provided for the PAKDD 2006 modeling competition consisted of summary data for each of 18,000 customers of an Asian mobile phone service provider. The data included customer demographics, a calling plan indicator, 6-month summaries of calling behavior, handset characteristics, summaries of billing amounts and late payment patterns, and other communication-related behavior, including Web, e-mail, and game usage. The training data came in the form of a flat file containing 252 columns, with 15,000 rows drawn from second generation (2G) customers and 3,000 rows drawn from 3G customers. In addition, a further 6,000 rows of prediction set data were provided in the same format, but with the 2G/3G flag suppressed. Essentially, the competition required the development of a classification model learned from the training set able to predict the 2G/3G class membership of the customers in the prediction set. However, some fine points regarding the competition require elaboration.

In predictive modeling of a binary (2 class) outcome, a number of performance criteria have been discussed extensively in the literature. For example, Caruana (2004) discusses cross-entropy (likelihood), the area under the receiver operating characteristic (ROC) curve, and classification accuracy in the context of the KDD2004 competition, and lift in a specified percentile was used as one performance criterion in the Duke/NCR Teradata 2002 churn modeling competition. In the PAKDD 2006 competition, the stated performance measure was classification accuracy, a metric that by itself appears to take no account of the ability of a model to properly rank order data from most probable to least probable 3G class membership. This competition had an important wrinkle, however. Classification accuracy was to be measured for the 3G class only. The competition organizers wanted to rule out the degenerate solution (all customers are 3G) as uninteresting, and also rule out what they termed "manipulated" solutions. A successful manipulated solution can be extracted from a model "that has strong rank ordering performance" by assigning the least probable customer to the 2G class and all others to the 3G class. This "solution" would have a high probability of yielding a perfect score on the 3G class, because even a moderately good model should be able to successfully identify a single 2G customer to place in the 2G class. Such a solution would presumably be disqualified as manipulated.

Less obviously manipulated solutions are possible, however. Given a good rank ordering of the customers by the probability of being 3G, a decision rule that assigns relatively few records to the 2G class should exhibit a high classifica-

tion accuracy rate for the 3G class. If the objective were average classification accuracy in the two classes, then the optimal decision boundary would be at P* = Prob(3G) = 1/6, assigning all customers with estimated Prob(3G)<1/6 to the 2G class, on the assumption that the population fraction of 3G is the same as observed in the training data, or 1/6. Finding a rationale to shift the decision boundary towards 0 would improve an entrant's chance of winning, provided that the solution did not strike the judges as manipulated. Possible rationales include weighting the costs of misclassification more heavily for 3G errors, perhaps on the basis of customer expenditures. We decided not to attempt this, however, because it is a form of manipulation. Instead, we used our classification tools employing unit costs and chose a decision threshold intended to maximize average within-class classification accuracy.

We were concerned that a competition entrant might successfully find and openly defend or even silently and surreptitiously deploy a "quasi-manipulated" winning strategy. Indeed, in the absence of a step-by-step review of winning entries, it would not be possible to tell if such hidden manipulation had been employed. (However, the top three winning outcomes were so close that it is reasonable to assume that all submitted nonmanipulated results.) Given these concerns as we prepared to enter the competition, we elected to make this competition into a speed contest. Our goal was to develop the best possible average accuracy classifier in a very short period of time. We managed to complete our data analysis in a single concentrated working day.

The organizers of this competition are to be commended for acquiring a substantial volume of real world customer data for public release. Such data can rarely be acquired without limitation. In this instance, the limitations pertain to the data fields made available, and to the descriptive information characterizing the data. We know that the data were drawn from the customer records of an Asian mobile phone provider, but we were not given the time period from which the data were drawn, and several valuable fields are provided with partial information only. For example, the nationality of the customer is listed as an uninformative numerical code, and similarly the countries most often called are also limited to numeric codes. We are provided with no information regarding the competitive landscape, the nature of the marketing campaigns for either 2G or 3G services, the pervasiveness of mobile phone use, or the overall 3G market share in the country or for the provider in question. These informational limitations severely restrict both the business value of any models developed and our ability to extract real-world insight into the workings of the marketplace or into consumer behavior.

What Can Be Learned from the Data?

The formal description of the competition data was confined to two pages. Therefore, we resorted to making plausible assumptions about its nature. We assumed that the data were gathered from a short time window. With 24,000 customers in total, such a sample could easily have been extracted from a single month's worth of retrospective account data. Our tests (reported hereafter) establish that the train and prediction sets were very likely drawn from the same customer population and from the same time period. From a business perspective the effort to predict 2G versus 3G in this context is of limited value. We would naturally expect that someone who goes to the trouble of acquiring a 3G handset and subscribing to a 3G plan would behave differently than someone who is content to use 2G hardware and services. Indeed, a powerful predictor of a customer being 3G is that they own a 3G-capable handset! From a business and marketing point of view, more valuable insights would be generated by studying the phone records of all customers at a point in time when 3G was not available with the goal of predicting

which 2G customers eventually migrated to 3G. However, the data available were gathered after the decision to go to 3G had been made. Thus, the modeling exercise is more properly thought of as a profiling effort in which a multivariate portrait of the behavioral differences between these two groups is extracted. We would agree that 2G customers who look like 3G customers (i.e., have a high estimated probability of being 3G) are prime candidates for an up-sell marketing campaign, but we would also want the service provider to investigate the primary reasons such customers have not already switched. The data made available for this competition have been generated via a classic self-selection process that can strongly bias the results of a model intended to predict the future probability of a 2G customer's migration to 3G. Extensive literature in statistics has been devoted to the selection bias topic, but is not one that can be addressed within the confines of this report (see, for example, Rosenbaum & Rubin, 1983).

DATA OVERVIEW

To get a better grasp of the data we grouped the available predictors into the following illustrative categories (not all variables are listed here):

Basic demographics: age, nationality, marital status, occupation

Handset related: model, manufacturer, duration owned, times changed

Calling plan related: plan ID, high end status, number of lines, line tenure

Payment behavior: delinquency, blacklisted, suspended, amount overdue

Special features: games, e-mail notification, WAP, GPRS, citiguide, picture xmit

Call behavior/bill detail: SMS, voice/data, inbound/outbound, 900 numbers,

peak/off-peak, roaming, forward, extranet, fixed/mobile, countries called

For the call behavior and bill detail variables, statistical measures summarizing the last 6 months were provided, including total, average, and standard deviations, and counts of certain events such as payment delinquency were included. While this represents substantial information it should be clear from this listing that a great deal of information was suppressed for both convenience and confidentiality. In a real-world consulting environment, a modeler would also have had access to postcode, as well as to month-by-month account details from which trends, minima, maxima, and other statistics could be extracted.

Conventional inspection of the data revealed no obvious errors or problems requiring repair prior to productive analysis. Missing values were evidently not a problem, either. While the variables recording the three most frequently called countries were heavily missing, this simply reflected the fact that most customers did not use their mobile phones to call out of country at all; therefore, we added new predictors to capture this information. The occupational code is missing in 63.41% of the training data, a genuine lack of information. The 11 remaining variables with missing values are listed in Table 1; certainly there

Table 1. Missing value prevalence by variable

Variable	%Missing
MARITAL_STATUS$	5.79%
CONTRACT_FLAG	5.07%
DAYS_TO_CONTRACT_EXPIRY	5.07%
PAY_METD$	4.92%
PAY_METD_PREV$	4.92%
HS_MODEL	2.94%
HS_MANUFACTURER	2.89%
AGE	2.71%
TOT_DEBIT_SHARE	1.51%
NATIONALITY	0.11%
GENDER	0.02%

is no reason to be concerned about the prevalence of missing values.

To ascertain the elementary predictive power in the data we conducted variable-by-variable statistical tests, using difference of means *t*-tests for continuous and binary flag variables, and chi-square tests for multi-level categorical predictors. The *t*-statistics for the difference of means tested are displayed in Figure 1, sorted in descending order with the X-axis listing the variable rank. (This diagram was inspired by Tusher, Tibshirani, & Chu, 2001.)

Conducting a simple *t*-test for each continuous or binary flag variable shows that almost all variables exhibit a *t*-statistic greater than two, and that 169 variables exhibit a *t*-statistic greater than three. Among the categorical predictors we find two overwhelmingly powerful predictors in handset type, (HS_MODEL, df=332 , Chi-square=6657.72), and calling plan (SUBPLAN, df=65, Chi-square=2002.33). Thus, we have an abundance of apparently useful predictors and should expect our final model to be quite accurate.

From a business decision support perspective we would find some of these predictors painfully obvious: a person who elects to purchase a 3G handset is probably (but not certainly) a 3G customer; a person who elects a 3G calling plan is probably (but not certainly) a 3G customer; a person who makes heavy use of interactive gaming is probably a 3G customer. While not surprising, these factors turn out to be key components of 2G/3G discrimination in the models that follow.

Train Set Versus Prediction Set Data

When a prediction set is made available along with training set data we always find it useful to ascertain the degree of similarity or difference between

Figure 1. Distribution of difference of means t-statistics: 2G versus 3G customers

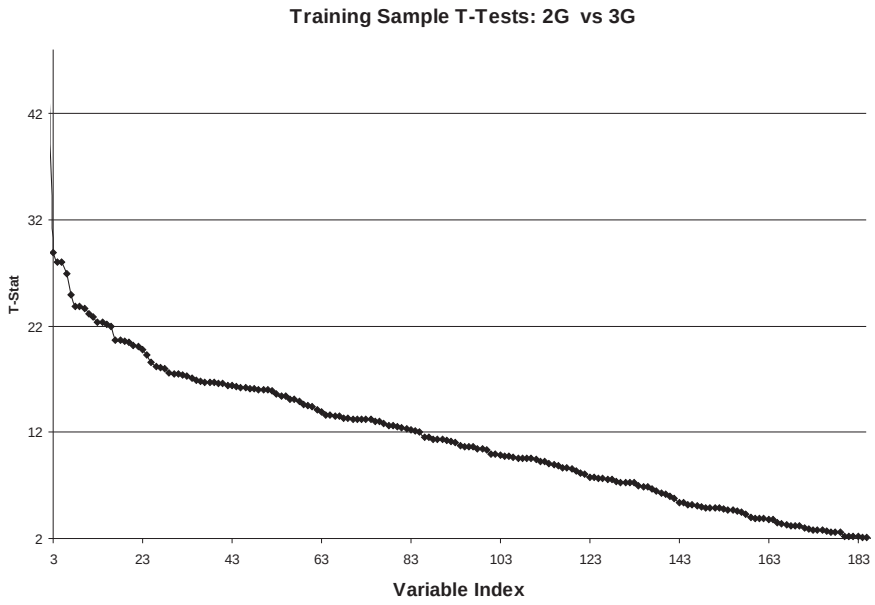
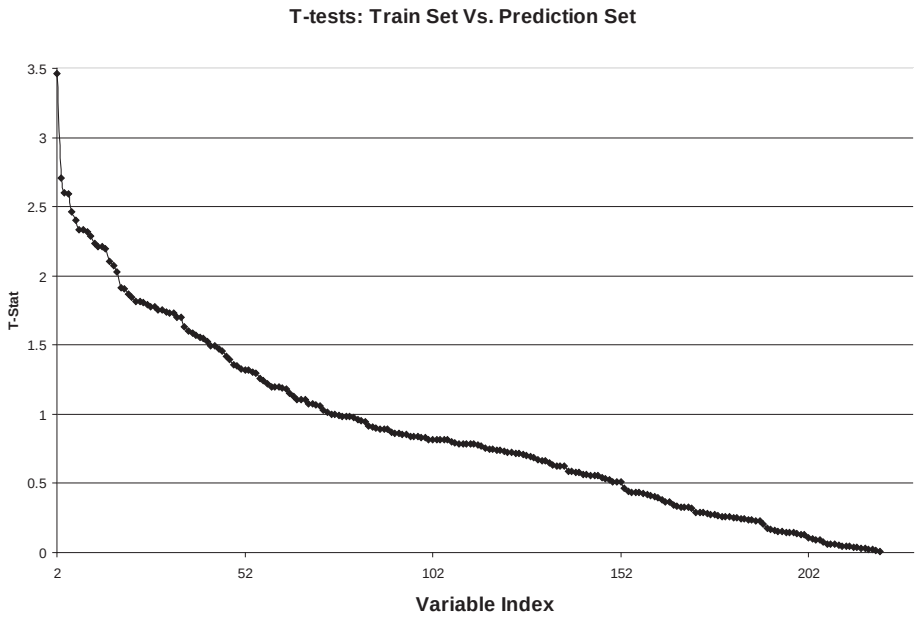


Figure 2. Distribution of difference of means t-statistics: Train versus prediction data



the two data sets. Comparisons can be conducted variable by variable and in a multivariate predictive model. Figure 2 displays the *t*-statistics from a test of train versus prediction set, where again the *t*-statistics are sorted in descending order and the X-axis lists the variable rank.

The good news here is that few substantial differences exist between the two samples. Only 16 variables display a *t*-statistic greater than 2, with the largest being 3.47. Given sample sizes of 18,000 train and 6,000 prediction set customers and 220 variables to test, one would expect to find 11 *t*-ratios greater than 2 due to chance alone. Among the categorical predictors only PAY_METD shows a marginally significant between-sample difference, with a *p*-value of .0458 Our conclusion is that there is no evidence that the train and prediction sets differ in any meaningful way.

Comparing Train and Prediction Sets via Modeling

For this test we concatenated the two data sets and used the sample indicator (Train_Predict) as the target in a multivariate modeling exercise. We used the Classification and Regression Tree (CART) (Breiman, Friedman, Olshen, & Stone, 1984; Steinberg & Colla, 1995) to automatically build this model and found virtual unpredictability: If we choose a record at random from one of these two sets it is not possible to predict which set the record came from. Figure 3 displays the results of the CART analysis. The upper panel contains

Figure 3. CART test of ability to discriminate between train and prediction data

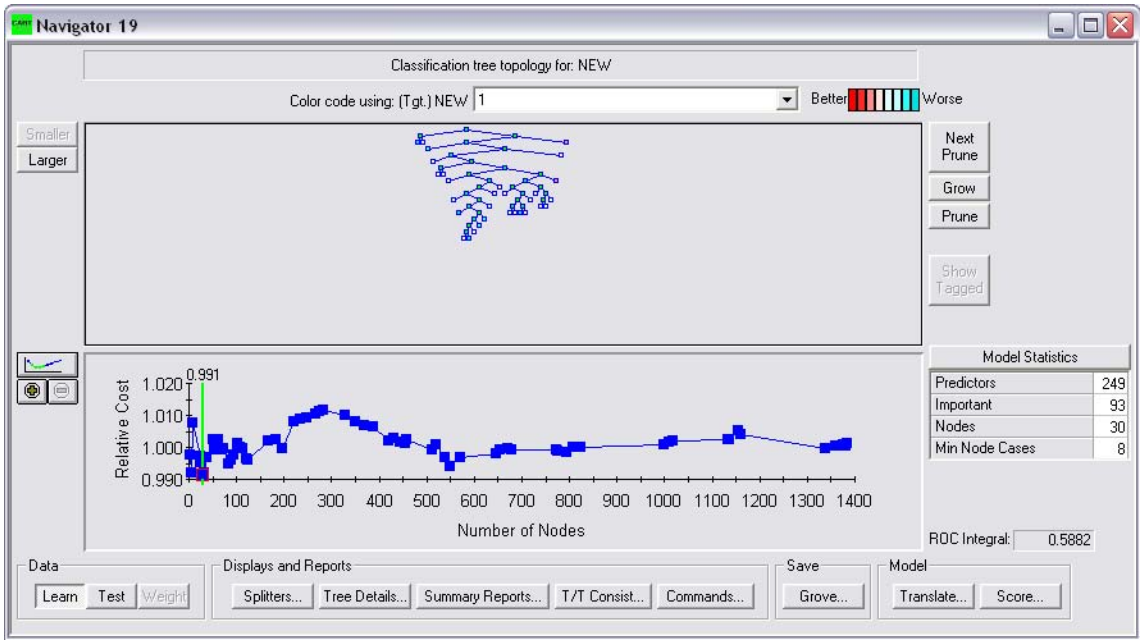


Table 2. Confusion matrix: CART model to discriminate between train and prediction data

Actual Class	Total Cases	Percent Correct	0 N=6084	1 N=5869
1	3,032	49.77	1,523	1,509
0	8,921	51.13	4,561	4,360
Total:	11,953.00			
Average:		**50.45**		
Overall % Correct:		**50.78**		

the tree topography for the optimal tree and the lower panel graphs the relative test error for the train versus predict set discrimination problem. The relative error never dips below .991, indicating an inability to discriminate between the two data sets. Table 2 contains the confusion matrix for the best of these models.

Confusion Matrix—Test Data—Count

Testing was conducted by dividing the data into equal-sized train and test partitions. The CART model shows negligible predictability at the optimal tree size, strongly supporting our hypothesis that both datasets come from the same population of customers.

Data Preparation: Compressing Categorical Variables

As we noted previously, the categorical predictors HS-MODEL and SUBPLAN are highly predictive of CUSTOMER_TYPE all by themselves. These predictors are somewhat awkward to use in their raw form because of the large number of levels in each, many of which have relatively low representation in the data. The HS_MODEL lists 332 different handset models in the training set, 60 of which contain only a single customer, and 182 of which list fewer than 10 customers each. Similarly, SUBPLAN lists 66 plans, 22 of which have fewer than 10 customers each. For convenience, interpretability, and reliability we elected to compress these predictors into new variables with substantially fewer levels.

To compress HS_MODEL and SUBPLAN we grew CART *probability* trees, using CUSTOMER_TYPE as the target and the variable to be compressed as the sole predictor. (Supervised binning via decision trees, for continuous predictors, is discussed by Dougherty, Kohavi, & Shamai, 1995—the extension to categorical predictors is quite natural.) We divided the data into equal learn/test partitions to assess model performance. Such a supervised compression group levels with similar mixes of 2G/3G customers. In

Figure 4. CART model to compress HS_MODEL

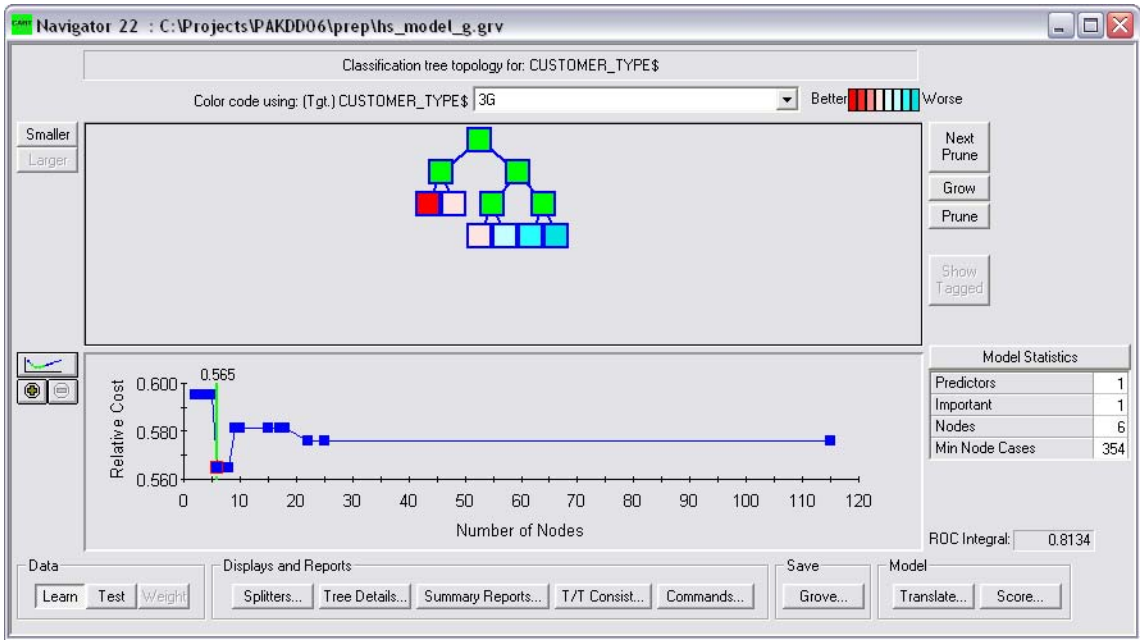


Figure 5. Lift ratios in CART terminal nodes for compressing HS_MODEL

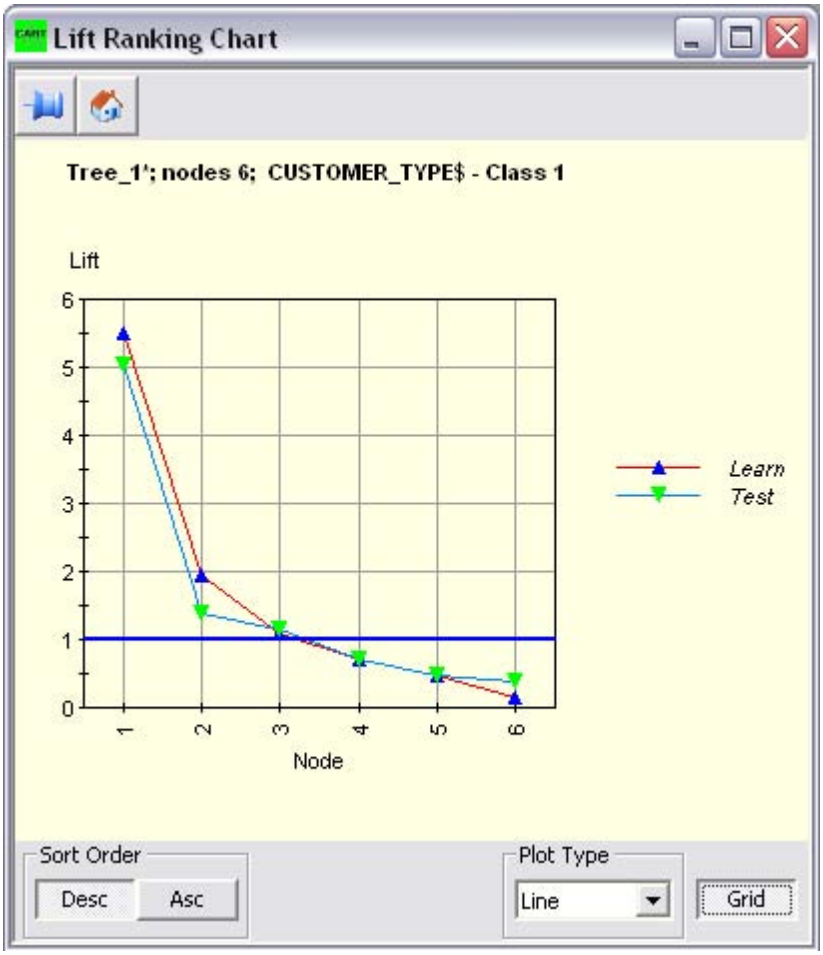


Figure 4 we see that HS_MODEL compresses nicely to just six groups; the 3G/2G lift ratios in each group (relative to the population baseline) are displayed in Figure 5.

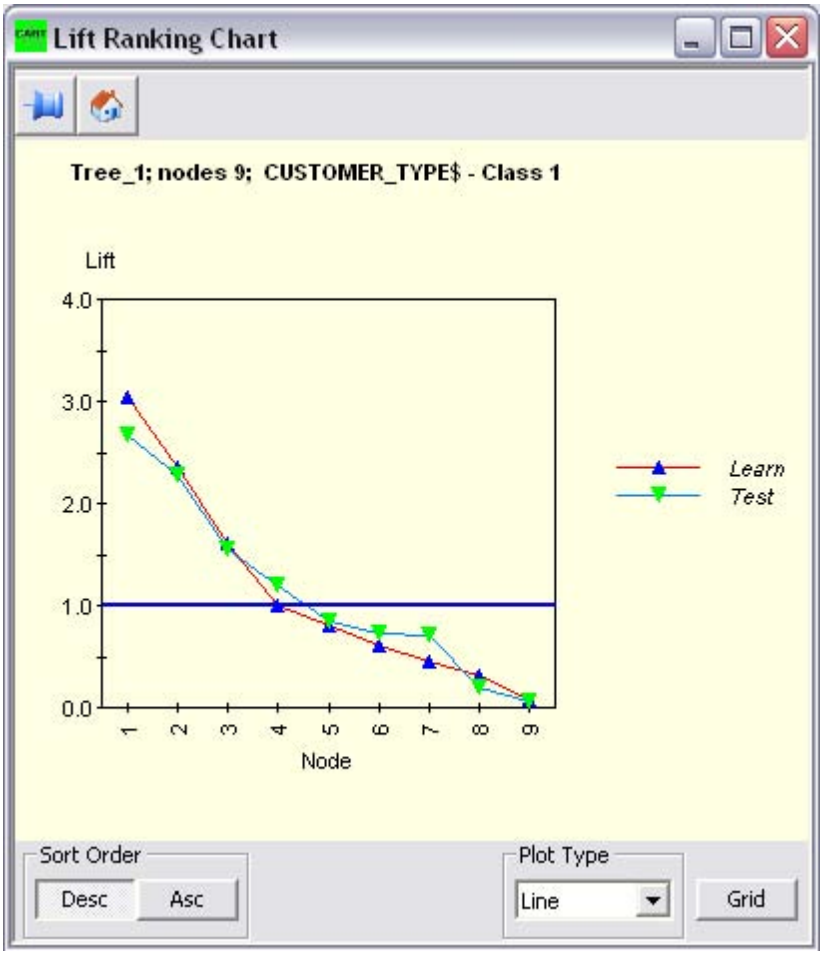
The same procedure was used to compress the SUBPLAN variable into nine separate groups. The lift ratios are presented in Figure 6.

MODELING

Single Tree CART Models

We began our analysis with CART trees to obtain a quick initial insight into the data via its visual displays and variable importance rankings. A display of the tree pruned back to seven nodes is

Figure 6. Lift ratios in CART terminal nodes for compressing SUBPLAN



shown in Figure 7, where a *black* terminal node indicates above-average probability of 3G (and thus assignment to the class 3G) and a *gray* terminal node indicates below-average 3G probability (and thus assignment to class 2G). The tree reveals that customers with older handsets tend to be 2G, as do those with more recent contracts (having longer durations before expiry). More game time is also associated with greater probability of 3G. While none of this is surprising, it is always reassuring to have core expectations ratified before moving on to more complex modeling methods.

The first CART models yielded cross-validated average class accuracies of about 81.28%, as shown in Table 3. The CART cross-validated area under the ROC curve was 0.8581.

Figure 7. Primary splitters in single CART tree to predict 2G/3G

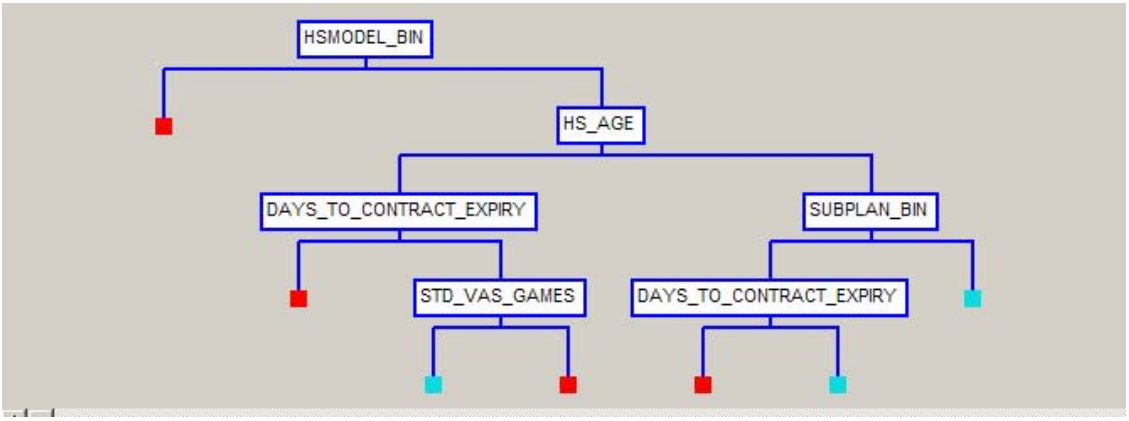


Table 3. Confusion matrix: CART model to discriminate between 2G and 3G customers. Cross-validation estimates

Actual Class	Total Cases	Percent Correct	2G N=13763	3G N=4237
2G	15,000	86.89	13,033	1,967
3G	3,000	75.67	730	2,270
Total:	18,000.00			
Average:		81.28		
Overall % Correct:		85.02		

CART Model Confusion Matrix—CV Test—Count

Splitting the data into 80% learn and 20% test partitions yields a similar class average accuracy of 80.12%, and an area under the ROC curve of .8451.

The most important variables as ranked by CART appear in Table 4.

Boosted Trees: TreeNet Models

We next moved on to stochastic gradient boosting models (Friedman, 2001), using the TreeNet™ (Salford Systems, 2005) commercial release of Friedman's MART™. Our primary reason for doing so was that tree ensembles often outperform single trees, and TreeNet in particular is able to extract small amounts of predictive information

Table 4. Variable importance rankings for CART model to discriminate between 2G and 3G customers where we see HS_MODEL ranked most important and SUBPLAN ranked third

Variable	Score
HS_MODEL	100.00
HS_AGE	69.17
SUBPLAN	22.95
DAYS_TO_CONTRACT_EXPIRY	13.23
HIGHEND_PROGRAM_FLAG	10.42
AVG_MINS_OB	10.00
AVG_NO_CALLED	9.04
AVG_MINS_MOB	8.73
AVG_MINS_OBPK	8.64
AVG_CALL_OB	8.60

from a large collection of predictors such as we have in this challenge. Also TreeNet tends to yield stable results in its variable importance rankings and predictive scores.

Using a slow learn rate, 241 predictors, and tuning the model to optimize for average class accuracy, we obtained the TreeNet summary report of Figure 8.

Figure 8. Summary report for TreeNet model displaying model performance versus number of trees

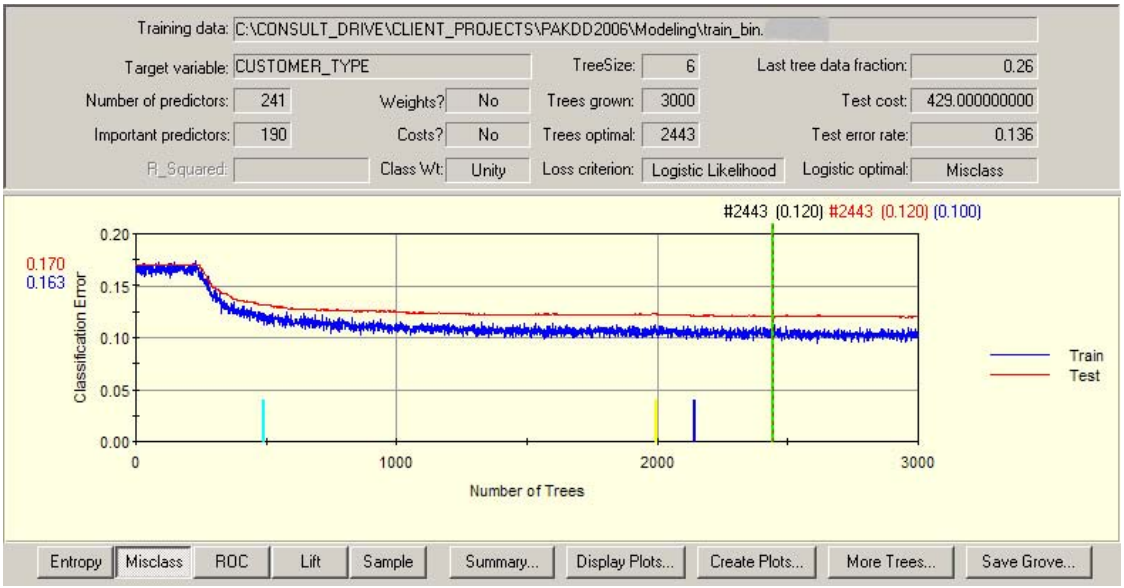


Table 5. Confusion matrix: TreeNet 6-node model for 2G versus 3G

Actual Class	Total Cases	Percent Correct	2G N=13763	3G N=4237
2G	2975	84.24	2506	469
3G	611	79.71	124	487
Total:	3586			
Average:		81.98		
Overall%Correct		83.46		

The default tree size of six nodes is adequate for capturing 3-way (and possibly some 4- and 5-way) interactions and yields in an area under the ROC curve of .907 on test data. Because this result is a good deal better than our best single CART tree, we were persuaded to use TreeNet scores as the basis of our classification scheme. In terms of average classification accuracy (Table 5) the model is only slightly better than the single CART tree.

TreeNet Confusion Matrix—20% Test—6-node Trees

To test whether interactions are needed at all in this model we re-ran our model restricting the tree

Figure 9. TreeNet Confusion Matrix—20% Test—2-node trees

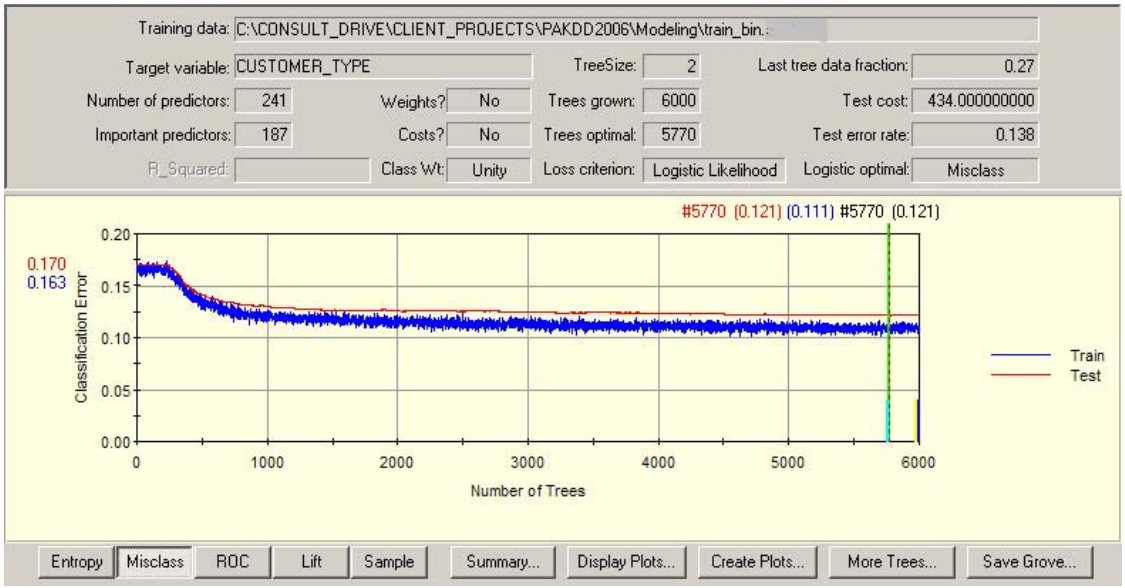


Table 6. Confusion matrix: TreeNet 2-node model for 2G versus 3G

Actual Class	Total Cases	Percent Correct	2G N=13763	3G N=4237
2G	2975	82.18	2445	530
3G	611	81.18	115	496
Total:	3586			
Average:		81.68		
Overall%Correct		82.01		

size to two nodes, or "stumps" models. Because each tree can involve, at most, one predictor, and because the final prediction is based on a sum of scores across all trees, the stumps model is constrained to be additive in the predictors. The model summary is shown in Figure 9. The confusion matrix appears in Table 6.

TreeNet Confusion Matrix—20% Test—2-node Trees

While this performance is not as good as those obtained from the 6-node trees, it is close enough to suggest that interactions are not playing a major role in this model. Nevertheless, we ran Friedman and Popescu's (2005) interaction tests to determine that DAYS_TO_CONTRACT_EXPIRY and HS_AGE are the only serious interaction candidates among the predictors.

To delve further into the details of the model it is worth looking at the distribution of model scores on the train dataset where CUSTOMER_TYPE is known. When the objective function is the logistic regression, TreeNet scores are literally predicted logits and a 0.0 score corresponds to a predicted probability of being a 3G of 0.50. The greater the score, the greater the predicted probability of 3G is (see Figure 10).

The heavy black curve with the highest peak represents 2G customers, the medium gray curve represents 3G customers, and the dark gray curve represents all customers.

The graph suggests reasonably good separation between the two target classes. The area of overlap mostly occurs in a narrow middle range of scores between -2.0 and -1.0, and the graph can be used to guide a trade-off between the true positive and false positive rates. For example, a decision threshold of –3.2 yields a 100% accuracy in predicting 3G customers and still isolates several hundred 2G customers in the leftmost spike. On the other hand, a threshold set at –1.7 keeps the true positive rate above 99% while still separating a very large group of 2G users on the left.

On the prediction set where the target is unknown the score distribution is seen in Figure 11.

As expected from our previous analysis of training and prediction set differences, the overall distribution of scores is virtually identical across the two data sets. From the prediction set scores graph one might conclude that the optimal score threshold to maximize the true positive 3G rate should be around –2.0. However, because such a low threshold would produce an unreasonably high false positive rate, our submitted predic-

Figure 10. Distribution of TreeNet scores among 2G and 3G customers in train data

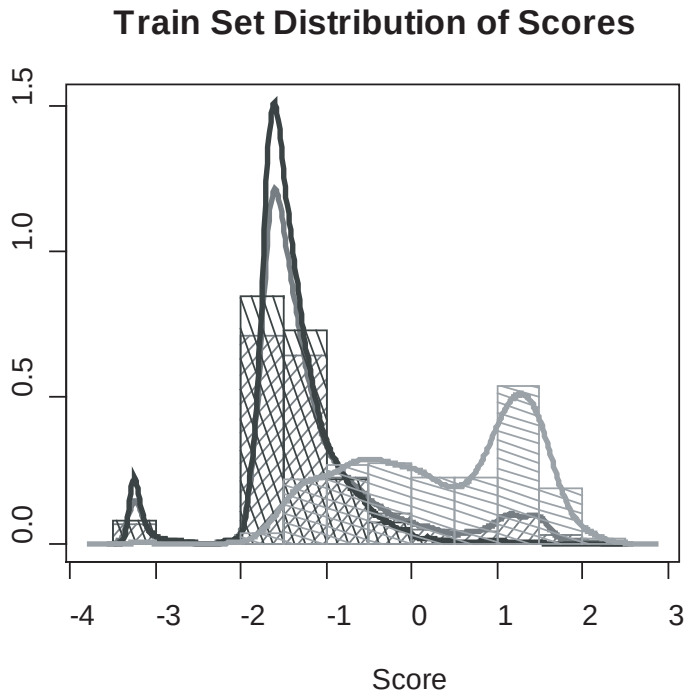
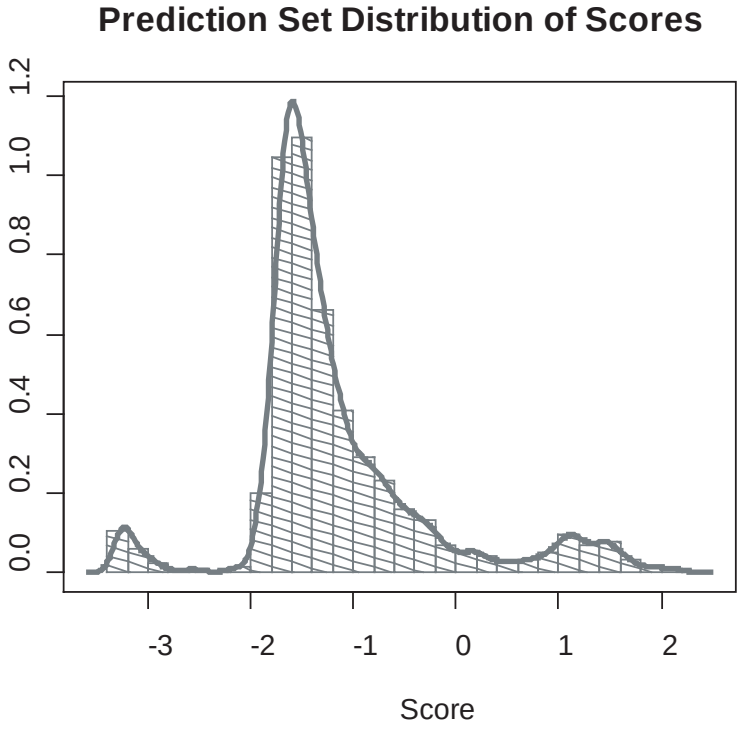


Figure 11. Distribution of TreeNet scores in prediction set for unlabeled customers



tions were based on the threshold of –1.1, which equalizes both true positive and true negative classification rates.

Model Simplification

For interpretation purposes, it is often useful to try to reduce the number of predictors that appear in the model. We used *variable shaving* to accomplish model simplification, fitting a series of progressively smaller models to the data. Each new model is specified by dropping the least important predictor from the previous model in a pattern similar to backwards stepwise regression.Figure 12 provides the results of the shaving process.

The graph indexes each model by the number of predictors, while the model accuracy is reported in terms of the area under the ROC curve. We observe that a model with eight predictors is sufficient to achieve an ROC not far below that of our "large scale" model.

Model Insights

Table 7 lists the relative importance of the eight predictors in the shaved TreeNet model (The larger 241 predictor model was used in our competitive submission.):

TreeNet Variable Importance

The grouped version of the HS_MODEL variable appears to be the most influential. Its additive contribution plot is shown in Figure 13.

Not surprisingly, the relative contribution of individual groups agrees with the results of the CART model used for grouping. It is clear that, on average, the phone models that appear in the first group are heavily favored by 3G customers, whereas phones appearing in the sixth group are on average favored by 2G customers.

Next on the importance list is the age of the handset and its impact on the probability of a customer being 3G, as illustrated in Figure 14.

Figure 12. ROC as a function of the number of variables removed. The integers superimposed over the curve represent the number of predictors left in the model.

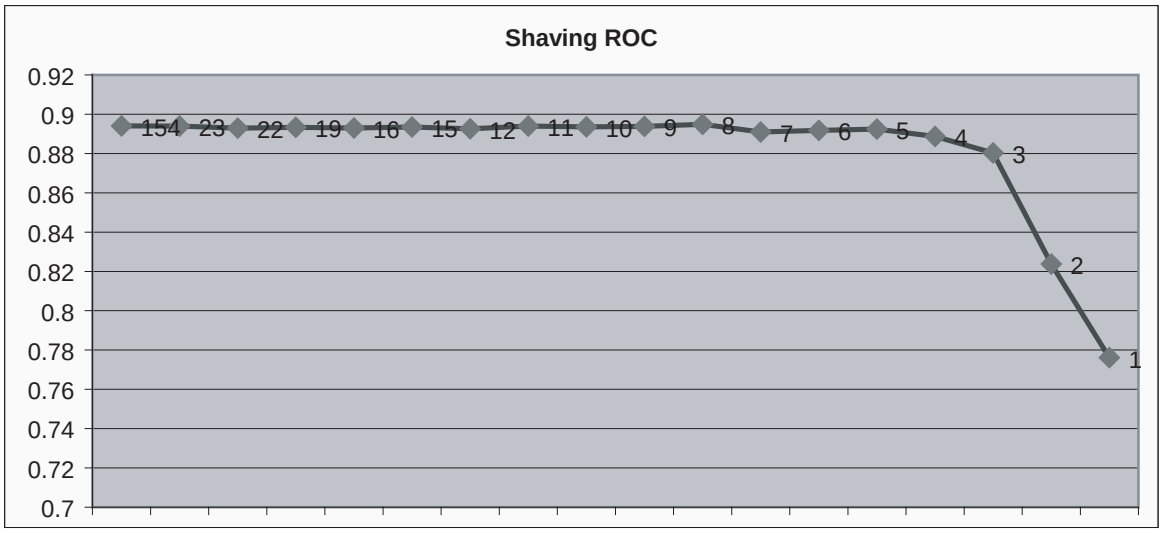


Table 7. TreeNet variable importance rankings for 2G versus 3G model

Variable	Score
HS_MODEL	100.00
SUBPLAN	23.90
HS_AGE	23.55
AVG_VAS_GAMES	13.84
AVG_VAS_GPRS	8.71
AVG_BILL_AMT	6.95
AVG_MINS_INTRAN	5.67
TOT_RETENTION_CAMP	4.02

Figure 13. TreeNet partial dependency plot: Contribution of HS_MODEL_G to predicted probability of 3G

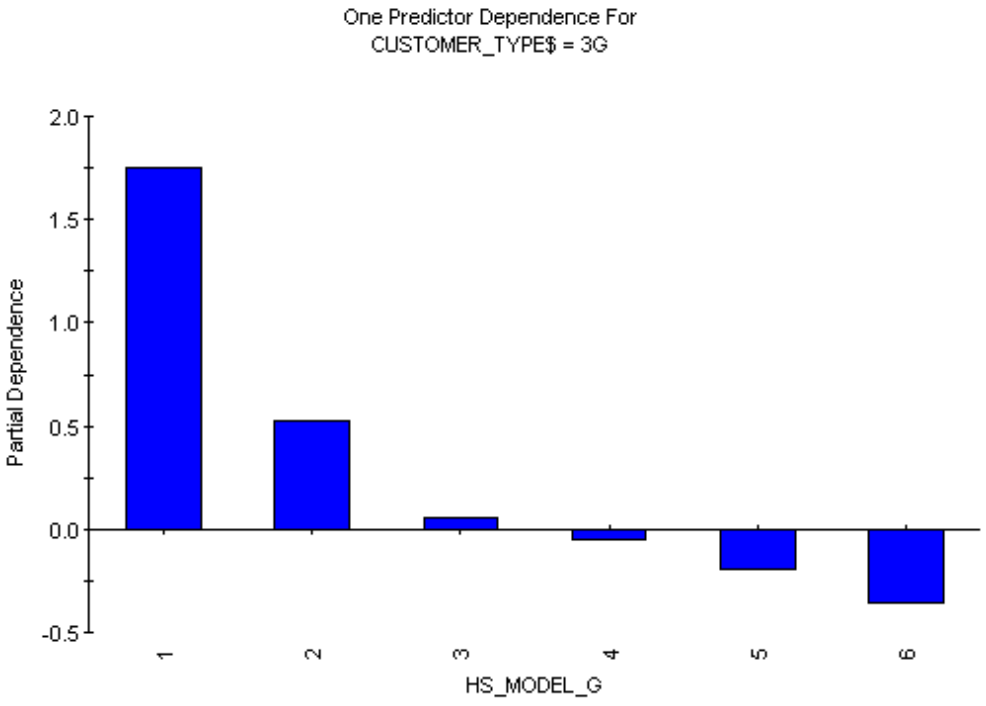
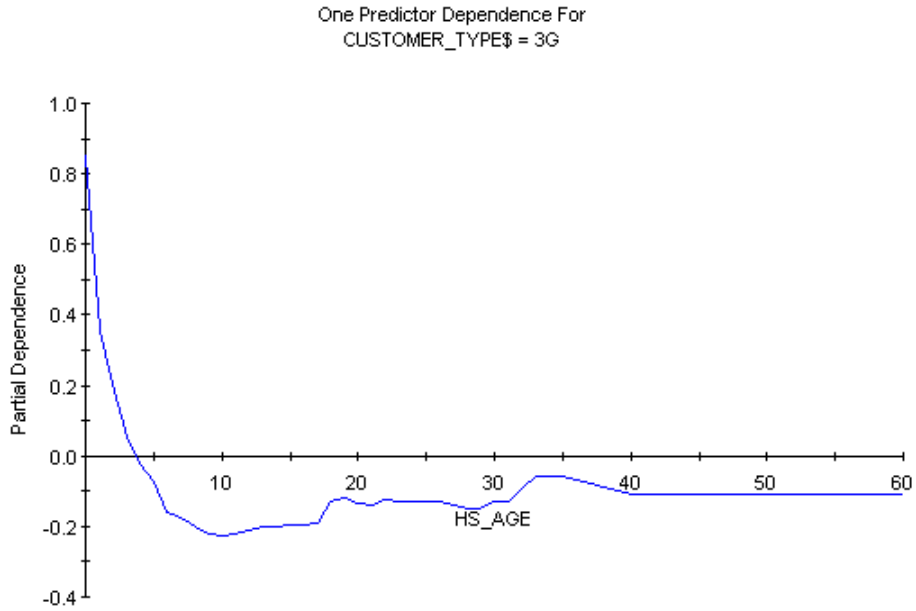


Figure 14. TreeNet partial dependency plot: Contribution of HS_AGE to predicted probability of 3G



This "partial dependency" plot is a standard TreeNet output and helps us understand the workings of key predictors. The newest handsets (with an age of near zero) are very likely to be owned by 3G customers, and the probability that a customer is 3G drops rapidly with every month that the handset has been owned. Once a handset is older than 10 months its age does not matter much as far as discrimination between 2G and 3G is concerned.

The next important predictor measures gaming activity AVG_VAS_GAMES (Average Games Utilization [Kb]) for the last 6 months (Figure 15).

This is essentially a step function, with anyone downloading more than 1 MB being far more likely to be 3G.

The next variable is a grouped version of the SUBPLAN (Figure 16; see the CART model described earlier).

Again, the relative contribution of each group agrees with the one suggested by the relevant CART model. Subplans comprising group 1 are more associated with 3G users whereas groups 8 and 9 are "closer" to 2G customers.

Not surprisingly, the next variable tells us that 3G customers are associated with larger monthly bills with a steep ramp rising from about 100 to about 275 units per month (Figure 17).

Figure 18 illustrates the importance of retention campaigns for 3G users. It appears that 3G users have been subjected to many more retention campaigns, though it is not possible to tell whether the retention campaigns induce 3G subscribership or that 3G status attracts campaigns.

Figure 15. TreeNet partial dependency plot: Contribution of AVG_VAS_GAMES to predicted probability of 3G

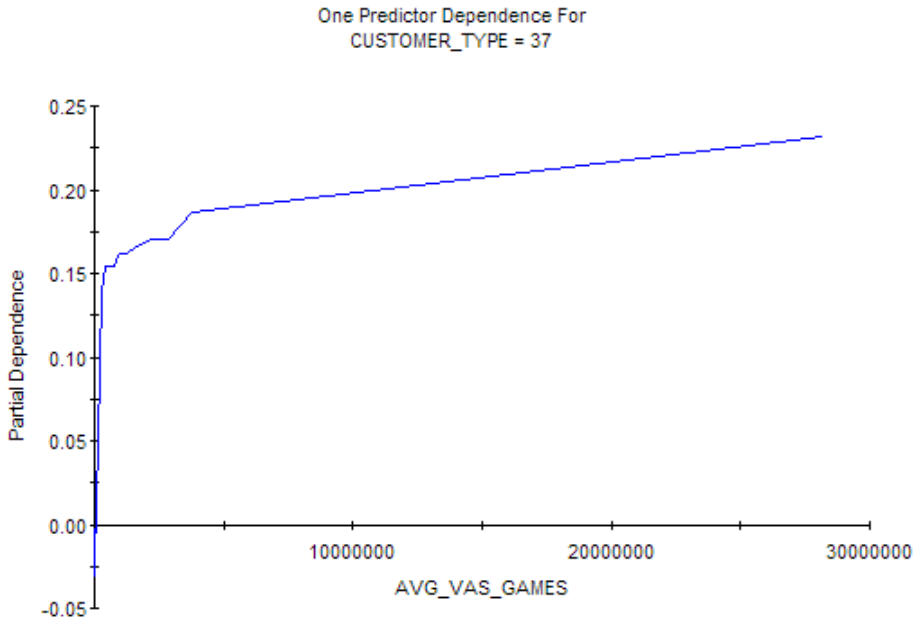


Figure 16. TreeNet partial dependency plot: Contribution of SUBGROUP_G to predicted probability of 3G

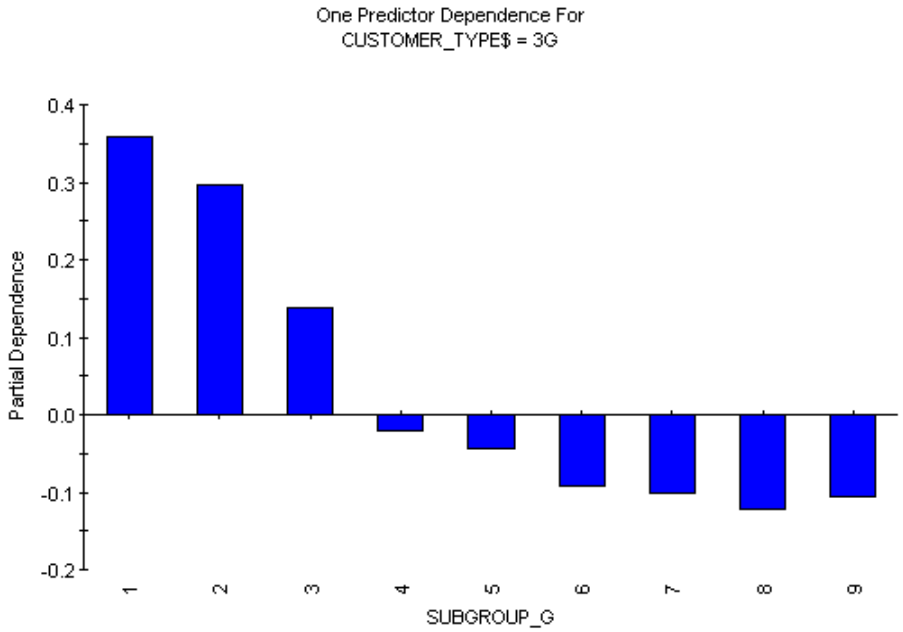


Figure 17. TreeNet partial dependency plot: Contribution of AVG_BILL_AMT to predicted probability of 3G

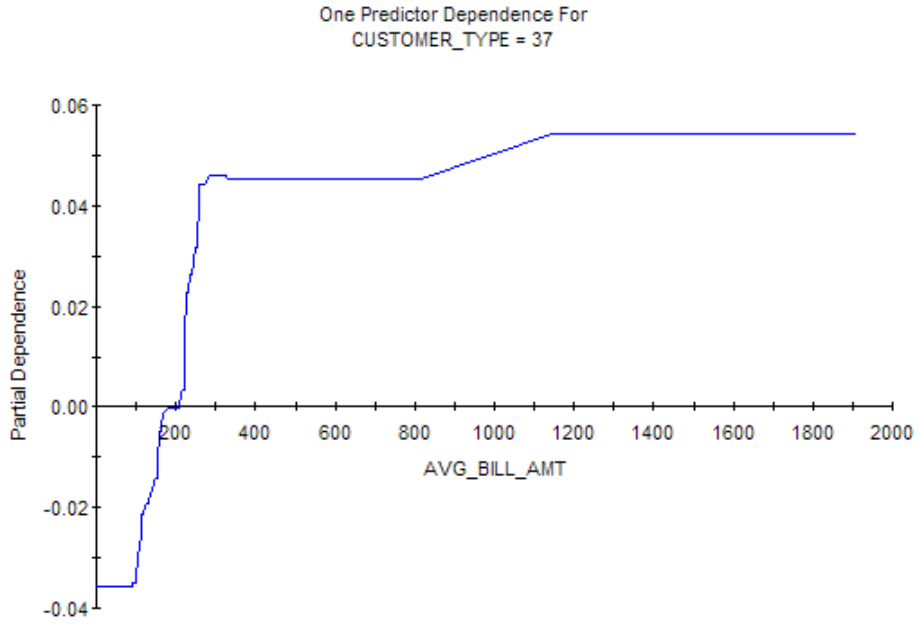


Figure 18. TreeNet partial dependency plot: Contribution of TOT_RETENTION_CAMP to predicted probability of 3G

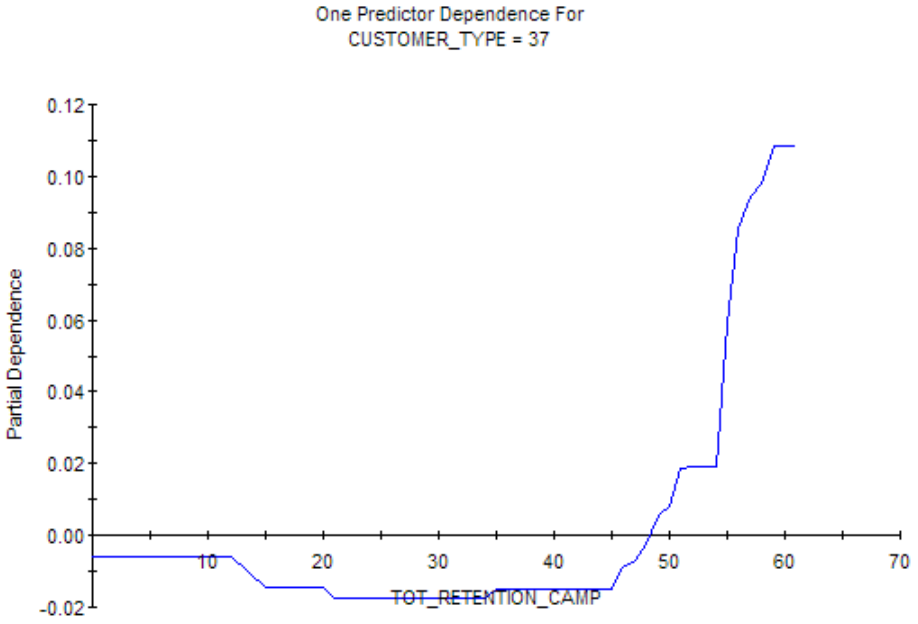


Figure 19. TreeNet partial dependency plot: Contribution of AVG_VAS_GPRS to predicted probability of 3G

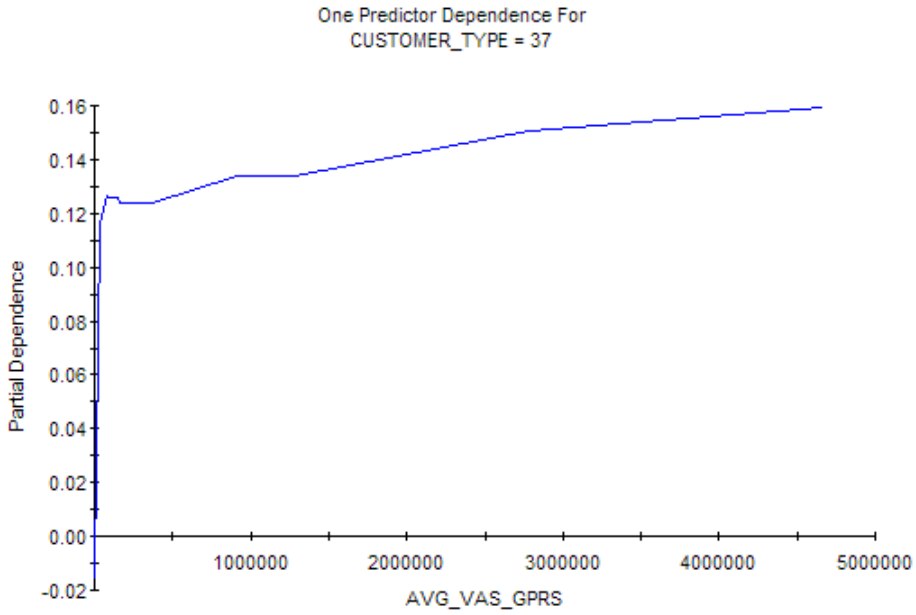
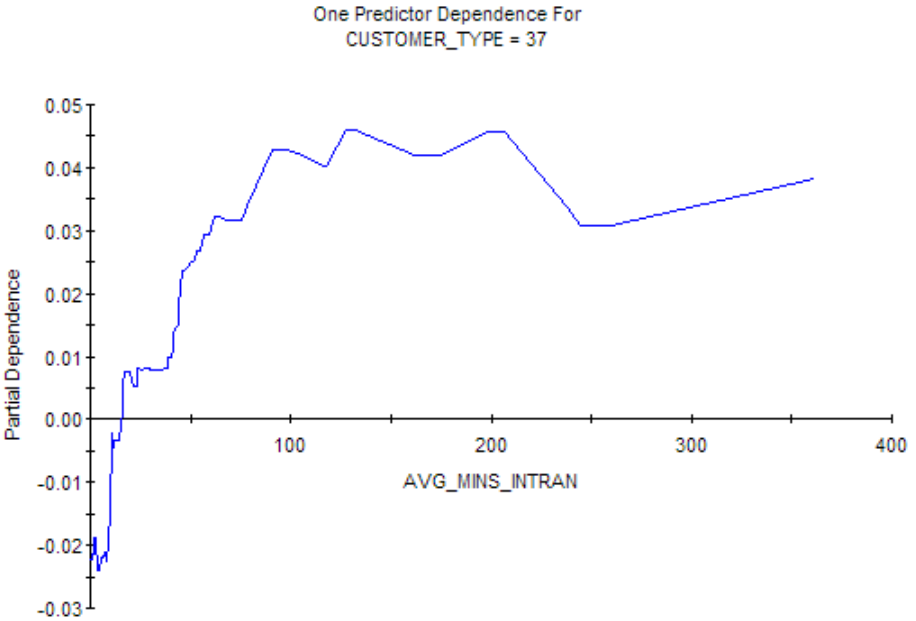


Figure 20. TreeNet partial dependency plot: Contribution of AVG_MINS_INTRAN to predicted probability of 3G



Similarly, GRPS utilization characterizes 3G consumers (Figure 19),again with a cutoff around 100 KB.

Finally, average intranet minutes are weakly associated with 3G users. Although the slope of the ramp is steep, the difference between the highest and lowest points on this curve is less than .07 (Figure 20).

A MODEL WITHOUT HS_MODEL OR SUBPLAN

Given the overwhelming importance of the HS_MODEL and SUBPLAN, along with the fact that these two predictors so closely proxy the customer's decision to subscribe to 3G services, we fit a model without these variables. We also suppressed other predictors that are obviously associated with 3G services, such as GAMES usage, and concentrated on predictors that have relevance to customers of all types, such as minutes used. One such model has a test area under the ROC of 0.78 and the following variable importance list (see Table 8).

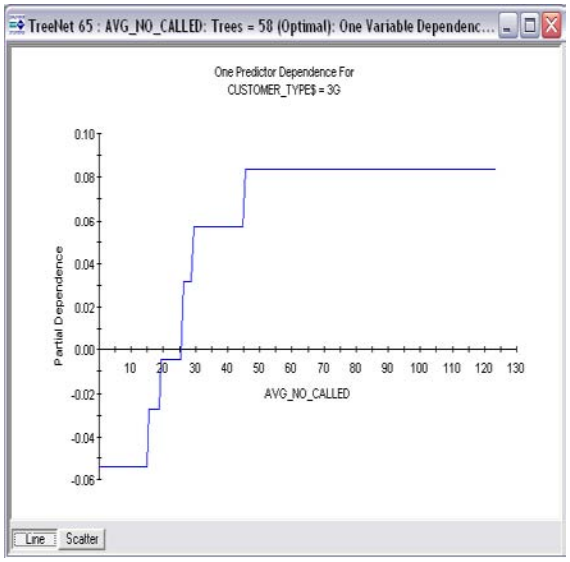
TreeNet Variable Importance

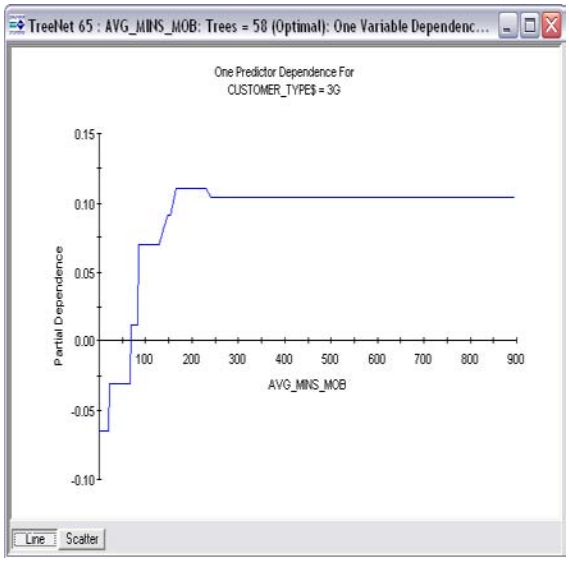
The collection of graphs in Figure 21 illustrate how these variables help us discriminate between

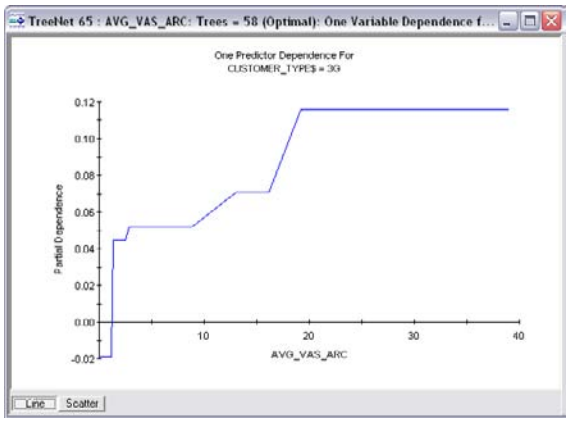
Table 8. TreeNet variable importance rankings for model excluding HS_MODEL and SUBPLAN

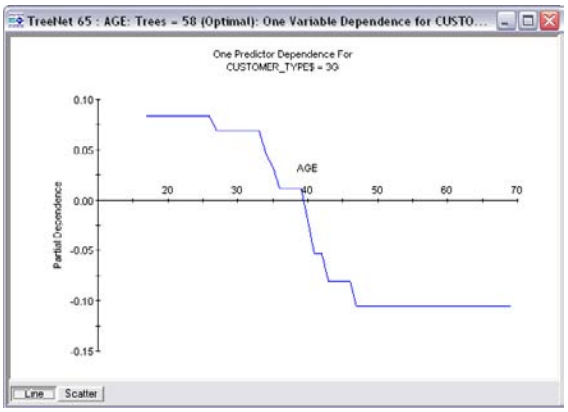
<table>
<tr><th>Variable</th><th>Score</th><th></th></tr>
<tr><td>AVG_NO_CALLED</td><td>100.00</td><td>||||||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>AVG_MINS_MOB</td><td>64.23</td><td>|||||||||||||||||||||||||||</td></tr>
<tr><td>AVG_VAS_ARC</td><td>58.92</td><td>||||||||||||||||||||||||</td></tr>
<tr><td>AGE</td><td>58.06</td><td>||||||||||||||||||||||||</td></tr>
<tr><td>LOYALTY_POINTS_USAGE</td><td>56.46</td><td>|||||||||||||||||||||||</td></tr>
<tr><td>TOT_DEBIT_SHARE</td><td>53.78</td><td>||||||||||||||||||||||</td></tr>
<tr><td>AVG_CALL_INTRAN</td><td>53.38</td><td>||||||||||||||||||||||</td></tr>
<tr><td>STD_VAS_ARC</td><td>49.67</td><td>||||||||||||||||||||</td></tr>
<tr><td>AVG_MINS</td><td>49.45</td><td>||||||||||||||||||||</td></tr>
<tr><td>LINE_TENURE</td><td>48.63</td><td>||||||||||||||||||||</td></tr>
<tr><td>AVG_CALL_OB</td><td>46.99</td><td>|||||||||||||||||||</td></tr>
<tr><td>VAS_AR_FLAG</td><td>45.73</td><td>|||||||||||||||||||</td></tr>
<tr><td>LST_RETENTION_CAMP</td><td>41.54</td><td>|||||||||||||||||</td></tr>
<tr><td>STD_PAY_AMT</td><td>40.47</td><td>||||||||||||||||</td></tr>
<tr><td>HIGHEND_PROGRAM_FLAG</td><td>39.25</td><td>||||||||||||||||</td></tr>
<tr><td>AVG_BILL_VOICE</td><td>38.98</td><td>||||||||||||||||</td></tr>
<tr><td>AVG_VAS_SMS</td><td>37.66</td><td>|||||||||||||||</td></tr>
<tr><td>VAS_CNND_FLAG</td><td>35.92</td><td>||||||||||||||</td></tr>
<tr><td>AVG_BILL_SMS</td><td>34.98</td><td>||||||||||||||</td></tr>
<tr><td>LOYALTY_POINTS</td><td>34.78</td><td>||||||||||||||</td></tr>
<tr><td>STD_NO_CALLED</td><td>33.90</td><td>||||||||||||||</td></tr>
<tr><td>AVG_VAS_GBSMS</td><td>33.89</td><td>||||||||||||||</td></tr>
<tr><td>AVG_MINS_OBOP</td><td>32.46</td><td>|||||||||||||</td></tr>
</table>

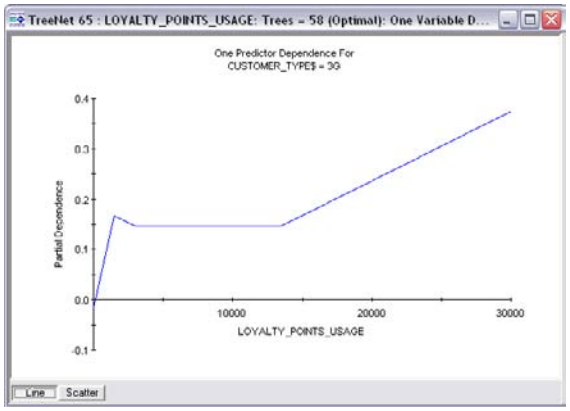
Figure 21. TreeNet partial dependency plots: Contribution of individual predictors to predicted probability of 3G











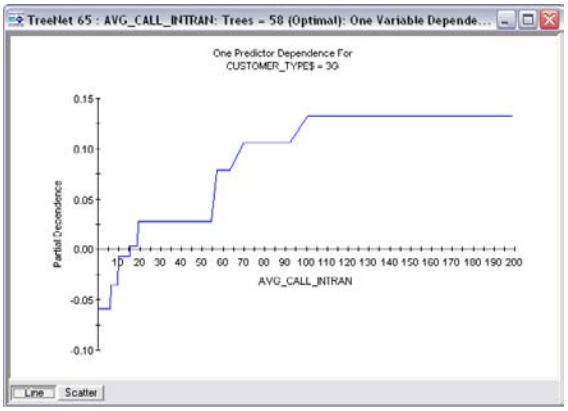
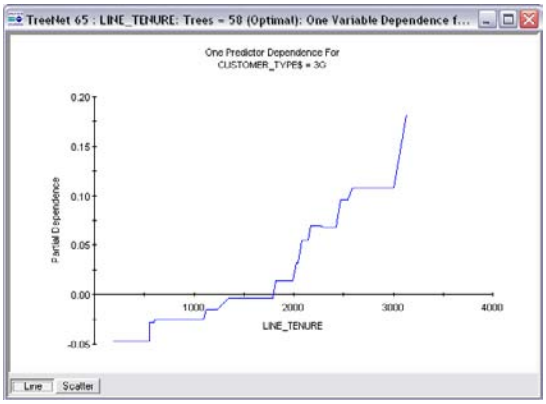
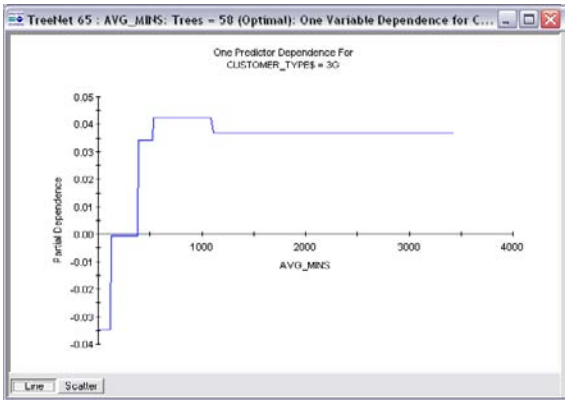


Figure 21. continued





the behavior of 2G and 3G customers (see Figure 21).

These suggest that all the top important variables (except for AGE) are positively associated with the 3G-user group.

DISCUSSION AND CONCLUSION

Data mining competitions are inherently limited by the data and information that the proprietary data owners are prepared to release. In the PAKDD 2006 competition, the limitations permitted a discrimination task to profile the differences between existing 2G and 3G customers at a given point in time. While some of these differences, such as the handset used, are relatively obvious, it was still possible to extract useful and nonobvious insights into the differences between these two customer segments. For example, it was surprising to see that 2G and 3G customers differ strongly along almost all 250 dimensions of the available data. At the same time, it was possible to produce a highly accurate discriminator between the two classes using only eight predictors. Given that the data correspond to historical choices customers have already made regarding which service plan to subscribe to, it is much harder to draw firm conclusions regarding the propensity of a 2G customer to switch to a 3G service. While existing 2G customers who behave very much like 3G customers in terms of their account profile seem likely candidates for a service upgrade, the data contain no information that might help us understand why they have elected to stay with a 2G service plan. Data tracking 2G customers over time and recording which of them switch to 3G eventually would provide more direct evidence on this question.

REFERENCES

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of

continuous features. In A. Priediris & S. Russell (Eds), *Machine learning: Proceedings of the Twelfth International Conference*. San Francisco: Morgan Kaufmann.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29,* 1189-1232.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Annals of Statistics, 28,* 337-407.

Friedman, J. H., & Popescu, B. (2005). *Predictive learning via rule ensembles.* CA: Stanford University, Department of Statistics.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning.* Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41-55.

Salford Systems. (2005). *TreeNet: Stochastic gradient boosting* (Version 2.0) [Computer software]. San Diego, CA: Author.

Steinberg, D., & Colla, P. (1995). *CART: Tree structured non-parametric data analysis.* Salford Systems.

Tusher, V., Tibshirani, R., & Chu, C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *PNAS,* 5116-5121.

This work was previously published in the International Journal of Data Warehousing and Mining, edited by D. Taniar, Volume 3, Issue 2, pp. 32-53, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.23
An Immune Systems Approach for Classifying Mobile Phone Usage

Hanny Yulius Limanto
Nanyang Technological University, Singapore

Tay Joc Cing
Nanyang Technological University, Singapore

Andrew Watkins
Mississippi State University, USA

ABSTRACT

With the recent introduction of third generation (3G) technology in the field of mobile communications, mobile phone service providers will have to find an effective strategy to market this new technology. One approach is to analyze the current profile of existing 3G subscribers to discover common patterns in their usage of mobile phones. With these usage patterns, the service provider can effectively target certain classes of customers who are more likely to purchase their subscription plans. To discover these patterns, we use a novel algorithm called Artificial Immune Recognition System (AIRS) that is based on the specificity of the human immune system. In our experiment, the algorithm performs well, achieving an accuracy rate in the range of 80% to 90%, depending on the set of parameter values used.

INTRODUCTION

A sample dataset of 24,000 mobile phone subscribers is used for this study, in which 20,000 records are known to be second generation (2G) network customers and the remaining are 3G network customers. The target field is the customer type (2G/3G). About 75% of the dataset have the target field available and are used for training purposes; while the remaining quarter of the dataset has a missing target field and are meant for prediction.

The objective of the classification is to correctly predict as many 3G customers as possible from the prediction set and obtain insights on the characteristics of the existing 3G customers to be used as reference for their marketing strategy.

To perform this task, we use a novel algorithm called the Artificial Immune Recognition System (AIRS) proposed by Watkins, Timmis, and Boggess (2004), which is based on the processes in biological immune systems.

OVERVIEW OF AIRS

The Human Immune System

The function of the human immune system is to identify and destroy foreign invaders (antigens) which are possibly harmful to the body. It does this through an innate and nonspecific response (mediated by macrophages) and also with an adaptive and specific response (mediated by lymphocytes). An innate response is not directed towards any specific antigens, but against any invaders that enter the body. The adaptive response is mediated mainly by two types of lymphocytes, B-cells and T-cells. The AIRS approach is modeled based on the behavior of B-cells, hence only the behavior of B-cells will be described here. On the surface of each B-cell are receptors that are capable of recognizing proteins of a specific antigen. Through costimulation and suppression of each other, similar B-cells form networks that can recognize similar antigens.

When antibodies on a B-cell bind with an antigen, the B-cell becomes activated and begins to proliferate. Thus, it means that only B-cells which are able to recognize the invading antigen will proliferate and produce clones (a process known as clonal selection). New B-cell clones are produced which are exact copies of the selected B-cells, but then undergo somatic hyper-mutation to generate a wider range of antibodies, so as to be able to remove the antigens from the body. A small quantity of B-cells remains in the system after the invading antigens have been removed. These B-cells act as an immunological memory to allow the immune system to produce a faster response to similar antigens that might re-infect the body in the future.

The AIRS Algorithm

Processes in biological immune systems have inspired the design of AIRS. An artificial recognition ball (ARB) is used to represent a set of *identical* B-cells. The ARBs in the system will compete for B-cells in order to survive (in the evolutionary sense); therefore an ARB with no B-cell will be removed from the system. A fixed number of B-cells are allowed in the system, so the least stimulated ARB will not be able to get any B-cells, and will therefore be removed from the system. We will at times refer to a B-cell as an ARB, only because an ARB is simply a representation of many B-cells of the same specification. When a new training data record (antigen) is presented to the system, each B-cell is cloned in proportion to how well it has matched the antigen according to the principle of clonal selection. The mutation rate used in the cloning process is *inversely* proportional to how well it matches the antigen. During the mutation process, new clones undergo a process of somatic hyper-mutation, where each attribute of the clones is varied slightly to provide a wider range of response to the training data record. Eventually, the clone with the best fit to the presented antigen will be retained as a memory cell. The memory cells are retained in the system to provide faster response should the system become re-infected with similar antigens.

AIRS relies heavily on finding the similarity (or difference) between a pair of customer records, therefore a proper distance measure needs to be defined. Hamaker and Boggess (2004) conducted a survey of distance measures that can be used in conjunction with AIRS. Based on this survey, we use the *heterogeneous value difference metric*

(HVDM) as our distance measure. The HVDM measure uses the *Euclidean distance* measure for numerical data fields and a *value difference metric* (VDM) for categorical data fields. Consider *x* and *y* as two customer records that are going to be compared, each having *G* data fields and having one of *C* possible classifications (where *C* = 2, either being 2G or 3G). All numerical fields are normalized to the range of [0, 1]. The function *HVDM*(*x*, *y*) returns the HVDM distance of *x* and *y* according to the formulas:

$$HVDM(x,y) = \frac{1}{\sqrt{G}}\sqrt{\sum_{g=1}^{G} hvdm(x_g, y_g)^2}$$

$$hvdm(x_g, y_g) = \begin{cases} \sqrt{vdm(x_g, y_g)} \\ \left| x_g - y_g \right| \end{cases}$$

g is categorical field
g is numerical field

$$vdm(x_g, y_g) = \sum_{c=1}^{C} (P(c \mid x_g) - P(c \mid y_g))^2$$

where $P(c|x_g)$ denotes the probability that a customer record has class *c* given the value x_g.

Based on our judgment, VDM is the most effective measure for categorical fields compared to other distance measures as presented in Hamaker and Boggess (2004) because it relies on the statistics observed in the training data, which more likely reflects the real-world situation. We will now present the algorithm (full details may be found inWatkins et al., 2004). The symbols that we will be using to define the algorithm are explained as follows:

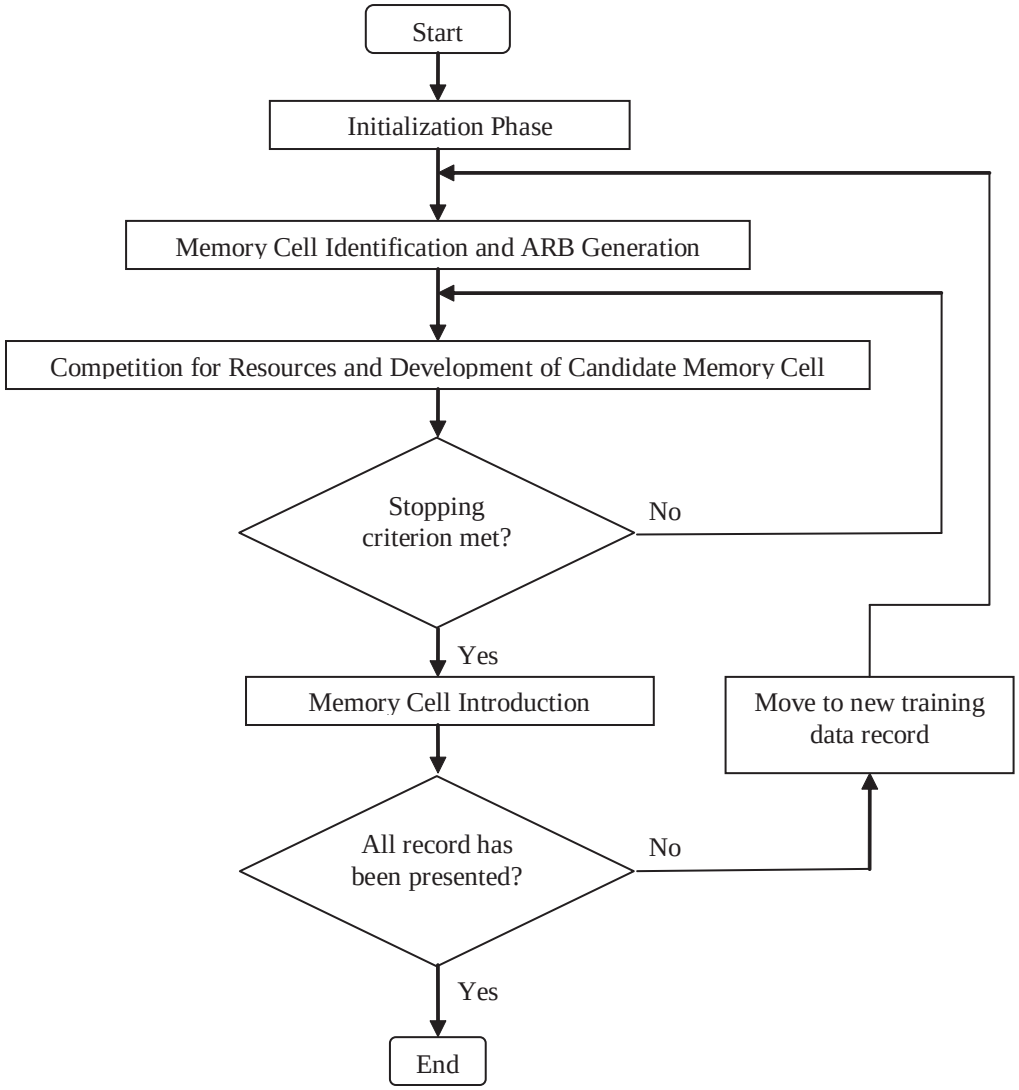
- There are *n* antigens, and there are *G* data fields and *one categorical* class in each antigen. The class of the antigen may take the value of $\{1, 2, \ldots, nc\}$, where *nc* is the number of possible values in the target class.
- *MC* denotes the set of memory cells. *mc* represents an individual member of *MC* and *mc.c* represents the class of the memory cell, while $mc.f_i$ represents the value of the i^{th} feature (data field) in the memory cell.
- *AB* denotes the set of ARBs and *ab* represents a single ARB. *ab.c* denotes the class and the i^{th} feature of an *ab* respectively.
- *ag* represents an antigen (training data record), and *ag.c* and $ag.f_i$ represent the class and value of the i^{th} feature (training data field) in the antigen, respectively.
- $MC_c \subseteq MC$ denotes the set of memory cells with the class *c*.
- *ab.stim* denotes the stimulation level of an ARB *ab*.
- *ab.resources* denotes the number of resources (B-cells) currently held by an ARB *ab*.

The user-defined parameter values we used in our experiments to control the behavior of the algorithm are presented as follows:

- *numEpochs*: Number of passes through the training data (= **3**);
- *numResources*: number of B-cells (resources) allowed in the system (= **200**);
- *clonalRate*: Number of mutated clones that a given ARB is allowed to produce. An ARB *ab* is allowed to produce at most (*clonalRate* * *ab.stim*) new clones. This product is also used to control how many B-cells will be allocated to an ARB (= **10**);
- *mutationRate*: A value between [0, 1] that determines the likelihood that a given feature or class of an ARB will be mutated (= **0.1**);
- *hypermutationRate*: number of mutated clones that a given memory cell is allowed to inject into the system. A memory cell *mc* injects at least (*hypermutationRate* * *clonalRate* * *mc.stim*) mutated clones (= **10**);

- *distance*: The distance metric used (= ***Heterogeneous Value Difference Metric***);
- *ATS*: affinity threshold scalar, a value between [0, 1] that will be multiplied with the affinity threshold (*AT*) to produce the cutoff value for memory cell replacement (= **0.2**);
- *stimThreshold*: a value between [0, 1] used as a stopping criterion for training on specific antigen (= **0.95**); and
- *k*: The parameters that specify how many nearest neighbors participate in the classification process (using *k*-nearest neighbor) (**= 15**);

Figure 1. Process flow of the AIRS training process



These parameter values were determined by performing multiple tests on the training data (using 10-fold cross validation) to verify their effectiveness. The training process of the algorithm is shown in Figure 1. Each phase will be discussed separately followed by the classification process.

The Initialization Phase

During this phase, all data preprocessing that is needed will be performed. Details of the data preprocessing involved will be elaborated in a later section. After the preprocessing, the affinity threshold (AT) of the system is calculated by averaging the distance between all training data. It should be noted that the affinity between two antigens should always be between 0 and 1, therefore

$$AT = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} affinity(ag_i, ag_j)}{{}^{n}C_2}$$

where *affinity*(*x*, *y*) will return the distance of the two antigens *x* and *y*, depending on which distance metric is used (HVDM for this specific case). Seeding of memory cells, when desirable, will also be performed in this phase.

Memory Cell Identification and ARB Generation

In this phase, the best matched memory cell is identified and ARBs are generated from the best matched memory cell that is identified. The first step is to identify the memory cell mc_{match} from $MC_{ag.c}$ (the set of memory cells with the same class label as the presented antigen *ag*) which has the highest stimulation. The level of stimulation between two records (antigens) *x* and *y* is defined as *stimulation*(*x*, *y*) = 1 – *affinity*(*x*, *y*). If $MC_{ag.c}$ = {} then mc_{match} = *ag* and *ag* is added to $MC_{ag.c}$.

After mc_{match} has been identified, *NumClones* number of clones of mc_{match} will be produced as ARBs. The number of clones is *NumClones* = (*hypermutationRate* * *clonalRate* * mc_{match}.*stim*) where *stim* denotes the current stimulation level. It should be noted that *NumClones* is proportional to the stimulation value of mc_{match} to the antigen. Each numerical feature in mc_{match} can only mutate in a constrained range centered at the original value. The width of the range is inversely proportional to the stimulation value *stim*. This implies that a heavily stimulated mc_{match} will produce clones that are closer to it because it has a limited range for mutation, which is reasonable since the original mc_{match} already matches well with the antigen; hence it is very likely that best-fitted B-cell can be found in the neighborhood of mc_{match}. A categorical feature does not have this constraint and can be mutated freely. All of the produced clones are added into the system as ARBs and will have the same class label as the original mc_{match}.

The pseudo code for this phase is shown in Figure 2. maxStim($MC_{ag.c}$, *ag*) returns a memory cell in the set of $MC_{ag.c}$ with the maximum stimulation to the antigen *ag*. mutate(mc_{clone}) mutates the clone of a memory cell according to the rule described previously. *AB* represents the set of ARBs (or set of representative B-cells) in the system, and |*AB*| denotes the size of the set *AB*.

Competition for Resources and Development of a Candidate Memory Cell

The objective of this phase is to find the ARB in the system (which is produced from the previous phase) which is the most stimulated when presented with an antigen. This ARB may then be retained as a memory cell. Recall that an ARB is a set of *identical* B-cells, therefore its characteristics can also be represented by a single cell. To find the most stimulated ARB, we allocate resources (B-cells) to all ARBs in the system based on their

Figure 2. Memory cell identification and ARB generation

```
if (MC_ag.c = {})
        mc_match = ag
        MC_ag.c = {mc_match}
else
        mc_match = maxStim(MC_ag.c, ag)
endif
mc_match.stim = stimulation(ag, mc_match)
NumClones = hypermutationRate * clonalRate * mc_match.stim
AB = {}
while ( |AB| < NumClones) do
        mc_clone = mc_match
        mc_clone = mutate(mc_clone)
        AB = AB ⋃ mc_clone
endwhile
```

stimulation level to the current antigen. An ARB *ab* is given (*ab.stim * clonalRate*) B-cells.

However, as B-cells allowed in the system are limited (set by the value of *numResources*), if the number of allocated B-cells exceeds the allowed number of B-cells; some B-cells need to be removed from the system. B-cells will be removed starting from the ARBs with the weakest stimulation, until the number of B-cells in the system equals to or are less than *numResources.* In this way, ARBs which are weakly stimulated by the antigen will be removed from the system. The surviving ARBs are examined by the algorithm to test whether they are stimulated enough by the antigen to cease further training. The average stimulation level *s* is computed by averaging the stimulation level of all surviving ARBs. If $s \geq stimThreshold$ then the stopping criteria is met.

Regardless of whether the stopping criterion is met or not, the surviving ARBs are then allowed to produce (*stim * clonalRate*) clones which undergo somatic hyper-mutation. The mutation rule is the same as the rule described in the previous section where numerical fields have constrained ranges for mutation. If the stopping criterion is met, the most stimulated ARB in the system is chosen as the candidate memory cell $mc_{candidate}$. The algorithm then proceeds to the next training phase, otherwise this phase is repeated.

The pseudo code for this training process is shown in Figure 3. minStim(*AB*) returns an ARB with minimum stimulation from the set of ARBs *AB.* avgStim(*AB*) returns the average stimulation value of ARBs in *AB. newAB* and *MU* are temporary storage locations for newly created ARBs.

Figure 3. Competition for resources and development of a candidate memory cell

```
resAlloc = 0
foreach (ab . AB) do
        ab.resources = ab.stim * clonalRate
        resAlloc = resAlloc + ab.resources
endfor
while (resAlloc > numResources) do
        numRemove = resAlloc – numResources
        ab_remove = minStim(AB)
        if (ab_remove.resources≤ numRemove)
                AB = AB – {ab_remove}
                resAlloc = resAlloc – ab_remove.resources
        else
                ab_remove.resources = ab_remove.resources – numRemove
                resAlloc = resAlloc – numRemove
        endif
endwhile
s = avgStim(AB)
newAB = AB
foreach (ab . AB) do
        NumClones = clonalRate * ab.stim
        MU = {}
        while ( |MU| < NumClones) do
                ab_clone = ab
                ab_clone = mutate(ab_clone)
                MU = MU ⋃ ab_clone
        endwhile
        newAB = newAB ⋃ MU
endfor
AB = newAB
if (s < stimThreshold)
        // repeat this phase
else
        //continue to next phase
endif
```

Memory Cell Introduction

In this phase, we will decide whether $mc_{candidate}$, identified in the previous phase as the ARB with highest stimulation, should be retained as a memory cell and whether $mc_{candidate}$ should replace mc_{match}. $mc_{candidate}$ is retained as memory cell if it is more stimulated compared to mc_{match} with respect to the current antigen. Furthermore, if the affinity between $mc_{candidate}$ and mc_{match} is less than the threshold ($AT * ATS$), mc_{match} is going to be replaced by $mc_{candidate}$. It means that if $\underline{mc}_{candidate}$ and mc_{match} have strong affinities (the distance between the two is sufficiently large), one memory cell cannot be used as a substitute for the other. If the affinity is weak, one memory cell can be regarded as a substitute of the other, and since $mc_{candidate}$ is more stimulated than mc_{match}, $mc_{candidate}$ is used as the memory cell.

Once the process is finished, the algorithm completes training for one antigen. If there are any other antigens (other data records) that need to be trained, the algorithm returns to train the new antigen. If all antigens have been trained, the training phase is finished and the produced memory cells can be used for classification.

Produced Classification Model

The algorithm produces a set of memory cells as a result of the training. In a clustering algorithm, each of the memory cells may be visualized as a cluster center. Using the parameter values specified before, we observed that the number of memory cells produced is only around 10% of the number of training data records.

Classification is performed using a weighted *k*-nearest neighbor approach; where *k* most stimulated memory cells have the right to vote for the presented antigen class. Since we have an imbalanced training data (3,000 3G customers against 15,000 2G customers), we can expect that there will be more memory cells representing the 2G customers compared to 3G customers; hence unweighted voting puts the 3G customers at a high probability of being incorrectly classified as 2G customers. To compensate for this, we set the ratio for a vote of 3G:2G = 5:1 (proportional to the ratio of training data available). Testing has shown that a weighted *k*-nearest neighbor approach performs more effectively compared to the unweighted approach.

As for the *k* value used to determine how many neighbors get to participate in the voting process, we used $k = 15$. By experimentation, we vary the value of *k* from $k = 1$ to 15 and observe that the solution quality increases. When *k* is higher than 15, there are some cases where the solution quality is reduced.

EXPERIMENTAL DESIGN

We are provided with 24,000 customer records, from which 18,000 records are to be used for training purposes, while the remaining 6,000 are for prediction. There is one categorical target field, which is customer type (2G/3G). Of the 18,000 training data records, 15,000 are 2G customers, and 3,000 are 3G customers. Each record in the dataset consists of 250 data fields. The data field can be categorized into two types; *numerical* and *categorical.*

Data Preprocessing

Before we supply the data into the AIRS algorithm, we need to preprocess the data. The data preprocessing that is done are normalizing numerical value, handling missing values, and converting categorical variables from literal strings to integer indices for easier processing.

For each numerical data field, we perform min-max normalization. First, we scan the data to find the minimum and maximum value of the field, then we normalize each value in this field to the

range of [0,1] by converting it to $\frac{val - \min}{\max - \min}$, where *val* is the value to be converted, *min* is the minimum value, and *max* is the maximum value.

For missing values that occur in the dataset, we simply replace them with a global value "MISSING," however, other strategies such as statistical regression might be more effective in handling the missing value.

For categorical data fields, we give an index to each possible value that might appear in the data and change the values from literal strings to indices. This has the advantage of more efficient processing because the algorithm only needs to deal with numbers instead of strings.

In addition, we also need to compute the conditional probability value $P(c|x_g)$ which is needed if we use HVDM distance measure. $P(c|x_g)$ denotes the probability that a customer record has class c given the value x_g.

We also note that in the dataset, there are some data fields in which only *one* value would appear in all customer records for the entire training data (e.g., **HS_CHANGE**, **TOT_PAS_DEMAND** for a numerical data field, and **VAS_SN_STATUS** for a categorical data field). We would not get any useful information from these fields since there is only *one* value that appears in the data and no other value to compare it to. Therefore, for these data fields, they can be safely removed from the data.

Training, Validation, and Parameter Tuning

To decide on which parameter values would yield the best results, we try several sets of parameter values. Validation is performed using 10-fold cross validation. We divide the training data randomly into 10 equal parts. For every run of the algorithm, nine parts will be used for training, and the remaining part will be used for validation purposes. We train the system with the training data, and we analyze the accuracy of the prediction on the validation set. By performing the validation methods repeatedly with different sets of parameter values, we can pick the set of parameter values which yield the best results. This is further illustrated in the next section.

RESULTS AND DISCUSSIONS

Obtained Results and Analysis

To perform the experiment, we perform a 10-fold cross validation method 10 times for each set of parameter values. The result shown in Table 1 is the average of the obtained values. Due to the time constraint, we only modify the number of training passes (*numEpochs*) and the value of *k* for classification using *k*-nearest neighbor.

As we can see from Table 1, as *k* increases, the accuracy rate generally increases, however, we also observe that sometimes, for *k* value higher than 15, the accuracy rate decreases. Therefore, we set the *k* value to 15. The same can be observed in the number of training passes, when the number of training passes is greater than 3, sometimes, the accuracy decreases. In addition, too many training passes increases the time involved in training the algorithm. Hence, we set the number of training passes to 3. Other values, such as the clonal rate, hyper-mutation rate, and number of allowed B-cells in the system are the default values for the algorithm which are initially used in the original source code by Hamaker and Watkins (2003).

Dataset Analysis

The discussion in this section will be based only on the training data, since we will need to know whether a record is incorrectly classified as false positive (classified as 3G although the actual class is 2G). Since our algorithm does not produce the characteristics of 3G customers, therefore the

Table 1. Accuracy rate of AIRS with varying sets of parameters

Number of training passes	*k* value for k-nearest neighbor			
	1	7	11	15
1	85.53%	86.54%	86.98%	87.22%
2	86.37%	87.33%	87.49%	87.82%
3	86.71%	87.47%	87.81%	88.08%
4	86.97%	87.69%	88.04%	88.26%
5	86.90%	87.81%	88.19%	88.32%

false positives obtained in our experiment might be a useful insight to decide the characteristics of customers that are likely to change their 2G subscription to 3G. If a false positive is found, it is possible that this record is closer to the characteristics of 3G customers rather than 2G customers, therefore, by analyzing the characteristics of all the false positives that are found, we can find some common characteristics that differentiates the 3G customers from the 2G customers.

To perform this experiment, we randomly divide the training data into 90% training set, and 10% testing set. After training the algorithm with the training set, we classify the testing set and obtain the false positives produced by the algorithm. This experimentation is repeated several times, and the statistics of the false positives obtained for each experiment is compared to the statistics of the entire training data whether there are a constant deviation between the statistics of the false positives and the entire data. For numerical data field, we compared the average and standard deviation, while for categorical data field, we compared the frequency count.

We find some almost constant, strong deviation between the statistics of false positives and the statistics of training data in the following fields. The field name uses the original name that is found in the dataset.

- **MARITAL_STATUS**: *More singles* can be found in the false positives compared to the entire data; this suggests that singles are more likely to purchase a 3G subscription.
- **OCCUP_CD**: *Less* customers with *other (OTH)* occupation is found in the false positives; therefore, customers with occupation {EXEC, POL, STUD, MGR, HWF, ENG, CLRC, SELF, GOVT, TCHR, SHOP, FAC, AGT, MED} (this occupation code are other codes that can be found in this data field) might be more likely to change to 3G.
- **HIGHEND_PROGRAM_FLAG**: *More* customers with *high end* programs are found in the false positives. This is expected since a 3G plan can also be considered a high-end program.
- **TOP1_INT_CD, TOP2_INT_CD, TOP3_INT_CD**: *Considerably more* customers in the false positives with values for this parameters set to *other than NONE*. It is probable that customers who make international calls are more attracted and able to afford a 3G subscription.

- **VAS_GPRS_FLAG**: *More* customers who own a *GPRS Data Plan* (code 1) can be found in the false positives. This is expected, since customers without GPRS are less likely to appreciate the improvement provided by 3G, and therefore, might not find a 3G subscription attractive.
- **LOYALTY_POINTS_USAGE**: Customers in the false positives set have *considerably higher average of loyalty points.* Loyal customers might be more easily persuaded to switch to a 3G subscription.
- **TOT_TOS_DAYS**: Customers in the false positives set have *a lot less average total temporarily on suspended* days compared to the entire data. Customers who are routinely using their mobile phone are more likely to purchase a 3G subscription.
- **AVG_CALL_FRW_RATIO**: *Less call forwarding* is utilized by customers in the false positive set. Customers who are carrying their mobile phone everywhere, therefore they do not need call forwarding services, are more likely to purchase a 3G subscription which would enable them to do more with their mobile set.
- **AVG_MIN_OBPK, AVG_MIN_OBOP**: *More minutes are spent in outbound calls in peak and off-peak period* by customers in the false positives set. Customers who make a lot of calls might be interested in the video phone feature provided by a 3G subscription.
- **AVG_MIN_FIX**: More minutes are spent by customers to call fixed line numbers in false positives set. Customers who are using their mobile phone actively are more likely to purchase a 3G subscription.
- **AVG_MIN_T1:** More top 1 minutes spent by customers in false positives set. Customers who are using their mobile phone actively are more likely to purchase a 3G subscription.
- **AVG_CALL_1900**: *Considerably more than 1,900 calls* by customers in the false positives set. Customers who are using their mobile phone for social use as well as for communication are more likely to purchase a 3G subscription.
- **AVG_REVPAY_AMT, REVPAY_FREQ**: *Slightly more reverse payment* in average for both the amount and the frequency can be detected in the false positives set.
- **CONTRACT_FLAG**: *Slightly more* customers *with contract* are found in the false positives. This may be caused by the price of 3G mobile phones that are still expensive; therefore, an attractive contract plan might attract more customers to switch to 3G.
- **AVG_VAS_QTUNE**: Customers in the false positives set *downloaded more quick tunes* than the average of the entire data. However, other services such as quick games, text, or pix do not exhibit any statistical deviation from the data.
- **AVG_VAS_GBSMS**: *More e-mail SMSs* are sent by customers in the false positives set. These customers might be interested in the functionality provided by 3G.

CONCLUSION

In this article, we described our approach in mining interesting patterns from the dataset of mobile phone customers to identify whether the customer is currently subscribed to a 2G or 3G subscription plan. To solve this problem we use the AIRS algorithm which operates using the principles of the human immune system. We experimented with different sets of parameter values to find the optimal values that are suitable for our purpose. Our experimentation shows that this algorithm is very effective; with proper parameter values it can achieve an accuracy rate near 90%. After a classification model was produced, we analyze

the training data for patterns found in the classification model. These patterns that we found can hopefully be used to identify the common characteristics of 3G and 2G customers. We also discuss how these patterns can assist a mobile phone service provider in marketing their 3G subscription plan.

REFERENCES

Goodman, D., Boggess, L., & Watkins, A. (2003, July). An investigation into the source of power for AIRS, An artificial immune classification system. In *Proceedings of the 2003 International Joint Conference on Neural Networks.*

Hamaker, J., & Boggess, L. (2004, June). Non-Euclidean distance measures in AIRS, an artificial immune classification system. In *Proceedings of the 2004 IEEE Congress on Evolutionary Computation* (pp. 1067-1073).

Hamaker, J., & Watkins, A. (2003). Artificial immune recognition system (AIRS) Java Source Code.

Marwah, G., & Boggess, L. (2002, September). Artificial immune system for classification: Some issues. In *Proceedings of 1st International Conference on Artificial Immune Systems (ICARIS)* (pp. 149-153).

Watkins, A. (2001, November). *AIRS: A resource limited artificial immune classifier.* Unpublished master's thesis.

Watkins, A., & Boggess, L. (2002, May). A new classifier based on resource limited artificial immune system. In *Proceedings of IEEE Congress on Evolutionary Computation* (pp. 1546-1551).

Watkins, A., & Timmis, J. (2002, September). Artificial immune recognition system (AIRS): Revisions and refinements. In *Proceedings of 1st International Conference on Artificial Immune Systems (ICARIS)* (pp. 173-181).

Watkins, A., Timmis, J., & Boggess, L. (2004, September). Artificial immune recognition system (AIRS): An immune-inspired supervised machine learning algorithm. *Genetic Programming and Evolvable Machines, 5*, 291-317.

This work was previously published in the International Journal of Data Warehousing and Mining, edited by D. Taniar, Volume 3, Issue 2, pp. 54-66, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.24
Mobile Ontologies:
Concept, Development, Usage, and Business Potential

Jari Veijalainen
University of Jyvaskyla, Finland

ABSTRACT

The number of mobile subscribers in the world is soon reaching the three billion mark. According to the newest estimates, majority of the subscribers are already in the developing countries, whereas the number of subscribers in the industrialized countries is about to stagnate around one billion. Because especially in the developing countries the only access to Internet are mobile devices, developing high quality services based on them grows in importance. Ontologies are an important ingredient towards more complicated mobile services and wider usage of mobile terminals. In this article, we first discuss ontology and epistemology concepts in general. After that, we review ontologies in the computer science field and introduce mobile ontologies as a special category of them. It seems reasonable to distinguish between two orthogonal categories, mobile domain ontologies and flowing ontologies. The domain of the former one is in some sense related with mobility, whereas the latter ones are able to flow from computer to computer in the network. We then discuss the creation issues, business aspects, and intellectual property rights (IPR), including patentability of mobile ontologies. We also discuss some basic requirements for computer systems architectures that would be needed to support the usage of mobile ontologies.

INTRODUCTION

The mobile subscriber base in the world is growing fast. The industry itself estimates that at the end of 2006, the number of subscribers reached 2.7 billion and it is expected that the number of subscribers will grow with 480 million during 2007 (Umsoy, 2007). The biggest growth will be in the developing countries, like India and China. There are over 130 3G WCDMA networks in 60 countries with 100 million subscribers. The latter number will grow to 170 million during 2007.

Low-cost WCDMA terminals (65 € a piece) are coming to the market. High Speed Packet Access (HSPA) with 3.6 Mbps downlink capacity is deployed in 51 countries in 93 networks and there were 128 devices on the market supporting HSDPA in March 2007 (Umsoy, 2007).

Digital convergence is tearing apart the old barriers between entertainment, media, telecom and computer industries, and all these industries are melting together into one huge industry. At the same time, the wireless operators are pondering their position on this market. They want to be more than bit pipes providing access to Internet for wireless terminals. Many operators think that they must provide better and more appealing services to the subscribers. How can appealing services be offered to mobile users both in developed and developing countries? These can be location-based or context-aware in a wider sense, or other services adapted to mobile Internet.

At the same time, the top models of mobile wireless terminals have reached capabilities of a laptop computer a few years ago with gigabytes of memory, programmability, fast processors, GPS receivers, text editors, calendars, e-mail clients, browsers, and so forth. Many have cameras and can record images and video with sound. Thus, users have begun to generate multimedia contents using these devices. Assuming that a user takes, for example, 2,000 digital photographs and some video clips in a year, there will be a substantial number, tens of thousands, even hundred thousand of these kinds of objects after 30-50 years. These are mostly relevant for the person himself or herself and for his friends and family members. Managing reasonably these emerging digital archives requires semantic metadata that cannot be generated fully automatically. Rather, user's help is needed (Sarvas, 2006). Also, storage space (approaching terabyte range for a life-time archive) is a problem and the stability of the formats used. Who guarantees that for example, the currently so popular JPEG-format would be supported in 2060? If the format originally used to store the images or video clips is not any more supported, what kind of automatic means are there to transform the contents into newer formats?

YouTube (Youtube, 2007) is currently one of the most known sites in the world where people can upload their video clips and other users can download them. Many of those videos have been produced by mobile handsets. Flickr (Flickr, 2007) offers sharing of photographs and a simple annotation in the form of tags. Not all material is suitable for distribution all over the world, though, for moral, legal, cultural, or privacy reasons. For instance, Flickr site does not allow sexually-oriented contents beyond a certain limit, although the tags "sex" and "sexy" are in use.

The above needs of individuals while managing and sharing digital contents are rather different from those of the companies offering various kinds of *mobile services.* Both can be satisfied in several ways. Perhaps the most sophisticated approach is to use *ontologies* in all these contexts. Because terminals are becoming more and more powerful over time, they can also be used to run complex computations, for example, inferences, required while using *formal ontologies* for various purposes. This is the main motivation behind this article.

In section 2, we discuss the concept of ontology and epistemology in general and in section 3, we discuss the concept of ontology in computer science field. In section 4, we discuss what should be understood by mobile ontologies. In section 5, we will turn our attention to the ontology creation issues in general and the peculiarities when creating mobile ontologies. In section 6, we discuss so-far largely ignored business and IPR issues related to mobile ontologies. Section 7 concludes the article.

ONTOLOGY AND EPISTEMOLOGY

"What is there? What exists?" This could be understood as the basic question of ontology,

that is, *study of being* or existence in philosophy. Different answers to this question were given over the course of history. One fundamental issue is, whether the deepest reality is ever changing and moving or whether it is stable and movement is just a human illusion After the Renaissance in Europe, one began to explore the nature in order to find the explanations from it itself by the methods of natural science. The starting point of the scientific inquiry is that there are no reasons or causes external to the nature and that the causes and true explanations can be found by interacting with the nature. Another starting point is that human perception of the nature is different from the nature itself, because otherwise we would know the essence of nature without further interactions and its exploration would be unnecessary. We adhere to the view that there are two different realities, one within the individual human *consciousness* and the other outside of it. Further, change and movement are real and change is actually the ultimate attribute of the realities. Stability is just transient.

The modern thinking and the concrete exploration of the nature have radically changed the answer to the ontological question and also the structure of modern societies. It has brought up such *concepts* as electrons, neutrons and other particles, atoms consisting of them and possibility to split atoms to gain energy, molecules and chemical industry, DNA, living cells, bio-industry and modern medicine, electro-magnetism, electronics and electrical industry including computer industry, relativity theory, galaxies, black holes, and so forth. In mathematics, one has developed axiomatic mathematics, formal logic and formal languages, paving the way to computing and formal ontologies.

The modern thinking has further divided the external reality into *physical* and *social* reality. Physical reality exists also without people, that is, it existed before any human being existed and will exist after human beings have vanished from the universe with their concepts and ideas, such as stability, gods, good life, good and bad, right and wrong. That is, physical reality is more fundamental than social reality, although admittedly also other views on the structure of physical reality and its relationship with the social reality have existed in the consciousness of the individuals and collectives over history.

Already the great Greek philosophers Plato and Aristotle posed the *epistemological* questions "What can we humans know of the reality?", "What is truth?", "What is knowledge?" These questions are always related with the basic ontological *world view that* covers all the elements considered basic by a certain group of people, such as nature or matter, human beings and perhaps various kinds of spirits or souls. *Concepts* and their relationships, *theories*, form the knowledge part of the contents of the consciousness and thus the above epistemological questions can be rephrased in terms of concepts and utterances about their relationship with the physical and social external reality. Which concepts refer to something existing and which relationships can be confirmed in the reality?

Coming back to the fundamental questions of ontology and epistemology and their relationships, our view is the following. A plausible solution to a consistent ontology and epistemology is that human beings develop their concepts using their mental, creative capabilities, while *interacting* with the physical reality and communicating with other people in a society, that is, while being part of the social reality. The *individual and collective needs* to develop and adopt new *concepts* emerge always in a certain social context. This means that the answer to the basic ontological question, "what is there", changes over time, represented in new concepts and their relationships. The concepts are developed by various groups of people in interaction with themselves and with the physical reality. During the modern times, these kind of activities produce ontology that is based on

the methods of natural and other sciences and explain the nature from and within itself, whereas in the past various kinds of spirits or gods were often included into the ontologies supported by various groups. Another object of the studies is the social reality, that is, human societies. In our view, macro- and microeconomic theories, as well as concepts of politics and social sciences belong to this sphere. The modern concepts and theories describing these are evidently widely different from those describing physical reality. Individuals adopt the views developed by previous generations and mediated by oral and/or written means and only some develop them further. In current societies different groups of people and even a single individual can have different ontological views that can be also contradictory.

In the modern separation of physical and social reality, the *artifacts* developed by human beings, such as houses, cars, fashion, mobile networks and terminals belong clearly to the physical reality created by human beings. But their existence or development cannot be explained in a similar manner from nature as development of, for example, mountains or oceans. Rather they are at the same time also results of human social activity. This duality should be reflected in the specific ontology describing them. There is an extensive discussion on a suitable ontology for these kind artifacts, that is, in Pohjola (2007). These kinds of discussions about the "true essence" of artifacts are relevant for ontologies in computer science, because they also are socially created artifacts that exist in the physical (digital) reality. What is a suitable "artifact ontology" for computerized ontologies is an interesting question, but beyond the scope of this article.

The general questions and issues above are relevant also for ontology development in the computer field. We point to them below at appropriate places.

ONTOLOGIES IN COMPUTER SCIENCE

In the computer science field, McCarthy introduced in 1980 the concept *environment's ontology* that contained a list of concepts involved in a problem (environment) and their meanings (Sanchez, 2007). Since then, the term has been associated with the representation of concepts and the usage has spread out to many fields of computer science. The often cited definition by Gruber reads "ontology is an explicit conceptualization of a domain", although in the abstract of the same paper it reads "A specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects — is called an ontology" (Gruber, 1993). The definition does not explicitly relate ontology to a specific group of people or its validity period. It has also been pointed out, for example, by Smith (2004) that "conceptualization" remains undefined in the definition.

In the sequel, we mean by ontology *a conceptualization of a domain created and shared by a group of human beings. The domain can also be called a possible world.* This definition stresses that an ontology must be understood by a group of people in the same way as far as it is possible; must be representable in an interpersonal form (using natural or formal language, pictures, etc.), and must refer to a common domain that the group agrees upon. The domain or possible world does not need to exist in the current physical or social reality, only in the consciousness of the group members, and as the corresponding externalized representation. Most often, though, an ontology is established with the goal of claiming that the domain indeed has had, has, or will have the structure, relationships and properties the ontology referring to the domain claims it to have. In this narrower sense, an ontology can be understood as a *body of knowledge describing*

some domain and as a representation vocabulary for it. In this capacity, it also makes sharing of knowledge among individuals and groups possible (Chandrasekaran, 1999). This view on ontology is narrower than the general definition above because knowledge should be truthful, that is, refer to reality it is to the best of our knowledge. One cannot say the same of an arbitrary possible world, or even about a possible future world. The fitness of an ontology as a carrier of knowledge is discussed, for example, in Smith (2006b) and more general discussion on the various aspects on ontologies can be found in Brewster (2007) and in the special issue of the International Journal of Human-Computer Studies.

Conceptualizations we have in mind are *terms* with natural and/or formal language definitions that *refer* to universals or instances in the physical or social reality or human consciousness, and the relationships between the terms. A fairly general way to construct ontology is expressed in LOA (2007): "ontology is assumed here as a semiotic object, including at least three objects: a graph (information object), a conceptualization (description), a semantic space (abstract). The semantic space refers to the 'formal' semantics of an ontology graph, while the conceptualization refers to its 'cognitive' semantics". For instance, "VW is-a car" means that there is a collection of physical artifacts called cars on Earth and some of them are manufactured by Volkswagen. The semantic space contains the set of cars and VW-cars are a subset of it. These can be represented by set-theoretic, inter-subjective notations, but the notations can only again refer to the real collection of physical cars by human mental act—or by computerized version thereof. The cognitive "car" refers to the culturally given connotations of a car as an artifact, vehicle, status symbol, and so forth, as individuals perceive them. Depending on the goal of the ontology, certain relationships with other concepts are included into the ontology graph, but not necessarily all. For example, the concept of car can only be related with vehicle (as a subclass), but other culturally relevant concepts are not included into the ontology. Thus, an ontology can be understood in the same way by different people, as concerns the concepts (signs), but these can have different connotations and thus the ontology is not understood in the same way by the people (Mancini, 2006).

Ontologies in computer science have been classified in several ways. The authors of vHeijst (1997) classify ontologies, according to their use, in *terminological, information, and knowledge modeling* ontologies, whereas Guarino (1998) divides them into *top-level, domain/task,* and *application* ontologies, based on their generality. The authors in Gomez-Perez (2003) classify ontologies into *lightweight* and *heavyweight* ontologies. The former include concepts, concept taxonomies, relationships between concepts, and properties describing the concepts. The latter enhance the former with axioms and constraints. D. Fensel introduces five categories differentiating between *metadata, domain, generic, representational,* and *method/task* ontologies (Fensel, 2004). Thus, we see that currently there is no commonly accepted single classification for ontologies.

Ontology development is an activity performed by a group of people. At the beginning of the development process, humans usually rely on written and spoken natural language and informal schemata and the end result can even remain in this informal form. The process of developing ontology can be from bottom-up or top-down, although both directions are usually mixed. The mixed version, where both directions are used simultaneously, is often called middle-out. In the bottom-up approach, the existing terms and concepts of a domain are taken as a starting point, definitions given and relationships between concepts established that reflect the relations in the domain. In the top-down approach, one can

start from an existing upper ontology and refine and enhance it with the concepts of the domain. Of course, the process of developing ontologies is rather complex. For instance, the borders of the domain can be unclear and the domain understood in a different way by participants at the beginning. Thus, as part of the ontology development, the exact domain can emerge and become shared by the people.

In general, the external representation of ontology is composed of symbols of one or several languages. It is a representation of the shared understanding of the group members of what the domain is and what is essential or interesting for the domain. As such, it is just a finite (information) object, carrying meanings to people, and it can be encoded into a bit string and stored on a computer. It means something only for those people who understand the languages used. A computer "understands" an ontology or part of it, if there is a portion represented with a formal enough language that can be interpreted by the computer. In other words, programs must run that use this portion as input and that compute results using it. Such ontologies that are composed using a formal language or contain a formal language portion are called *formal ontologies,* others are called informal *ontologies.*

The formal languages used are usually subsets of first-order predicate logic, such as description logic or frame logic (Staab, 2004), but also (extended) UML and ER-notation have been proposed to be used as a formal ontology language (Sanchez, 2007). *Automatic reasoning* in computers based on the axioms and inference rule(s) of the logic become thus possible. Generally, any formal language that can be given an operational, computable semantics would be a possible candidate for an ontology description language, as long as it is easy for a human being to represent concepts and their relationships, as well as restrictions that are typical of the domain. In Smith (2006), it is stated that such a resulting ontology should be *intelligible*, that is, understandable by other persons that did not develop it, with a reasonable effort. Often, formal ontologies are not especially intelligible; check the proposed ontologies, for example, at WSMO (2007) or LOA (2007). Therefore, formal ontologies usually contain also portions that contain descriptions of the concepts and relationships in natural language. Based on the informal and formal parts one can also ask, whether on ontology is internally *coheren*t, that is, do the informal and formal part specifies the same domain, the same relationships and axioms.

Another set of requirements presented by Smith (2006a) requires *openness:* "ontology should be open and available to be used by all potential users without any constraint, other than (1) its origin must be acknowledged and (2) it should not to be altered and subsequently redistributed except under a new name… In addition the ontology should be (3) explained in ways which make its content intelligible to human beings, and (4) implemented in ways which make this content accessible to computers." It is rather clear that open ontologies in the above sense would be usable by anybody and those who develop them could not sell them and thus directly take the benefit. Indirectly, it might be possible, though. The last point (4) would namely mean that all ontologies should be flowing (or native) and thus mobile in one sense (see below). In this form, they could possibly be integrated or downloaded into larger computer systems and used by software. A business opportunity might exist here both for the developers of the ontology or for the third parties (see the next section).

MOBILE ONTOLOGIES

What would then be mobile ontologies? While answering this question, we are actually constructing a further classification of ontologies. The term mobile could be added to almost any class above, obtaining such terms as mobile top-level, mobile domain, mobile application, mobile metadata, or

mobile terminological ontology. But what would they mean? We gave a tentative characterization for the term mobile ontology in Veijalainen (2006), and refined it in the preliminary version of this article (Veijalainen, 2007). According to it, a mobile ontology can conceptualize a mobile domain, that is, the possible world the ontology is conceptualizing must be related with mobility. On the other hand, an ontology can itself be mobile, that is, its digital representation can move from one node to another among networked computers, or it can move physically with the terminal. These two aspects are largely orthogonal.

What should be understood by a mobile domain in this context? Let us start from bottom up. In the current business practice and also in the scientific literature, the term mobile is used in many different ways. One speaks, for example, about mobile networks, mobile applications, mobile users, mobile terminals, and so forth. Often, the term mobile refers only indirectly to physical movement, and the reference is primarily to wireless and other technologies that make physical movement possible or that can be used while on the move. In some contexts, the term mobile could be replaced with the term wireless to emphasize that the central issue dealt with is wireless communication technology that facilitates physical movement of the terminal during service delivery. Seen from the service accessibility point of view, mobility in service delivery refers to the possibility to *deliver the services anytime anywhere*. Where the user happens to be and whether the user moves or not or is not of importance for the service delivery.

How should "mobile" or by "mobility" be defined in the context of the mobile ontology? If we think that these terms are somehow directly or indirectly related with the physical movement of the human user or movement of the ontology representation either from one computer to another or within the computer around, we can conceptualize the movement as *context change*. We can ask, what a user or ontology would need in a *new context* and in which cases she/it would need (portions of) her existing context? In general, a context can always be related with a particular physical place, but there can be different relevant contexts attached to a place. The context can change even if the user sits in her chair in the office, once speaking to (VoIP) phone, once having a face-to-face meeting with some people in her office, once wanting to concentrate alone on her work. The context usually changes, though, if the user is roaming to another country, and it could change as a result of the micro-mobility, at least from the network infrastructure perspective.

The user context and its management can be taken care of mainly by the terminal with the help of the user, but it often requires support also from the network infrastructure. From the network infrastructure point of view, "mobility" of a terminal or person can refer to at least five different aspects (cf. (Puttonen, 2006)):

1. A person changes the terminal in use (often as a result of physical movement, or context change)
2. A terminal changes its point of attachment to the network (often as a result of the physical movement with its owner; cf. hand-over)
3. Application is migrated from one network node to another (cf. mobile agents)
4. An on-going session is moved from one terminal to another (cf. 1)
5. Services available for a subscriber at one network location are offered at a new location the subscriber physically moves to (cf. 2; context transfer)

The above view on mobile domain and mobility is mainly a context change issue from the network infrastructure point of view. Although this is an important domain, it does not cover all mobile domains. Keeping this in mind, we can refine the concept of mobile ontologies further. We introduce the shorthand notion *md-ontology* for mobile *domain ontology*, where the domain

is related with mobility (see below). On the other hand, if ontology (representation) can move from one computer network node to another, such ontology is called *flowing or fl-ontology*. Those ontologies that are stored (by manufacturers or operators, etc.) into terminals or other devices and move physically with them, but cannot flow into them from the network, are called *native* or *nt-ontologies*. Notice that we allow the native ontologies to be read, that is, flow out, from the terminal or other wireless device. In addition, all mobile ontologies can be divided into formal and informal ontologies as discussed above. These two aspects are orthogonal and thus, informal-formal and flowing-native divide the set of mobile ontologies into pair-wise disjoint subsets. Are there any ontologies that are not mobile in any sense? Yes, those whose domain is not mobile (say bioinformatics (OBO, 2007)) and which are neither installed into terminals by manufacturers nor can be downloaded into them later for whatever reason.

All the aspects, 1-5 above, seen from the network infrastructure's point of view, can be a domain of a md-ontology. In addition, there can be further md-ontologies that model, for example, physical movements of objects on earth. In that case, the ontology would support, for example, tracking applications where the physically moving objects are not primarily users, but physical objects that are tracked, such as trucks or parcels—or terminals or RFID chips mounted on them. Ontologies describing essential concepts for this application domain (such as tracking, positioning, trace, trajectory, velocity, etc.) can also be regarded as mobile domain ontologies.

Ontologies that help in anywhere/anytime service delivery to mobile terminals are also typical md-ontologies. They do not need to be installed at the terminals, but can reside also at other network components, like servers providing the anywhere/anytime services. This subcategory might contain md-ontologies that support content format transformations and mobile Web service descriptions. Another case is described in Massimo (2007) where the mobile device accesses a tag in its vicinity to which information and/or a Web service is provided from a server. The (ontology-based) service description is dynamically loaded into the device and, for example, information on movies or a movie ticket can be provided to the device as a result of the service invocation.

Fl-ontologies are a special case of 3 if we consider ontologies to be a special kind of software or belonging to a software package moving in the network from one node to another. They have many commonalities with mobile agents, and one might argue that such ontologies are as such actually pieces of software moving around in the network. This view is valid, if the ontology can be directly interpreted or compiled at the receiving node and appropriate actions taken by the computer. An example of this is presented in Khusraj (2005), where a user interface for a Web service is generated from a semantic description at a mobile device. The ontologies might also be carried by mobile agents as part of their state. Another possible scenario is software package that is downloaded from the network and is installed at the terminal. Its functionality might be guided by an ontology component that travels with it or that is downloaded separately.

Any ontology that is stored at a home location or by manufacturer into a device can be considered an nt-ontology. The significance of this concept is that such an ontology could have been designed to a certain environment or context (e.g., by an operator or a manufacturer) and it might not work properly in other environments or contexts the terminal roams to. But it might also be upper-level ontology that works everywhere and does not need to be changed, based on the location. Some nt-ontologies might be invisible to the outside world, but some of them might flow out, that is, they might be readable from outside. This might a possibility for future digital tags that can tell what to store and how to access it.

Point 5 above is challenging from the usage and development point of view of such services that are based on ontologies. Challenge concerning point 3 is heterogeneity and autonomy, but also the applicable business model that makes transfer of (formal) ontologies profitable, or at least possible.

What is the relationship between informal-formal and flowing-native ontologies? Informal ontologies can flow "more easily" than formal ones, because they are only based on natural language and perhaps on some commonly available drawing tool or text editor formats, including XML-based ones, whereas the formal ontologies need rigorous execution (reasoning) environment and in a heterogeneous environment in addition powerful mediators that translate from one formal language to another (Roman, 2005; Euzenat, 2007). The formal ontologies usually also have an XML encoding for transmission purposes, so that at this basic level they can be transported in computer networks or over a wireless short-range link from node to node. In these cases, the enhancement of an ontology with another might happen and the consistency issues are of importance.

DEVELOPING MOBILE ONTOLOGIES

Developing ontologies can be performed by various actors. Known alternatives are research groups, research projects, and larger organizations who have hired specialists to do the work for themselves, or coordinate the work paid by large companies and government organizations. Further, developer communities can engage in the development of ontologies, in a similar fashion as some groups develop free software. ISO is an example of a large international organization that has developed a standard ISO 15926 (ISO, 2003) that could be used as an upper ontology. Barry Smith has checked the quality of the ontology, though, and comes to the conclusion that it is not an ontology at all due to many flaws (Smith, 2006b). Because it can be downloaded from the ISO site as a pdf-file, it would be an fl-ontology, but not a formal one.

Another example of an existing ontology is, for example, "high-level mobile ontology" developed by the SPICE project (Zhanova, 2006). It is developed for the mutual understanding of the project. It is also exploiting many other existing ontologies and related specifications, such OMA's UAp (OMA, 2006) and FOAF vocabulary (FOAF, 2007). The ontology specifies the concept of service and subtypes thereof, device and subtypes thereof, mobile access network and subtypes thereof, person (a physical thing), user groups, configuration, location, context, contents and contents types, and so forth. It is an md-ontology in our classification that can be used to describe mobile infrastructures. It is a formal upper-level ontology that could be used by various terminals and by the service delivery sites. It is developed by a research consortium.

Is developing mobile ontologies any different from other ontologies? If it is question of md-ontologies, the same challenges are encountered as while developing ontologies for other domains. The only difference might be that some md-ontologies are inherently related with changes in the domain, and change is a general challenge for ontologies. As concerns methodologies, Sanchez (2007) reviews shortly several proposed ontology development methodologies. It seems that any of them could be used, but it is too early to say whether various md-ontologies would have specialties that would require modifications to the methodologies, or perhaps new developing methodologies to be developed. In any case, Fl-ontologies, especially formal ones, present additional challenges, both organizational and technical. When an ontology crosses an organizational border, the receiving organization should understand and accept the world view its developers had and also *trust* the sending side. Neither is by no means generally given between autonomous organizations. Detect-

ing the discrepancies between different world views is easy in the case of intelligible ontologies (cf. above), because they explicate the views of the developers for other humans. Perceiving the world view is more difficult in the case of usual software, because any piece of software crossing an organizational border carries with it a world view of the developers, but it is often represented only in the code and detected only when using the software. The same might be true for large formal ontologies that cannot be understood by humans without a big effort and high technical skills. And of course, if the ontology comes as data part of mobile agent or software package, the problem is essentially the same as for any moving software. If we think that pure ontologies or mobile agents would be downloaded by ordinary mobile users, without special technical skills, the problem is aggravated. They could not check the quality of the ontology, mobile agent or piece of software, but rather they could only trust the provider.

The technical challenges are mostly software engineering challenges in the mobile environment for the flowing ontologies. They are similar to downloading software from the network. There must be a platform support for correctly installing and interpreting the ontology. What the support exactly looks like depends at least partially on the formalism used to represent the ontology and the nature of the ontology (top-level vs. application, lightweight vs. heavyweight, etc.). A requirement peculiar to ontologies is that if the ontology is to be integrated with an existing one, or existing ones, then the resulting ontology should be at least *consistent*. This means that no contradictions can be inferred from valid input definitions (Gomez-Perez, 2004). There should be tools to check this at least at the provider site, but perhaps also at the terminal. Like in the case of other software entities, there should be versioning mechanisms and valid configuration schemes for mobile ontologies. Using them, compatibility relationships could be expressed between different ontologies and their versions.

According to Gomez-Perez (2004), ontologies should also be evaluated for *completeness* (everything meant to be there is explicit or can be inferred, and each definition is complete), *conciseness* (no redundancy), *expandability* (how easy is it to add new definitions and knowledge without altering the set of already guaranteed properties), and *sensitiveness* (how small changes alter the well-defined properties already guaranteed). All these are relevant attributes for md- and fl-ontologies and should be taken into consideration when developing and updating them.

There are many emerging formal ontologies in various domains, such as biomedicine, law, software engineering, and so forth, (LOA, 2007; OBO, 2007; Ontomed, 2006). Some of these could be perhaps used as part of mobile ontologies.

The problem of generating metadata for private photographs taken by camera phones has been investigated in Sarvas (2006). The author concluded that this area is a special domain and requires metadata—or ontology—that is different from the commercial multimedia material. He also discovered in user tests that people are not willing to insert the metadata at spot (who are on the picture, what was the social situation the picture was taken in, etc.), because it tends to disturb the photographing act.

In jpg-metadata there is already some information, such as camera type, shutter, date, and so forth, but coordinates where the picture was taken should be added. This would require GPS or other satellite receiver to be integrated to or accessible for the terminal. Currently, the trend is that high-end terminals begin to include a GPS receiver. To the coordinates, one could attach the name of the place automatically when precise enough geographic information was made available. Automatic face recognition might be possible for acquainted people in the picture, if once done and the resulting meta-information stored at the terminal. The hard parts are the semantically high level concepts, like "situation" or "social context" where the picture was taken. This should

be described by the user textually or orally and transformed into text by speech recognition. Speech recognition could be also used to identify people in the picture if they happened to speak when the picture was taken. In Khusraj (2005), the concept of semantic cache was proposed that keeps track of the places that the user has visited lately. These can be directly included into the metadata and used by the inference engine.

A general solution to ontology development is to reuse existing pieces of ontology. In Salminen (2005) the authors present an architecture that allows a mobile terminal user to develop, organize, and share the digital contents and the associated metadata. The device manufacturer would provide a software that is able to store and interpret the metadata, and evidently, a native (meta)ontology. The user can then enhance this with flowing components or develop himself or herself new ontologies.

Although the ontologies conceptualizing low-level mobile protocol layers or device characteristics are important for the end-to-end service delivery, still modeling the service, context, and contents domains is the key for the proliferation of nice services. Could the vast end-user population help here? What are the minimal skills and minimal knowledge required? If we are speaking of informal taxonomies, such as folksonomies (Koivunen, 2006; Wikipedia, 2007), the requirements are not so high. They can be generated by ordinary users and are used to tag all kinds of digital objects. A collection of such tags is usually not a genuine conceptualization of a domain, that is, an ontology in a strict sense. This is because tagging can be contradictory and usually lacks hierarchical relationships, such as "is a", "part-of", and so forth. Collectively, generated tags can still be used to organize private photographs either at Web sites like Flickr (Flick, 2007) or on private devices. What kind of tools would be needed at the mobile terminals to support generation of remote tags? It seems that inserting photos to the above Flickr site directly from camera phones and tagging them is possible already now.

Folksnomies are developed by a user community. The approach bears resemblance to the development of open source software. Whereas open software cannot be developed by unskilled people, folksnomies can be developed by people who cannot write code. How big is the distance from folksonomies to ontologies and further to proper formal ontologies that could be used by programs as input? Folksonomies can be used at least as some kind of starting point to develop ontologies for certain domains bottom up, but it is evident that developing even an informal ontology and especially a formal ontology requires expert work. For instance, the Flickr tags are just a collection of individual words and the frequency they occur in metadata of the photographs reflects the frequency of the situations the pictures were taken in. A rather common tag is for instance "Canon" referring to the camera manufacturer, and "Wedding" referring to a situation the picture was taken in. One might argue that not even every person who is able to write, for example, Java programs would be able to compile a formal ontology. The latter requires different skills than programming.

ONTOLOGY SHOPPING

The organization or a group of individuals that develops an ontology, or fraction of it, primarily owns it. Some capable experts might be interested in developing ontologies in the same spirit as they are ready to write articles to Wikipedia or open source software. Still, if formal ontologies would become directly usable by mobile terminals, then it would make sense to develop them as commercial activity and pay for the work. This is currently hardly envisaged in the research literature. Probably, because developing ontologies, especially formal ones, requires high skills and is currently still research activity.

What legal status would ontologies have? They are scientific or technical works and fall under the Berne Convention that also regulates software ownership and rights (WIPO, 2007). According to Berne convention, a copyright holder is inherently the creator of the work. The rights cover economic and moral rights and the creator has thus the right to demand economic compensation when the work is performed or distributed to audience. The creator can also sell his or her rights to another person or company totally or in part. Those countries that have joined the Berne Convention treat the rights uniformly, although the convention allows exceptions. The local legislation may, for instance, automatically move the copyrights of an employee to the employer in some cases. For instance, the rights to (production) software, if produced by employees as part of their normal duties at a Finnish university, are moved by the law to the university without further action and without additional compensation. The copyright protection holds of course to original enough works, not works that would infringe the rights of other copyright owners. Another question is whether ontologies would be software or other scientific of technical works. The informal ontologies could hardly be considered as software, but the formal ones could be interpreted as software if they can be executed in a computer like any other program. Some formal ontologies have been copyrighted as software, such as Ontomed ontology (Ontomed, 2006). This can be inferred from the fact that the ontology includes a GPL-like license at the beginning and categorizes the contents as "software". In any case, all kind of ontologies are material that is protected by copyright if developed in a country that has joined Berne Convention (cf. WIPO, 2007).

Whereas copyright grants legal and moral rights to the owner, it does not protect the ideas presented in the work. That is, if somebody buys a book, where a construction of a machine a process, a business model, and so forth, is presented, the copyright owner neither can prohibit the buyer to construct the machine or implement the process or business model, nor can demand any economic compensation for the possible yield. *Patents* protect the innovations and give the exclusive rights to the patent owner to reap the economic benefits from the innovation for a limited period of time (e.g., 20 years) in the jurisdiction the patent has been granted in (e.g., USA, EU, Japan). A patent is a public document. After the patent has expired, the innovation can be copied and developed further by others and economic benefits reaped freely.

Could ontologies be *patented?* In this case, there is evidently a difference between informal and formal ontologies. The former cannot be patented, only copyrighted and usage perhaps licensed in some cases. If formal ontologies are considered to be software, then in some countries (such as USA) they might be patentable even as such, or at least as part of some larger software system or other technical context. The patentability seems to be largely open, though, because the exact nature of them as creative work is not established yet..

Digital music, videos and software can be bought and used in the mobile terminals. Full-fledged formal ontologies are also machine-process able and could become at least in theory, separate objects of trade. Their behavior and usage is quite close to a of piece software. They need some kind of interpreter in order to be usable, in a similar fashion as Java byte code. But can formal ontologies be considered as software in the legal sense? Or are they treated rather like other digital contents? This is largely open, but in both cases ontologies could become intangible goods—or information society services as the EU jargon proposes.

There are not many examples of ontology shopping. ISO has developed an international standard ISO 15926 (IS, 2003) that is also considered by some groups to be an upper-level ontology (cf. above). As a standard, it has a price tag of CHF

258 (EUR 156). It is protected by copyright that denies its further distribution. Thus, it evidently could not be used in computerized ontology applications without licensing—if somebody managed to make it executable in computers, that is, to formalize it properly.

In order for ontology market to be economically viable, several conditions must be met. They are rather similar as those for the other mobile contents or mobile application software. Some device manufacturers could, for example, implement some functionality of the devices using ontology-based technology and thus increase the attractiveness of the devices. The devices could have an ontology (reasoner) engine installed by the manufacturer, in a similar manner as some now have a Java virtual machine or a media player. This engine could then be used to process ontologies in various contexts. The engine could also be licensed and installed later, and necessary ontologies could then be downloaded and processed using it.

The manufacturer could install some nt-ontologies into the device that would not need to be md-ontologies. A typical example would be an ontology for digital pictures that would make their indexing and retrieval possible for the user or other digital contents (Salminen, 2005).

Mobile md-ontologies might also come with the device, but the user could order various md-ontologies as he or she needs, over the network or on a memory stick/memory card. So, those md-ontologies that are bought during the usage should be designed as fl-ontologies. Technical problems are similar to those of downloading software from the network. Perhaps it makes sense to develop "source code form" and some kind of "byte code" for the mobile ontologies as well, in analogy to Java, or one could also develop the interpreter in Java. Source code format should have XML encoding, like OWL-S and WSML (WSMO, 2007) have.

While selling or otherwise distributing private contents, privacy becomes an issue. The actual contents, but also metadata can contain information that a person might not want to share with a larger audience, although this would be shareable within, for example, family. This has been addressed, for example,. in Sarvas (2006) and Salminen (2005).

CONCLUSION

We have discussed in this article mobile ontologies; what should be understood by them, how they could be used, who would develop them, and why. The concept is still under development and the term "mobile" can refer to two rather different aspects. On one hand, the domain of the ontology can be related to mobility (md-ontology), on the other hand the interpersonal representation of the ontology can move (flow) from node another in the network (fl-ontology) or move with the terminal (native or n-ontology). In the latter cases, the domain of the ontology can be anything. The definition of mobile ontology should address these aspects. We suggest the following definition: *If the domain of an ontology is related with mobility or it can be mounted or downloaded to and used at a mobile terminal, or both, th*en *it is a mobile ontology.* Informal ontologies can flow more easily, as far as technical constraints are considered, whereas formal ontologies that have usually a portion consisting of first-order logic expressions have more difficulties in crossing heterogeneous and autonomous system borders. The latter kind of ontologies facilitate formal reasoning, and further automatic processing, whereas informal ontologies can only be applied by humans. Organizational autonomy is also an important issue in the scenario, where ontologies are dynamically downloaded to mobile terminals or incorporated into the ICT infrastructure of an organization. Because ontologies are difficult to understand and inspect, an organization, not to speak about a usual user, must trust the ontology provider.

Mobile formal ontologies might become objects of trade, if they turn out to be useful enough to justify the investment in developing a software infrastructure and terminal and ontology base for them. Business models for these might look similar to those of the mobile software and both native and flowing ontologies could be used. Free ontologies are also possible, in a similar manner as free software, and in analogy, open mobile ontologies are also possible.

Quality of ontologies is an emerging theme and the identified attributes like consistency, completeness, conciseness, expandability, and sensitiveness are important to evaluate for mobile ontologies. It is evident that if ontologies are downloaded to a terminal, the providers should indicate with which earlier versions of which other ontologies a certain ontology is compatible. Compatibility statements must evidently consider all quality attributes above.

Intellectual property rights in the context of ontologies seem similar to scientific pieces of work or software and are thus primarily governed by copyright laws. Only formal ontologies could be viewed as software and as patentable is some countries, but the entire area of intellectual property rights in the context of ontologies is largely open. The solution might have some ramifications to the possible mobile ontology market, though, as we have seen in software business in the context of free/open software.

ACKNOWLEDGMENT

The author wishes to thank the reviewers of this article and those of the preliminary version (Veijalainen, 2007) that appeared in Proceedings of the MoSO2007 workshop, available at IEEE Digital Library.

REFERENCES

Brewster, C., & O'Hara, K. (2007). Knowledge representation with ontologies: Present challenges—Future possibilities. *International Journal of Human-Computer Studies 65,* 563-568.

Chandrasekaran, B., Josephson, J., & Benjamins, V. (1999). What are ontologies, and why do we need them?. *IEEE Intelligent Systems, 14*,(Jan.-Feb.), 20-26,

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching.* Berlin/Heidelberg: Springer-Verlag.

Fensel, D. (2004). *Ontologies, a silver-bullet for knowledge management and electronic commerce,* 2nd ed. Berlin: Springer Verlag.

Flickr Virtual Community. (2007). www.flickr.com.

Friend-of-Friend Virtual Community. (2007). www.foaf-project.org

Folksonomies. (2007). www.wikipedia.org/folksonomy

General Formal Ontology. (2007). Retrieved May 30, 2007 from http://www.onto-med.de/ontologies/gfo.owl

Gomez-Perez, A. (2004). *Ontology evaluation*. Ch. 13 in [StSt2004].

Gomez-Perez, A., Fernadez-Lopez, M., & Corcho, O. (2003). *Ontological engineering.* London: Springer Verlag.

Gruber, T. (1993). A translation approach to portable ontology specification. *Knowledge Acquisition, 5*(2), 199-220. Retrieved October 14, 2007 from http://tomgruber.org/writing/ontolingua-kaj-1993.pdf

Guarino, N. (1998). Formal ontology and information systems. In: N. Guarino (Ed.), *Formal*

ontology in information systems. Proceedings of FOIS'98 (pp. 3-15). Trento, Italy, Amsterdam: IOS Press. http://www.loa-cnr.it/Papers/FOIS98.pdf

Herre, H., Heller, B., Burek, P., Hoehndorf, R., Loebe, F., & Michalek, H. (2007). *General formal ontology (GFO); A foundational ontology integrating objects and processes. Part I, v.1.0.1.* Retrieved May 30, 2007 from http://www.onto-med.de/en/theories/gfo/part1-drafts/gfo-part1-v1-0-1.pdf

ISO. (2003). *Industrial automation systems and integration — Integration of life-cycledata for process plants including oil and gas production facilities —Part 2: Data Model.* International standard, 1ST ed. www.iso.org

Khusraj, D., & Lassila, O. (2005). Ontological approach to generating personalized user interfaces for Web services. In: Y. Gil, et al. (Eds.), *ISWC 2005, LNCS 3729* (pp. 916-927). Berlin/Heidelberg: Springer.

Koivunen, M-R. (2006). Annotea and semantic-Web-supported collaboration. Retrieved October 31, 2007 from http://kmi.open.ac.uk/events/usersweb/papers/01_koivunen_final.pdf

Laboratory for Applied Ontologies. (2007). *LOA site*. Retrieved October 31, 2007 from http://wiki.loa-cnr.it/index.php/Main_Page; http://www.loa-cnr.it/ontologies/EVAL/oQual.owl

The OBO foundry. (2007). Retrieved October 31, 2007 from http://obofoundry.org/

Ontomed. (2006). *Ontomed ontology.* Retrieved October 31, 2007 from http://www.onto-med.de/ontologies/gfo.owl

Open Mobile Alliance. (2006). *User Agent profile ,Version 2.* Retrieved February 28, 2007 from http://www.openmobilealliance.org/release_program/docs/UAProf/V2_0-20060206-A/OMA-TS-UAProf-V2_0-20060206-A.pdf

Peterson, E. (2006). Beneath the metadata; Some philosophical problems with folksonomy. *D-Lib Magazine 12*(11). Retrieved May 31, 2007 from http://www.dlib.org/dlib/november06/peterson/11peterson.html

Pohjola, P. (2007). *Technical artefacts, an ontological investgation of arfacts.* Jyväskylä Studies in Education, Psychology and Social Research, Report No. 300. Retrieved October 14, 2007 from http://dissertations.jyu.fi/studeduc/9789513927561.pdf

Paoluccci, M., Broll, G., Hamard, J., Rukzio, E., Wagner, M., & Schmidt, A. (2008). *Bringing semantic services to real-world objects.* In this issue.

Puttonen, J. (2006). *Mobility management in wireless networks.* Doctoral Thesis. Jyväskylä Studies in Computing # 69, University of Jyvaskylä, Jyväskylä, Finland.

Roman, D., Keller, U., Lausen, H., deBruijn, J., Lara, R., Stollberg, M., et al. (2005). Web service modeling ontology. *Applied Ontology, 1*(1), 77-106.

Sanchez, D., Cavero, J., & Martinez, E. (2007). The road towards ontologies. Ch. 1. in *Ontologies: A handbook of principles, concepts and applications in information systems.* New York, NY: Springer Verlag.

Salminen, I., Lehikoinen, J., Huuskonen, P. (2005). Developing and extensible metadata ontology. In: W. Tsai & M. Hamza (Eds.), *Procedings of the 9th IASTED Intl. Conference on Software Engingeering and Applications (SEA)* (pp. 266-272). .Phoenix, AZ: ACTA Press.

Sarvas, R. (2006). *Designing user-centric metadata for digital snapshot photography.* Doctoral Dissertation, Helsinki University of Technology, Department of Computer Science and Engineering/Soberit, and Helsinki Institute for Informa-

tion Technology (HIIT)/HUT. http://lib.tkk.fi/Diss/2006/isbn9512284448/isbn9512284448.pdf

Zhanova, A. (Ed.). (2006). *Deliverable 3.1, Ontology definition for the DCS and DCS resource description, User rules. EU-IST SPICE project.* Retrieved February 28, 2007 from http://www.ist-spice.org/documents/D3.1_061017_v1_final_bis.pdf

Smith, B. (2004). Beyond concepts, or: Ontology as reality representation systems. In: A. Varzi & L.Vieu (Eds.), *Proceedings of the 3rd International Conference on Formal Ontology in Information, Systems , Turin 4-6 (FOIS 2004)* (pp. 73-84). Amsterdam: IOS Press. Retrieved May 29, 2007 from http://ontology.buffalo.edu/bfo/BeyondConcepts.pdf

Smith, B. (2006a). Against fantology. In: M. Reicher & J. Marek (Eds.), *Experience and analysis (pp.* 153-170).

Smith, B. (2006b). Against idiosyncrasy in ontology development. In: B. Bennett & C. Fellbaum (Eds.), *Formal ontology in information systems: Proceedings of the 4th International Conference (FOIS 2006). Frontiers in Artificial Intelligence and Application, 150.* New York, NY: IOS Press. Retrieved October 14, 2007 from http://ontology.buffalo.edu/bfo/west.pdf

Staab, S., & Studer, R. (Eds.). (2004). *Handbook on ontologies.* Berlin-Heidelberg, Germany, New York, NY: Springer Verlag.

Umsoy, M. (2007). *3GSM Congress 2007 Notes*. Retrieved February 27, 2007 from http://cartagena-capital.com/pdfs/3gsm_congress_2007_notes.pdf

van Heijst, G., Schreiber, A., & Wielinga, B. (1997). Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies, 46*(2-3), 183-292.

Veijalainen, J., Nikitin, S., & Törmälä, V. (2006). Ontology-based semantic Web service platform in mobile environments. *Proceedings of Mobile Ontologies Workshop.* Nara, Japan. http://csdl2.computer.org/persagen/DLPublication.jsp?pubtype=p&acronym=MDM

Veijalainen, J. (2007). Developing mobile ontologies; who, why, where, and how?. *Mobile Services-oriented Architectures and Ontologies Workshop (MoSO 2007).* Mannheim, Germany.

World Intellectual Property Organisation. (2007). *Berne convention*. Retrieved April 3, 2007 fromhttp://www.wipo.int/treaties/en/ip/berne/

Web Services Modeling Ontology. (2007). http://www.wsmo.org/

Youtube Virtual Community. (2007). www.youtube.com

This work was previously published in the International Journal on Semantic Web & Information Systems, edited by A. Sheth, Volume 4, Issue 1, pp. 20-34, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.25
Web Mining System for Mobile-Phone Marketing

Miao-Ling Wang
Minghsin University of Science & Technology, Taiwan, ROC

Hsiao-Fan Wang
National Tsing Hua University, Taiwan, ROC[1]

ABSTRACT

With the ever-increasing and ever-changing flow of information available on the Web, information analysis has never been more important. Web text mining, which includes text categorization, text clustering, association analysis and prediction of trends, can assist us in discovering useful information in an effective and efficient manner. In this chapter, we have proposed a Web mining system that incorporates both online efficiency and off-line effectiveness to provide the "right" information based on users' preferences. A Bi-Objective Fuzzy c-Means algorithm and information retrieval technique, for text categorization, clustering and integration, was employed for analysis. The proposed system is illustrated via a case involving the Web site marketing of mobile phones. A variety of Web sites exist on the Internet and a common type involves the trading of goods. In this type of Web site, the question to ask is: If we want to establish a Web site that provides information about products, how can we respond quickly and accurately to queries? This is equivalent to asking: How can we design a flexible search engine according to users' preferences? In this study, we have applied data mining techniques to cope with such problems, by proposing, as an example, a Web site providing information on mobile phones in Taiwan. In order to efficiently provide useful information, two tasks were considered during the Web design phase. One related to off-line analysis: this was done by first carrying out a survey of frequent Web users, students between 15 and 40 years of age, regarding their preferences, so that Web customers' behavior could be characterized. Then the survey data, as well as the products offered, were classified into different demand and preference groups. The other task was related to online query: this was done through the application of an information retrieval technique that responded

to users' queries. Based on the ideas above the remainder of the chapter is organized as follows: first, we present a literature review, introduce some concepts and review existing methods relevant to our study, then, the proposed Web mining system is presented, a case study of a mobile-phone marketing Web site is illustrated and finally, a summary and conclusions are offered.

LITERATURE REVIEW

Over 150 million people, worldwide, have become Internet users since 1994. The rapid development of information technology and the Internet has changed the traditional business environment. The Internet has enabled the development of Electronic Commerce (e-commerce), which can be defined as selling, buying, conducting logistics, or other organization-management activities, via the Web (Schneider, 2004). Companies are finding that using the Web makes it easier for their business to communicate effectively with customers. For example, Amazom.com, an online bookstore that started up in 1998, reached an annual sales volume of over $1 billion in 2003 (Schneider, 2004). Much research has focused on the impact and mechanisms of e-commerce (Angelides, 1997; Hanson, 2000; Janal, 1995; Mohammed, Fisher, Jaworski, & Paddison, 2004; Rayport & Jaworski, 2002; Schneider, 2004). Although many people challenge the future of e-commerce, Web site managers must take advantage of Internet specialties which potentially enable their companies to make higher profits and their customers to make better decisions. Given that the amount of information available on the Web is large and rapidly increasing, determining an effective way to help users find useful information has become critical. Existing document retrieval systems are mostly based on the Boolean Logic model. Such systems' applications can be rather limited because they cannot handle ambiguous requests. Chen and Wang (1995) proposed a knowledge-based fuzzy information retrieval method, using the concept of fuzzy sets to represent the categories or features of documents. Fuzzy Set Theory was introduced by Zadeh (1965), and is different from traditional Set Theory, as it uses the concept of membership functions to deal with questions that cannot be solved by two-valued logic. Fuzzy Set Theory concepts have been applied to solve special dynamic processes, especially those observations concerned with linguistic values.

Because the Fuzzy concept has been shown to be applicable when coping with linguistic and vague queries, Chen and Wang's method is discussed below. Their method is based on a concept matrix for knowledge representation and is defined by a symmetric relation matrix as follows:

$$A = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_n \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots & A_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \end{array}$$

$$0 \le a_{ij} \le 1,\ 1 \le i \le n,\ 1 \le j \le n \qquad \textbf{(1)}$$

where n is the number of concepts, and a_{ij} represents the relevant values between concepts A_i and A_j with $a_{ii} = 1, \forall\ i$. It can be seen that this concept matrix can reveal the relationship between properties used to describe objects, which has benefits for product identification, query solving, and online sales development. For effective analysis, these properties, determined as the attributes of an object, should be independent of each other; however this may not always be so. Therefore a transitive closure matrix A^* must be obtained from the following definition.

Definition 1: Let A be a concept matrix as shown in Equation (1), define:

$$A^2 = A \otimes A$$
$$= \begin{bmatrix} \vee_{i=1,\cdots,n}(a_{1i} \wedge a_{i1}) & \vee_{i=1,\cdots,n}(a_{1i} \wedge a_{i2}) & \cdots & \vee_{i=1,\cdots,n}(a_{1i} \wedge a_{in}) \\ \vee_{i=1,\cdots,n}(a_{2i} \wedge a_{i1}) & \vee_{i=1,\cdots,n}(a_{2i} \wedge a_{i2}) & \cdots & \vee_{i=1,\cdots,n}(a_{2i} \wedge a_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \vee_{i=1,\cdots,n}(a_{ni} \wedge a_{i1}) & \vee_{i=1,\cdots,n}(a_{ni} \wedge a_{i2}) & \cdots & \vee_{i=1,\cdots,n}(a_{ni} \wedge a_{in}) \end{bmatrix} \quad \textbf{(2)}$$

where $\otimes$ is the max-min composite operation with "$\vee$" being the maximum operation and "$\wedge$" being the minimum operation. If there exists an integer $p \le n - 1$ such that $A^p = A^{p+1} = A^{p+2} = ...$, $A^* = A^p$ is called the Transitive Closure of the concept matrix A.

Matrix A^* is an equivalent matrix which satisfies reflexive, symmetric and transitive properties.

To identify each object by its properties, a document descriptor matrix D is constructed in the following form:

$$D = \begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_m \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots & A_n \end{array} \\ \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} \end{array}$$
$$0 \le d_{ij} \le 1 \quad \textbf{(3)}$$

where d_{ij} represents the degree of relevance of document D_i with respect to concept A_j and m is the number of documents in general terms. By applying the max-min composite operation $\otimes$ to D and A^*, we have matrix $B = D \otimes A^* = [b_{ij}]_{m \times n}$ where b_{ij} represents the relevance of each document D_i with respect to a particular concept A_j.

The implication of this approach for Web mining is: when we classify objects by their properties, if we can also cluster people according to their properties and preferences, then when a query is made, matching a user's properties to retrieval of the corresponding concept matrices of each cluster will speed up online response.

Clustering is fundamental to data mining. Clustering algorithms are used extensively, not only to organize and categorize data, but also for data compression and model construction. There are two major types of clustering algorithms: hierarchical and partitioning. A hierarchical algorithm produces a nested series of patterns with similarity levels at which groupings change. A partitioning algorithm produces only one partition by optimizing an objective function, for example, squared-error criterion (Chen, 2001). Using clustering methods, a data set can be partitioned into several groups, such that the degree of similarity within a group is high, and similarity between the groups is low. There are various kinds of clustering methods (Chen, 2001; Jang, Sun, & Mizutani, 1997; Wang, Wang, & Wu, 1994). In this study, we applied the forward off-line method in order to group people according to their properties and preferences.

The *c*-Means algorithm (Tamura, Higuchi, & Tanaka, 1971) also called Hard *c*-Means (HCM), is a commonly used objective-clustering method, which finds the center of each cluster and minimizes the total spread around these cluster centers. By defining the distance from each datum to the center (a measure of Euclidean distance), the model ensures that each datum is assigned to exactly one cluster. However, in this case in contrast to the HCM, there is vague data and elements may belong to several clusters, with different degrees of belonging. For such situations, Bezdek (1973) developed an algorithm called the Fuzzy *c*-Means (FCM) algorithm for fuzzy partitioning, such that one datum can belong to several groups with degrees of belonging, specified by membership values between 0 and 1. Obviously, the FCM is more flexible than the HCM, when determining data related to degrees of belonging.

Because of the vague boundaries of fuzzy clusters, Wang et al. (1994) showed that it is not

sufficient to classify a fuzzy system simply by minimizing the within-group variance. Instead, the maximal between-group variance also had to be taken into account. This led to a Bi-objective Fuzzy c-Means Method (BOFCM) as shown below, in which the performance of clustering can be seen to be improved:

$$\text{(BOFCM):}\quad \begin{aligned} &Min \quad Z(U;V)=\sum_{i=1}^{c}\sum_{k=1}^{K}(\mu_{ik})^2\|x_k-v_i\|^2 \\ &Max \quad L(U;V)=\sum_{i=1}^{c}\sum_{j<i}\|v_i-v_j\|^2 \\ &\text{subject to} \quad \sum_{i=1}^{c}\mu_{ik}=1, \quad \forall k=1,...,K \\ &\qquad \mu_{ik}\geq 0, \quad \forall i=1,..,c, k=1,...,K \end{aligned} \qquad \textbf{(4)}$$

where c is number of the clusters, $\|\cdot\|$ is an inner product norm, x_k, $k = 1, ..., K$, denote K elements, v_i, $i = 1, ..., c$ is the center of Cluster i and μ_{ik}, $i = 1, ..., c$; $k = 1, ..., K$ are the membership values of x_k belonging to Cluster i.

THE WEB MINING SYSTEM

When looking at Web site development, it needs to be appreciated that Web users can be quite capricious, given the multitude of Web sites available on the Internet. The question arises: how should a Web site be set up so that it provides the right product information to the right customers? And specifically, how can the query response time be speeded up?

Clustering types of users with their preferences is one solution. From the above discussion, it can be seen that we may use the BOFCM algorithm for this purpose. After introducing the weights α, β, $\alpha+\beta = 1$ for the objectives, Model (4) was transformed into a single objective nonlinear problem. The following lemmas provide the basis for solving it:

Lemma 1 (Wang et al., 1994): The solution set of Model (4) can be found by:

$$v_i=\frac{\beta\sum_{k=1}^{K}\mu_{ik}^2x_k}{\beta\sum_{k=1}^{K}\mu_{ik}^2-\alpha c}-\frac{\alpha\sum_{s=1}^{c}\dfrac{\beta\sum_{k=1}^{K}\mu_{sk}^2x_k}{\beta\sum_{k=1}^{K}\mu_{sk}^2-\alpha c}}{\left(\beta\sum_{k=1}^{K}\mu_{ik}^2-\alpha c\right)\left(1+\alpha\sum_{s=1}^{c}\dfrac{1}{\beta\sum_{k=1}^{n}\mu_{sk}^2-\alpha c}\right)} \qquad \textbf{(5)}$$

$$\mu_{ik}=\frac{\left(\dfrac{1}{\|x_k-v_i\|^2}\right)^{1/f-1}}{\sum_{s=1}^{c}\left(\dfrac{1}{\|x_k-v_s\|^2}\right)^{1/f-1}}. \qquad \textbf{(6)}$$

Then, the solution procedures can be summarized as:

Step 1. Fix c. Give the initial membership value of each datum to each cluster, that is, the membership matrix $U^{(0)} = [\mu_{ik}^{(0)}]$ is constructed. Assign an allowed perturbed value τ and set $\delta = \tau$, f=2, and l=0.

Step 2. Calculate α and β by using $U^{(l)}$,

$$\alpha=\min_i\left\{\frac{\sum_{k=1}^{K}(\mu_{ik}^{(0)})^2}{\sum_{k=1}^{K}(\mu_{ik}^{(0)})^2+c-1}\right\}-\delta$$

and

$$\beta=\max_i\left\{\frac{c-1}{\sum_{k=1}^{K}(\mu_{ik}^{(0)})^2+c-1}\right\}+\delta.$$

Step 3. Calculate $\{v_i^{(l)}\}$with Equation (5) and $U^{(l)}$.

Step 4. Calculate the new membership matrix $U^{(l+1)}$ by using $\{v_i^{(l)}\}$ and Equation (6), if $x_k \neq v_i^{(l)}$; else set

$$\mu_{jk}^{(l+1)} = \begin{cases} 1, \text{for} & j = i \\ 0, \text{for} & j \neq i \end{cases}.$$

Step 5. Let $\delta = \tau + \delta$ and go to Step 2 if $\{m_k^{(l+1)}\}$ does not satisfy

$$\alpha < \min_i \left\{ \frac{\sum_{k=1}^{K} (\mu_{ik}^{(l+1)})^2}{\sum_{k=1}^{K} (\mu_{ik}^{(l+1)})^2 + c - 1} \right\}$$

and

$$\beta > \max_i \left\{ \frac{c-1}{\sum_{k=1}^{K} (\mu_{ik}^{(l+1)})^2 + c - 1} \right\}.$$

Step 6. Calculate $\Delta = \left\| \tilde{U}^{(l+1)} - \tilde{U}^{(l)} \right\|$. If $\Delta > \varepsilon$ set l=l+1 and go to Step 3. If $\Delta \leq \varepsilon$ stop.

From the above analysis, we can obtain the clustered data within each center.

To speed up the process, the documents can also be grouped according to their degrees of similarity, as defined by Jaccard's coefficient as follows:

$$r_{ij} = \frac{\sum_{s=1}^{m} \min\left[b_{is}, b_{js}\right]}{\sum_{s=1}^{m} \max\left[b_{is}, b_{js}\right]}, \quad 0 \leq b_{is}, b_{js} \leq 1, \tag{7}$$

where r_{ij} is the similarity between document D_i and document D_j. b_{iS}, b_{jS} from matrix B are the relevant values with respect to documents D_i, D_k and documents D_j, D_k. So we can obtain the document fuzzy relationship matrix R:

$$R = \begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_m \end{array} \begin{array}{c} \begin{array}{cccc} D_1 & D_2 & \cdots & D_m \end{array} \\ \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} \end{array} \tag{8}$$

Again a transitive closure R* of R must be obtained. Then by defining an acceptable level of λ by the mean of the upper triangular matrix R*, i.e.,

$$\lambda = \frac{\sum_{i=1}^{m-1} \sum_{j>i}^{m} r_{ij}}{\binom{m(m-1)}{2}},$$

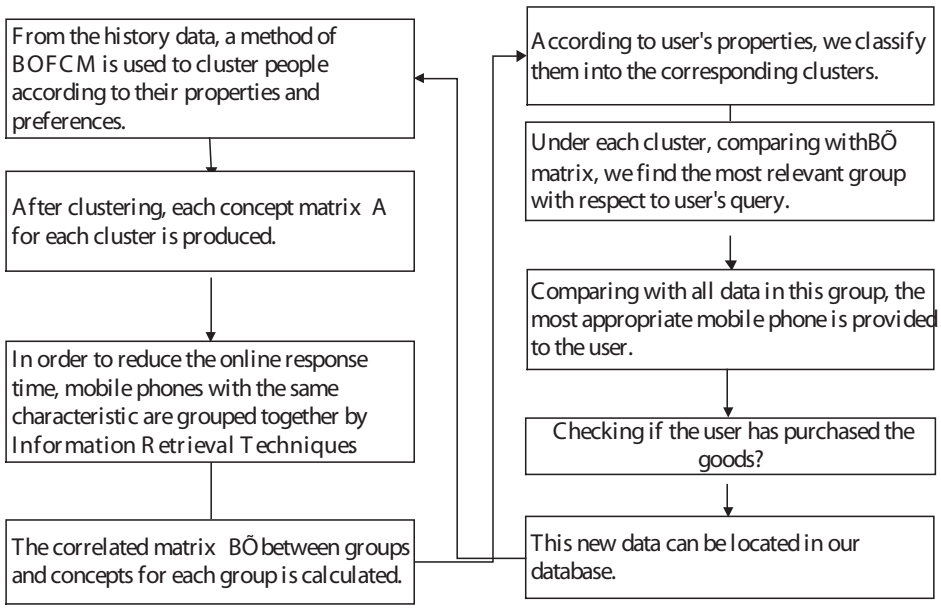
we have an λ-threshold partition of documents into clusters. Based on the document descriptor of each document, we can obtain a cluster-concept matrix B':

$$B' = \begin{array}{c} \\ Group\ 1 \\ Group\ 2 \\ \vdots \\ Group\ u \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots & A_n \end{array} \\ \begin{bmatrix} b'_{11} & b'_{12} & \cdots & b'_{1n} \\ b'_{21} & b'_{22} & \cdots & b'_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b'_{u1} & b'_{u2} & \cdots & b'_{un} \end{bmatrix} \end{array},$$

where u is the number of clusters of documents. **(9)**

With the results of above off-line analysis, a user can take advantage of the clustered documents to improve response time when making an online query. By comparing the matrix B' with the user's query vector, the most relevant cluster(s) are selected. Then, by searching the documents within the selected cluster(s), the documents may

Figure 1. Framework of the Web mining system



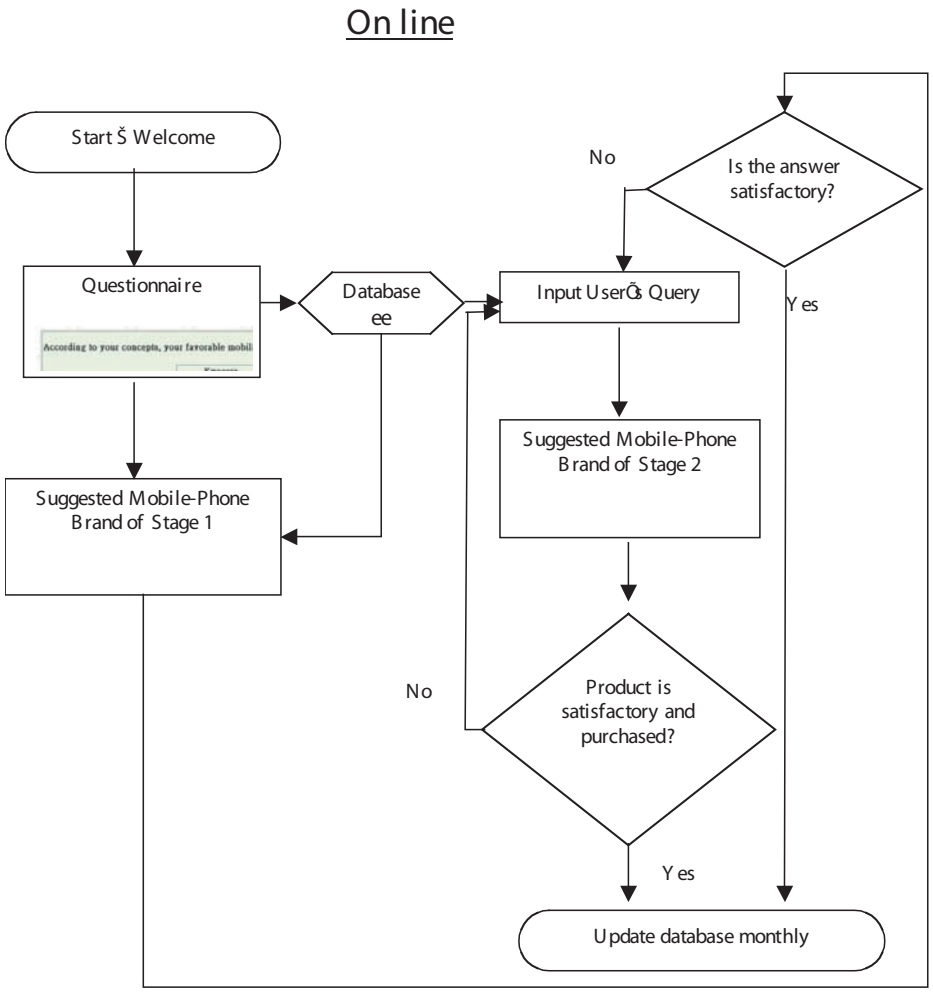
be retrieved more efficiently. The framework of the proposed Web mining system (Lin, 2002), with both online and off-line operations, is shown in Figure 1.

A CASE STUDY OF THE WEB MINING SYSTEM

In order to demonstrate the proposed system, a Web site, called veryMobile-Phone (http://203.68.224.196/verymobile/), was constructed in Chinese, in order to catch the behavior of local customers in Taiwan. The online operation procedure, based on the information provided from an off-line established database, is shown in Figure 2.

This initial database was established based on a survey of 800 individuals. The respondents were full- and part-time students, ranging from 15 to 35 years of age, at the Minghsin University of Science and Technology and full-time students, ranging from 20 to 30 years of age, at the Tsing Hua University. A total of 638 questionnaires were returned. After deleting invalid questionnaires, we had 562 valid responses. In this questionnaire, personal data, such as *Sex — male or female*; *Age — under 15, 16~18, 19~22, 23~30 or over 30*; *Education — senior high school, college, university, masters, Ph.D. or others*; *Average Income — none, under NTS 18,000, 18,000~30,000, 30,000~45,000 or over 45,000* were recorded, along with their preferences in purchasing a mobile phone, with features made up of A_1:*brand*, A_2:*external*, A_3:*price*, A_4:*service*, A_5:*function*, A_6:*ease of use*, A_7:*special offer*, etc. Via the BOFCM, users were classified into c=4 groups. The mobile phones, in stock, were also grouped by their

Figure 2. Flow diagram of verymobile-phone system



features, according to the concepts defined for information retrieval. Below, we demonstrate how the proposed mechanism can be used to suggest the appropriate mobile phone brand for each user, by responding to his or her query, based on his or her features and preferences.

Off-Line Phase

The off-line analysis is used mainly to establish the initial database features, including user categories and preferences, as well as mobile phone clusters. The users' data were grouped by applying the

BOFCM. Four clusters were obtained and stored. For each group of users, the concept matrix was calculated, as shown below, to describe the preference relationships obtained among the mobile phones features:

$A1 =$

	brand	external	price	service	function	ease of use	special offer
brand	1	0.11	0.12	0.1	0.07	0.11	0
external	0.11	1	0.1	0.08	0.06	0.09	0
price	0.12	0.12	1	0.09	0.07	0.11	0
service	0.1	0.08	0.09	1	0.05	0.1	0
function	0.07	0.06	0.07	0.05	1	0.06	0
ease of use	0.11	0.09	0.11	0.1	0.06	1	0
special offer	0	0	0	0	0	0	1

;

$A2 =$

	brand	external	price	service	function	ease of use	special offer
brand	1	0.12	0.15	0.1	0.09	0.15	0
external	0.12	1	0.11	0.08	0.07	0.1	0
price	0.15	0.11	1	0.1	0.08	0.12	0
service	0.1	0.08	0.1	1	0.06	0.09	0
function	0.09	0.07	0.08	0.06	1	0.07	0
ease of use	0.15	0.1	0.12	0.09	0.07	1	0
special offer	0	0	0	0	0	0	1

;

$A3 =$

	brand	external	price	service	function	ease of use	special offer
brand	1	0.13	0.15	0.12	0.1	0.13	0.05
external	0.13	1	0.11	0.1	0.08	0.1	0.04
price	0.15	0.11	1	0.12	0.09	0.12	0.05
service	0.12	0.1	0.12	1	0.08	0.1	0.04
function	0.1	0.08	0.09	0.08	1	0.08	0.03
ease of use	0.13	0.1	0.12	0.1	0.08	1	0.04
special offer	0.05	0.04	0.05	0.04	0.03	0.04	1

;

$A4 =$

	brand	external	price	service	function	ease of use	special offer
brand	1	0.12	0.14	0.12	0.08	0.13	0.04
external	0.12	1	0.11	0.1	0.07	0.11	0.03
price	0.14	0.11	1	0.11	0.08	0.12	0.04
service	0.12	0.1	0.11	1	0.07	0.11	0.04
function	0.08	0.07	0.06	0.07	1	0.08	0.03
ease of use	0.13	0.11	0.12	0.11	0.08	1	0.04
special offer	0.04	0.03	0.04	0.04	0.03	0.04	1

.

Taking Cluster 2 as an example, the transitive closure of the concept matrix $A2$ is shown in the following analysis:

$A2^* =$

	brand	external	price	service	function	ease of use	special offer
brand	1	0.12	0.15	0.1	0.09	0.15	0
external	0.12	1	0.12	0.1	0.09	0.12	0
price	0.15	0.12	1	0.1	0.09	0.1	0
service	0.1	0.1	0.1	1	0.09	0.1	0
function	0.09	0.09	0.09	0.09	1	0.09	0
ease of use	0.15	0.12	0.15	0.1	0.09	1	0
special offer	0	0	0	0	0	0	1

In the meantime, the document descriptor matrix was generated by 14 mobile-phone brands versus 7 concepts:

$D =$

	brand	external	price	service	function	ease of use	special offer
BenQ	0	0.14	0.29	0	0.14	0	0.43
ALCATEL	0.13	0.38	0.25	0	0.06	0	0.19
Sony ERICSSON	0.17	0.27	0.08	0.06	0.27	0.04	0.12
Kyocera	0	0.80	0.07	0	0.07	0	0.07
Mitsubishi	0.20	0.40	0	0	0.20	0	0.20
MOTOROLA	0.17	0.30	0.11	0.03	0.21	0.06	0.12
NEC	0.17	0.39	0.22	0.04	0.17	0	0.00
NOKIA	0.22	0.27	0.19	0.02	0.21	0.03	0.06
Panasonic	0.10	0.54	0.10	0	0.17	0	0.08
PHILIPS	0	0.33	0.50	0	0	0	0.17
SAGEM	0.14	0.21	0.14	0.07	0.11	0	0.32
SIEMENS	0.04	0.17	0.25	0	0.33	0	0.21
BOSCH	0	0	0.67	0	0	0	0.33
Others	0	0	1.00	0	0	0	0.00

To obtain the document-concept matrix, the D and $A2^*$ matrix was composed.

$B2 = D \otimes A2^* =$

	brand	external	price	service	function	ease of use	special offer
BenQ	0.15	0.14	0.29	0.10	0.14	0.15	0.43
ALCATEL	0.15	0.38	0.25	0.10	0.09	0.15	0.19
Sony ERICSSON	0.17	0.27	0.15	0.10	0.27	0.15	0.12
Kyocera	0.12	0.80	0.12	0.10	0.09	0.12	0.07
Mitsubishi	0.20	0.40	0.15	0.10	0.20	0.15	0.20
MOTOROLA	0.17	0.30	0.15	0.10	0.21	0.15	0.12
NEC	0.17	0.39	0.22	0.10	0.17	0.15	0.00
NOKIA	0.22	0.27	0.19	0.10	0.21	0.15	0.06
Panasonic	0.12	0.54	0.12	0.10	0.17	0.12	0.08
PHILIPS	0.15	0.33	0.50	0.10	0.09	0.15	0.17
SAGEM	0.14	0.21	0.14	0.10	0.11	0.14	0.32
SIEMENS	0.15	0.17	0.25	0.10	0.33	0.15	0.21
BOSCH	0.15	0.12	0.67	0.10	0.09	0.15	0.33
Others	0.15	0.12	1.00	0.10	0.09	0.15	0.00

From the matrix $B2$, the relationship between each mobile phone is calculated.

$R2 =$

	BenQ	ALCATEL	Sony ERICSSON	Kyocera	Mitsubishi	MOTOROLA	NEC	NOKIA	Panasonic	PHILIPS	SAGAM	SIEMENS	BOSCH	Others
BenQ	1.00	0.65	0.57	0.37	0.58	0.58	0.53	0.56	0.45	0.61	0.74	0.70	0.69	0.43
ALCATEL	0.65	1.00	0.68	0.58	0.81	0.73	0.77	0.67	0.65	0.79	0.69	0.70	0.56	0.42
Sony ERICSSON	0.57	0.68	1.00	0.51	0.79	0.93	0.71	0.84	0.65	0.61	0.67	0.75	0.45	0.37
Kyocera	0.37	0.58	0.51	1.00	0.57	0.54	0.56	0.51	0.77	0.48	0.47	0.40	0.32	0.28
Mitsubishi	0.58	0.81	0.79	0.57	1.00	0.84	0.77	0.77	0.72	0.65	0.68	0.68	0.47	0.34
MOTOROLA	0.58	0.73	0.93	0.54	0.84	1.00	0.76	0.86	0.70	0.65	0.69	0.70	0.46	0.37
NEC	0.53	0.77	0.71	0.56	0.77	0.76	1.00	0.78	0.71	0.63	0.55	0.60	0.42	0.42
NOKIA	0.56	0.67	0.84	0.51	0.77	0.86	0.78	1.00	0.64	0.60	0.62	0.67	0.44	0.40
Panasonic	0.45	0.65	0.65	0.77	0.72	0.70	0.71	0.64	1.00	0.54	0.55	0.51	0.36	0.31
PHILIPS	0.61	0.79	0.61	0.48	0.65	0.65	0.63	0.60	0.54	1.00	0.60	0.61	0.70	0.56
SAGEM	0.74	0.69	0.67	0.47	0.68	0.69	0.55	0.62	0.55	0.60	1.00	0.67	0.61	0.36
SIEMENS	0.70	0.70	0.75	0.40	0.68	0.70	0.60	0.67	0.51	0.61	0.67	1.00	0.56	0.41
BOSCH	0.69	0.56	0.45	0.32	0.47	0.46	0.42	0.44	0.36	0.70	0.61	0.56	1.00	0.66
Others	0.43	0.42	0.37	0.28	0.34	0.37	0.42	0.40	0.31	0.56	0.36	0.41	0.66	1.00

Then the transitive closure of the concept matrix *R*2 can be obtained, as shown below, which is an equivalent matrix that can be used for clustering, according to the desired level of similarity, λ.

$R2^* =$

	BenQ	ALCATEL	Sony ERICSSON	Kyocera	Mitsubishi	MOTOROLA	NEC	NOKIA	Panasonic	PHILIPS	SAGAM	SIEMENS	BOSCH	Others
BenQ	1.00	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.74	0.70	0.70	0.66
ALCATEL	0.70	1.00	0.81	0.72	0.81	0.81	0.78	0.81	0.72	0.79	0.70	0.75	0.70	0.66
Sony ERICSSON	0.70	0.81	1.00	0.72	0.84	0.93	0.78	0.86	0.72	0.79	0.70	0.75	0.70	0.66
Kyocera	0.70	0.72	0.72	1.00	0.72	0.72	0.72	0.72	0.77	0.72	0.70	0.72	0.70	0.66
Mitsubishi	0.70	0.81	0.84	0.72	1.00	0.84	0.78	0.84	0.72	0.79	0.70	0.75	0.70	0.66
MOTOROLA	0.70	0.81	0.93	0.72	0.84	1.00	0.78	0.86	0.72	0.79	0.70	0.75	0.70	0.66
NEC	0.70	0.78	0.78	0.72	0.78	0.78	1.00	0.78	0.72	0.78	0.70	0.75	0.70	0.66
NOKIA	0.70	0.81	0.86	0.72	0.84	0.86	0.78	1.00	0.72	0.79	0.70	0.75	0.70	0.66
Panasonic	0.70	0.72	0.72	0.77	0.72	0.72	0.72	0.72	1.00	0.72	0.70	0.72	0.70	0.66
PHILIPS	0.70	0.79	0.79	0.72	0.79	0.79	0.78	0.79	0.72	1.00	0.70	0.75	0.70	0.66
SAGEM	0.74	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	1.00	0.70	0.70	0.66
SIEMENS	0.70	0.75	0.75	0.72	0.75	0.75	0.75	0.75	0.72	0.75	0.70	1.00	0.70	0.66
BOSCH	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	1.00	0.66
Others	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	1.00

In our system, a default value of λ is defined by taking the mean value of all elements of the upper triangle. That is, λ = 0.73 is the clustering threshold of $R2^*$ and with such λ-cut operation, $R2^*$ can be transformed into a 0/1 matrix.

$R2^*_{\lambda=0.73} =$

	BenQ	ALCATEL	Sony ERICSSON	Kyocera	Mitsubishi	MOTOROLA	NEC	NOKIA	Panasonic	PHILIPS	SAGAM	SIEMENS	BOSCH	Others
BenQ	1	0	0	0	0	0	0	0	0	0	1	0	0	0
ALCATEL	0	1	1	0	1	1	1	1	0	1	0	1	0	0
Sony ERICSSON	0	1	1	0	1	1	1	1	0	1	0	1	0	0
Kyocera	0	0	0	1	0	0	0	0	1	0	0	0	0	0
Mitsubishi	0	1	1	0	1	1	1	1	0	1	0	1	0	0
MOTOROLA	0	1	1	0	1	1	1	1	0	1	0	1	0	0
NEC	0	1	1	0	1	1	1	1	0	1	0	1	0	0
NOKIA	0	1	1	0	1	1	1	1	0	1	0	1	0	0
Panasonic	0	0	0	1	0	0	0	0	1	0	0	0	0	0
PHILIPS	0	1	1	0	1	1	1	1	0	1	0	1	0	0
SAGEM	1	0	0	0	0	0	0	0	0	0	1	0	0	0
SIEMENS	0	1	1	0	1	1	1	1	0	1	0	1	0	0
BOSCH	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Others	0	0	0	0	0	0	0	0	0	0	0	0	0	1

In consequence, 5(*u*=5) groups of mobile-phone types can be obtained from the 14 available brands, as follows:

Group 1 = {BenQ, SAGEM},

Group 2 = {ALCATEL, Sony ERICSSON, Mitsubishi, MOTOROLA, NEC, NOKIA, PHILIPS, SIEMENS},

Group 3 = {Kyocera, Panasonic},

Group 4 = {BOSCH},

Group 5 = {Others}

Based on the document descriptor of each document, we obtained the group-concept matrix *B*2', which extensively reduces the data dimension and thus speeds up the information retrieval process.

$B2' =$

	brand	external	price	service	function	ease of use	special offer
Group 1	0.15	0.17	0.21	0.10	0.13	0.15	0.38
Group 2	0.17	0.31	0.23	0.10	0.20	0.15	0.13
Group 3	0.12	0.67	0.12	0.10	0.13	0.12	0.08
Group 4	0.15	0.12	0.67	0.10	0.09	0.15	0.33
Group 5	0.15	0.12	1.00	0.10	0.09	0.15	0.00

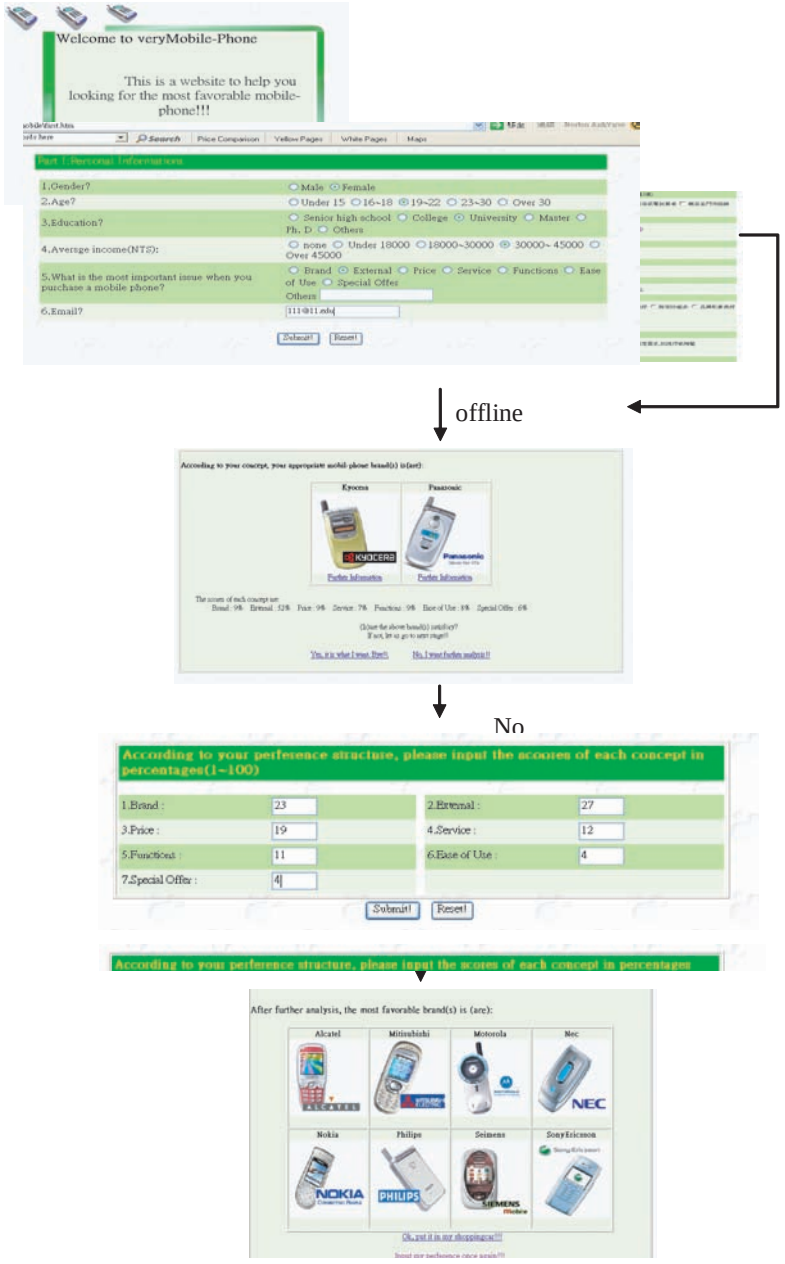
With the same procedure, we can calculate the document-concept matrices *B1, B3, B4* for each set of clustered users, respectively; this clustering information is also stored in the database. This completes the off-line phase.

Online Phase

If a Web user wants to buy a mobile phone, and signs into the Web site, he or she is asked to provide basic data. If, say, the user is *female*, *22 years old*, *university* educated and earns *NTS 30,000~45,000 Income*, and emphasized *external* as her top preference in purchasing a mobile phone, then this information will allow the sys-

tem to classify the user into user-cluster 2, and with lexicographic ordering of the components, corresponding to the concept "*external*" of *B*2', the system will provide { Kyocera, Panasonic } of *Group* 3 with the scores of each concept in percentages of (9,50,9,7,10,9,6). The correspond-

Figure 3. Result of the case study



ing scores come up with *brand*: 12, *external*: 67, *price*: 12, *service*: 10, *function*: 13, *ease of use*: 12, and *special offer*: 8 for reference.

If she is not satisfied, the system will ask for her preference structure with reference to the above scores. If she replies with the scores of (23, 27, 19, 12, 11, 4, 4), comparing vector Q with the matrix *B'*, we can find that the most compatible group of mobile phone is the second one, *Group* 2 = {ALCATEL, Sony ERICSSON, Mitsubishi, MOTOROLA, NEC, NOKIA, PHILIPS, SIEMENS} and then suggest that this user purchase the most relevant mobile phone. The result, shown below, has been translated into English for ease of understanding (see Figure 3). Different types of users map into different users' clusters and the system provides the most appropriate information corresponding to different clusters of documents. For example, if a *male* user, *18 years old*, *college* educated, with *no Income*, and an emphasis on *function*, he would be classified into Cluster 4. The documents would be grouped as *Group* 1: {BenQ, SAGEM}; *Group* 2: {ALCATEL, Sony ERICSSON, Kyocera, Mitsubishi, MOTOROLA, NEC, NOKIA, PHILIPS, SIEMENS}, *Group* 3: {Panasonic}, *Group* 4: {BOSCH} and *Group* 5: {Others}. The system will provide {Panasonic} with the scores of each concept in percentages of (11, 13, 18, 8, 25, 9, 15). Furthermore, if he is not satisfied, after entering a new set of scores, the system will provide a new suggestion.

If the users referred to above purchased the mobile phones recommended, their data would be used to update the database, otherwise the database will not be changed. Due to the billing system, such updating processes would be carried out once a month.

SUMMARY AND DISCUSSION

Internet technology has developed rapidly in recent years and one of the primary current issues is how to effectively provide information to users. In this study, we utilized a data mining information retrieval technique to create a Web mining system. Since existing retrieval methods do not consider user preferences and hence do not effectively provide appropriate information, we used an off-line process to cluster users, according to their features and preferences, using a bi-criteria BOFCM algorithm. By doing so, the online response time was reduced in a practical test case when a user sent a query to the Web site. The case study in this chapter (a service Web site selling mobile phones) demonstrated that by using the proposed information retrieval technique, a query-response containing a reasonable number, rather than a huge number, of mobile phones could be provided which best matched a users' preferences. Furthermore, it was shown that a single criterion for choosing the most favorable mobile-phone brand was not sufficient. Thus, the scores provided for the suggested group could be used as a reference for overall consideration. This not only speeds up the query process, but can also effectively support purchase decisions. In system maintenance, counterfeit information causes aggravation for Web site owners. Our proposed system updates the database only if the purchase action is actually carried out, which reduces the risk of false data. Further study into how a linguistic query may be transformed into a numerical query is necessary to allow a greater practical application of this proposal.

REFERENCES

Angelides, M. C. (1997). Implementing the Internet for business: A global marketing opportunity. *International Journal of Information Management*, *17*(6), 405-419.

Bedeck, J. C. (1973). *Fuzzy mathematics in pattern classification*. Unpublished doctoral dissertation, Applied Mathematics Center, Cornell University, Ithaca.

Chen, S. M., & Wang, J. Y. (1995). Document retrieval using knowledge-based fuzzy information retrieval techniques. *IEEE Transactions on Systems, Man and Cybernetics, 25*(5), 793-802.

Chen, Z. (2001). *Data mining and uncertain reasoning.* New York: John Wiley.

Hanson, W. (2000). *Principles of Internet marketing.* Sydney: South-Western College.

Janal, D. S. (1995). *Online marketing handbook: How to sell, advertise, publicize and promote your product and services on Internet and commercial online systems.* New York: Van Nostrand Reinhold.

Jang, J. S., Sun, C. T., & Mizutani, E. (1997). *Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence*. Upper Saddle River, NJ: Prentice-Hall.

Lin, C. L. (2002). *Web mining based on fuzzy means for service web site*. Unpublished master's dissertation, Tsing Hua University, Taiwan.

Mohammed, R. A., Fisher, R. J., Jaworski, B. J., & Paddison, G. J. (2004). *Internet marketing - Building advantage in a networked economy.* Boston: McGraw Hill / Irwin.

Schneider, G. P. (2004). *Electronic commerce: The second wave.* Australia: Thomson Learning.

Rayport, C., & Jaworski, H. (2002). *E-commerce marketing: Introduction to e-commerce.* Boston: McGraw Hill / Irwin.

Tamura, S., Higuchi, K., & Tanaka, K. (1971). Pattern classification based on fuzzy relations. *IEEE Transactions on Systems, Man and Cybernetics, 1,* 61-66.

Wang, H. F., Wang, C., & Wu, G. Y. (1994). Multicriteria fuzzy C-means analysis. *Fuzzy Sets and Systems, 64,* 311-319.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338-353.

ENDNOTE

[1] This study is supported by National Science Council, Taiwan, ROC, with project number NSC 91-2213-E-007-075.

This work was previously published in Business Applications and Computational Intelligence, edited by K. Voges and N. Pope, pp. 113-130, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.26
Semantic Web Services and Mobile Agents Integration for Efficient Mobile Services

Vasileios Baousis
University of Athens, Greece

Vassilis Spiliopoulos
University of the Aegean and National Centre of Scientific Research "Demokritos", Greece

Elias Zavitsanos
University of the Aegean and National Centre of Scientific Research "Demokritos", Greece

Stathes Hadjiefthymiades
University of Athens, Greece

Lazaros Merakos
University of Athens, Greece

ABSTRACT

The requirement for ubiquitous service access in wireless environments presents a great challenge in light of well-known problems like high error rate and frequent disconnections. In order to satisfy this requirement, we propose the integration of two modern service technologies: Web Services and Mobile Agents. This integration allows wireless users to access and invoke semantically enriched Web Services without the need for simultaneous, online presence of the service requestor. Moreover, in order to improve the capabilities of Service registries, we exploit the advantages offered by the Semantic Web framework. Specifically, we use enhanced registries enriched with semantic information that provide semantic matching to service queries and published service descriptions.

Finally, we discuss the implementation of the proposed framework and present our performance assessment findings.

INTRODUCTION

Efficient execution of wireless applications is of paramount importance due to the highly dynamic wireless network conditions. Link outages occur in a near-stochastic pattern, thus, rendering the execution of applications quite tedious and uncertain. Research on mobile computing has for a long time focused on this specific aspect of wireless application engineering (Pour, 2006). In this article, we adopt the mobile agent paradigm in order to overcome the difficulties discussed above. Surely, this is not the first time that mobile agents are proposed as the vehicle for the implementation of wireless/mobile applications. Their autonomic nature and wide spectrum of characteristics render the specific technological platform a great enabler for the emerging ubiquitous computing paradigm.

Mobile computing is not the only development that significantly impacts the computer industry nowadays. Service-oriented architectures (SOA) are gradually changing the contemporary structure of the Internet and become a key facilitator for electronic commerce applications and related application domains. We try to incorporate both the discussed technologies into our wireless/mobile computing framework. Mobile agents are dispatched by mobile terminals in order to efficiently and safely satisfy the specific computing needs of their nomadic owner. After securing the autonomicity characteristic in order to progress the required task without the need for the mobile terminal to be constantly online, we try to minimize the service-related tasks. Our prime concern lies on the exact identification of the services to be executed at the demand of the user and minimize potential waste of time on unwanted invocations. The accuracy of the service inquiry mechanism has to be improved to really boost the mobile agent and service-oriented architecture. To expedite the service querying procedure and simplify the querying semantics, we employ a semantically enriched service registry. A precise definition of the user's requirement is mapped to existing services through a semantically enriched registrar.

In this article, we introduce a novel framework for dynamic discovery and integration of semantically enriched Web Services (WS) with Mobile Agents (MA). The proposed framework is mostly intended for wireless environments where users access Semantic Web Services (SWS) in the fixed network (the terms Web Service (WS) and Semantic Web Service (SWS) are used interchangeably within this article). This framework enhances the fixed network with the intelligence needed to dispatch the service requests of the wireless user in an efficient, reliable and transparent manner. The proposed approach enables users to execute multiple services with minimum interaction, without the requirement of being online during their entire session. Additionally, the proposed framework provides better fixed network utilization since unnecessary communication overhead is avoided and reliable delivery of the service results is provided.

The rest of this article is structured as follows. In section 2, we provide some background knowledge about the implemented technologies, whereas section 3, we discuss relevant prior work. In section 4, we present an overview of the proposed architecture. Section 5 studies the performance of the proposed framework and presents the results. Finally, section 6 concludes the article.

BACKGROUND KNOWLEDGE

In this section, we briefly describe the two technologies that are integrated in our proposed framework, namely Web Services and Mobile Agents.

Web services (WS) provide a loosely coupled infrastructure for service description, discovery and execution. In the traditional WS model, service requestors find the appropriate service by placing a request to the service registry, often implemented with universal description, discovery and integration (UDDI), obtain the result(s)—public interfaces of the chosen service(s) (expressed in Web services description language - WSDL) and, finally, send simple object access protocol (SOAP) messages to WS provider(s).

The main problems experienced in these interactions are:

- UDDI guarantees syntactic interoperability, and does not provide a semantic description of its content. UDDI is characterised for its lack of semantic description mechanisms, such as semantic interoperability, explicit semantic models to understand the queries and inference capabilities. UDDI service discovery is performed primarily by service name (keyword matching), but not by service attributes/capabilities. UDDI tModels may be regarded as a vocabulary where service descriptions are unstructured and intended for human comprehension. Different services with the same capabilities can thus be categorized in different business categories.
- WSDL is XML-based and used to specify the interface of a WS. It describes the information being exchanged (structure of the SOAP messages), how this information is being exchanged via interactions with the WS (transport protocols) and where the WS is located. However, WSDL does not contain any information about the capabilities of the described service and as such service discovery based on service capabilities or semantics cannot be performed.

Several efforts have been made to address the lack of expressiveness in WSDL in terms of semantic description that fall into the area of the Semantic Web (SW). SW is a vision in which Web pages are augmented with semantic information and data expressed in an unambiguous manner and can be understood and interpreted by machine applications and humans alike (Berners-Lee, 2001). This requires means to represent the semantics of the exchanged data so that it could be automatically processed. This requirement is met with the use of ontologies. Ontologies facilitate knowledge sharing among heterogeneous systems, through explicit formal specifications of the terms used in a knowledge domain and relations among them (Gruber, 1993). Ontologies are machine-understandable and, as such, a computer can process data, annotated with references to ontologies. Through the knowledge encapsulated in the ontology, a computer can deduce facts from the originally provided data. The use of ontologies enables systems to share common understanding of the structure of information and reuse of domain knowledge, make domain assumptions explicit and separate domain knowledge from the operational knowledge.

Currently, several upper ontologies (terminology in the form of an ontology) have been proposed for Web Service description. The first was DAML-S (McIlraith, 2003), which was based on DAML+OIL ontology language. When DAML+OIL evolved to the widely accepted OWL (Web Ontology Language) family of languages, DAML-S was replaced by OWL-S (OWL-S, 2007). Still, OWL-S does not constitute a commonly accepted description language; there are also other languages proposed such as WSDL-S (Verma, 2006), WSMO (Roman, 2005) and SWSO (SWSL Committee, 2007). All these languages differ in terms of expressiveness, complexity and tool support.

OWL-S, which is adopted in our work, has well-defined specifications by the W3C (World Wide Web) consortium (OWL-S, 2007) and is widely accepted by the scientific community. OWL-S ontology implicitly defines message types

(as input/output types of processes) in terms of OWL classes, which allows for a rich, class-hierarchical semantic foundation. Specifically, OWL-S models the Web services via a three-part ontology: (i) a service profile describes what the service requires from users and what it gives them; (ii) a service model specifies how the service works; and (iii) a service grounding provides information on how to use the service.

With OWL-S, SWS are described in an unambiguous manner allowing for a potential service requestor to place a capability search in a service registry rather than a keyword search in UDDI registries. Registries that offer such capability search functionalities are called Semantic Web Registries (SWR).

The most representative matching techniques used are detailed in Tsetsos (2007) and are summarised below:

- **Semantic capability matching:** The basic idea is that an advertised service matches a requested one, when all the inputs (respectively outputs) of a requested service are matched by the inputs (respectively outputs) of the advertised service. For this purpose, description logics (DL) reasoning services are exploited for inferring relations between ontology concepts.
- **Multi-level mapping:** Matching is performed in many levels, not only between input and output descriptions. Service categories or other custom service parameters (e.g., OoS) may be exploited. The result is a more efficient ranking of the matched services.
- **DL matching with service profile ontologies:** Each service and query is represented as ontologies following the DL formalization. Hence, a DL reasoner is utilized for placing the query concept in its proper position in each service ontology description (e.g., as a sub-concept). Then specific rules are applied for computing the degree of relevance between the query and each service description.
- **Information-retrieval-based:** In this category, vector space techniques (Raghavan, 1986) are utilized for locating the most related service to a provided query.
- **Graph-based approaches:** Ontologies representing services are transformed into directed graphs and various algorithms accomplish the matching between such graphs.

There is a plethora of tools that provide Semantic Web Services functionalities (e.g., OWLS-MX (OWLS-MX, 2007) and TUB OWLSM (OWLSM, 2007), each one implementing a portion of the above matching techniques. In our work, we adopted the OWL-S/UDDI Matchmaker tool (Paolucci, 2002; Srinivasan, 2004; OWL-S/UDDI Matchmaker Web Interface, 2007), which mainly implements techniques from the first aforementioned matching technique.

Mobile agent technology is one of the most promising technologies for communicating and managing functional components comprising a mobile service (Lange, 1998; Wooldridge, 2002). A MA has the unique ability to autonomously transport itself from one system to another. The ability to travel allows a MA to move to a system that contains an entity (-ies) with which the agent wishes to interact and take advantage of being in the same host or network with the collaborating entity. MAs can operate synchronously and asynchronously, and are equipped with the appropriate intelligence and knowledge to dynamically accomplish their task without user interaction. MAs are not trying to replace traditional ways of communication but to enhance the functionality and operation of the involved service entities. Researchers agree that MAs are not always the best solution and a combination of the MA, client-server and remote execution paradigms delivers

the best performance with respect to network operation metrics like bandwidth, response time, and scalability.

RELATED WORK

In this section, we provide an overview of the related work performed in the areas of semantic WS and multi-agent systems and, especially, on research activities that integrate these two technologies.

In Ishikawa (2004; 2004b), BPEL (business process execution language) is used to form simple rules to describe MA physical behaviours (e.g., migration and cloning). Such simple rules are separated from the integration logic, allowing for addition or change of physical behaviours without modification of the BPEL description. This separation is considered helpful in dealing with the dynamic environment of WS, however, the discussed framework supports actions only in case of predefined events. The implemented rules do not consider dynamic events that might be generated during WS invocation and MA roaming. Moreover, and importantly, directory services and multicast protocols are assumed pre-existing and not discussed. The discussed framework refers only to interactions occurring among MA and WS without considering the interactions of the MA and service registries that have equal importance in such a system. Finally, the system description does not include any implementation, hence benchmarking is not considered.

There are several proposed models that adopt BPEL4WS (Business Process Execution Language for Web Services) as a specification language for expressing the social behaviour of multi-agent systems and adapt to changing environment conditions (Bulher, 2003;2004). Moreover, in Montanari (2003; 2003b), the authors propose a policy-based framework for flexible management and dynamic configurability of agent mobility behaviour in order to reduce code mobility concerns and support rapid mobile code-based service provisioning. Policies specify when, where, how and the parts of the agent that will perform a given task (e.g., migrating to a host and invoke a service). However, these models do not provide for the semantic description of the WS involved in their systems. Other proposed frameworks adopt DAML-S for describing the WS, thus, allowing for service capability search and matching (Kagal, 2002; Gibbins, 2004). However, the proposed systems are intended for fixed networks, and problems related to the wireless environments are not considered.

An agent-based approach for composite mobile WS is proposed in Zahreddine (2005), where three methods for compositions are discussed: parallel, sequential and a hybrid of these two. The service composition scenario is that a user with a wireless device places a request to execute a WS and a MA executes the service on behalf of the user by moving to the service registry, query the registry, get service description (in WSDL), and finally invoke the service. Service execution, depending on the WS itself, is performed with one of the aforementioned composition methods. This approach does not consider semantic information describing the involved WS, thus services are selected by simple keyword queries to the UDDI registry. Additionally, it does not include mechanisms to decide which composition approach to follow. Integration depends upon the nature of the WS (if the service is a composition of other services it must be accessed sequentially, if not, then in parallel). A similar approach is proposed in Cheng (2002), with the difference that a personal and a service agent are used to perform the task of the MA described in the previously mentioned approach.

Moreover, in the research literature, it has proposed the use of Asynchronous Web Services (AWS) in order to access WS with asynchronous interaction. AWS can be used where the standard

Web Service Business Model has some limitations, as described in Brambilla (2004): a. when service time is expected to be too long, b. when response time is not predictable, and c. when users may not be continuously online. The most common way to achieve asynchronous calls to Web Services is by using a correlation or conversation ID (Huang, 2003; Brambilla, 2004). This unique ID is assigned initially by the Web Service providers to each Web service transaction and it is passed in each exchanged message between the conversational parties. This way, the client is able to perform correlation and to retrieve application data related to current conversation. The drawback though of such an approach is the production of possible mismatches. Specifically, if multiple asynchronous Web Service calls happen in the context of a single conversation, responses might not be able to be unambiguously related to their requests (Brambilla, 2004).

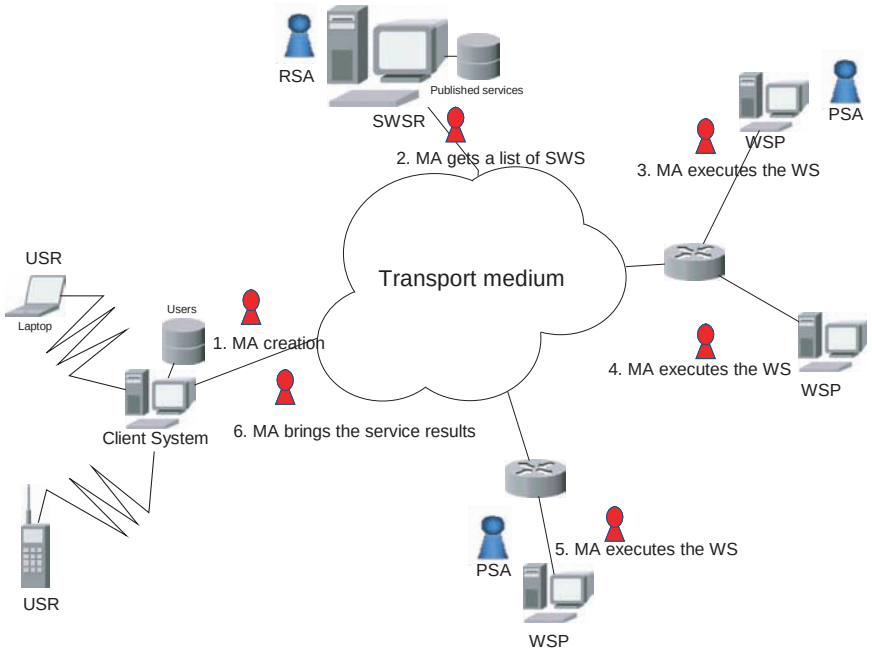
FRAMEWORK ARCHITECTURE

The proposed framework consists of the mobile user that uses SWS, the MA representing the user in the fixed network, the service registry and the SWS provider. The last two entities are implemented as stationary agents. According to the service implementation scenario (Figure 1), a mobile user accesses the proposed system and places service requests specifying some criteria. Subsequently, the system creates a MA (step 1) that migrates to the registry to find the WS that best meets the user requirements (step 2). Service registry allows for a capability search to be performed, since it is enriched with semantic information. The MA, after acquiring the WS listing and technical details, migrates to service provider(s), invokes the WS, collects the results (steps 3-5) and returns to the service requestor to deliver the results to the mobile user (step 6). In the presented scenario, the SWS that matched the service request were three thus MA migrates and invokes these three services (steps 3-5). If the service request matched more than three services during the step 2, the MA would migrate to all these matched WS (Figure 1 would include more steps). The advantages of this scenario is that the MA has the necessary intelligence to invoke only the best matched service(s) and unnecessary service invocations are avoided leading to better network utilization, and the wireless user is not required to be online and may obtain the results on future time.

In the proposed framework, the route of the agent may vary, depending on the service requestor preferences and the network topology. As explained below, the user may dynamically force his MA to send its clones to the providers, invoking the services in parallel, rather than serially migrate to each one. Moreover, the user may force the MA to implement different service execution strategies (e.g., execute all services locally or remotely, change timeout limit), during its itinerary and execution of service(s).

Our framework consists of the following functional components: (1) User service requestor (USR) who is the user that invokes a SWS, and the client system, the system in the fixed network that provides user access to the SWS, (2) mobile agent which is the representative of the user in the fixed network (3) provider stationary agent (PSA) which is a stationary agent that resides in the host offering a certain WS (its implementation is optional), (4) registry stationary agent (RSA) which is a stationary agent that acts as a broker between the MA and the service registry (its implementation is optional), (5) Semantic Web services registry (SWSR), the registry where the service providers advertise their services, and (6) Web service provider (WSP) which provides the WS to interested users. Their structure and functionalities are described below. In the end of this section, we provide a service implementation scenario presenting all possible supported service invocation alternatives.

Figure 1. Service implementation scenario



User Service Requestor (USR)

USR is the client that invokes a WS. USR logs into the client system, which communicates with the agent platform using IIOP (Internet Inter-ORB Protocol). The agent platform is responsible for creating and handling MA, according to user specifications. The client system is implemented in JSP/Servlet technology, and many users can be accommodated without having java runtime environment (JRE) or the MA platform (MAP) installed on their device. The only requirement is a browser to access the client system.

The client system offers services to clients like: account creation, user login/logout, service invocation policies profile editing, and control of existing agents. Moreover, the administrator is allowed to add/remove/edit user properties/profiles. Finally, users' service invocation policy profiles are serialised and stored into the server's database that enables the seamless and transparent provision of services.

The proposed framework is in addition able to communicate with mobile devices that are capable of hosting JADE/LEAP (Lightweight Extensible Agent Platform) (JADE, 2007). LEAP is an extension of JADE that enables MAs to be executed on wireless devices with limited processing capabilities. In such a case the MA is spawned on the mobile device, gathers the user preferences/specifications either from this device or from the client system. The behaviour of the system and of the device created the MA is exactly the same.

Mobile Agent (MA)

The MA is the representative of the user in the fixed network and is capable of roaming, finding and executing services and delivering results to the user. The MA may also spawn clones that execute the selected WS in parallel to minimize the total processing time. Clones can migrate and invoke simultaneously the chosen WS and return to the service requestor with the results. The MA has the following components: (1) data state, (2) code, (3) migration and cloning policies, (4) matching engine, and, (5) policy management component (Figure 2).

The proposed MA architecture is based on that the logic of the MA has to be separated from its implementation, enabling the modelling of the MA to be platform independent. That is to say, the physical behaviours of the MA are portable to any MAP (Jade, Grasshopper, etc.). Below, we describe the components of a MA.

The data component contains the information collected by the MA from the SWS invocations. Several compression algorithms may be applied in order to reduce the size of the collected information. The migration and cloning policies component specifies the autonomous behaviour of the MA. It should be noted that the social behaviour of the MA (migration, cloning) is separated from integration logic and code implementation. This separation is accomplished with user's specified invocation policies that govern the behaviour of the MA, being external and independent of its code and integration with the WS. Moreover, the matching engine component is responsible for post-processing the service registry query results, that is, confirm the availability of the service providers prior to agent migration.

The policy management component is responsible for the MA external communication and the transparent installation of policies into the agent's repository.

As shown in Figure 3, the policy management component provides four services, namely communication, update handler, specification and policy repository. Policy repository contains

Figure 2. Mobile agent structure

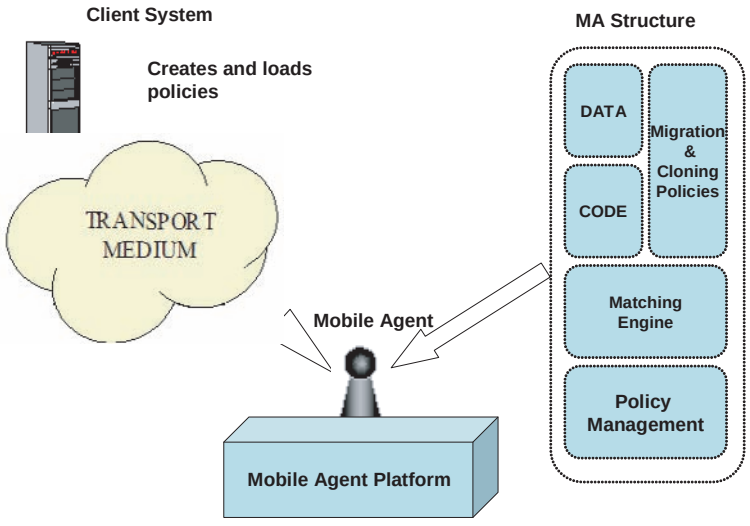
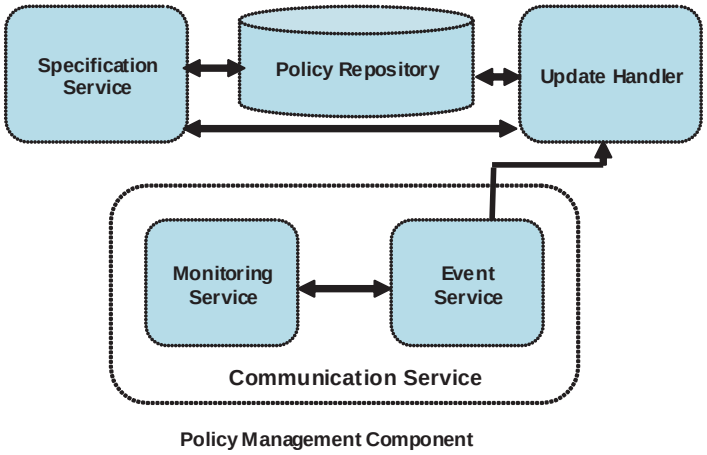


Figure 3. MA policy management component



the user preferences and policies that govern the behaviour of the MA. Communication service enables the MA to interact with the client and other network entities. Such functionality is achieved through the monitoring service which filters the messages coming from the client system and through the event service which handles events concerning policy changes. When a policy change occurs, the update handler is notified to update the policy repository. Specification service is responsible for fulfilling this task.

The agent's policies determine its physical behaviour while roaming in the network and executing WS. Currently, the MA considers the policies (Table 1), which are Boolean and numerical variables. Agent policies are expressed in XML and stored in a serialised format into the client system database. For each registered user there is an associated policies file, to provide personalized WS access.

Provider Stationary Agent (PSA)

PSA is a stationary agent that resides in the host offering a certain WS. Its purpose is to wrap the functionality of the WS. The PSA is created and maintained by the service provider. PSA communicates with the service providers through protocols specified for WS invocation and interaction (e.g., SOAP). When the MA migrates to a host offering a WS with a PSA, it obtains the results through the PSA. This communication is performed with agent-to-agent protocols (either Remote Method Invocation or exchanging FIPA/ACL messages using a FIPA/Message Transport Protocol-MTP (FIPA, 2007)), instead of the resource consuming SOAP. In this approach, the MA need not be SOAP fluent, thus leading to a lightweight implementation. It should be noted that this implementation maintains the platform independence as far as it concerns the SWS provider. This is due to the fact that the PSA wraps

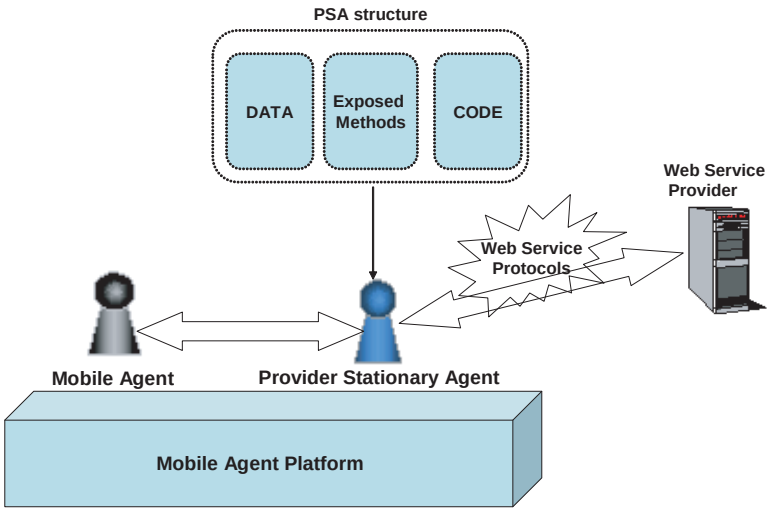
Table 1. Policy names and their respective meaning

Policy name	Type	Description
<Migrating> and <Cloning>	Boolean	MA's ability to migrate to another host and spawn clones respectively.
<retryTimes>	Numerical	The number of attempts that MA will perform when a WS is unavailable.
<timeBetweenReattempts>	Numerical	The time that MA will wait between consecutive reattempts.
<suspendWhenFinished>	Boolean	States if the user wishes (dis)-connected operation and what the MA should do when returns to the Client System (suspend its state and wait user to connect back or to deliver immediately the service results).
<rollBackBehaviour>	Boolean	Specifies in a case of failure if a roll back solution will be followed.
<maxNumberOfHits>	Numerical	The maximum number of services to be invoked.
<minNumberOfResults>	Numerical	The minimum number of results when searching the semantically enriched service registry (it is accomplished through the similarity level that is returned from the semantic engine that enables the system to always return a result, even though it does not always satisfy completely the request).
<pingServer>	Boolean	States that the MA should check if the targeted service provider is alive, before MA starts the migrating process to this host.
<migrateToServer>	Boolean	Specifies if the service will be invoked locally or remotely.
<remoteCall>	Boolean	MA invokes the chosen services using SOAP/RPC (remotely from other host) without migrating to each provider.
<callThroughStationary>	Boolean	Indicates if communication between WS and MA will take place with or without the Provider's Stationary Agent (PSA).
<HitAllServices>	Boolean	Forces the MA to invoke all retrieved services from service registry.
<cloneToServer>	Boolean	Enables the agent to decide whether to serially migrate to each located service provider or sent clones to accomplish the task in parallel and return service results to their parent agent and then are self destroyed.
	Boolean	States that user wishes to retrieve results in a future time, by reconnecting to the Client System.
<timeBetweenReattempts>	Numerical	The time that MA will wait between consecutive reattempts.

the SWS functionality, and is also SOAP fluent and exposes the same SWS's functionality in a native form to the MA.

Figure 4 presents the structure of a PSA. PSA interface exposes the available methods of the SWS as they are described in OWL-S. PSA consists of two parts: (1) its data state, and, (2) its code. PSA methods are multi-threaded to accommodate and simultaneously serve multiple MAs.

Figure 4. Provider stationary agent logic



Registry Stationary Agent (RSA)

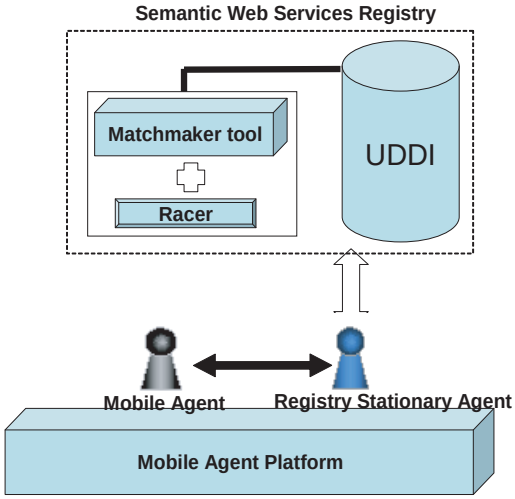
RSA is a stationary agent that acts as a broker between the MA and the service registry (Figure 5). RSA implements part of the registry's functionality and serves MA's requests. By using a RSA in the WS registry, MA does not have to be aware of the implementation-specific functionalities of the registry. Thus, different service registries can be used as long as RSA acts between WS registry and MA. The proposed framework can be used with different registries that are currently available (e.g., ebXML (ebXML, 2007), OWLS-MX (OWLS-MX, 2007), and TUB OWLSM (OWLSM, 2007)).

Semantic Web Services Registry (SWSR)

The SWSR (Figure 5) consists of the RSA, the matchmaking tool and the UDDI registry. The matchmaker (OWL-S/UDDI Matchmaker Web Interface, 2007) is a tool which enhances the UDDI server by adding capability-based discovery. In combination with Racer (RACER, 2007), it processes the ontologies expressed in OWL. Service advertisements are first processed by the UDDI server, and if any semantic information is contained by them, they are passed to the OWL-S matchmaking engine. Finally, the engine processes service queries and returns the results to the UDDI server, which in turn, communicated with the requesting service client.

The matching algorithm used by Matchmaker to match a service request to a service advertisement is based on matching all the outputs of the first to the outputs of the latter, and all the inputs of the latter to the inputs of the first. The matching degree (between I/O of a request and I/O of an advertisement) depends on the correlation of the domain ontology concepts associated with these I/O. Matchmaker specifies four matching degrees

Figure 5. Semantic WS registry



(in decreasing order of matching importance): Exact, plugin, subsumes, and fail. The query language used in the registry is the standard query language of Racer that has its basis on LISP. It is powerful and has more functionalities than standard OWL query languages.

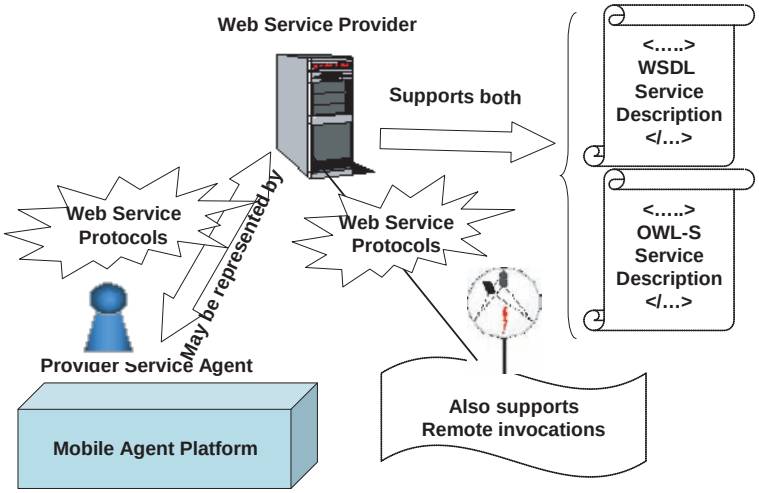
Matchmaker is a tool that integrates seamlessly with registries such as UDDI. In our system, we used a local implementation of UDDI, called jUDDI (jUDDI, 2007). JUDDI is a Web application for Apache Tomcat. The matchmaker tool is responsible for the mapping of the OWL-S service description to JUDDI. Matchmaker is plugged in JUDDI and is available in two versions, a Web-based and a standalone version. The standalone version provides a matching engine and a client API for invoking this engine. In our framework, we used the standalone version of Matchmaker. An extensive description of matchmaker can be found in Paolucci (2002) and Sycara (2004).

Web Service Provider (WSP)

The WSP provides the WS to interested clients. It maintains a description of the WS expressed in WSDL and OWL-S. Figure 6 depicts the WSP and their supported functionalities. Service invocation by the MA depends on the OWL-S description of the service. In our framework, service invocation by MA is performed either directly or through the PSA. In the direct access case, the agent has to be SOAP fluent, a fact that increases the size of the MA when moving over the network. Inside the OWL-S description of the WS, it is indicated if a PSA wraps the functionality of the service to allow the roaming MA to interact with the PSA instead of the SWS.

As mentioned above, OWL-S is used to enhance the expressiveness of WSDL in terms of semantic information. For this reason, in our framework, WS are described both in WSDL and

Figure 6. Web service provider



OWL-S. WSDL is used to describe the technical details (information included in the service grounding) and OWL-S is used to specify the input and output ontologies, thus, enabling an advanced service capability search (service profile and model). Upon retrieval of the desired services from the registry, the WSDL description is used to find the necessary definitions for its successful invocation.

As already mentioned, the SWS provider can expose a PSA to act as his delegate and interact with the user's MA. This is revealed to the MA through the OWL-S description. If this is not the case, the MA infers that no PSA is offered and the service should be accessed directly.

Service Usage Description

In this section, a functional description of the proposed framework is provided, through a service scenario. According to this scenario, a USR needs to find and invoke a certain WS by using a mobile device. Therefore, he/she connects to the client system, the platform front-end. After a successful registration, the USR sets the desired criteria for the WS. The user also defines the MA service invocation policies and forces the MA to follow a certain policy while roaming throughout the network. Subsequently, a MA is created, equipped with the user's unique ID, service invocation and agent behavioural policies, to represent the user in the fixed network and dispatch his service requests. The aforementioned policies are passed to the MA in XML format and stored into its policy repository, which remind the update handler that he has the authority to change these policies, according to the messages that the event service may receive from the USR or other network entities. The MA, after creation, migrates to the SWSR. The SWSR provides SWS descriptions and allows service capability search. When the MA arrives at the service registry, it communicates with the

RSA, which queries the registry on behalf of the MA. RSA finds the service(s) that meet the user needs and delivers them to the MA, which decides on the next step according to its specified service invocation and agent behavioural policies.

The MA may follow several WS invocation alternatives and these are listed below:

1. Poll the servers where the services are located to check their availability, in order to migrate only to those that are alive. In this way, the MA is released from the burden of migrating to a malfunctioning remote server. This strategy improves the overall performance of the framework by avoiding unnecessary migrations.
2. Try to invoke the services from remote and not migrate to the provider. Remote invocation or migration of MA is specified in the MA policies. Specifically, depending on the size of the MA or the distance between its current location and the location of the provider, it might be preferable not to migrate, but remotely invoke the WS.
3. Migrate to the WSP and collaborate with the PSA. The MA invokes the service and obtains the results through the PSA.
4. Migrate to the WSP and directly invoke the WS. This option requires the MA to carry additional code libraries. The implementation of the WSP is much simpler and straightforward since there is no change in the traditional WS implementation model.
5. Finally, to send clones to each WSP, instead of migrating serially to each one. This scenario results to a parallel invocation of WSs where each MA clone invokes one WS. In this way, the overall service invocation time is reduced in comparison to the previous service invocation alternatives.

All these service invocation alternatives are decided at runtime through the user's specified service invocation and agent behavioural policies. When the MA(s) have collected the results, there are two options depending on the selected policies:

1. When the MA invokes all the services, it migrates back to the client system. If the user is logged in the system, the MA passes the results to the user. Otherwise, the MA waits for the user to login and ask for the service results.
2. When the MA clones have been used for service invocation, they return to the client system and deliver service results to the father MA. After this interaction, the MA clones are destroyed. Consequently, the father MA delivers the services results to the user in a similar way to the previous case.

When the USR obtains the results, he may ask the MA to repeat one of the above scenarios by changing, if necessary, its policies, or he may cancel the execution of the agent. The USR may also, at any time, search for the agent, instruct him to return or cancel its execution at runtime.

A practical example of the proposed framework usage could be to book a trip from a place A to a place B and probably specifying some preferences on each action (e.g., the flight to have an intermediate stop to location C). The user requests this service by specifying his preferences and a MA fulfils this request. The MA has the intelligence to query the Semantic Web Service Registry and with the help of the semantic matchmaking capability of the registry to retrieve the most accurate service (that meets the requirements of the user), to invoke this service and provide synchronously or asynchronously the results to the user without his/her on-line presence. The semantic expression of the WS to the registry and the unambiguous matching of the user's criteria with the capabilities of the available SWS, leads to as many accurate results as possible, as well as maximization of the recall.

Maximization of the matching performance is of paramount importance, since it might be possible that the exact service does not exist in the registry catalogue, but still the most accurate result has to be retrieved. The matching algorithm that is used in the registry ensures this fact: it defines a flexible matching mechanism based on the OWL's subsumption mechanism. The degree of match between the request and the available services depends on the match between the concepts of the two ontologies. Specifically, the matching mechanism relies on a semantic matching between concepts, rather than a syntactic one. Let us consider the practical example mentioned above. The user wants to book a "flight" from location A to location B, and the registry contains a service that does not match exactly with the user's request in the sense that in the service advertisement the output is specified as "trip", as shown in Figure 7. Although there is no exact match between the output of the request and the advertisement, the matching algorithm recognizes a match, since "trip" subsumes "flight". This is a clear advantage over a simple string matching-based UDDI registry.

Figure 8 illustrates the semantic matching process that ensures efficient service retrieval since compares concepts that are unambiguous specified on three levels, service profile, model, and grounding.

If traditional methods of WS invocation are followed, this booking would be performed as follows: the user would browse to a UDDI registry, request all the WSs that provide a flight booking service and get the results. Due to the lack of semantics in the service registry (UDDI does not supports for WS semantic annotation) and to keyword search that is performed in such registries, the user would obtain WSs that provide booking services, probably either irrelevant WSs or services that are not classified according to the relevance of the query. As a result, the user would need to sequentially or randomly invoke each service till he finds a service that best meets his requirements. This interaction requires the online presence of the user during the whole interaction.

Figure 7. Semantic matching

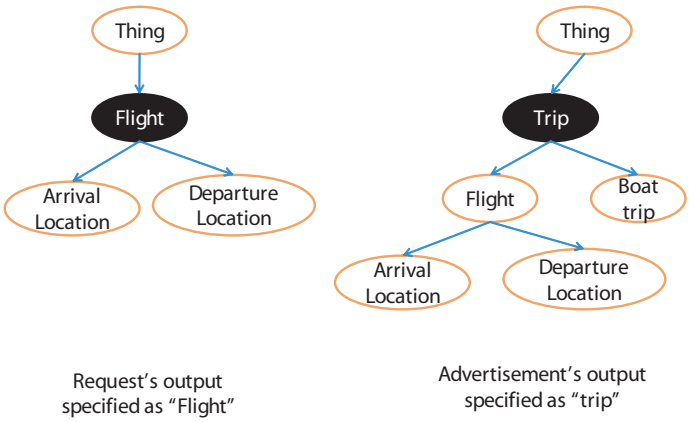
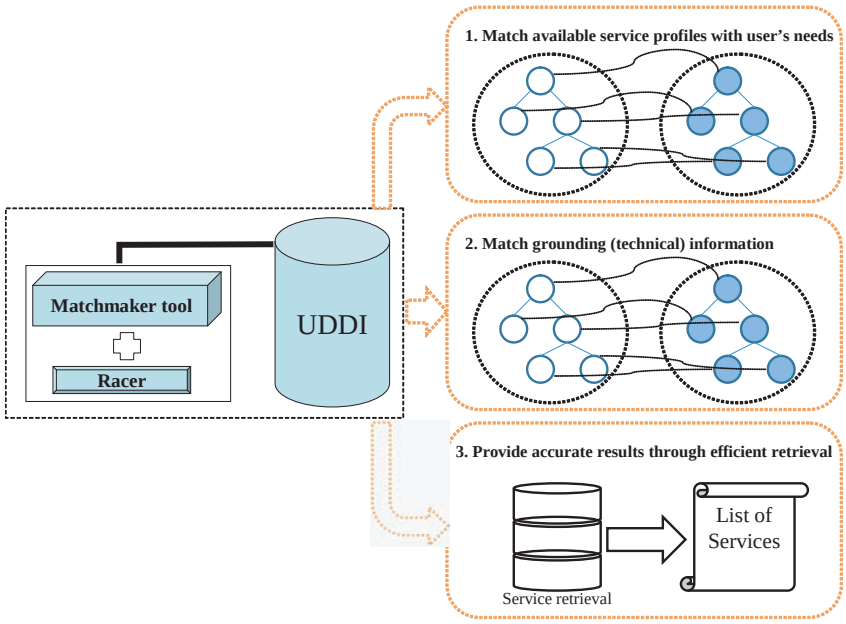


Figure 8. Web service retrieval



PERFORMANCE EVALUATION

In this section, we discuss the performance evaluation and present the results of the proposed system. Specifically, we compare the performance of our framework against the traditional business model of WS provision. In the following description, the term "conventional WS Business Model" (WSBM), refers to the model where a user requests a service to be executed and the system dispatches (either automatically or with user intervention) the request by discovering the appropriate service(s) from the service registry, and then, sequentially, invokes these WS, receives and forwards/presents to the user the service results. All communication among the involved network entities is performed with SOAP. Moreover, in our framework the mobile agents are implemented on JADE (JADE, 2007) MA platform. We have developed and tested the following system:

a. A WS system implemented with the "Conventional WS Business Model" (WSBM).
b. Our framework (Semantic Web services and mobile agents) (SWS& MA)

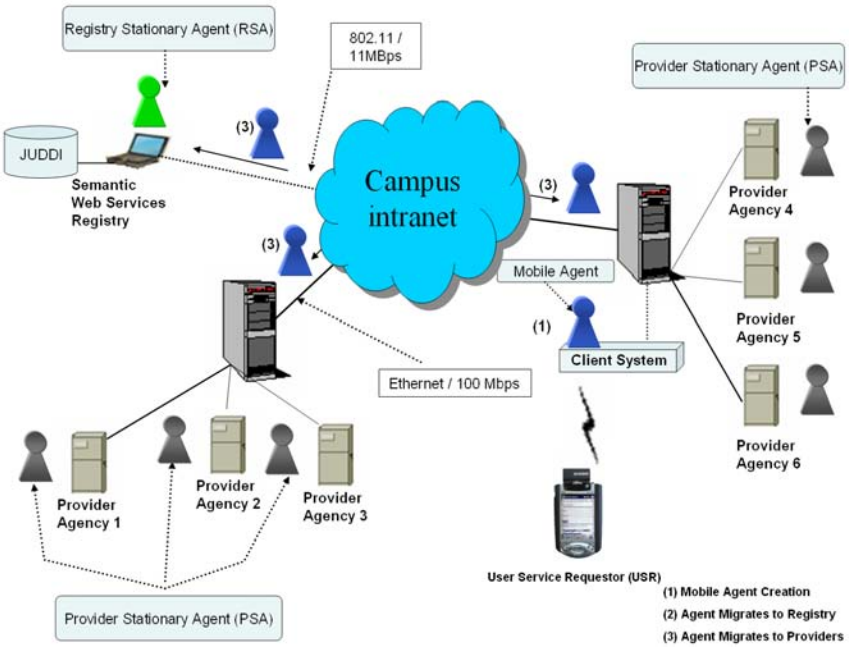
The SWS logic implemented in our experiments is as follows: the SWS have an extensive service description, stating unambiguously their capabilities in OWL-S. This description is published in the registry (SWSR). However, the SWS internal functionality is fairly simple, returning a pre-specified data volume subject to the service request. In our trials, these service results are 1

KB, 10 KB, 100KB and 1 MB. Moreover, six SWS have been implemented and distributed in the testing network.

In the performance evaluation scenario, a user requests a service, specifies his/her preferences and each of the above systems dispatches this request to the service registry. The service registry in the WSBM is a simple local UDDI providing a keyword service search on each service request, whereas in the SWS&MA system the registry is offering a service capability search to the placed service requests. In our evaluation, the description of SWS had small differences in the OWL-S descriptions. As a result, in the WSBM system, the service search to the UDDI registry had an average of three matches per service search/request. Contrary to WSBM system, in the SWS&MA system the MA had the necessary intelligence and knowledge to filter the results from the semantic registry and invoke only a SWS where its semantic description matched the service request and user's preferences. Consequently, in the WSBM system, we considered the average time this system requires to execute a service and we multiplied that by three (the average service results from the registry), whereas in SWS&MA system we consider the average time that is needed to invoke only a SWS. Moreover, in the SWS&MA system, the average time need was used from all the system variations to execute a SWS. These system variations are: (a) a system that uses MA cloning, (b) a system that uses PSA, and (c) a system that uses both MA cloning and PSA.

The testing platform we used is depicted in Figure 9. The system is a LAN that is composed of two workstations and a portable PC, all connected to the Internet through University's MAN.

Figure 9. Performance evaluation network topology



Below, we elaborate on the metrics that we adopted in order to assess the performance of the two systems. In Equation (1), total service time (TST_{MA}) (for the SWS&MA platform) is the sum of registry interaction time (*RIT*), migration of MA to a service provider time (MSPT) and the interaction time with this service provider (ITSP):

$$TST_{MA} = RIT + MSPT + ITSP \quad (1)$$

In the WSBM system, Equation (1) has the form:

$$TST_{\text{WSBM}} = RIT + \left\lfloor \frac{N}{2} \right\rfloor * \overline{ITSP} \quad (2)$$

where the $\overline{ITSP}$ is defined as:

$$\overline{ITSP} = N^{-1} * \sum_{i=1}^{N} ITSP_i \quad (3)$$

In (2) $ITSP_i$ is the time between service request submission and service results reception.

In Figure 10, the results of the proposed system performance evaluation and comparison against a system implemented using the Conventional WS Business Model are presented. More specifically, the average time needed to execute three services for the WSBM, is plotted against the time required to invoke only one SWS in the SWS&MA for each service result size (1 KB, 10 KB, 100KB and 1 MB). We observe that the TST in the SWS&MA system is approximately half the TST in the WSBM system, irrespective of the service results size. It should be noted that the RIT in our system is considerably greater than the WSBM system, and this explains that the TST of the SWS&MA is half and not the one third (or even smaller) of the TST of the WSBM system. The high RIT of the proposed SWS&MA framework is attributed to the specific semantic registry implementation and might be less if other semantic registry is used (e.g., OWLS-MX (OWLS-MX, 2007), and TUB OWLSM (OWLSM, 2007).

CONCLUSION

In this article, we presented a framework that provides wireless access to WS using MA to

Figure 10. Total service time (TST) vs. service result size

find and execute WS in the fixed segment. The WS are semantically enriched and are expressed in OWL-S. Furthermore, the proposed system adopts an enhanced WS registry enriched with semantic information that provides semantic matching between service requests submitted and the service description published to them. The advantages of the presented system are: (1) users may invoke a set of services with only one interaction with the fixed network (post the request and receive the results), (2) users do not have to be connected during service discovery and invocation; the results of such operations are downloaded to their mobile devices after their network session re-establishment, (3) service invocations are performed locally or according to the user's specified policies, and unnecessary information is not transmitted over the network leading to better resource utilization, (4) the framework ensures the delivery of the service results to the user, (5) the MA dynamic behaviour improves system robustness and fault tolerance, (6) new services, agents, users and service registries can be easily integrated to the framework, thus, providing an expandable, open system.

Future work includes the study of agent mobility for SWS dynamic invocation and composition that takes network events into account. Network events (e.g., node failures, overloading) occurring while the service invocation is underway, may force the MA to dynamically reschedule its itinerary. The MA will implement routing algorithms that generate itineraries by considering network information published in the WS description, network status and topology.

REFERENCES

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American,* 2001.

Buhler, P., et al. (2003). Adaptive workflow = Web services + agents. *Proceedings of the International Conference on Web Services (ICWS03).* Las Vegas, NV.

Buhler, P., & Vidal, J. (2004). Enacting BPEL4WS specified workflows with multi-agent systems. *Proceedings of the Workshop on Web Services and Agent-Based Engineering (WSABE04).* New York, NY.

Brambilla, M., Ceri, S., Passamani, M., & Riccio, A. (2004). Managing asynchronous Web services interactions. *Proceedings of the IEEE International Conference on Web Services (ICWS).*

Cheng, S., et al. (2002). A new framework for mobile Web services. *Proceedings of the Symposium on Applications and the Internet (SAINT'02w).* Nara City, Japan.

ebXML. (2007). Retrieved June 1, 2007 from http://www.ebxml.org

FIPA. (2007). *Foundation for the intelligent physical agents.* Retrieved June 1, 2007 from http://www.fipa.org.

Gibbins, N., Harris, S., & Shadbolt, N. (2004). Agent-based Semantic Web services. *Journal of Web Semantics, 1.*

Gruber, T. (1993). A translation approach to portable ontology specification. *Knowledge Acquisition, 5.*

Huang, Y., & Chung, J. (2003). A Web services-based framework for business integration solutions. *Electronic Commerce Research and Applications, 2*(1), 15-26.

Ishikawa, F., Tahara, Y., Yoshioka, N., & Honiden, S. (2004b). Behavior descriptions of mobile agents for Web services integration. *Proceedings of the IEEE International Conference on Web Services* (ICWS) (pp. 342-349). San-Diego, CA.

Ishikawa, F., Yoshioka, N., Tahara, Y., & Honiden, S. (2004). Mobile agent system for Web services integration in pervasive networks. *Proceedings of*

the International Workshop on Ubiquitous Computing (IWUC) (pp. 38-47). Porto, Portugal.

JADE. (2007). *Java agent development environment.* Retrieved June 1, 2007 from http://jade.tilab.com

jUDDI. (2007). *Open source Java implementation of the universal description, discovery, and integration (UDDI) specification for Web services.* Retrieved June 1, 2007 from http://ws.apache.org/juddi/

Kagal, L., et al. (2002). Agents making sense of the semantic Web. *Proceedings of the First International Workshop on Radical Agent Concepts, (WRAC).* McLean, VA.

Lange, D., & Oshima, M. (1998). *Programming and deploying Java mobile agents with aglets.* Addison-Wesley.

Li, K., Verma, K., Mulye, R., Rabbani, R., Miller, J., & Sheth, A. (2006). Designing semantic Web processes: The WSDL-S approach. In: J. Cardoso & A. Sheth (Eds.), *Semantic Web services, processes and applications.* Springer-Verlag.

McIlraith, S., & Martin, D. (2003). Bringing semantics to Web services. *IEEE Intelligent Systems*, *18*(1), 90-93.

Montanari, R., Tonti, G., & Stefanelli, C. (2003). A policy-based mobile agent infrastructure. *Proceedings of the 3rd IEEE International Symposium on Applications and the Internet Workshops (SAINT03) IEEE Computer Society Press.* Orlando, FL.

Montanari, R., Tonti, G., & Stefanelli. C. (2003). Policy-based separation of concerns for dynamic code mobility management. Proceedings of the 27th International Computer Software and Applications Conference, (COMPSAC'03). Dallas, TX: IEEE Computer Society Press.

OWL-S. (2007). *OWL Web ontology language for services (OWL-S).* Retrieved June 1, 2007 from http://www.w3.org/Submission/2004/07/

OWLSM. (2007). *The TUB OWL-S Matcher.* Retrieved June 1, 2007 from http://kbs.cs.tu-berlin.de/ivs/Projekte/owlsmatcher/index.html

OWLS-MX. (2007). *Hybrid OWL-S Web Service Matchmaker.* Retrieved June 1, 2007 from http://www.dfki.de/~klusch/owls-mx/

OWL-S/UDDI Matchmaker Web Interface. (2007). Retrieved June 1, 2007 http://www.daml.ri.cmu.edu/ matchmaker/

Paolucci, M., Kawamura, T., Payne, T., & Sycara, K. (2002). Semantic matching of Web services capabilities. *Proceedings of the International Semantic Web Conference (ISWC).* Sardinia, Italy.

Pour, G., & Laad, N. (2006). Enhancing the horizons of mobile computing with mobile agent components. *Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR'06)* (pp. 225-230).

RACER. (2007). *DL reasoner.* Retrieved June 1, 2007 from http://www.racer-systems.com

Raghavan, V., & Wong, S. (1986). A critical analysis of vector space model for information retrieval. *JASIS, 37*(5), 279-287.

Roman, D., Keller, U., Lausen, H., de Bruijn, J., Lara, R., Stollberg, M., et al. (2005). Web service modeling ontology. *Applied Ontology*, *1*(1), 77-106.

Srinivasan, N., Paolucci, M., & Sycara, K. (2004). Adding OWL-S to UDDI, implementation and throughput. *Proceedings of the First International Workshop on Semantic Web Services and Web*

Process Composition (SWSWPC). San Diego, CA.

SWSL Committee. (2007). *Semantic Web services framework (SWSF).* Retrieved June 1, 2007 from http://www.daml.org/services/swsf

Sycara, K., Paolucci, M., Ankolekar, A., & Srinivasan, N. (2004). Automated discovery, interaction and composition of semantic Web services. *Journal of Web Semantics, 1.*

Tsetsos, V., Anagnostopoulos, C., & Hadjiefthymiades, S. (2007). Semantic Web service discovery: Methods, algorithms and tools. In: J. Cardoso (Ed.), *Semantic Web services: Theory, tools and applications.* Hershey, PA: IGI Publishing.

Wooldridge, M. (2002). *An introduction to multiagent systems.* John Wiley & Sons.

Zahreddine, W., & Mahmoud, Q. (2005). An agent-based approach to composite mobile Web services. *Proceedings of the 19th IEEE International Conference on Advanced Information Networking and Applications (AINA05).* Taipei, Taiwan.

This work was previously published in the International Journal on Semantic Web & Information Systems, edited by A. Sheth, Volume 4, Issue 1, pp. 1-19, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.27
Advanced Resource Discovery Protocol for Semantic-Enabled M-Commerce

Michele Ruta
Politecnico di Bari, Italy

Tommaso Di Noia
Politecnico di Bari, Italy

Eugenio Di Sciascio
Politecnico di Bari, Italy

Francesco Maria Donini
Università della Tuscia, Italy

Giacomo Piscitelli
Politecnico di Bari, Italy

INTRODUCTION

New mobile architectures allow for stable networked links from almost everywhere, and more and more people make use of information resources for work and business purposes on mobile systems. Although technological improvements in the standardization processes proceed rapidly, many challenges, mostly aimed at the deployment of value-added services on mobile platforms, are still unsolved. In particular the evolution of wireless-enabled handheld devices and their capillary diffusion have increased the need for more sophisticated service discovery protocols (SDPs).

Here we present an approach, which improves Bluetooth SDP, to provide m-commerce resources to the users within a piconet, extending the basic service discovery with semantic capabilities. In particular we exploit and enhance the SDP in

order to identify generic resources rather than only services.

We have integrated a "semantic layer" within the application level of the standard Bluetooth stack in order to enable a simple interchange of semantically annotated information between a mobile client performing a query and a server exposing available resources.

We adopt a simple piconet configuration where a stable networked zone server, equipped with a Bluetooth interface, collects requests from mobile clients and hosts a semantic facilitator to match requests with available resources. Both requests and resources are expressed as semantically annotated descriptions, so that a *semantic distance* can be computed as part of the ranking function, to choose the most promising resources for a given request.

STATE OF THE ART

Usually, resource discovery protocols involve a requester, a lookup or directory server and finally a resource provider. Most common SDPs, as service location protocol (SLP), Jini, UPnP (Universal Plug aNd Play), Salutation or UDDI (universal description discovery and integration), include registration and lookup of resources as well as matching mechanisms (Barbeau, 2000).

All these systems generally work in a similar manner. Basically a client issues a query to a directory server or to a specific resource provider. The request may explicitly contain a resource name with one or more attributes. The lookup server—or directly the resource provider—attempts to match the query pattern with resource descriptions stored in its database, then it replies to the client with discovered resources identification and location (Liu, Zhang, Li, Zhu, & Zhang, 2002).

These discovery architectures are based on some common assumptions about network infrastructure under the application layer in the protocol stack. In particular, current SDPs usually require a continuous and robust network connectivity, which may not be the case in wireless contexts, and especially in the ad-hoc ones. In fact in such environments, network consistence varies continuously and temporary disconnections occur frequently, so bringing to a substantial decrease traditional SDP performances (Chakraborty, Perich, Avancha, & Joshi, 2001).

Actually, there are several issues that restrain the expansion of advanced wireless applications, among them, the variability of scenarios. An ad-hoc environment is based on short-range, low power technologies like Bluetooth (Bluetooth, 1999), which grant the peer-to-peer interaction among hosts. In such a mobile infrastructure there could be one or more devices providing and using resources but, as a MANET is a very unpredictable environment, a flexible resource search system is needed to overcome difficulties due to the host mobility. Furthermore, existing mobile resource discovery methods use simple string-matching, which is largely inefficient in advanced scenarios as the ones related to electronic commerce. In fact, in these cases there is the need to submit articulate requests to the system to obtain adequate responses (Chakraborty & Chen, 2000).

With specific reference to the SDP in the Bluetooth stack, it is based on a 128-bit universally unique identifier (UUID); each numeric ID is associated to a single service class. In other words, Bluetooth SDP is code-based and consequently it can handle only exact matches. Yet, if we want to search and retrieve resources whose description cannot be classified within a rigid schema (e.g., the description of goods in a shopping mall), a more powerful discovery architecture is needed (Avancha, Joshi, & Finin, 2002). SDP should be able to cope with non-exact matches (Chakraborty & Chen, 2000), and to provide a ranked list of discovered resources, computing a distance between each retrieved resource and the request after a matchmaking process.

To achieve these goals, we exploit both theoretical approach and technologies of semantic Web

vision and adapt them to small ad-hoc networks based on the Bluetooth technology (Ruta, Di Noia, Di Sciascio, Donini, & Piscitelli, 2005).

In a semantic-enabled Web—what is known as the semantic Web vision—each available resource should be annotated using RDF (RDF Primer, 2004), with respect to an OWL ontology (Antoniou & van Harmelen, 2003). There is a close relation between the OWL-DL subset of OWL and description logics (DLs) (Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2002) semantics, which allows the use of DLs-based reasoners in order to infer new information from the one available in the annotation itself.

In the rest of the article we will refer to DIG (Bechhofer, 2003) instead of OWL-DL because it is less verbose and more compact: a good characteristic in an ad-hoc scenario. DIG can be seen as a syntactic variant of OWL-DL.

THE PROPOSED APPROACH

In what follows we outline our framework and we sketch the rationale behind it. We adopt a mobile commerce context as reference scenario.

In our mobile environment, a user contacts via Bluetooth a zone resource provider (from now on *hotspot*) and submits her semantically annotated request in DIG formalism. We assume the zone server—which classifies resource contents by means of an OWL ontology—has previously identified shopping malls willing to promote their goods and it has already collected semantically annotated descriptions of goods. Each resource in the m-marketplace owns an URI and exposes its OWL description.

The *hotspot* is endowed with a *MatchMaker* [in our system we adapt the MAMAS-tng reasoner (Di Noia, Di Sciascio, Donini, & Mongiello, 2004)], which carries out the matchmaking process between each compatible offered resource and the requested one measuring a "semantic distance." The provided result is a list of discovered resources matching the user demand, ranked according to their degree of correspondence to the demand itself.

By integrating a semantic layer within the OSI Bluetooth stack at service discovery level, the management of both syntactic and semantic discovery of resources becomes possible. Hence, the Bluetooth standard is enriched by new functionalities, which allow to maintain a backward compatibility (handheld device connectivity), but also to add the support to matchmaking of semantically annotated resources. To implement matchmaking and ontology support features, we have introduced a *semantic service discovery* functionality into the stack, slightly modifying the existing Bluetooth discovery protocol.

Recall that SDP uses a simple request/response method for data exchange between SDP client and SDP server (Gryazin, 2002). We associated unused classes of 128-bit UUIDs in the original Bluetooth standard to mark each specific ontology and we call this identifier *OUUID* (*ontology universally unique identifier*). In this way, we can perform a preliminary exclusion of supply descriptions that do not refer to the same ontology of the request (Chakraborty, Perich, Avancha, & Joshi, 2001). With OUUID matching we do not identify a single service, but directly the context of resources we are looking for, which can be seen as a class of similar services. Each resource semantically annotated is stored within the *hotspot* as resource record. A 32-bit identifier is uniquely associated to a semantic resource record within the *hotspot*, which we call *SemanticResourceRecordHandle*. Each resource record contains general information about a single semantic enabled resource and it entirely consists of a list of resource attributes. In addition to the *OUUID* attribute, there are *ResourceName*, *ResourceDescription*, and a variable number of *ResourceUtilityAttr_i* attributes (in our current implementation 2 of them). *ResourceName* is a text string containing a human-readable name for the resource, the second one is a text string including the resource description expressed in

DIG formalism and the last ones are numeric values used according to specific applications. In general, they can be associated to context-aware attributes of a resource (Lee & Helal, 2003), as for example its price or the physical distance it has from the *hotspot* (expressed in metres or in terms of needed time to get to the resource). We use them as parameters of the overall *utility function* that computes matchmaking results.

In particular, to allow the representation and the identification of a semantic resource description we introduced in the data representation of the original Bluetooth standard two new *data element type descriptor*: OUUID and DIG text string. The first one is associated to the type descriptor value 9 whereas to the second one corresponds the type descriptor value 10 (both reserved in the original standard). We will associate 1, 2, 4 byte as valid size for the first one and 5, 6, 7 for the DIG text string.

Since the communication is referred to the peer layers of the protocol stack, each transaction is represented by one request Protocol Data Unit (PDU) and another PDU as response. If the SDP request needs more than a single PDU (this case is frequent enough if we use semantic service discovery) the SDP server generates a partial response and the SDP client waits for the next part of the complete answer.

By adding two SDP features *SDP_OntologySearch* (request and response) and *SDP_SemanticServiceSearch* (request and response) to the original standard (exploiting not used PDU ID) we inserted together with the original SDP capabilities further semantic-enabled resource search functions (see Table 1).

The transaction between service requester and *hotspot* starts after ad-hoc network creation. When a user becomes a member of a MANET, she is able to ask for a specific service/resource (by submitting a semantic-based description). The generic steps, up to response providing, for a service request are detailed in the following:

1. The user searches for a specific ontology identifier by submitting one or more

Table 1. List of PDU IDs with corresponding descriptions

PDU ID	**Description**
0x00	Reserved
0x01	SDP_ErrorResponse
0x02	SDP_ServiceSearchRequest
0x03	SDP_ServiceSearchResponse
0x04	SDP_ServiceAttributeRequest
0x05	SDP_ServiceAttributeResponse
0x06	SDP_ServiceSearchAttributeRequest
0x07	SDP_ServiceSearchAttributeResponse
0x08	SDP_OntologySearchRequest
0x09	SDP_OntologySearchResponse
0x0A	SDP_SemanticServiceSearchRequest
0x0B	SDP_SemanticServiceSearchResponse
0x0C-0xFF	Reserved

$OUUID_R$ she manages by means of her client application
2. The *hotspot* selects OUUIDs matching each $OUUID_R$ and replies to the client
3. The user sends a service request (*R*) to the *hotspot*
4. The *hotspot* extracts descriptions of each resource cached within the *hotspot* itself, which is classified with the previously selected $OUUID_R$
5. The *hotspot* performs the matchmaking process between *R* and selected resources it shares. Taking into account the matchmaking results, all the resources are ranked with respect to *R*
6. The *hotspot* replies to the user.

It is important to remark that basically all the previous steps are based on the original SDP in Bluetooth. No modifications are made to the original structure of transactions, but simply we differently use the SDP framework. In what follows we outline the structure of the SDP PDUs we added within the original framework to allow semantic resource discovery.

The first one is the *SDP_OntologySearchRequest* PDU. Their parameters are shown in Table 2.

The *OntologySearchPattern* is a data element sequence where each element in the sequence is a OUUID. The sequence must contain at least 1 and at most 12 OUUIDs, as in the original standard. The list of OUUIDs is an ontology search pattern. The *ContinuationState* parameter maintains the same purpose of the original Bluetooth (Bluetooth, 1999).

The *SDP_OntologySearchResponse* PDU is generated by the previous PDU. Their parameters are reported in Table 3.

The *TotalOntologyCount* is an integer containing the number of ontology identifiers matching the requested ontology pattern. Whereas the *OntologyRetrievedPattern* is a data element sequence where each element in the sequence is a OUUID matching at least one sent with the *OntologySearchPattern*. If no OUUID matches

Table 2. SDP_OntologySearchRequest PDU parameters

PDU ID	**parameters**
0x08	- ***OntologySearchPattern*** - *ContinuationState*

Table 3. SDP_OntologySearchResponse PDU parameters

PDU ID	**parameters**
0x09	- ***TotalOntologyCount*** - ***OntologyRetrievedPattern*** - *ContinuationState*

the pattern, the *TotalOntologyCount* is set to 0 and the *OntologyRetrievedPattern* contains only a specific OUUID able to allow the browsing by the client of all the OUUIDs managed by the *hotspot* (see the following *ontology browsing* mechanism for further details). Hence the pattern sequence contains at least 1 and at most 12 OUUIDs.

The *SDP_SemanticServiceSearchRequest* PDU follows previous PDU. Their parameters are shown in Table 4.

The *SemanticResourceDescription* is a data element text string in DIG formalism representing the resource we are searching for; *ContextAwareParam1* and *ContextAwareParam2* are data element unsigned integers. In our case study, which models an m-marketplace in an airport terminal, we use them respectively to indicate a reference price for the resource and the hour of the scheduled departure of the flight. Since a generic client interacting with a *hotspot* is in its range, using the above PDU parameter she can impose—among others—a proximity criterion in the resource discovery policy.

The *SDP_SemanticServiceSearchResponse* PDU is generated by the previous PDU. Their parameters are reported in Table 5.

The *SemanticResourceRecordHandleList* includes a list of resource record handles. Each of the handles in the list refers to a resource record potentially matching the request. Note that this list of service record handles does not contain header fields, but only the 32-bit record handles. Hence, it does not have the data element format. The list of handles is arranged according to the relevance order of resources, excluding resources not compatible with the request. The other param-

Table 4. SDP_SemanticServiceSearchRequest PDU parameters

PDU ID	**parameters**
0x0A	- ***SemanticResourceDescription*** - ***ContextAwareParam1*** - ***ContextAwareParam2*** - *MaximumResourceRecordCount* - *ContinuationState*

Table 5. SDP_SemanticServiceSearchResponse PDU parameters

PDU ID	**parameters**
0x0B	- *TotalResourceRecordCount* - *CurrentResourceRecordCount* - ***SemanticResourceRecordHandleList*** - *ContinuationState*

eters maintain the same purpose of the original Bluetooth (Bluetooth, 1999).

In all the previous cases, the error handling is managed with the same mechanisms and techniques of Bluetooth standard (Bluetooth, 1999).

Notice that each resource retrieval session starts after settling between client and server the same ontology identifier (OUUID).

Nevertheless if a client does not support any ontology or if the supported ontology is not managed by the *hotspot*, it is desirable to discover what kind of merchandise class (and then what OUUIDs) are handled by the zone server without any a priori information about resources. For this purpose we use the *service browsing* feature (Bluetooth, 1999) in a slightly different fashion with respect to the original Bluetooth standard, so calling this mechanism *ontology browsing.* It is based on an attribute shared by all semantic enabled resource classes, the *BrowseSemanticGroupList* attribute which contains a list of OUUIDs. Each of them represents the browse group a resource may be associated with for browsing.

Browse groups are organized in a hierarchical fashion, hence when a client desires to browse a *hotspot* merchandise class, she can create an *ontology search pattern* containing the OUUID that represents the *root browse semantic group.* All resources that may be browsed at the top level are made members of the *root browse semantic group* by having the root browse group OUUID as a value within the *BrowseSemanticGroupList* attribute.

Generally a *hotspot* supports relatively few merchandise classes, hence all of their resources will be placed in the root browse group. However, the resources exposed by a provider may be organised in a browse group hierarchy, by defining additional browse groups below the root browse group.

Having determined the goods category and the corresponding reference ontology, the client can also download a DIG version of it from the *hotspot* as *.jar* file [such a file extension—among other things—also allows a total compatibility with the Connected Limited Device Configuration (CLDC) technology].

Also notice that since the proposed approach is fully compliant with semantic Web technologies, the user exploits the same semantic enabled descriptions she may use in other Semantic Web compliant systems (e.g., in the Web site of a shopping mall). That is, there is no need for different customized resource descriptions and modelling, if the user employs different applications either on the Web or in mobile systems. The syntax and formal semantics of the descriptions is unique with respect to the reference ontology and can be shared among different environments.

In e-commerce scenarios, the match between demand and supply involves not only the description of the good but also data-oriented properties. It would be quite strange to have a commercial transaction without taking into account price, quantity, and availability, among others. The demander usually specifies how much she is willing to pay, how many items she wants to buy, and the delivery date. Hence, the overall match value depends not only on the distance between the (semantic-enabled) description of the demand and of the supply. It has to take into account the description distance with the difference of (the one asked by the demander and the other proposed by the seller), quantity, and delivery date. The overall utility function combines all these values to give a global value representing the match degree.

Also notice that, in m-commerce applications, in addition to "commercial" parameters also context-aware variables should influence matching results. For example, in our airport case study, we consider the price difference but also the physical distance between requester and seller to weigh the match degree. The distance becomes an interesting value since a user has a temporal deadline for shopping: the scheduled hour of her flight. Hence, a resource might be chosen also according to its proximity to the user.

We will express this distance in terms of time to elapse for reaching the shop where a resource is, leaving from the *hotspot* area. In such a manner the *hotspot* will exclude resources not reachable by the user while she is waiting for boarding and it will assign to resources unlikely reachable (farther) a weight smaller than one assigned to easily reachable ones.

The above approach can be further extended to other data-type properties.

The utility function we used depends on:

- p_D : price specified by the demander
- p_O : price specified by the supplier
- t_D : time interval available to the client
- t_O : time to reach the supplier and come back, leaving from the *hotspot* area
- *s_match*: score computed during the semantic matchmaking process, computed through *rankPotential* (Di Noia, Di Sciascio, Donini, & Mongiello, 2004) algorithm.

$$u(s\_match, p_D, p_O, t_D, t_O) =$$

$$\frac{s\_match}{2} + \frac{\tanh\frac{t_D - t_O}{\beta}}{3} + \frac{(1+\alpha)p_D - p_O}{6(1+\alpha)p_D} \quad (1)$$

Notice that p_D is weighted by a (1+α) factor. The idea behind this weight is that, usually, the demander is willing to pay up to some more than what she originally specified on condition that she finds the requested item, or something very similar. In the tests we carried out, we find α=0.1 and β=10 are values in accordance with user preferences. These values seem to be in some accordance with experience, but they could be changed according to different specific considerations.

RUNNING EXAMPLE

A simple example can clarify the rationale of our setting. Here we will present a case study analogous to the one presented in Avancha, Joshi, and Finin (2002), and we face it by means of our approach.

Let us suppose a user is in a duty free area of an airport, she is waiting for her flight to come back home and she is equipped with a wireless-enabled PDA. She forgot to buy a present for her beloved little nephew and now she wants to purchase it from one of the airport gift stores.

In particular she is searching for a learning toy strictly suitable for a kid (she dislikes a child toy or a baby toy) and possibly the toy should not have any electric power supply.

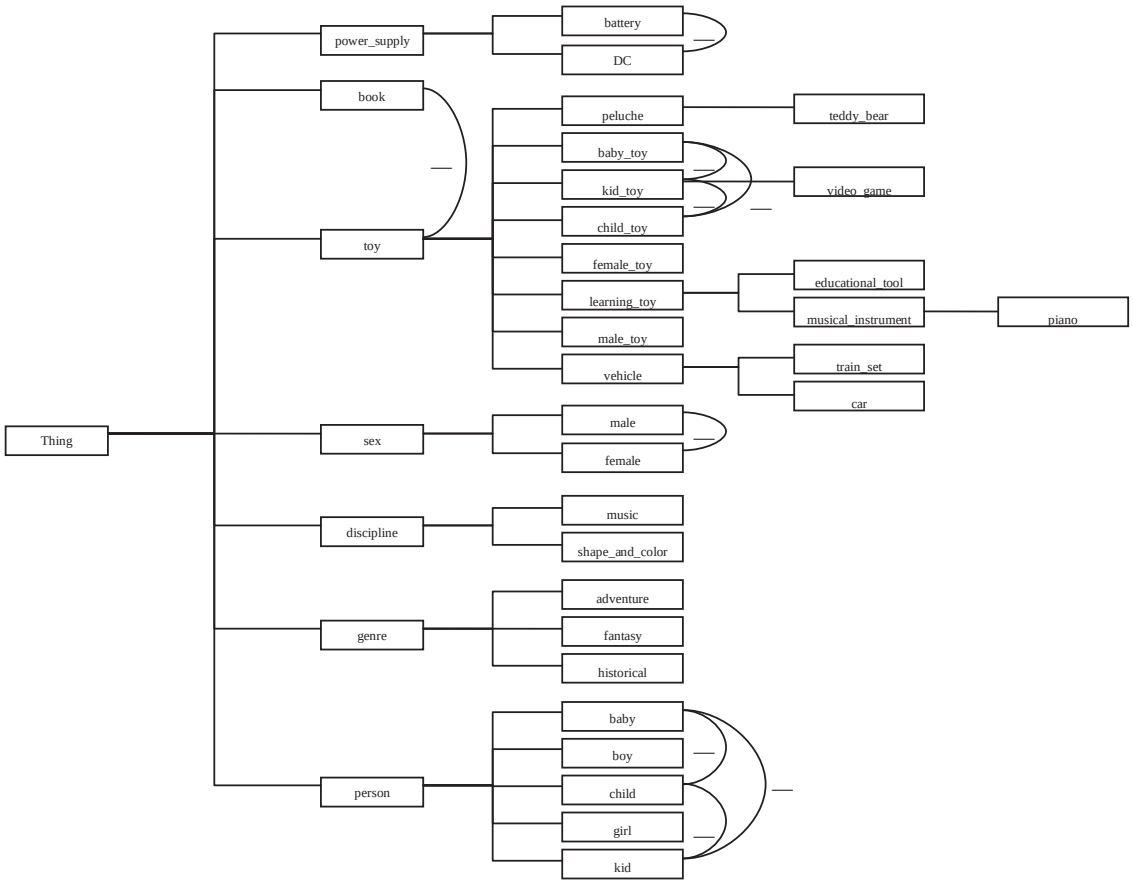
Clearly this request is too complex to be expressed by means of standard UUID Bluetooth SDP mechanism. In addition, non-exact matches between resource request and offered ones is highly probable and the on/off matching system provided by the original standard in this case could be largely inefficient.

Hence both the semantic resource request and offered ones can be expressed in a DIG statement exploiting DL semantics and encapsulated in an SDP PDU.

The *hotspot* equipped with MAMAS reasoner collects the request and initially selects supplies expressed by means of the same ontology shared with the requester. Hence a primary selection of suitable resources is performed. In addition, the matchmaker carries out the matchmaking process between each offered resource in the m-marketplace and the requested one measuring a "semantic distance" (Colucci, Di Noia, Di Sciascio, Donini, & Mongiello, 2005). Finally the matchmaking results are ranked and returned to the user.

A subset of the ontology used as a reference in the examples is reported in Figure 1. For the sake of simplicity, only the class hierarchy and disjoint relations are represented.

Figure 1. The simple toy store ontology used as reference in the example



Let us suppose that after the *hotspot* selects supplies, its knowledge base is populated with the following individuals whose description is represented using DL formalism:

- *Alice_in_wonderland*. Price 20$. 5 min from the *hotspot*:
  book ⊓ ∀has_genre.fantasy
- *Barbie_car*. Price 80$. 10 min from the *hotspot*:
  car ⊓ ∀suggested_for.girl ⊓ ∀has_power_supply. battery
- *classic_guitar*. Price 90$. 17 min from the *hotspot*:
  musical_instrument ⊓ ∀suitable_for.kid ⊓ (≤ 0 has_power_supply)
- *shape_order*. Price 40$. 15 min from the *hotspot*:
  educational_tool ⊓ ∀suitable_for.child ⊓ ∀stimulates_to_learn. shape_and_color

- *Playstation*. Price 160$. 28 min from the *hotspot*:
  video_game ⊓ ∀has_power_supply.DC
- *Winnie_the_pooh*. Price 30$. 15 min from the *hotspot*:
  teddy_bear ⊓ ∀suitable_for.baby

On the other hand, the request ***D*** submitted to the system by the user can be formalized in DL syntax as follows:

learning_toy ⊓ ∀suggested_for.boy ⊓ ∀suitable_for.kid ⊓ (≤ 0 has_power_supply)

In addition she imposes a reference price of 200$ (p_D=200) as well as the scheduled departure time as within 30 minutes (t_D=30).

In Table 6 matchmaking results are presented. The second column shows whether each retrieved resource is compatible or not with request *D* and, in case, the *rankPotential* computed result. In the fourth column, matchmaking results are also expressed in a relative form between 0 and 1 to allow a more immediate semantic comparison among requests and different resources and to put in a direct correspondence various rank values.

Finally in the last column results of the overall utility function application are shown.

Notice that the semantic distance of the individual *classic_guitar* from *D* is the smaller one; then the system will recommend to the user this resource first. Hence the ranked list returned by the *hotspot* is a strict indication for the user about best available resources in the airport duty free piconet in order of relevance with respect to the request. Nevertheless a user can choose or not a resource according to her personal preferences and her initial purposes.

After having selected the best resource, the server of the chosen virtual shop will receive a connection request from the user PDA with its connection parameters and in this manner the transaction may start. The user can provide her credit card credentials, so that when she reaches the store, her gift will be already packed. This final part of the application is not yet implemented, but it is trivially achievable exploiting the above SDP infrastructure.

CONCLUSION AND FUTURE WORK

In this article we have presented an advanced semantic-enabled resource discovery protocol for m-commerce applications. The proposed approach aims to completely recycle the basic functionalities of the original Bluetooth service discovery

Table 6. Matchmaking results

demand – supply	**compatibility (y/n)**	**score**	**s_match**	**u(·)**
D - Alice_in_wonderland	n	-	-	-
D - Barbie_car	y	7	0.364	0.609
D - classic_guitar	y	3	0.727	0.748
D - shape_order	n	-	-	-
D - Playstation	y	5	0.546	0.378
D - Winnie_the_pooh	n	-	-	-

protocol by simply adding semantic capabilities to the classic SDP ones and without introducing any change in the regular communication work of the standard. A matchmaking algorithm is used to measure the semantic similarity among demand and resource descriptions.

Future trends of the proposed framework aim to create a more advanced DSS to help a user in a generic m-marketplace. Under investigation is the support to creation of P2P small communities of mobile hosts where goods and resources are advertised and opinions about shopping are exchanged (Avancha, D'Souza, Perich, Joshi, & Yesha, 2003). If a user decides to "open" her shopping trolley sharing information she owns (purchased goods, discounts, opinion about specific vendors or products) the system will insert her in a buyer mobile community where she can exchange information with other users.

Another future activity focuses on strict control of the good advertising. In an m-marketplace, the system will send to various potential buyers best proposals about their interests.

We intend to implement a mechanism to advertise goods or services in a more direct and personalized fashion. From this point of view, an additional feature of the system is oriented to the user profiling extraction and management (Prestes, Carvalho, Paes, Lucena, & Endler, 2004; Ruta, Di Noia, Di Sciascio, Donini, & Piscitelli, 2005; von Hessling, Kleemann, & Sinner, 2004). Without imposing any explicit profile submission to the user, the system could collect her preferences by means of previously submitted requests (Ruta, Di Noia, Di Sciascio, Donini, & Piscitelli, 2005); that is, by means of the "history" of the user in the m-marketplace.

REFERENCES

Antoniou, G., & van Harmelen, F. (2003). Web ontology language: OWL. In *Handbook on Ontologies in Information Systems.*

Avancha, S., D'Souza, P., Perich, F., Joshi, A., & Yesha, Y. (2003). P2P m-commerce in pervasive environments. *ACM SIGecom Exchanges*, *3*(4), 1-9.

Avancha, S., Joshi, A., & Finin, T. (2002). Enhanced service discovery in Bluetooth. *IEEE Computer*, *35*(6), 96-99.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., & Patel-Schneider, P. (2002). *The description logic handbook*. Cambridge: Cambridge University Press.

Barbeau, M. (2000). Service discovery protocols for ad hoc networking. *Workshop on Ad-hoc Communications (CASCON '00).*

Bechhofer, S. (2003). *The DIG description logic interface: DIG/1.1.* Retrieved from http://dlweb.man.ac.uk/dig/2003/02/interface.pdf.

Bluetooth specification document. (1999). Retrieved from http://www.bluetooth.com.

Chakraborty, D., & Chen, H. (2000). Service discovery in the future for mobile commerce. *ACM Crossroads*, *7*(2), 18-24.

Chakraborty, D., Perich, F., Avancha, S., & Joshi, A. (2001). Dreggie: Semantic service discovery for m-commerce applications. In *Workshop on Reliable and Secure Applications in Mobile Environment.*

Colucci, S., Di Noia, T., Di Sciascio, E., Donini, F. M., & Mongiello, M. (2005). Concept abduction and contraction for semantic-based discovery of matches and negotiation spaces in an e-marketplace. *Electronic Commerce Research and Applications*, *4*(4), 345-361.

Di Noia, T., Di Sciascio, E., Donini, F. M., & Mongiello, M. (2004). A system for principled matchmaking in an electronic marketplace. *International Journal of Electronic Commerce*, *8*(4), 9-37.

Gryazin, E. (2002). *Service discovery in Bluetooth*. Retrieved from http://www.hpl.hp.com/techreports/2002/HPL-2002-233.pdf.

Lee, C., & Helal, S. (2003). Context attributes: An approach to enable context awareness for service discovery. In *Symposium on Applications and the Internet (SAINT '03)* (pp. 22-30).

Liu, J., Zhang, Q., Li, B., Zhu, W., & Zhang, J. (2002). A unified framework for resource discovery and QoS-aware provider selection in ad hoc networks. *ACM Mobile Computing and Communications Review, 6*(1), 13-21.

Prestes, R., Carvalho, G., Paes, R., Lucena, C., & Endler, M. (2004). Applying ontologies in open mobile systems. In *Workshop on Building Software for Pervasive Computing OOPSLA'04.*

RDF Primer-W3C Recommendation. (2004, February 10). Retrieved from http://www.w3.org/TR/rdf-primer/

Ruta, M., Di Noia, T., Di Sciascio, E., Donini, F.M., & Piscitelli, G. (2005). Semantic based collaborative P2P in ubiquitous computing. In *IEEE/WIC/ACM International Conference Web Intelligence 2005 (WI '05)* (pp. 143-149).

von Hessling, A., Kleemann, T., & Sinner, A. (2004). Semantic user profiles and their applications in a mobile environment. In *Artificial Intelligence in Mobile Systems 2004.*

KEY TERMS

Description Logics (DLs): A family of logic formalisms for knowledge representation. Basic syntax elements are concept names, role names, and individuals. Intuitively, concepts stand for sets of objects, and roles link objects in different concepts. Individuals are used for special named elements belonging to concepts. Basic elements can be combined using constructors to form concept and role expressions, and each DL has its own distinct set of constructors. DL-based systems are equipped with reasoning services: logical problems whose solution can make explicit knowledge that was implicit in the assertions.

M-Marketplace: Virtual environment where demands and supplies (submitted or offered by users equipped with mobile devices) encounter each other.

Ontology: An explicit and formal description referred to concepts of a specific domain (classes) and to relationships among them (roles or properties).

Piconet: Bluetooth-based short-range wireless personal area network. A Bluetooth piconet can host up to eight mobile devices. More piconets form a *scatternet.*

Service Discovery Protocol (SDP): It identifies the application layer of an OSI protocol stack and manages the automatic detection of devices with joined services.

Semantically Annotated Resource: any kind of good, tangible or intangible (e.g., a document, an image, a product or a service) endowed of a description that refers to a shared ontology.

Semantic Matchmaking: The process of searching the space of possible matches between a request and several resources to find those best matching the request, according to given semantic criteria. It assumes that both the request and the resources are annotated according to a shared ontology.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 43-50,

Chapter 7.28
Improving Mobile Web Navigation Using N-Grams Prediction Models

Yongjian Fu
Cleveland State University, USA

Hironmoy Paul
Cleveland State University, USA

Namita Shetty
Cleveland State University, USA

ABSTRACT

In this article, we propose to use N-gram models for improving Web navigation for mobile users. N-gram models are built from Web server logs to learn navigation patterns of mobile users. They are used as prediction models in an existing algorithm which improves mobile Web navigation by recommending shortcuts. Our experiments on two real data sets show that N-gram models are as effective as other more complex models in improving mobile Web navigation.

INTRODUCTION

Wireless users of the Web grow rapidly as more and more mobile devices such as PDAs, mobile phones and pagers are now equipped with browsing capabilities. Many current Web sites are optimized for desktop, broadband clients, and deliver content poorly for mobile devices due to display size and bandwidth. Moreover, the associated cost will prohibit maintaining two versions of a site, one for wired users and the other for mobile users. A viable solution is adaptive Web sites (Perkowitz & Etzioni, 1997). An adaptive Web site dynamically changes its contents or structure based on browsing activities.

Following the idea of adaptive Web sites, Anderson, Domingos, and Weld (2001) proposed shortcuts to improve mobile Web navigation. A shortcut is a dynamic link that provides a shorter path with fewer clicks for users to reach their desired pages. A shortcut to a destination page is dynamically created and inserted into the next page a user is going to browse. If that destination page is the one in which the user is interested, the user can access the page by following the shortcut. For example, assume a browsing session consists of A-B-C-D-E-F-G, where each letter represents a page. After browsing pages A and B, if a shortcut to G is created and inserted into page C, the user can follow the shortcut to reach G, without going through intermediate pages D, E, and F. The critical question is how to find shortcuts that are useful with only part of the session known. A shortcut C → H, for example, is useless in the previous example.

In order to provide useful shortcuts, Web usage mining techniques are employed. User browsing patterns are extracted from Web server logs. These patterns are built into prediction models that can be used to predict user browsing behaviors. Given a partial session, such prediction models will compute what other pages in which the user may be interested. These predictions are used to create and recommend shortcuts for mobile users.

A critical component in this approach is the prediction model. The model should be as accurate as possible with as little information about the session as possible. An accurate shortcut found earlier in a session is more worthwhile than one found close to the end of a session. Moreover, the prediction model should be easy to build and use. In their MINPATH algorithm, Anderson et al. (2001) used Markov models, which proved to be accurate (Anderson et al., 2001). However, those models require Web graphs. In this article, we propose to use a simpler prediction model, N-gram, for learning user browsing patterns.

N-grams are well known and are widely used in speech and text processing applications. Researchers have found that accuracy increases with N, the order of N-grams. For example, 4-grams are more accurate than 3-grams, which is turn is more accurate than 2-grams. Though accuracy increases with higher values of N, it requires a larger number of training sessions to have a well trained N-gram model.

An N-gram based prediction model for Web browsing patterns is proposed by Su et al. (2000). The N-gram model has several advantages over other prediction models. It is simple, robust, and easy to use. Besides, N-gram does not use a Web graph. In our study, the same N-gram model with a slightly different lookup operation is used. Moreover, its effectiveness in improving mobile Web navigation is examined.

In our approach, first, Web server logs are preprocessed to identify sessions. A session is conceptually a single visit. The sessions are then used to train an N-gram model. A revised version of MINPATH algorithm, MINCOST, is proposed to find shortcuts. MINCOST uses a different function in calculating the saving and ranking of shortcuts. Our approach has been implemented and evaluated against two real data sets from NASA and EPA Web servers. Our experiments show that the N-gram prediction model is as effective as more sophisticated models in recommending useful shortcuts.

The article is organized as follows. The second section discusses related work in Web usage mining, adaptive Web sites, and MINPATH algorithm. Our approach is presented in Section 3. Experimental results with two data sets are reported in the fourth section The fifth section concludes the article and gives some future research direction.

RELATED WORK

We briefly discuss Web usage mining techniques and its applications in adaptive Web sites and mobile Web navigation.

Web Usage Mining

Web usage mining refers to the mining of Web server logs to find interesting patterns in Web usage. Web server logs are preprocessed to find sessions. Conceptually, a session is a single visit to a Web site by a user. A session is represented by the pages browsed in that visit. From sessions, many Web usage patterns can be extracted, including associations, frequent paths, and clusters.

Association rules represent correlations among objects, which were first proposed to capture correlations among items in transactional data. If a session is viewed as a transaction, association rule mining algorithms can be employed to find associative relations among pages browsed (Mobasher, Cooley, & Srivastava, 1999a; Yang, Zhang, & Li, 2001). Using the same algorithms, we may find frequent paths traversed by many users (Frias-Martines & Karamcheti, 2002).

If each Web page represents a dimension, a session can be represented as a vector in the page space. Sessions can be clustered based on their similarity in the page space. In other words, sessions containing similar pages will be grouped into clusters (Fu, Sandhu, & Shih, 1999; Shahabi, Zarkesh, Adibi, & Shah, 1997).

Adaptive Web Sites

Perkowitz and Etzioni (1997) proposed adaptive Web site as a solution to the problem of Web navigation. An adaptive Web site is a Web site that automatically or semi-automatically adapts it structure based on user browsing. They proposed creating dynamic links using Web usage mining techniques. Koutri, Daskalaki, & Avouris, (2002) gives an overview of techniques for adaptive Web sites.

Anderson et al. (2001) argued that an adaptive Web site is especially interesting for mobile Web navigation. Due to limited display size, computing and storage capability, and network bandwidth, Web sites developed mainly for desktops deliver content poorly to mobile devices. To better serve the needs of mobile Web users, they proposed building Web site "personalizers" that observe the behavior of Web visitors and automatically customize and adapt Web sites for each mobile visitor. The MINPATH algorithm as described in Section 2.3 epitomizes their approach.

The MINPATH algorithm tries to improve the mobile Web user browsing experience by suggesting useful shortcuts. MINPATH finds shortcuts by using a learned model of user behavior to estimate the savings provided by shortcuts. The shortcuts are dynamically inserted into the page that the user will browse next. For example, after a user browsed pages A-B-C, MINPATH may provide a shortcut D->K in the next page D. It uses a prediction model that learns the user browsing behaviors to find the best shortcuts. If the user follows this shortcut, the user session becomes A-B-C-D-K. Assuming that without shortcut, the user session would contain pages A-B-C-D-E-F-G-H-I-J-K, the shortcut results in a saving of six pages or links.

MINPATH Algorithm

The MINCOST algorithm works as follows Anderson et al. (2001). Given a sequence of pages, called prefix, it uses prediction models to find the possible next pages and their probabilities. These pages and their savings are added to a shortcut list. For each page found by the prediction model, it is appended to the sequence, forming a suffix, which is used to find more shortcuts following the same procedure. This process continues recursively until the length of the suffix exceeds the depth bound or the probability of the predicted page becomes less than the probability threshold. Once the recursive part is over, the shortcuts found are sorted in descending order based on savings and the best shortcuts are returned. The number of shortcuts to be returned are user specified.

To estimate savings of a shortcut, MINPATH counts the number of links saved by following the

shortcut. For example, if a prefix is A-B-C, and a suffix is A-B-C-D-E-F-G, the expected saving for a shortcut D-> G is two.

There are two threshold parameters, depth bound and probability threshold; used in the MINPATH algorithm to limit searches for shortcuts. Depth bound represents the maximum length of the suffix. Probability threshold represents the minimum value of page probability.

MINPATH algorithm uses Markov models (Deshpande & Karypis, 2000; Sarukkai, 2000). Though accurate these models are complex and require Web graphs as a part of their implementation. There is a need to find less complex prediction models with comparable accuracy. Besides, MINPATH does not consider page size in estimating saving.

THE PROPOSED APPROACH

There are three steps in our approach to improve navigation for mobile Web users. First, Web server logs are preprocessed to extract sessions. Second, an N-gram prediction model is built from these sessions. Third, an algorithm that extends MINPATH, called MINCOST, recommends shortcuts based on the N-gram model and the current browsing sequence.

Server Log Preprocessing

A record in a server log file contains raw browsing data, such as the IP address of the user, date and time of the request, URL of the page, the return code of the server, and the size of the page, if the request is successful. Since such records are in chronological order, they do not provide much meaningful information about user browsing. The Web server log files are transformed into a set of sessions. A session represents a single visit of a user. The following procedure is used to transform a server log file into sessions (Mobasher, Cooley, & Srivastava, 1999b).

1. Records about image files (.gif, .jpg, etc.) and unsuccessful requests (return code belonging to the 4XX series) are filtered out.
2. Requests from the same IP address are grouped into a session. A timeout of *max-idle* is used to decide the end of a session, i.e., if the same IP address does not occur within a period of *max-idle* seconds, the current session is closed. Subsequent requests from the same IP address will be treated as a new session.

In our experiments, we used a value of 1800 seconds for *max-idle*, which is very common in Web usage mining studies.

Prediction Model Learning

Once we are done with preprocessing of Web logs, the next step is to build prediction models to predict the navigation patterns of Web users. If $P_1, P_2, P_3, \ldots, P_i$ are the pages browsed by the user so far, the prediction models will try to predict the next page, P_{i+1}.

We use an N-gram based prediction model. An N-gram is a substring of N characters, each character from an alphabet. The order of an N-gram is defined as N, the number of characters in the N-gram. In the context of this work, the alphabet is the set of URLs of Web pages on the Web server. An N-gram is a sequence of URLs.

After Web server logs are preprocessed into a number of sessions, an N-gram prediction model can be built as follows:

1. Each session is decomposed into a set of overlapping, subsequent paths of length N.
2. These paths are entered into an N-gram table T as N-grams.
3. For each path in T, the next pages right after it and their occurrences, in all the sessions, are recorded.

4. The probabilities of the next pages for all paths are calculated from the occurrences of all possible next pages.

For example, given a log file consisting of the following sessions, a 3-gram model for prediction can be built as follows:

A, B, C, D, H
B, C, D, G

The first session is decomposed into two 3-grams: "A, B, C" and "B, C, D." The second session is decomposed into one 3-gram: "B, C, D." For 3-gram "A, B, C," the next page would

Table 1. A 3-gram prediction table

3-gram	Predicted next page	Probability of next page
A, B, C	D	1.0000
B, C, D	G	0.5000
B, C, D	H	0.5000

Figure 1. Algorithm for constructing N-gram models

```
Input:
        L: sessions from Web server logs.
        N: order of N-gram
Output:
        T: N-gram prediction table
        // for an N-gram P and a predicted next page C, cell T[P, C] stores the probability.
Procedure:
        Begin
        For i = 1 to |L| do    // for every session
                S = L[i];       // the i-th session
                For j =1 to |S| do        // |S| represents the number of pages in session S
                        If (|S| - j)> N  // Find a sub-string of length N starting at j
                                P = sub-string (S, j, N);    // the j-th N-gram
                                                //sub-string returns N consecutive pages in S,
                                                // starting from j-th page,
                                C = sub-string(S, j+N, 1); // Find the next page
                                T [P, C] = T [P, C] + 1;    // increment count of (N-gram, next page) pair
                        End If
                End For
        End For
        For each [P, C] in T

                T [P, C] = T [P, C]  /  Σ_C(T[P, C]);  // convert count into probability

        End For
        Return T;
        End
```

be D, while for 3-gram "B, C, D," the next page may be G or H. The complete N-gram table, T, is given in Table 1.

The N-gram table T is used as our prediction model. Given an observed path, it is matched against N-grams in T. The predicted next pages and their corresponding probabilities of the matching entries in T are returned for shortcut recommendations.

The algorithm for constructing an N-gram model is given in Figure 1. It is modified from the algorithm from Su et al., (2000) such that it stores all possible next pages for an N-gram, rather than the most probable one only.

This algorithm has a time complexity of O(|L| * |S|) where |L| is the number of sessions and |S| is the length of sessions. The time of the algorithm is dominated by the first for loop. Its outer loop runs |L| times and its inner loop runs |S| times, which gives the time complexity of O(|L| * |S|).

An N-gram based prediction model is proposed in Su et al. (2000), which evaluates the model's accuracy without a specific application. In our study, the N-gram model is evaluated on its effectiveness in improving mobile Web navigation. To suit our application, the model is modified so that a lookup operation for an N-gram will return all predictions, instead of just the most probable one.

Shortcut Recommendation

The MINPATH algorithm ranks shortcuts based on their expected savings. In computing expected savings, MINPATH considers only the number of links saved. We modified MINPATH to reflect expected saving in total cost.

Given a prefix $< P_0, P_1, P_2,....,P_i >$, of a session $< P_0, P_1, P_2,....,P_n >$, a shortcut to P_n , i.e. , P_{i+1}-> P_n, can be added to page P_{i+1}. In MINPATH, the expected saving of the shortcut is calculated as the product of the number of links saved by following that particular shortcut and the probability that P_n is the destination page. Instead, we compute the expected saving as a cost function of page size and page probability.

The cost of browsing a page P_k, cost(P_k), is composed of the times to download the page, to view the page, and to click the link for next page. In other words, if a page is skipped by following a shortcut, the times to download, view, and click the page, are saved. The time to view and click is constant, but the time to download depends on page size. A parameter, Δ, is introduced to represent the download cost, which can be thought as time for transmission a unit of data. By introducing Δ, download time is separated from view and click time. The value of Δ is determined by network bandwidth and congestion. Since the cost function is relative, i.e., only the ratio of download time to view and click time matters, the view and click time is normalized to 1 in the cost function and Δ is adjusted accordingly.

$$\text{Cost}(P_k) = \text{download time} + \text{view and click time} = \text{Size}(P_k)*\Delta+1 \quad (1)$$

where $\text{Size}(P_k)$ is the size of page P_k in kilobytes.

The saving of a shortcut is the sum of the costs of pages skipped.

$$\text{Expected savings } (P_{i+1} \text{-> } P_n) = P_s * \{\sum_{k=i+2}^{N-1} \text{Cost}(P_k)\} =$$

$$P_s * \{\sum_{k=i+2}^{N-1} [\text{Size}(Pk)*\Delta+1\ ]\} \quad (2)$$

where P_s is the probability that page P_n is the destination page.

In our experiments, we found Δ has little effect on results, because pages have similar sizes. Its value is fixed at 0.5 assuming an effective bandwidth of 1 kbps and view and click time of 2 seconds. When pages have quite different sizes, Δ may have an impact on results.

To differentiate, we call the modified algorithm MINCOST since it takes into consideration cost for downloading. MINPATH considers only savings in view and click time, while MINCOST considers both view and click time and download time. It is easy to see that MINCOST is a generalization of MINPATH. When Δ is set to 0, the cost function in (2) degrades into the number of links saved which is used in MINPATH.

PERFORMANCE EVALUATION

The MINCOST algorithm is implemented in the C programming language and experiments are performed to evaluate the efficiency of our approach to improve mobile Web navigation with real datasets. The experiments are run on a PC with a 2.66 GHz Intel Pentium 4 processor, a memory of 512 MB, and running Windows XP professional.

Datasets

Two datasets are used in our experiments, the NASA dataset and the EPA dataset. The NASA dataset was collected from July 1, 1995 through July 31, 1995 for a total of one month's requests from the NASA server at Kennedy Space Center. The EPA dataset was collected from 23:53:25 August 29, 1995 through 23:53:07 August 30, 1995 for a total of 24 hours of requests from the EPA server at Research Triangle Park, NC.

The NASA and EPA datasets are converted into sessions as described in the third section. Table 2 gives a summary of the datasets.

Performance Measurements

The efficiency of the MINCOST algorithm is evaluated using average cost saved and percentage of average cost saved. The total cost of pages in the initial sessions and the total cost of pages in the final sessions after MINCOST are calculated using the definition in Equation 1. The difference between these two gives the total cost saved. The total cost saved is averaged over all sessions giving the average cost saved. The percentage of average cost saved is the average cost saved as a percentage of the average cost.

Total Cost Saved = Total Cost without MINCOST - Total Cost with MINCOST

Average Cost = Total Cost without MINCOST / Total Sessions

Table 2. NASA and EPA datasets summary

	NASA DATASET	EPA DATASET
Total Log Records	3,461,612	47,748
Total Sessions	132539	2074
Unique URLs	768	1821
Average Session Length (Number of Pages in a Session)	3.134	4.222
URL For Download	http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html	http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html

Average Cost Saved = Total Cost Saved/Total Sessions

Average Cost Saved (%) = (Average Cost Saved / Average Cost) * 100

Experimental Parameters

The average cost saved and the percentage of average cost saved are measured with respect to probability threshold, depth bound, number of shortcuts, and the order of the N-gram. In each experiment, we vary one parameter while keeping others to their default values. The results are reported from a 10 fold cross-validation. The entire dataset is divided into ten equal portions. Each portion is used as the testing set for MINCOST, while the remaining portions are used as training set for building the N-gram model as described in the third section. The results are averaged over these ten runs.

The parameters and their default values are given in Table 3.

The user behavior when provided with shortcuts is simulated by making two assumptions. First, it is assumed that when presented with one or more shortcuts that lead to destinations along the user's session, the user will select the shortcut that leads farthest along the session. Second, when no shortcuts lead to pages in the user's session, the user will follow the next link in the session.

Experimental Results from NASA Dataset

The results from experiments on NASA data set are presented in this section. Similar results are obtained from experiments on EPA data set, and are thus omitted.

Figures 2 and 3 show the average cost saved and percentage of average cost saved with respect to probability threshold, respectively. It is observed that both measures increase with decrease in probability threshold and stabilize around 0.0080. This is because there are not many shortcuts with significant savings after this value.

From Figures 2 and 3, it is obvious that the two performance measures, average cost saved and percentage of average cost saved, react similarly to changes in probability threshold. It is not a coincidence. Other experiments also reveal the

Table 3. Parameters and their default values

PARAMETER	DEFAULT VALUE	DESCRIPTION
Probability Threshold	0.006	Minimum probability of a shortcut
Depth Bound	5	Maximum length of suffix
Number of Shortcuts	3	Number of top shortcuts recommended
N	2	Order of N-gram model

Figure 2. Average cost saved vs. probability threshold

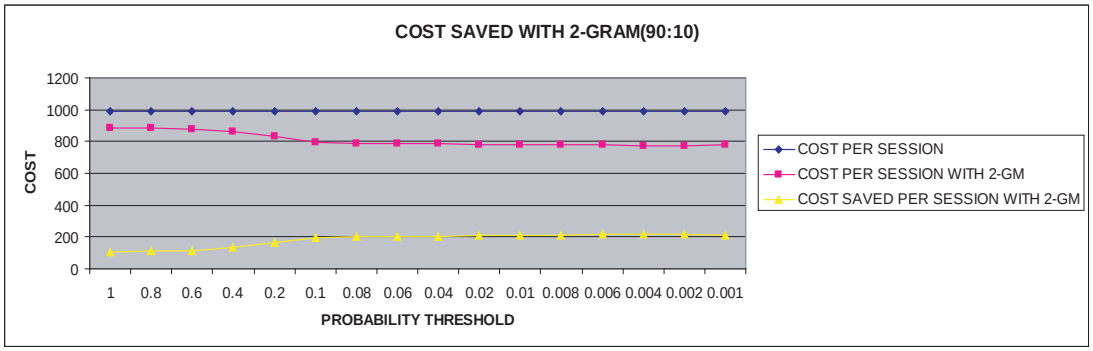
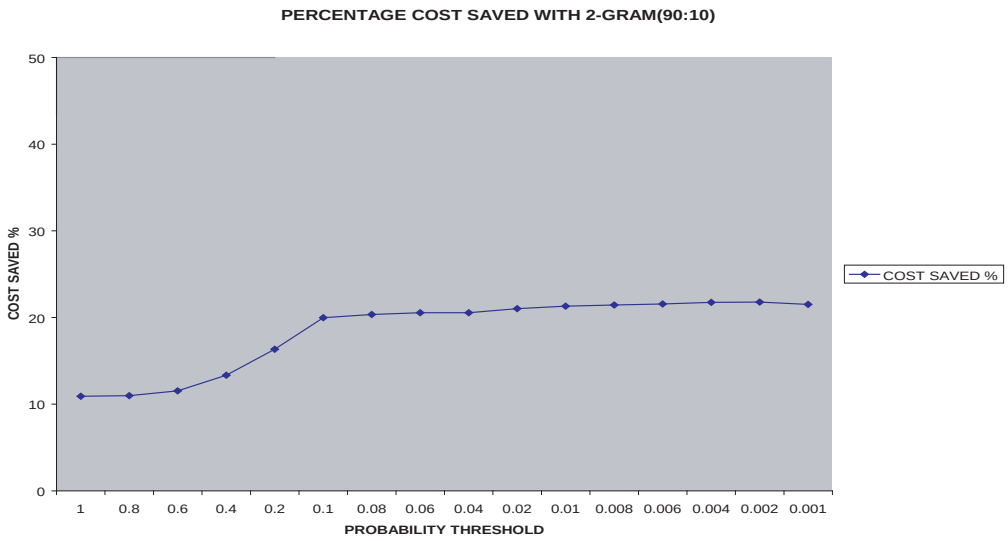


Figure 3. Percentage of average cost saved vs. probability threshold



high correlation of these two measures. In the rest of this section, only the results for percentage of average cost saved are presented.

As shown in Figure 4, the percentage of average cost saved is not affected by depth bound except when it increases from 1 to 2. Just like MINPATH, MINCOST is more sensitive to probability threshold than depth bound. Depth bound has a larger impact when the probability threshold is large. However, for most reasonable probability

Figure 4. Percentage of average cost saved vs. depth bound

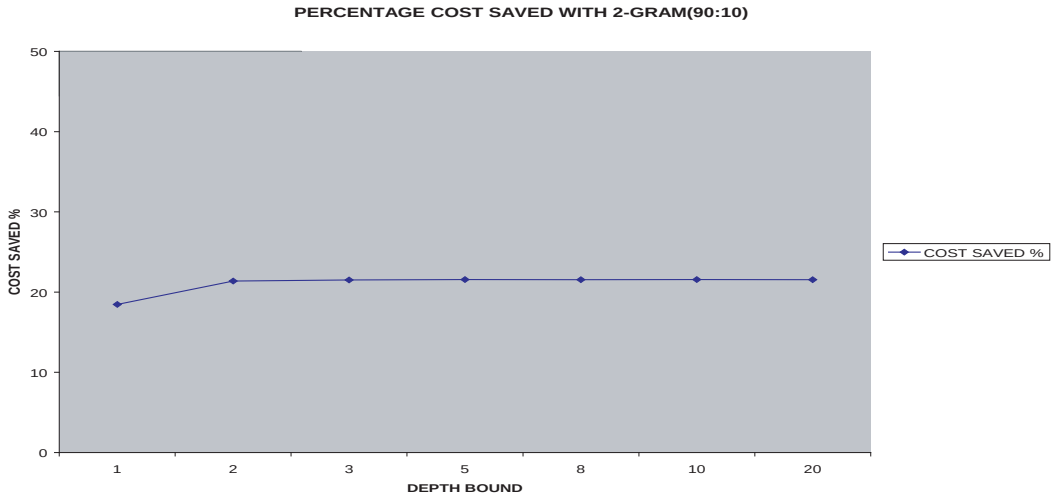


Figure 5. Percentage of average cost saved vs. number of shortcuts recommended

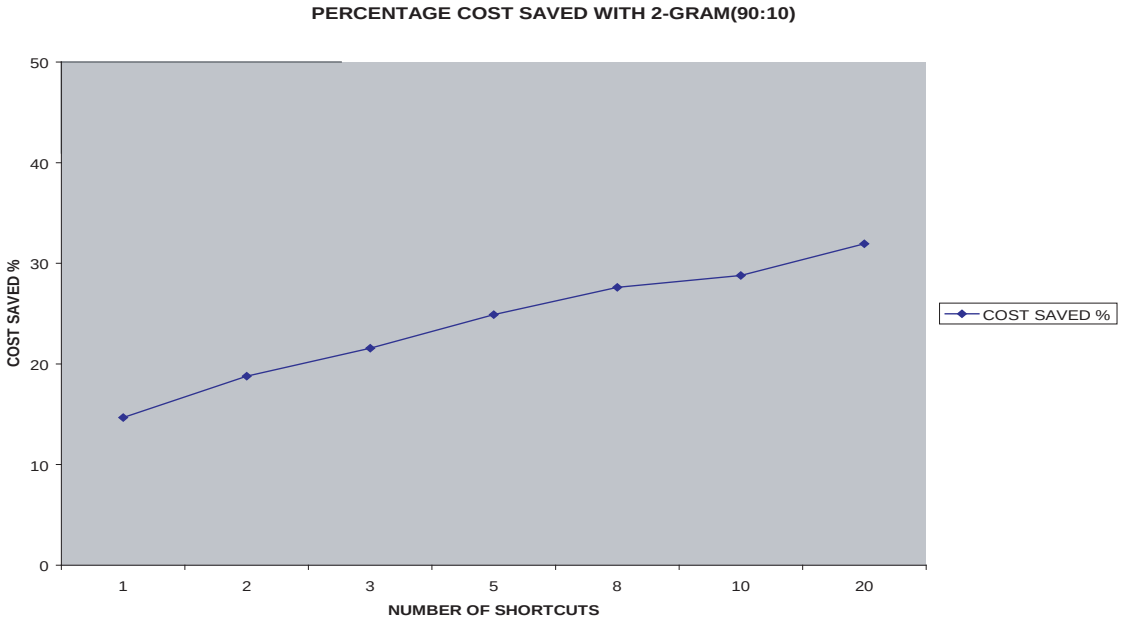
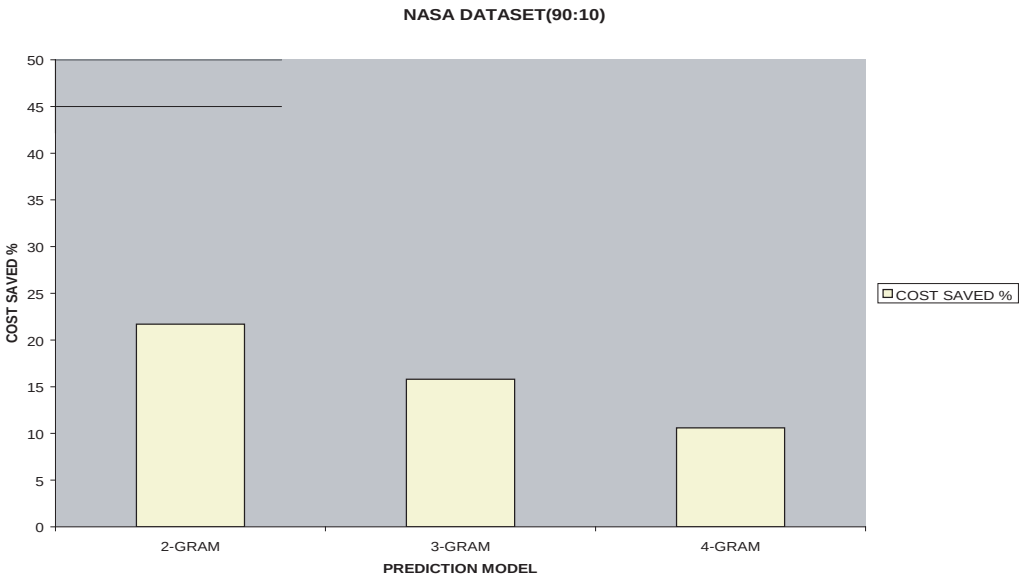


Figure 6. Percentage of average cost saved vs. N



thresholds, depth bound's effect is overshadowed by that of probability threshold.

As expected, the percentage of average cost saved increases with the number of shortcuts recommended by MINCOST, as shown in Figure 5. The increase in percentage of average cost saved is especially significant when the number of shortcuts increases from 1 to 2 and 2 to 3. However, the increase is much smaller for larger numbers of shortcuts, for example, when the number of shortcuts increases from 10 to 20. More importantly, it is not possible to display many shortcuts due to screen size limitations of mobile devices. We select a default value of 3 even though it does not provide the maximum savings. Besides, too many shortcuts may confuse users.

From Figure 6 we see that the percentage of average cost saved decreases with N. We conclude that best results are obtained with the 2-gram model. This is somewhat surprising. As discussed in Section 3.2, N-gram model's accuracy increases with N. The explanation for Figure 6 is that as N increases, N-grams become less applicable, i.e., fewer predictions are available for a given prefix, which results in fewer shortcuts.

Observations and Discussions

The percentage of average cost saved ranges from 19.4% to 23.0% for the NASA dataset and from 9.6% to 37.1% for the EPA dataset. The number of shortcuts recommended has the biggest impact among the parameters followed by the probability threshold, while depth bound has little impact. The best prediction model is the 2-gram model followed by 3-gram model and 4-gram model.

These results are comparable to results from the MINPATH algorithm with N-gram model.

Moreover, the percentage of savings are comparable to these reported by Anderson et al., (2001), though a different data set was used. This demonstrates that N-gram models work as well as other models.

CONCLUSIONS AND FUTURE WORK

In this article, we proposed to use a simple prediction model, N-gram, for improving mobile Web navigation. Our approach is implemented and experimented with two real datasets. Experimental results show that N-gram is as effective as more complex models used in other research in predicting useful shortcuts. An interesting finding is that 2-gram works better than 3-gram and 4-gram in predicting useful shortcut. Higher order N-grams require more training and are less applicable.

In the future, we plan to use mixed N-gram models as a prediction model. Multiple N-gram models of various N used simultaneously for suggesting the best shortcuts.

It will also be interesting to mix Web content mining and Web usage mining techniques. For example, the destination page of a session is predicted by looking at current browsing sequence as well as its contents.

Another interesting research topic is to compare N-gram models with models that learn from their errors such as neural networks.

REFERENCES

Anderson, C. R., Domingos, P., & Weld, D. S. (2001). Adaptive web navigation for wireless devices. In *Proceedings of IJCAI-01 Workshop.* Seattle, WA.

Deshpande, M. & Karypis, G. (2000). *Selective markov models for predicting Web-page accesses.* (Tech. Rep. No. 00-056). University of Minnesota, IN.

Frias-Martinez, E., & Karamcheti, V. (2002). A prediction model for user access sequences, In *Proceedings of the Workshop on Web Mining for Usage Patterns and User Profiles.*

Fu, Y., Sandhu, K., & Shih, M. (1999). *Clustering of Web users based on access patterns.* International Workshop on Web Usage Analysis and User Profiling. San Diego, CA.

Koutri, M., Daskalaki, S., & Avouris N. (2002). Adaptive interaction with Web Sites: An overview of methods and techniques. In *Proceedings of the 4th International Workshop on Computer Science and Information Technologies.*Patras, Greece.

Mobasher, B., Cooley, R., & Srivastava, J. (1999a). Automatic personalization based on web usage mining. *Communications of the ACM, 43*(8), 142-151.

Mobasher, B., Cooley, R. & Srivastava, J. (1999b). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems, 1*(1), 5-32.

Perkowitz, M., & Etzioni, O. (1997). Adaptive Web sites: An AI challenge. In *Proceedings of IJCAI-97 Workshop.* Nagoya, Japan.

Sarukkai, R. (2000). Link prediction and path analysis using markov chains. *Computer Networks, 33*(1-6), 377-386.

Shahabi, C., Zarkesh, A., Adibi, J. & Shah, V. (1997). *Knowledge discovery from users Web-page navigation.* In Workshop on Research Issues in Data Engineering, Birminghan, UK.

Su, Z., Yang, Q., Lu, Y. & Zhang, H. (2000). WhatNext: A prediction system for Web requests using n-gram sequence models. In *Proceedings of First International Conference on Web Information Systems and Engineering Conference*. Hong Kong, China.

Yang, Q., Zhang, H. H., & Li, T. (2001). Mining Web logs for prediction models in WWW caching and prefetching. In *Proceedings of KDD-01 Workshop.* San Francisco, CA.

This work was previously published in the International Journal of Intelligent Information Technologies, edited by V. Sugumaran, Volume 3, Issue 2, pp. 51-64, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.29
A Study on the Performance of IPv6-Based Mobility Protocols:
Mobile IPv6 vs. Hierarchical Mobile IPv6

Ki-Sik Kong
Korea University, Republic of Korea

Sung-Ju Roh
Technology R&D Center, LG Telecom Co., Republic of Korea

Chong-Sun Hwang
Korea University, Republic of Korea

ABSTRACT

The performance of IP mobility protocols is highly dependent on the change of mobile nodes' (MNs') mobility and traffic-related characteristics. Therefore, it is essential to investigate the effects of these characteristics and to conduct an in-depth performance study of these protocols. In this paper, we introduce a novel analytical approach using a continuous-time Markov chain model and hierarchical network model for the performance analysis of IPv6 mobility protocols: Mobile IPv6 (MIPv6) and Hierarchical Mobile IPv6 (HMIPv6). According to these analytical models, we derive the location update costs (i.e., binding update costs plus binding renewal costs), packet tunneling costs, and total signaling costs, which are generated by an MN during its average domain residence time, when MIPv6 or HMIPv6 is deployed under the same network architecture, respectively. In addition, based on these derived costs, we investigate the effects of various parameters, such as the average speed of an MN, binding lifetime period, the ratio of the network scale, and packet arrival rate, on the signaling costs generated by an MN under MIPv6 and HMIPv6. Moreover, we conduct the performance comparison between these two protocols by showing the relative total signaling costs under the various conditions. The analytical results show that as the average speed of an MN gets higher and the binding lifetime period is set to the larger value or as its packet arrival rate gets lower, the total signaling cost generated by an MN during its average domain residence time under HMIPv6 will get relatively lower than that under

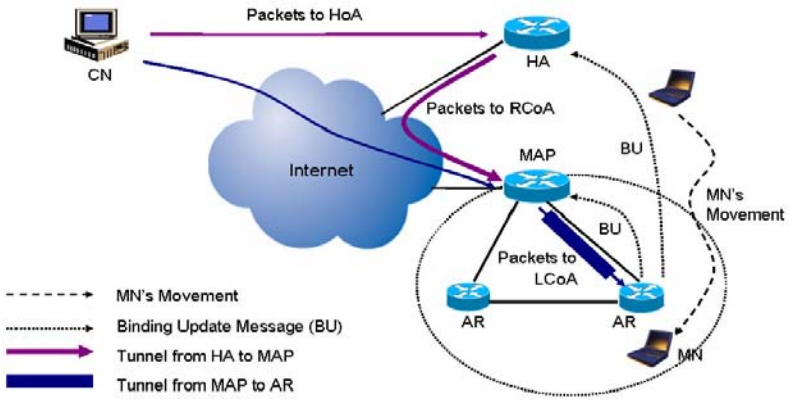
MIPv6, and that under the reverse conditions, the total signaling cost under MIPv6 will get relatively lower than that under HMIPv6.

INTRODUCTION

The demand for anywhere, anytime high-speed Internet access has been a driving force for the increasing growth and advances in wireless/mobile communication and portable devices. As a consequence, these trends have prompted research into mobility support in networking protocols. The IETF Mobile IP working group has proposed Mobile IPv4 (MIPv4) (Perkins, 1997, 2002) as a main protocol for supporting IP mobility. However, MIPv4 has some problems, such as triangle routing, security, and limited IP address space. Thus, based on the next generation Internet protocol, IPv6, Mobile IPv6 (MIPv6) (Johnson, 2004) has been developed by the IETF with new functionalities. In MIPv6, when a mobile node (MN) moves from one subnet to another, it acquires a care-of-address (CoA) by stateless address autoconfiguration (Johnson, 2004). After configuring the CoA, the MN registers the association between the CoA and its home address (HoA) by sending a binding update (BU) message to its home agent (HA) or correspondent node (CN). Although MIPv6 solves some drawbacks addressed in MIPv4, it still has a problem, nevertheless. That is, MIPv6 handles local mobility of an MN in the same way as it handles global mobility. As a result, an MN sends the BU message to its HA and its CN each time it changes its point of attachment, regardless of locality. Such an approach may cause excessive signaling traffic, especially for MNs with relatively high mobility or long distance to their HAs or CNs. In addition, this approach is not scalable, since the signaling traffic generated by the MNs can become quite overwhelming as the number of the MNs increases. In order to overcome these drawbacks, Hierarchical Mobile IPv6 (HMIPv6) (Castelluccia, 2000; Soliman, 2004) has been proposed to accommodate frequent mobility of the MNs and to reduce the signaling load in the backbone networks. In addition, handoff performance may be improved by reducing handoff latency. HMIPv6 introduces a new entity, the mobility anchor point (MAP), which works as a proxy for the HA in a foreign network. When an MN moves into a network controlled by a new MAP, it configures two CoAs: a regional care-of-address (RCoA) and an on-link care-of-address (LCoA). The RCoA is an address on the MAP's subnet, whereas the LCoA is an address configured to the MN's current point of attachment. Figure 1 shows the basic operation of HMIPv6. When the MN first enters a MAP domain, it sends the BU message to the MAP, the HA, and, potentially, to the CNs. When an MN changes the subnets within a same MAP domain, it only sends the BU message to the MAP. In other words, if the MN changes its current LCoA within a MAP domain, it only needs to register its LCoA with the MAP. The RCoA does not change, as long as the MN moves within the same MAP domain. This makes the MN's mobility transparent to the HA and the CNs. However, this does not imply any change to the periodic binding renewal (BR) message that an MN has to send to the HA and the CN, and now an MN additionally should send it to the MAP. In addition, since the MAP acts as a local HA, it receives all packets on behalf of the MNs that it is serving and tunnels the received packets to the MN's current LcoA.

Generally, the performance of IP mobility protocols may vary widely, depending on the change of MNs' mobility and traffic-related characteristics (Campbell, 2002). Our previous work on HMIPv6 (Kong, 2004) also revealed that such characteristics are the crucial factors that significantly may affect the signaling load on the Internet. Therefore, it is essential to analyze and evaluate the IP mobility protocols under various conditions, and more in-depth study on these protocols should be considered as the first step

Figure 1. Hierarchical Mobile IPv6



to design a more efficient IP mobility management scheme.

The remainder of this paper is organized as follows. In the second section, we briefly describe the existing works on the performance analysis of IP mobility protocols. In the third section, we introduce the user mobility model and network model for the performance analysis of IP mobility protocols. Then, the various cost functions under MIPv6 and HMIPv6 are analytically derived. In the fourth section, we investigate the results of the third section by applying the various numerical examples. Finally, conclusion and future work are given in the fifth section.

RELATED WORK

Although a lot of researches for the performance analysis of IP mobility protocols have been proposed, most of them, have been simulation-based approaches. Recently, several analytical approaches for IP mobility protocols have been proposed. This section focuses on describing the existing works on the analytical approaches for the performance analysis of IP mobility protocols.

In terms of performance analysis of HMIPv6, there is some difference between the work in Castelluccia et al. (2000) and our work. Castelluccia et al. (2000) presented a hierarchical mobility architecture that separates local mobility and global mobility for the mobility management that is hierarchical, flexible, and scalable. But they focused mainly on evaluating the signaling bandwidth according to the BU emission frequency. As already introduced and studied in location management for PCS networks (Akyildiz, 1999), in order to evaluate the efficiency of IP mobility management protocol, the tradeoff relationship between the location update cost and the packet tunneling cost has to be taken into consideration in terms of the total signaling cost (Xie, 2002). Nevertheless, in Castelluccia et al. (2000), they did not consider the signaling overhead generated by the packet tunneling. Moreover, when the network administrators or designers consider the deployment of either MIPv6 or HMIPv6, they should understand fully how various mobility and traffic-related parameters may have an effect on the overall system performance. While the work in Castelluccia et al. (2000) made a good contribution toward introducing a new IP mobil-

ity protocol, they just showed few of the effects and relations of the various mobility and traffic related parameters.

Woo (2003) investigated the performance of MIPv4 regional registration. The performance measures used were registration delay and the CPU processing overheads loaded on the agents to handle mobility of the MNs. Through the investigation, the effectiveness of adopting MIPv4 regional registration in the wide-area wireless network in terms of reducing the CPU processing overheads on the HA and lowering the signaling delay was observed.

Costa et al. (2002) compared the handover latency resulting from the handover procedures in MIPv6, FMIPv6, and HMIPv6. They conducted that the best option to get better performance was to implement both HMIPv6 and FMIPv6. In this work, they computed handover latency based on the length of the path among the MN, the HA, and the CN.

Pack et al. (2003) proposed an analytic model for the performance analysis of HMIPv6 in IP-based cellular networks, which is based on the random walk mobility model. Based on this model, they formulated location update cost and packet delivery cost. Then, they analyzed the impact of cell residence time on the location update cost and the impact of user population on the packet delivery cost. Although their analytical model is well-defined, they did not take the periodic BR and the effect of binding lifetime period into account, which may have a significant effect on the total signaling cost. In addition, their analysis was only investigated on the performance of HMIPv6, and their analysis about the packet delivery cost under HMIPv6 is not likely to be the pure extra signaling bandwidth consumption generated by the packet tunneling but the network bandwidth consumption, including the data traffic as well as the signaling traffic.

However, from the viewpoint of IP mobility management, the consideration of the extra signaling bandwidth consumption (not including the data traffic) generated during the processes of the location update and the packet tunneling should be taken into account (Xie, 2002).

In contrast to the related literature mentioned, we perform in-depth analysis and more integrated comparison between MIPv6 and HMIPv6 in terms of the total signaling cost. Moreover, while the previous analyses did not consider either the periodic BR or the extra packet tunneling, our work considers both of them for the analysis. Also, we present a novel analytical approach based on a continuous-time Markov chain and a simplistic hierarchical network model. Based on these models, we analytically derive the binding update cost, binding renewal cost, packet tunneling cost, and total signaling cost generated by an MN during its average domain residence time under MIPv6 and HMIPv6. Then, based on these derived costs, we investigate the effects of various mobility and traffic-related parameters on these costs. In addition, we conduct the performance comparison between MIPv6 and HMIPv6, and evaluate the conditions where the performance gain between the two protocols is larger or smaller in terms of the total signaling cost. The aim of this paper is not to determine which protocol performs better but to evaluate the performance that can be expected for each protocol, broaden our understanding of the various parameters that influence the performance, and help in the network design decision.

ANALYTICAL FRAMEWORK FOR THE ANALYSIS OF IPv6-BASED MOBILITY PROTOCOLS

First, we introduce our user mobility and network models to evaluate the performance of MIPv6 and HMIPv6. Then, for the analysis, we analytically derive the various cost functions under MIPv6 and HMIPv6.

User Mobility Model

Figure 2 indicates a state transition diagram of a continuous-time Markov chain model, which describes the BU process of an MN. The state of a continuous-time Markov chain, i ($i \geq 0$), is defined as the number of subnets where an MN has stayed within the given domain. Also, state *0* represents that the MN stays outside of the given domain. The state transition $a_{i,i+1}$ ($i \geq 1$) represents an MN's movement rate to an adjacent subnet within the given domain, and the state transition $a_{0,1}$ represents an MN's movement rate to a subnet within a given domain from another domain. On the other hand, $b_{i,0}$ ($i \geq 1$) represents an MN's movement rate to another domain from a given domain. In addition, for the analysis of the signaling costs generated by an MN during its average domain residence time, we assume that an MN moves out of a given domain within the maximum of finite K movements.

On the other hand, in order to obtain an MN's movement rates, we assume a fluid flow mobility model (Baumann, 1994). The model assumes that the MNs are moving at an average speed of v, and their movement direction is uniformly distributed over $[0, 2\pi]$, and that all the subnets are of the same rectangular shape and size and form together a contiguous area. The parameters used for the analysis of an MN's movement rates are summarized as follows.

- γ : The movement rate for an MN out of a subnet
- λ : The movement rate for an MN out of a subnet within the given domain
- μ : The movement rate for an MN out of domain

According to Baumann (1994), the movement rate γ for an MN out of a subnet is derived as

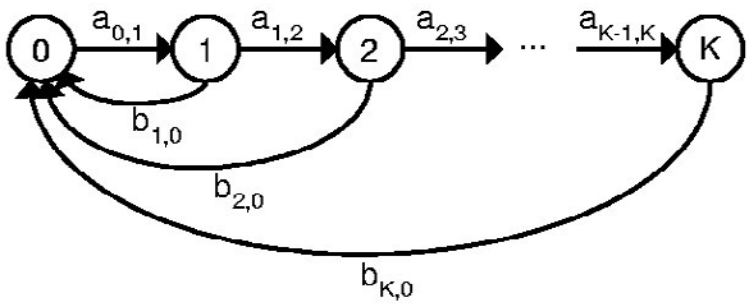
$$\gamma = \frac{4v}{\pi\sqrt{S}} \tag{1}$$

where S is the subnet area. We assume that a domain is composed of N equally subnets. Therefore, the movement rate μ for an MN out of a domain is

$$\mu = \frac{4v}{\pi\sqrt{NS}} \tag{2}$$

Note that an MN that moves out of a domain also will move out of a subnet. So, the movement

Figure 2. State transition diagram of a continuous-time Markov chain



rate λ for an MN out of a subnet within the given domain is obtained from Equations (1) and (2):

$$\lambda = \gamma - \mu = (1 - \frac{1}{\sqrt{N}})\gamma \tag{3}$$

Therefore, in Figure 2, we get $a_{i,i+1} (i \geq 1) = \lambda$ and $a_{0,1} = b_{i,0} = \mu$, respectively.

On the other hand, we assume π_i to be the equilibrium probability of state *i*. Thus, we can obtain the following equations from state transition diagram shown in Figure 2.

$$\mu\pi_0 = \mu(\pi_1 + \pi_2 + ... + \pi_K) \tag{4}$$

$$\mu\pi_{i-1} = (\lambda + \mu)\pi_i, \qquad i = 1 \tag{5}$$

$$\lambda\pi_i = (\lambda + \mu)\pi_{i+1}, \qquad 2 \leq i \leq K - 1 \tag{6}$$

$$\lambda\pi_{i-1} = \mu\pi_i, \qquad i = K \tag{7}$$

On the other hand, by the law of total probability, the sum of the probabilities of all states is 1. Thus,

$$\pi_0 + \pi_1 + \pi_2 + \cdots + \pi_K = \sum_{i=0}^{K} \pi_i = 1 \tag{8}$$

By substituting Equation (8) into Equation (4), we can obtain the equilibrium probability of state *0*, π_0. Thus,

$$\pi_0 = \frac{1}{2} \tag{9}$$

Finally, using Equations (5), (6), (7), and (9), π_i can be expressed as

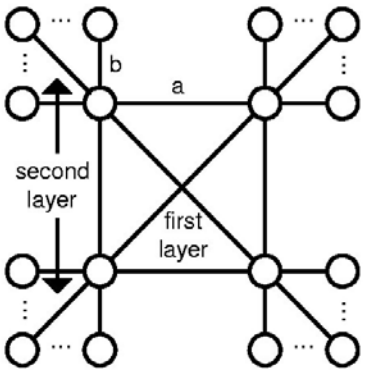
$$\pi_i = \begin{cases} \frac{1}{2} & if \quad i = 0 \\ \frac{1}{2}(\frac{\mu}{\lambda+\mu})(\frac{\lambda}{\lambda+\mu})^{i-1} & if \quad 1 \leq i \leq K-1 \\ \frac{1}{2}(\frac{\lambda}{\lambda+\mu})^{i-1} & if \quad i = K \end{cases} \tag{10}$$

Network Model

Similar to Ihara (2000), we assume a simplistic two-layer hierarchical network model given in Figure 3. The first layer has a mesh topology, which consists of *M* nodes. Each first layer node is a root of a *N*-ary tree with depth of 1. We assume that the HA and the access routers (ARs) are all second layer nodes, and that each domain is composed of all the second layer nodes under the same first layer node. For simplicity, the CN, MN, and HA are assumed to be located in the different domain. Also, we define the domain size *(N)* as the number of all the second layer nodes under the same first layer node. For the performance analysis of the two protocols under the same network architecture, we assume that when HMIPv6 is considered, the functionality of the MAP is placed on the first layer node. In this model, the link hops between the first layer nodes is *a*, and the link hops between the first and second layer nodes in the same domain is *b*, respectively. On the other hand, we assume that the link hop between the CN and the CN's default AR is zero. We also do not consider the transmission cost over the wireless link.

By adjusting the ratio of the network scale in the network model shown in Figure 3, we can investigate the effects of the distance among the CN, the HA, and the MN. Let the ratio of the network scale *(r)* be $0.1 < r = b/a < 1$, and $b = 3$ are assumed. Thus, the large value of *r* means that the MN is located close to the HA or the CN, and

Figure 3. Two-layer hierarchical network mode



the small value of *r* means that the MN is located far away from the HA or the CN.

Cost Analysis

According to our user mobility and network models given in the previous subsections, we analytically derive the location update costs (i.e., the binding update costs plus binding renewal costs), packet tunneling costs, and total signaling costs generated by an MN during its average domain residence time under MIPv6 and HMIPv6. There are two kinds of BU messages under MIPv6 and HMIPv6. That is, the one occurs from an MN's subnet crossing, and the other occurs when the binding is about to expire. In this paper, we use *binding update (BU) message* to refer to the former, and *binding renewal (BR) message* to refer to the latter in order to differentiate these two kinds of binding related messages. Also, we use *BAck message* to refer to the binding acknowledgment message from the HA or the MAP.

For the analysis, several parameters and assumptions given in the previous subsections are used.

Hierarchical Mobile IPv6

Based on our mobility model given in the previous subsection, the average binding update cost under HMIPv6 (U_{Hmip}) can be derived as

$$U_{Hmip} = \pi_0 \times (U_m + U_h + \delta U_c) + (\Phi(K) - 1) \times U_m \quad (11)$$

where U_m, U_h, and U_c are the binding update costs to register with the MAP, the HA, and the CN under MIPv6 and HMIPv6, respectively. Based on our network model and assumptions given in the previous subsection, U_m, U_h, and U_c are expressed as $2S_{bu}b$, $2S_{bu}(a + 2b)$, and $S_{bu}(a + 2b)$, respectively. Here, S_{bu} represents the signaling bandwidth consumption generated by a BU/BR/BAck message. Note that the HA and the MAP must return BAck message to the MN. In addition, for simplicity, we assume that the binding related messages only are sent alone in a separate packet without being piggybacked and do not consider the BAck message from the CN.

On the other hand, $\Phi(K)$ means the average number of the subnets that an MN stays within a given domain, which is derived from the continuous-time Markov chain in Figure 2 and can be expressed as follows.

$$\Phi(K) = \pi_1 + 2\pi_2 + 3\pi_3 + \cdots + K\pi_K = \sum_{i=1}^{K} i\pi_i \quad (12)$$

On the other hand, for the calculation of the signaling costs generated by performing location update with the CNs, we roughly define the ratio of an MN's average binding time for the CNs to its average domain residence time δ as the following:

$$\delta = \frac{\sum_{i=1}^{n} C_i}{n\Delta} \quad (13)$$

where C_i represents the binding time for the *i*-th CN, which has been recorded in an MN's binding update list during its average domain residence time, and *n* means the number of all the CNs recorded in the MN's binding update list during its average domain residence time. Also, Δ means an MN's average domain residence time.

On the other hand, let the binding lifetime periods for the MAP, the HA, and the CN under HMIPv6 be T_m, T_h, and T_c, respectively. From Equations (1) and (2), an MN's average subnet residence time is derived as

$$\frac{\pi\sqrt{S}}{4v}.$$

Thus, the average rate of sending the BR message to the MAP under HMIPv6 while an MN stays in a subnet is

$$\lfloor\frac{\pi\sqrt{S}}{4vT_m}\rfloor.$$

Similarly, the average rates of sending the BR message to the HA and the CN under HMIPv6 during an MN's average domain residence time become

$$\lfloor\frac{\Phi(K)\pi\sqrt{S}}{4vT_h}\rfloor$$

and

$$\delta\times\lfloor\frac{\Phi(K)\pi\sqrt{S}}{4vT_c}\rfloor,$$

respectively. Consequently, the average binding renewal cost under HMIPv6 (R_{Hmip}) can be derived as follows.

$$R_{Hmip}=U_m\Phi(K)\times\lfloor\frac{\pi\sqrt{S}}{4vT_m}\rfloor+U_h\times\lfloor\frac{\Phi(K)\pi\sqrt{S}}{4vT_h}\rfloor+\delta U_c\times\lfloor\frac{\Phi(K)\pi\sqrt{S}}{4vT_c}\rfloor \quad (14)$$

Thus, the average location update cost under HMIPv6 (L_{Hmip}) generated by the binding related messages is

$$L_{Hmip} = U_{Hmip} + R_{Hmip} \quad (15)$$

Let the probability that the CN has a binding cache entry for an MN be *q*. Then, the average packet tunneling cost under HMIPv6 (T_{Hmip}) can be derived as follows.

$$T_{Hmip}=\frac{\Phi(K)\pi\sqrt{S}}{4v}\times p\times\{qD_{dir}+(1-q)D_{indir}\} \quad (16)$$

where *p* is the average packet arrival rate per hour for an MN. Also, D_{dir} and D_{indir} are the tunneling cost generated by a direct packet delivery (not intercepted by the HA) and the tunneling cost generated by delivering a packet routed indirectly through the HA under HMIPv6, respectively. According to our network model and assumptions, D_{dir} and D_{indir} can be derived as $S_{pt}b$ and S_{pt} (a + $2b$), respectively. Here, S_{pt} represents the additional signaling bandwidth consumption generated by tunneling per packet. Finally, the total signaling cost (C_{Hmip}) under HMIPv6 can be expressed as follows.

$$C_{Hmip} = L_{Hmip} + T_{Hmip} \quad (17)$$

Mobile IPv6

Similar to Equation (11), the average binding update cost under MIPv6 (U_{Mip}) can be derived as follows.

$$U_{Mip} = \pi_0 \times (U_h + \delta U_c) + (\Phi(K) - 1)\ (U_h + \delta U_c) \quad (18)$$

Table 1. Default parameter value

Parameter	Type	Value
N	Domain Size	64
S	Subnet Size	10 Km^2
T	Binding Lifetime	0.3 hour
v	An average speed of an MN	10 Km/hour
S_{bu}	The signaling bandwidth generated by a BU/BR/Back message	68 byte
S_{pt}	The signaling bandwidth generated by tunneling per packet	40 byte

Let the binding lifetime period for the HA and CN under MIPv6 be $\overline{T}_h$ and $\overline{T}_c$, respectively. Then, the average binding renewal cost under MIPv6 (R_{Mip}) can be derived as follows.

$$R_{Mip} = U_h \Phi(K) \times \lfloor \frac{\pi\sqrt{S}}{4v\overline{T}_h} \rfloor + \delta U_c \Phi(K) \times \lfloor \frac{\pi\sqrt{S}}{4v\overline{T}_c} \rfloor \quad (19)$$

Therefore, the average location update cost under MIPv6 (L_{Mip}) generated by the binding related messages is

$$L_{Mip} = U_{Mip} + R_{Mip} \quad (20)$$

On the other hand, the average packet tunneling cost in MIPv6 (T_{Mip}) can be derived as follows.

$$T_{Mip} = \frac{\Phi(K)\pi\sqrt{S}}{4v} \times p \times \{q\overline{D}_{dir} + (1-q)\overline{D}_{indir}\} \quad (21)$$

where $\overline{D}_{dir}$ and $\overline{D}_{indir}$ are the tunneling cost generated by a direct packet delivery and the tunneling cost generated by delivering a packet routed indirectly through the HA under MIPv6, respectively. According to our network model and assumptions, $\overline{D}_{dir}$ and $\overline{D}_{indir}$ can be derived as *0* and $S_{pt}b(a + 2b)$, respectively. Finally, the total signaling cost (C_{Mip}) under MIPv6 can be expressed as follows.

$$C_{Mip} = L_{Mip} + T_{Mip} \quad (22)$$

NUMERICAL RESULTS

In this section, we compare the performance of MIPv6 and HMIPv6 based on the various cost functions derived in the previous subsection. For the analysis, we first investigate the effects of various parameters, such as the average speed of an MN, binding lifetime period, ratio of the network scale, packet arrival rate, and the probability that the CN has a binding cache for an MN. Then, we evaluate the conditions where the performance gain between the two protocols is the largest or the smallest in terms of the total signaling cost. The performance measure used is the signaling bandwidth consumption per packet multiplied by the number of link hops that the packet traverses during an MN's average domain residence time (i.e., *bytes* × *number of link hops* / *average domain residence time*).

For the analysis, we set the default values of p, q, r, and δ to be 100, 0.7, 0.2, and 0.1, respectively. Also, we assumed that K is equal to the domain size, 64. The other default parameter values for the performance analysis are given in Table 1. Most parameters used in this analysis are set to typical values found in Woo (2003) and Ramjee (2002). The size of a BU/BR/BAck message is equal to the size of an IPv6 header (40 bytes) plus the size of a BU extension header (28 bytes), so 68 bytes (Castelluccia, 2000). In addition, the additional signaling bandwidth consumption generated by tunneling per packet is equal to the size of IPv6 header, so 40 bytes. According to Ihara (2000), the binding lifetime periods (i.e., $\overline{T_h}$, T_h, $\overline{T_c}$, T_c, and T_m) under MIPv6 and HMIPv6 are assumed to be all the same and denominated as T.

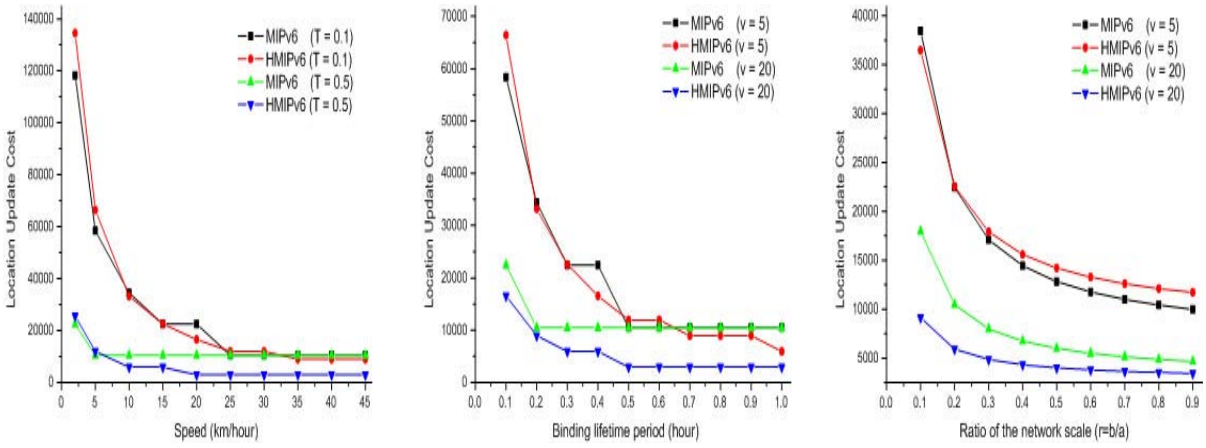
Analysis of Location Update Costs

Figure 4(a) indicates the effect of v on the average location update costs under MIPv6 and HMIPv6. As already mentioned, the location update cost consists of the binding update cost generated by an MN's mobility and the periodic binding renewal cost generated by the expiration of the binding lifetime period. The results presented in Figure 4(a) show that the location update costs under MIPv6 and HMIPv6 decrease as v increases. Also, this phenomenon gets larger as T is set to the larger value. When an MN is not moving fast, most of the signaling traffic is generated by the periodic BR messages. However, as the average speed of an MN increases, the number of the periodic BR messages decreases, and the BU messages generated by an MN's mobility dominate most of the signaling traffic. Note here that the location update cost under MIPv6 remains the same when v exceeds 5 km/hour for T = 0.5 hour in Figure 4(a). This is due to the fact that as the average speed of an MN increases, it moves to an adjacent subnet before the BR message occurs. In other words, that is the case that the average subnet residence time of an MN is shorter than the binding lifetime period. Therefore, when v exceeds 5 km/hour for T = 0.5 hour, the average location update cost is equal to the average binding update cost, and the average binding renewal cost is 0.

Figure 4(b) indicates the effects of T on the average location update costs under MIPv6 and HMIPv6. As shown in Figure 4(b), the average location update costs of the two protocols are decreased as T gets larger. The results shown in Figure 4(b) indicate that the binding lifetime period has a significant effect on the location update cost. Generally, the smaller the value of binding lifetime period is, the larger the binding renewal cost gets. Therefore, too much small value of binding lifetime period may result in significant signaling load throughout the networks. On the other hand, the larger the value of binding lifetime period is, the larger the binding cache entry size at the mobility agents gets. Thus, this may result in an increase of the binding cache lookup time and memory consumption at the mobility agents. In practice, the value of binding lifetime period must be specified in the implementation of MIPv6 and HMIPv6. Therefore, further study on the effects of binding lifetime period needs to be investigated to achieve the best performance.

Figure 4(c) indicates the effects of r on the average location update costs under MIPv6 and HMIPv6. Evaluating the ratio of the network scale enables the effect of the layer in which the HA or the CN is located or the effect of the network scales of each layer to be estimated. In Figure 4(c), the relationship $0.1 < r = b/a < 1$ and $b = 3$ are assumed to hold. Thus, the large value of r means that the MN is located close to the HA or the CN, and the small value of r means that the MN is located far away from the HA or the CN. As shown in Figure 4(c), the cost gain of HMIPv6 over MIPv6 gets larger as r decreases. However, for v = 5 km/hour, as r increases, the gain of MIPv6 over HMIPv6 gets larger. When v is fixed, an increase of r results in the relative reduction of

Figure 4. Analysis of location update costs: (a) Effects of v (top left), (b) Effects of T (top right), and (c) Effects of r (bottom)



(a) *(b)* *(c)*

the gain of HMIPv6, and thus, the gain of MIPv6 relatively increases. This phenomenon becomes more prominent as *v* decreases.

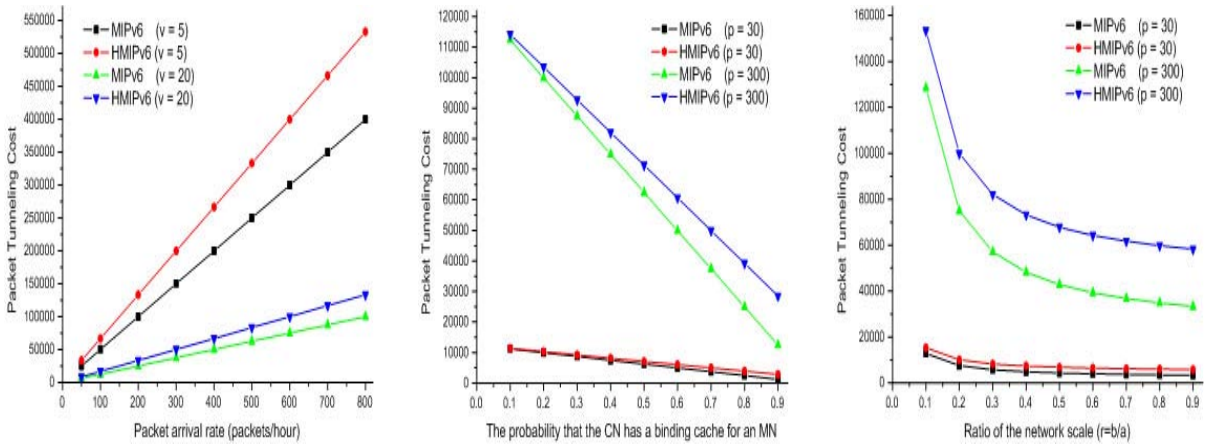
Analysis of Packet Tunneling Costs

Figure 5(a) indicates the effects of *p* on the average packet tunneling costs under MIPv6 and HMIPv6. The results shown in Figure 5(a) indicate that the packet tunneling costs are linearly increased as *p* increases. Figure 5(a) also indicates that a slowly moving MN is more affected by the packet arrival rate when the packet arrival rate is fixed. This is due to the fact that the average domain residence time of a slowly moving MN is longer than that of a fast moving MN, and thus, the packet tunneling cost of a slowly moving MN is larger than that of a fast moving MN. Also, the cost gap between the MIPv6 and HMIPv6 is due to the additional packet tunneling from the MAP to the AR under HMIPv6.

Figure 5(b) indicates the effects of *q* on the average packet tunneling costs. The results shown in Figure 5(b) indicate that the packet tunneling costs are linearly decreased as *q* increases. This is due to the fact that as *q* increases, the number of packets routed indirectly through the HA to the MAP decreases. In addition, the cost gap between the two protocols is due to the additional packet tunneling from the MAP to the AR under HMIPv6, and this gap increases as *q* does. Note that the cost gap occurs only from the difference between D_{dir} and $\overline{D}_{dir}$, and that D_{indir} and $\overline{D}_{indir}$ do not affect the cost gap between the two protocols (refer to the third section).

Figure 5(c) indicates the effects of *r* on the average packet tunneling costs. The results shown in Figures 5(b) and (c) indicate that when the packet arrival rate is low, the costs of the two protocols are almost the same, since the additional packet tunneling at the MAP is relatively negligible. In

Figure 5. Analysis of packet tunneling costs: (a) Effects of p (top left), (b) Effects of q (top right), and (c) Effects of r (bottom)



(a) *(b)* *(c)*

contrast, when the packet arrival rate is high, the cost gap gets larger.

Analysis of Total Signaling Costs

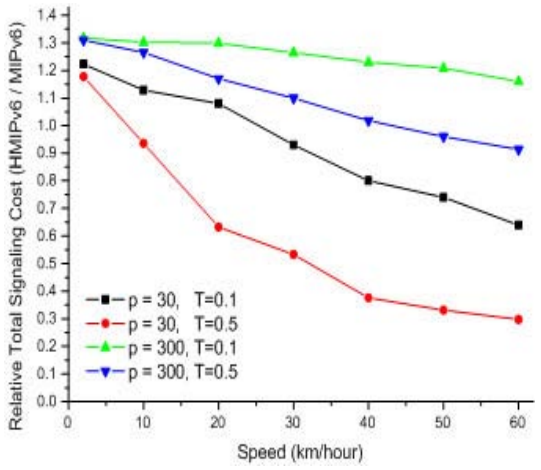
In this subsection, we evaluate the total signaling cost gain between MIPv6 and HMIPv6 by comparing the relative total signaling costs of the two protocols under the four parameter sets given in Table 2. This evaluation is investigated to show the variation in the relative total signaling costs of MIPv6 and HMIPv6 according to the change of *v*, *p*, and *T*. For the analysis, we define the relative total signaling cost of HMIPv6 as the ratio of the total signaling cost under HMIPv6 to that under MIPv6. Therefore, a relative total signaling cost of 1 means that the total signaling costs under the two protocols are exactly the same. As already mentioned, in order to evaluate the efficiency of IP mobility management protocols, the tradeoff relationship between the location update cost and the packet tunneling cost has to be taken into consideration (Xie, 2002). In this subsection, we will show the efficiency of MIPv6 and HMIPv6 by comparing the relative total signaling costs of the two protocols under the four parameter sets given in Table 2.

Figure 6 indicates the following several facts: As *v* decreases, the signaling cost gain of MIPv6

Table 2. Parameter set

set	*p*	*T*
1	30	0.1 hour
2	30	0.5 hour
3	300	0.1 hour
4	300	0.5 hour

Figure 6. Analysis of relative total signaling costs



over HMIPv6 gets larger, and this trends go into the reverse as *v* increases. In addition, under the same value of *v*, the cost gain of MIPv6 over HMIPv6 gets larger as *p* increases or as *T* is set to the smaller value. HMIPv6 aims to reduce the number of the BU messages in the backbone network by using the MAP. However, this does not imply any change to the periodic BR messages that an MN has to send to the HA and the CN, and the MN rather should send it to the MAP additionally. In addition, since the MAP acts as a local HA in HMIPv6, it receives all packets on behalf of the MNs it is serving and should tunnel the received packets to the MN. In other words, from the viewpoint of the signaling bandwidth consumption inside the domain, the signaling bandwidth consumption inside domain under HMIPv6 is generally larger than that under MIPv6, because of the additional periodic BR messages to the MAP and the packet tunneling at the MAP. In addition, this phenomenon becomes more prominent as an MN stays for a considerable time within a domain (i.e., when *v* is very small) or the incoming packet arrival rate is high. This is due to the fact that the domain residence time and packet arrival rate are proportional to the binding renewal cost and packet tunneling cost, respectively. These are the reasons why the gain of MIPv6 over HMIPv6 gets larger as *v* and *T* decrease or *p* increases. On the other hand, the cost gain of HMIPv6 over MIPv6 gets larger as *v* increases, and this phenomenon becomes more prominent as *p* decreases or as *T* is set to the larger value. The results shown in Figure 6 verify these facts. Note that when *v* is greater than or equal to 7 km/hour in Figure 6, the parameter set 2 is the case where the gain of HMIPv6 over MIPv6 is the largest under the four parameter sets, and the parameter set 3 is the case where the gain of MIPv6 over HMIPv6 is the largest, as shown in Figure 6.

CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel analytical approach for the performance analysis of IPv6 mobility protocols. In order to evaluate the signaling load of the two protocols, we designed a user mobility model and network model. Then, based on these two models, we analytically derived the location update costs, packet tunneling costs, and total signaling costs generated by an MN during its average domain residence time under MIPv6 and HMIPv6. In addition, we investigated the effects of various mobility and traffic-related parameters on each derived costs. The analytical results provide a deep understanding of the overall performance of MIPv6 and HMIPv6, and demonstrate that as the average speed of an MN gets higher and the binding lifetime period is set to the larger value, or as its packet arrival rate gets lower, the total signaling cost generated by an MN during its average domain residence time under HMIPv6 will get relatively lower than that under MIPv6, and that under the reverse conditions, the total

signaling cost under MIPv6 will get relatively lower than that under HMIPv6.

Our future research directions are as follows. First, we intend to extend our analytical performance study to the other protocols, such as Fast Handovers for MIPv6 (FMIPv6) and Fast Handovers for HMIPv6 (FHMIPv6). Second, based on the modeling technique and results shown in this paper, we will focus on designing the adaptive mobility management scheme that enables an MN to selectively switch its mobility management scheme in HMIPv6 networks in order to minimize the total signaling cost according to its mobility/traffic history (Kong, 2005), and these works are underway.

REFERENCES

Akyildiz, I. et al. (1999). Mobility management in next-generation wireless systems. *IEEE Proceedings Journal, 87*(8), 1347-1385.

Baumann, F., & Niemegeers, I. (1994). An evaluation of location management procedures. In *Proceedings of the UPC94*, (pp. 359-364).

Campbell, A. et al. (2002). Comparison of IP micromobility protocols. *IEEE Personal Communications*, 72-82.

Castelluccia, C. (2000). HMIPv6: A hierarchical mobile IPv6 proposal. *ACM Mobile Computing and Communications Review, 4*(1), 48-59.

Costa, X., Schmitz, R., Hartenstein, H., & Liebsch, M. (2002). A MIPv6, FMIPv6 and HMIPv6 handover latency study: Analytical approach. In *Proceedings of the IST Mobile and Wireless Telecommunications.*

Ihara, T., Ohnishi, H., & Takagi, Y. (2000). Mobile IP route optimization method for a carrier-scale IP network. In *Proceedings of the ICECCS 2000*, (pp. 11-14).

Johnson, D., & Perkins, C. (2004). *Mobility support in IPv6*. IETF RFC 3775.

Kong, K., Roh, S., & Hwang, C. (2004). Signaling load of hierarchical mobile IPv6 protocol in IPv6 networks. In *Proceedings of the PWC 2004, LNCS* (Vol. 3260, pp. 440-450). Berlin: Springer-Verlag.

Kong, K., Roh, S., & Hwang, C. (2005). History-based auxiliary mobility management strategy for hierarchical mobile IPv6 networks. In *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Special Issue on Multi-Dimensional Mobile Information Networks* (accepted for publication).

Pack, S., & Choi, Y. (2003). Performance analysis of hierarchical mobile IPv6 in IP-based cellular networks. In *Proceedings of the PIMRC 2003*, (pp. 2818-2822).

Perkins, C. (1997). Mobile IP. *IEEE Communications Magazine*, 84-99.

Perkins, C. (2002). *IP Mobility Support for IPv4*. IETF RFC 3344.

Ramjee, R. et al. (2002). HAWAII: A domain-based approach for supporting mobility in wide-area wireless networks. *IEEE/ACM Transactions on Networking, 10*(3), 396-410.

Soliman, H., Castelluccia, C., Malki, K., & Bellier, L. (2004). *Hierarchical mobile IPv6 mobility management (HMIPv6).* Retrieved from *http://draft-ietf-mipshop-hmipv6-04.txt*

Woo, M. (2003). Performance analysis of mobile IP regional registration. *IEICE Transactions on Communications, E86-B*(2), 472-478.

Xie, J., & Akildiz, I. (2002). A novel distributed dynamic location management scheme for minimizing signaling costs in mobile IP. *IEEE Transactions on Mobile Computing, 1*(3), 163-176.

This work was previously published in the International Journal of Business Data Communications and Networking, edited by M. Khosrow-Pour, Volume 1, No. 4, pp. 38-51, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.30
A Novel Fuzzy Scheduler for Mobile Ad Hoc Networks

S. Shanmugavel
Anna University, India

C. Gomathy
Deemed University, India

ABSTRACT

As mobile computing gains popularity, the need for ad hoc routing also continues to grow. In mobile ad hoc networks, the mobility of nodes and error prone nature of the wireless medium pose many challenges, including frequent route changes and packet losses. Such problems increase the packet delays and decrease the throughput. To meet with the dynamic queuing behaviour of Ad hoc networks, to provide QoS and hence to improve the performance, a scheduler can be used. This chapter presents a novel fuzzy based priority scheduler for mobile ad-hoc networks, to determine the priority of the packets. The performance of this scheduler is studied using GloMoSim and evaluated in terms of quantitative metrics such as packet delivery ratio, average end-to-end delay and throughput.

INTRODUCTION

A mobile ad hoc network is a cooperative engagement of mobile hosts or routers connected by wireless links. In the performance evaluation of a protocol, for an ad hoc network, the protocol should be tested under realistic conditions with representative data traffic models and realistic movement of mobile users. In order to thoroughly simulate a new protocol for an ad hoc network, it is very essential to use a mobility model that accurately represents the mobile nodes (MNs). MNs within an ad hoc network move from location to location. A mobility model should attempt to mimic the movements of the real MNs. Currently, there are two types of mobility models used in simulations of ad hoc networks: traces and synthetic models (Camp, Boleng, & Davies, 2002; Lin, Noubir, & Rajaraman, 2004). Traces are those mobility patterns that are observed in real-life systems.

Traces provide accurate information when they involve a large number of participants and a long observation period, but privacy issues will prohibit the collection and distribution of such statistics, and new environments cannot be easily modeled. Hence, in these situations, synthetic models are used. They realistically represent MNs without the use of traces. We consider here three of the synthetic models—namely, random walk, random way point, and random direction mobility models (Bettsetter, 2001).

The random walk mobility model is a widely used mobility model and, in this, the current speed and direction of MN is independent of its past speed and direction. It has a memory-less mobility pattern, because it retains no knowledge containing its past location and speed values. Here, we encounter unrealistic generation of movements such as sudden stopping, sharp turning, and completely random wandering.

The random waypoint mobility model includes pause times between changes in direction and speed. An MN begins by staying in one location for a certain period of time (Jardosh, 2003; Camp et al., 2002). Once this time expires, the MN chooses a random destination in the simulation area and a speed that is uniformly distributed between minspeed and maxspeed. The MN then travels towards the newly chosen destination at the selected speed. Upon arrival, the MN pauses for a specified time period before starting the process again. This is also a widely used model. The RWP model is similar to the random walk model if pause time is zero.

The random direction mobility model is a revised version of random walk, and it ensures that every node is assigned the same speed throughout the entire simulation. After a random direction is chosen in the range 0 to 2π, an MN begins moving. If the MN reaches a grid boundary, it bounces off the simulation border with an angle determined by the incoming direction. The MN then continues along this new path.

The choice of a mobility model can have a significant effect on the performance of an ad hoc network protocol. The performance of random walk, random waypoint, and random direction mobility models are compared. Dynamic source routing (DSR) protocol is chosen to be the routing protocol (Royer & Toh, 1999; Das, Castaneda, Yan, & Sengupta, 1998; Das, Perkins, & Royer, 2001). It determines the routes on demand. Here, the packet carries the full route that the packet should be able to traverse in its header.

DSR is chosen since it performs well in many performance evaluations of unicast protocols.

The performance metrics—namely, packet delivery ratio, end-to-end delay, average hop count, and protocol overhead—are used for comparison of these mobility models. The results prove that the random waypoint mobility model has the highest packet delivery ratio, lowest end-to-end delay, and lowest hop count (Camp et al., 2002). The random direction mobility model has the highest average hop count, highest end-to-end delay, and lowest packet delivery ratio since each MN moves to the border of the simulation area before changing its direction. The performance of the random walk model falls between these two. Hence to conclude, the random waypoint mobility model is used in many prominent simulation studies of ad hoc network protocols since it is flexible and it creates realistic mobility patterns for the way people might move in.

Research in the area of ad hoc networks has focused mainly on the routing protocols that decide the routing of packets hop by hop as efficiently as possible and medium access control (MAC), which indicates how to share the medium efficiently. But there is little focus towards the queuing dynamics in the nodes of the networks and on the effects scheduling algorithms in the queues of the nodes. Hence, we believe that choice of scheduling algorithm will certainly improve the performance of the ad hoc network. Here, the different scheduling algorithms and the network's

effect on mobile communication with the random waypoint mobility model are discussed.

We also design and analyze the performance of a fuzzy logic-based priority scheduler (FLPS), which combines the metrics and computes the crisp value of priority. The fuzzy algorithm for finding the priority of the packet based on some attributes of the packets is devised and coded in C language. The C code is linked with GloMoSim and is tested. It is found that the proposed fuzzy scheduler provides improved packet delivery ratio, reduced average end-to-end delay, and increased throughput when tested with various unicast routing protocols under different mobility conditions.

SCHEDULING ALGORITHMS

The ad hoc networks produce unique queuing dynamics due to the possible frequent transmission of control packets due to mobility, multi-hop forwarding of packets, and multiple roles of nodes as routers, sources, and sinks of data. The selection of the scheduling algorithms for mobile ad hoc networks is highly dependent on the queuing dynamics. These algorithms determine which queued packet to process next, and they have significant impact on the end-to-end performance. A scheduler for an ad hoc network is required to schedule the packets to reach the destination quickly, which are at the verge of expiry. The scheduler is positioned between the routing agent and the MAC layer. Without scheduling, packets will be processed in FIFO manner, and hence there are more chances that either more packets may be dropped or may not meet the quality of service (QoS) target.

Generally, in all algorithms, high priority is given to control packets. Different drop policies are used for data and control packets when the buffer is full. When the incoming packet is a data packet, the data packet is dropped. If it is a control packet, the last enqueued data packet is dropped. If all packets are control packets, the incoming control packet is dropped.

There are many scheduling algorithms proposed in literature. No priority scheduling services both control and data packet in FIFO order. Priority scheduling gives high priority to control packets, and data packets are serviced in FIFO order. When considering the effect of setting priorities to data packets, these schedulers give high priority to control packets (Chun & Baker, 2002). Their differences are in assigning priorities among data queues. Weighted hop and weighted distance scheduling methods use the distance metrics. Weighted hop scheduling gives higher weight to data packets that have fewer remaining hops to traverse. If the packet has fewer remaining hops, then it has to reach the destination quickly. The data packets can be stored in round-robin fashion. The remaining hops to traverse can be obtained from packet headers. In weighted distance scheduling, physical distance is used. It is also a weighted round-robin scheduler. It gives higher weight to data packets, which have shorter geographic distances. The remaining distance is the distance between a chosen next hop and a destination. Round-robin scheduling maintains per-flow queues. The flow can be identified by a source and destination pair. Here each flow queue is allowed to send one packet at a time in a round-robin fashion. In greedy scheduling scheme, each node sends its own data packets before forwarding those of other nodes (Luo, Lu, & Bhargavan, 2000). The other nodes' data packets are serviced in FIFO order. Two other schedulers are earliest deadline first (EDF) and virtual clock (VC) (Kanodia, Li, Sabharwal, Sadeghi, & Knightly, 2002). In EDF, a packet arriving at time t and having delay bound d has a deadline t + d. In virtual clock, a packet with size L of a flow with service rate r has priority index L/r plus the maximum of current time t and priority index of the flow's previous packet.

In these priority scheduling algorithms we considered, the parameters used to find the priority of packets are: remaining hops to traverse, remaining distance, per-flow queues, greediness of nodes, delay bound, and service rate. With the thorough study of ad hoc networks, and the above mentioned scheduling algorithms, it is found that a number of metrics can be combined into a single decision so as to find the crisp value of the priority of packets. Our solution to determine the priority index of the packets utilizes the fuzzy logic concept (Gomathy & Shanmugavel, 2004). It deals with the imprecise and uncertain information of the network parameters since ad hoc network is dynamic in nature. This is advantageous in the target system because the fuzzy logic system is flexible and capable of operating with imprecise data and hence can be used to model nonlinear functions with arbitrary complexity. The fuzzy inference process works in three stages: fuzzification, rule evaluation, and defuzzification. In the first stage the parameters of the system are fed into a fuzzifier, which transforms the real-time measurements into fuzzy sets. The second stage applies a set of fuzzy rules onto fuzzy input in order to compute fuzzy outputs. Finally, outputs are translated into crisp values.

QoS Provisioning

QoS provisioning is becoming a critical issue in designing wireless ad hoc networks due to the necessity of providing multimedia applications in such networks. These are typically delay sensitive and have high bandwidth requirements. It is a challenging task since the wireless channel is shared among adjacent hosts and network topology changes. Typical metrics for providing QoS include delay, loss rate bandwidth, and so forth. Here, in the design of scheduler, end-to-end delay and delivery ratio of packets are considered to analyze the performance of ad hoc networks and thus to provide QoS to the networks.

FUZZY LOGIC-BASED PRIORITY SCHEDULER (FLPS)

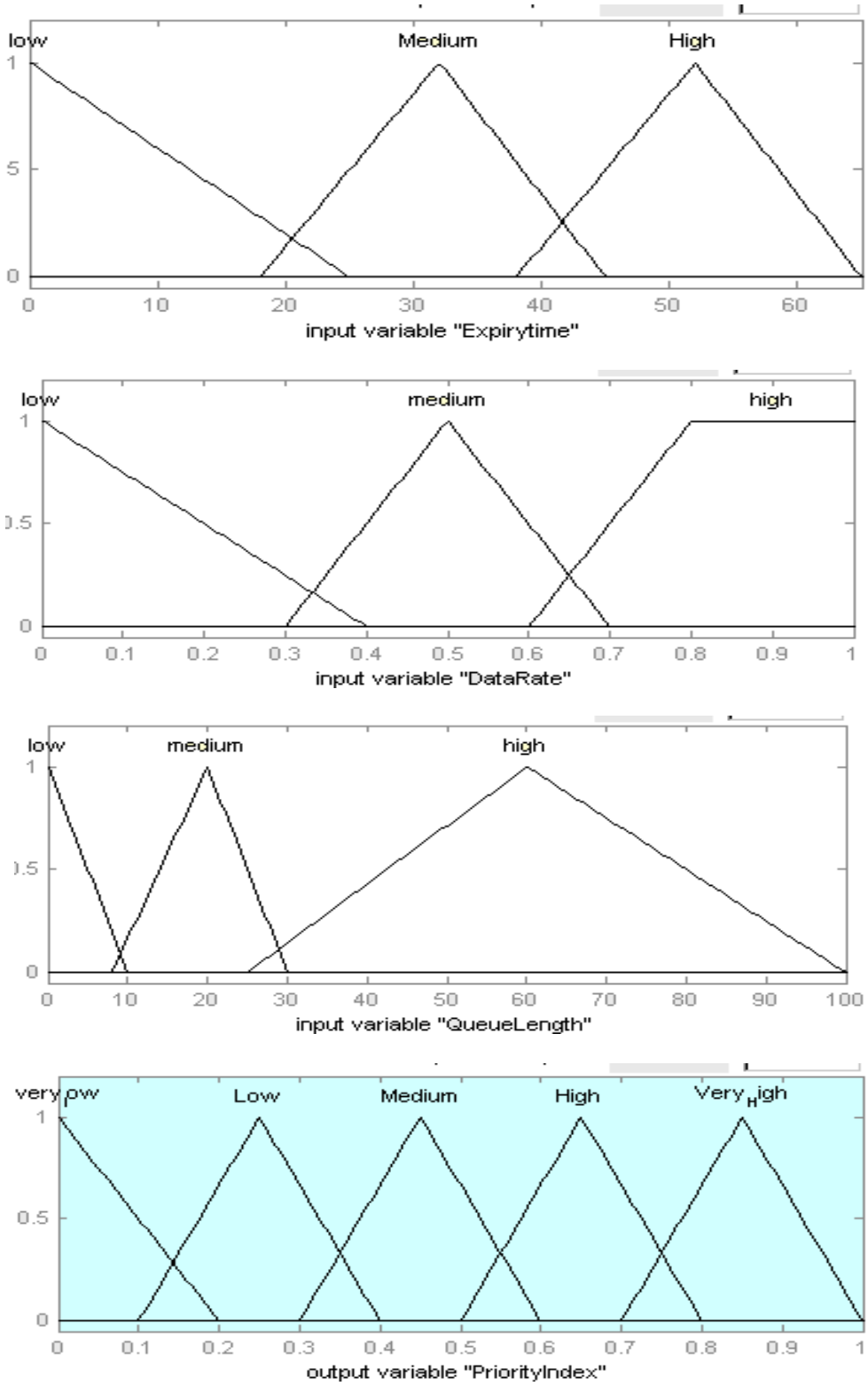
Fuzzy logic (FL) is a kind of artificial intelligence technology, which has the capability of mimicking how the field experts would make decisions. In networking area, there are a variety of traffic mixes, different traffic types, and long- and short-lived traffic flows. FL can incorporate networking expert knowledge to generate sensible solutions.

In order to improve the performance of scheduling algorithms, a mechanism using FL has been proposed to find the value of priority index of packets. The fuzzy logic-based priority scheduler aims to determine the priority index of data packets based on three input variables: data rate, expiry time of the packets, and queue length of the nodes. With these inputs, the fuzzification, rule evaluation, and defuzzification are performed. The following section describes the three processes.

Fuzzification

The three inputs are translated into fuzzy sets. The fuzzy sets contain elements that have a varying degree of membership in a set. Therefore, it is different from an ordinary set, where elements will only be considered as members of a class if they have full membership in that class. For example, if the expiry time is considered in an ordinary set, then it can only be either low or high and not both simultaneously, whereas in a fuzzy set, it can be classed as quite low, not so high, or medium and high. This indicates that the element in the fuzzy set can have membership in more that one set. The membership values are obtained by mapping the values obtained for a particular parameter onto a membership function, which will be used to determine the system outputs. This function is a curve or line that defines how each data or value is mapped onto a membership

Figure 1. Membership functions of input and output variables



value. We define here what are low L, medium M, and high H for each fuzzy set. Then by mapping the position of current input value onto the graph of the membership function, the input is allocated with a membership value in each set ranging from 0 to 1. Fuzzification of output parameter priority index is also performed, and five linguistic terms are attached to it. Figure 1 shows the membership function of the input and output variables for the FLPS.

Rule Evaluation

This stage involves feeding the fuzzy sets into an inference engine, where a set of fuzzy rules is applied. Fuzzy rules are usually defined as a set of possible scenarios in the form of if/then rules which determines the value of the priority index. Table 1 provides the summary of the decision-making logic. The first rule can be interpreted as: if the expiry time is low and data rate is low and queue length is low, then priority index is low. This indicates that, if expiry time of packets is low, it shows that packets are at the verge of expiry. And even if the data rate and number of packets in queue is low, the priority index is set to be low, so as to enable the packets to reach the destination quickly. The index value if very low indicates that packets are attached with high priority and will be scheduled immediately. If the index is very high, then packets are with lowest priority, and will be scheduled only after all high-priority packets are scheduled (Gomathy & Shanmugavel, 2004).

The rationale behind setting rules is that if the expiry time is low and both data rate and queue length are high, packets are attached with a very low priority index and hence possess high priority. Whereas if expiry time is high and both data rate and queue length are low, packets are attached

Table 1. Fuzzy rule base (D=data rate; Q=queue length)

Q D	L	M	H
Expiry Time – Low			
L	L	L	VL
M	VL	VL	VL
H	L	VL	VL
Expiry Time – Medium			
L	M	M	L
M	M	M	L
H	M	M	M
Expiry Time – High			
L	VH	VH	H
M	H	M	M
H	H	H	M

with higher priority index and hence possess low priority. These rules are then applied to fuzzy inputs and return the fuzzy outputs.

Defuzzification

At this stage the resultant fuzzy decision sets have to be converted into precise quantities. There exist several heuristic defuzzification methods such as max criterion, mean of maximum, and center of area or centroid method. In the FLPS, we consider the weighted average method of defuzzification to find the crisp output. The weighted average defuzzification technique can be expressed as:

$$x^* = \frac{\sum_i w_i m_i}{\sum_i w_i}$$

where x^* is the defuzzified output, m^i is the membership of the output of each rule, and w_i is the weight associated with each rule.

Example

Consider the scenario when the packets have an expiry time of 20 seconds, queue length of node the packet reaches is 50, and normalized data rate is 0.66 (normalized with respect to the channel capacity of 2 Mbps). For these set of inputs, the priority index is calculated as follows, which is done in three stages.

- **Fuzzification:** As seen from the figure, the expiry time of 20 seconds is fuzzified into low expiry with a degree of 0.2 and medium expiry with degree of 0.07. Queue length is fuzzified into high queue with degree 0.7142. Similarly, the data rate is fuzzified into medium rate with degree 0.2 and high rate with degree 0.3.
- **Rule Evaluation:** Now a series of if/then rules which are provided in the table are applied in order to determine the fuzzy outputs. An example of firing of rules is shown below.
  - If expiry time is low and data rate is medium and queue length is high, then priority index is very low.
  - If expiry time is medium and data rate is medium and queue length is high, then priority index is low.
  - If expiry time is low and data rate is high and queue length is high, then priority index is very low.
  - If expiry time is medium and data rate is high and queue length is high, then priority index is medium.

  Since the rules are connected by an AND operation, we calculate the minimum function—that is, min {0.2, 0.3, 0.7142} = 0.2, and we cut the fuzzy set very low of the output parameter priority index at this minimum level. Similar steps are done to determine the index of other rules. The four results of output overlap and they yield overall result.
- **Defuzzification:** The results are still a fuzzy set. Therefore we have to choose the representative crisp value for getting the final output. For this purpose, the weighted average method of defuzzification is used which yields a crisp value of P = 0.175. This value of P indicates that the packets are attached with high priority and will be scheduled immediately.

PERFORMANCE EVALUATION

The simulation for evaluating the proposed fuzzy scheduler is implemented using GloMoSim library. First, the input variables used in fuzzy logic C code are identified. Then the calculated priority index is used for scheduling the data packet. By this way of scheduling, the packets that are about to expire or the packets in highly congested

queues are given first priority for sending. As a result of this, the number of packets delivered to the client node, the end-to-end delay of the packet transmission, and the throughput improve.

The inputs to the fuzzy system are identified by a complete search of the GloMoSim environment. The input expiry time is the variable TTL, which is present in the network layer of the simulator. TTL stands for "time to live". If the packet suffers excessive delays and undergoes multi-hop, its TTL falls to zero. As a result of this, the packet is dropped. If this variable is used as an input to the scheduler for finding the priority index, a packet with a very low TTL value is given the highest priority. Hence due to this, the dropping of packets experiencing multi-hops gets reduced. The next input to the scheduler is the data rate of transmission and it is normalized. The third input to the scheduler is the queue length of the node in which the packet is present. If the packet is present in a highly crowded node, it suffers excessive delays and gets lost. So, such a packet is given a higher priority and hence it gets saved.

The priority index is calculated with the inputs obtained from the network layer. This is then added to the header associated with the packet. Hence whenever the packet reaches a node, its priority index is calculated and it is attached with it. Each node has three queues. Each queue in the node is sorted based on the priority index, and the packet with the lowest priority index (i.e., packet with the highest priority) is scheduled next, when the node gets the opportunity to send. By this method of scheduling, the overall performance increases.

Simulation Environment and Methodology

The simulation package GloMoSim is used to analyze and evaluate the performance of the proposed fuzzy scheduler. The GloMoSim (GLObal MObile information system SIMulator) provides a scalable simulation environment for wireless network systems. It is designed using the parallel discrete event simulation capability provided by PARSEC (PARallel Simulation Environment for Complex Systems) (Bargodia et al., 1999). It is a C-based simulation language developed by the parallel computing laboratory at UCLA (n.d.) for sequential and parallel execution of discrete event simulation model.

A network of mobile nodes is modeled and placed randomly within a 1000x1000-meter area. There were no network partitions throughout the simulation. Each simulation is executed for 600 seconds of simulation time. Transmission range is chosen to 250 meters. Multiple runs with different seed values were conducted for each scenario, and collected data was averaged over those runs. A free space propagation model was used in our experiments. A traffic generator was developed to simulate CBR sources. Data sessions with randomly selected sources and destinations were simulated. Each source transmits data packets at a minimum rate of four packets/second and a maximum rate of 10 packets/second. The data payload was chosen to be 512 bytes/second.

Performance Metrics

The *packet delivery ratio* is the ratio of the number of data packets actually delivered to the destinations to the number of data packets supposed to be received.

The *average end-to-end delay* indicates the end-to-end delay experienced by packets from source to destination. This includes the route discovery time, the queuing delay at node, the retransmission delay at the MAC layer, and the propagation and transfer time in the wireless channel.

Throughput is measured in bytes per second and serves as an effective performance metric.

The performance of the network with the fuzzy code (FLPS) and without the code (WS) is studied under and the routing protocols used in the simulator. The results are shown in Table 2 (Gomathy & Shanmugavel, 2004), showing that

Table 2.

PACKETS DELIVERED - FOR UNICAST PROTOCOLS, WITH FLPS AND WS

Routing Protocol		Packets Delivered	
		FLPS	WS
1.	AODV	33155	21818
2.	DSR	34503	21676
3.	WRP	32183	28373

THROUGHPUT - FOR UNICAST PROTOCOLS WITH FLPS AND WS

Routing Protocol		Throughput	
		FLPS	WS
1.	AODV	263347	233513
2.	DSR	269890	164680
3.	WRP	263841	225320

END TO END DELAY - FOR UNICAST PROTOCOLS WITH FLPS AND WS

Routing Protocol		Average End to end delay	
		FLPS	WS
1.	AODV	0.97	1.127
2.	DSR	0.09	1.49
3.	WRP	0.3	0.571

the proposed scheduler works well with the three routing protocols.

Scheduler Performance with Different MAC Layer Protocols: IEEE 802.11, CSMA, MACA

Experiments were performed to check the performance of the scheduler with different MAC layer protocols such as IEEE 802.11, CSMA, and MACA. In the IEEE 802.11 protocol, each node maintains the scheduling table by overhearing all the RTSs and CTSs transmitted by other nodes within its broadcast range. Here an acknowledgment (ACK) of transmission is required after successful reception of data packet. In CSMA, if the transmission medium is in use, the node waits. It is limited by the hidden and exposed

Figure 2. Packet delivery ratio vs. number of nodes

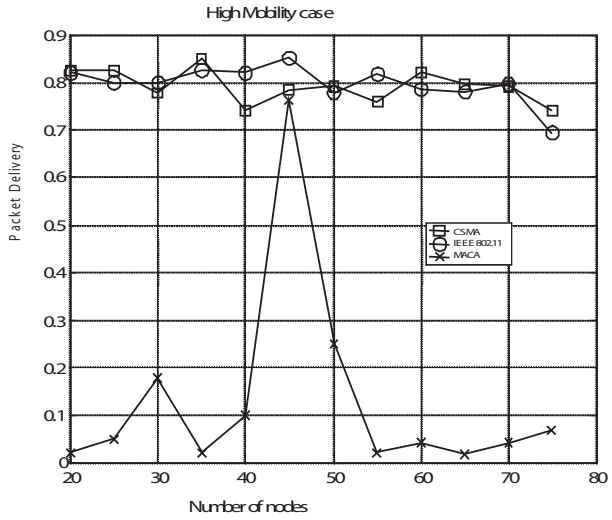
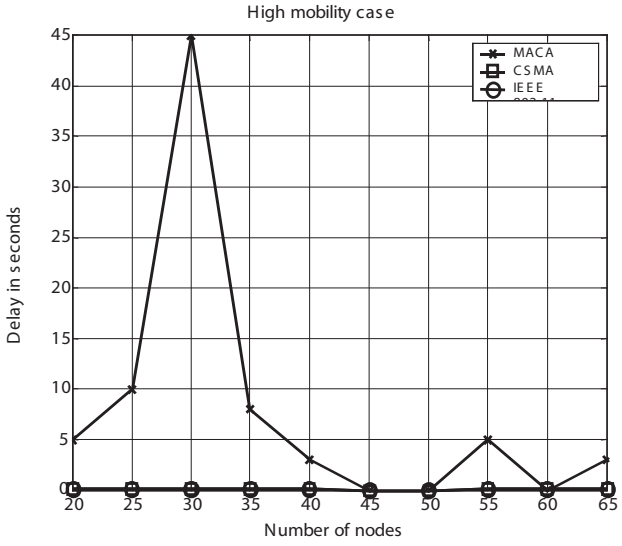


Figure 3. Average end-to-end delay as a function of number of nodes



terminal problem. This can be solved by the use of RTS/CTS dialogue for collision avoidance. Hence IEEE 802.11 always shows a better performance as seen in Figures 2 and 3, compared to CSMA and MACA protocols (Gomathy & Shanmugavel, 2004). The collision avoidance mechanism in IEEE 802.11 aids in reducing the number of collisions, and hence more data packets reach the destination. Also in an exposed terminal scenario, both CSMA and MACA present poor performance behavior.

Scheduler Performance with Mobility: Two Nodes Transmitting at the Same Time to the Same Node

When two different nodes transmit at the same time to the same node with CSMA, less than half of the number of total packets is received by the receiving node due to collision. This scenario is presented for both MACA and IEEE 802.11 protocols. It is seen that a better behavior is obtained with IEEE 802.11. When used along with the scheduler, the performance with respect to throughput, packet delivery ratio, and delay improves further. The results are proved by experimenting with mobility changes under random waypoint condition in GloMoSim and are plotted. It is clear from the results that a fuzzy scheduler performs well with two nodes transmitting to the same node. Packet delivery ratio of the network with scheduler improves by 2-5%, as the mobility of the nodes varies from low to high range. The results are again verified for varieties of combinations of nodes, and results are averaged out (Gomathy & Shanmugavel, 2004). Similarly, there is an increase in throughput. There is a marked reduction in delay, which measures as low as 0.02 seconds under high mobility.

Variation in Network Size

In this simulation, the node mobility is set at 1m/s and network traffic load is made relatively heavy. The routing protocol is chosen to be AODV. Now the impact of node density on scheduler performance is studied. The packet delivery ratio is much improved as compared with that of one without scheduler, as seen in Figure 4. It is also seen that for lighter loads, the inclusion of the

Figure 4. Packet delivery ratio as a function of network size

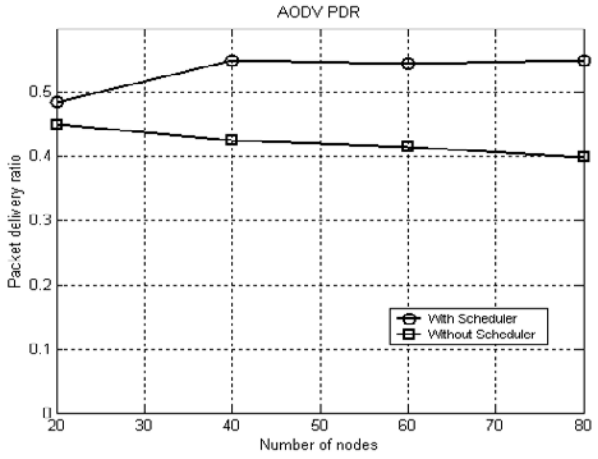
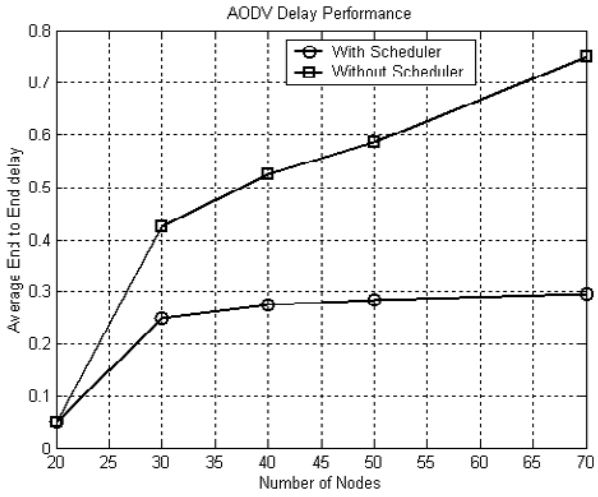


Figure 5. Average end-to-end delay vs. network size



scheduler does not provide much improvement. But as traffic load increases, marked improvement is noticed in the performance. The average end-to-end delay performance proves that the end-to-end delay improves by 0.4 seconds when scheduler is included, as seen in Figure 5. This performance is achieved by the crisp calculation of the priority index including TTL as one of the inputs.

Variation in Mobility

In this simulation, each node is moved constantly with a predefined speed. The random waypoint mobility model is chosen for this study. The node movement speed or the mobility of the nodes is varied from 0 m/s to 10 m/s. The number of nodes is set as 30 and the routing protocol is selected to be DSR. From the results in Figure 6, it is evident that the packet delivery ratio is at the higher side for the network with scheduler. Even though the delivery ratio reduces as mobility approaches 10 m/s, there is always an increase of 10% in the performance of fuzzy scheduler (Gomathy & Shanmugavel, 2004)

In the fuzzy scheduler, there is a slight degradation in performance as the number of nodes increases above 70. This is due to the increase in number of hops the packets have to take to reach the destination. But still, the end-to-end delay is much smaller compared to that of the network without the scheduler. It can be inferred from Figure 7 that the fuzzy scheduler provides a superior performance in terms of the end-to-end delay. As the mobility varies from 0-10 m/s, the fuzzy scheduler provides an end-to-end delay reduced by around 0.1 sec. to 0.2 sec. The performance of the fuzzy scheduler is also tested by varying the pause time, for RWP mobility model, using the AODV routing algorithm.

The results prove that the network performs better when FPLS is included. Increasing pause time results in smoother transitions and hence

Figure 6. Packet delivery ratio vs. mobility

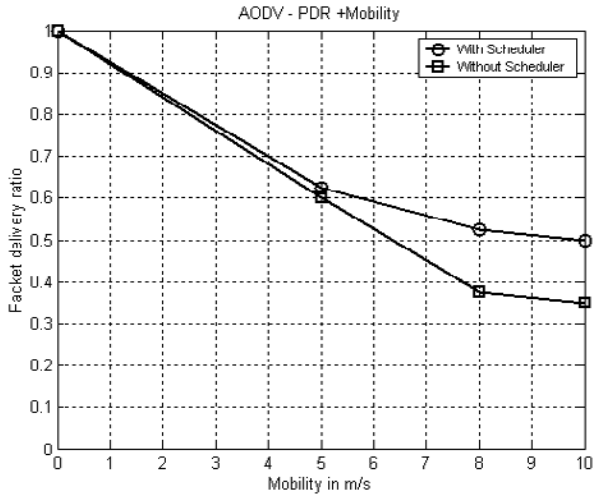
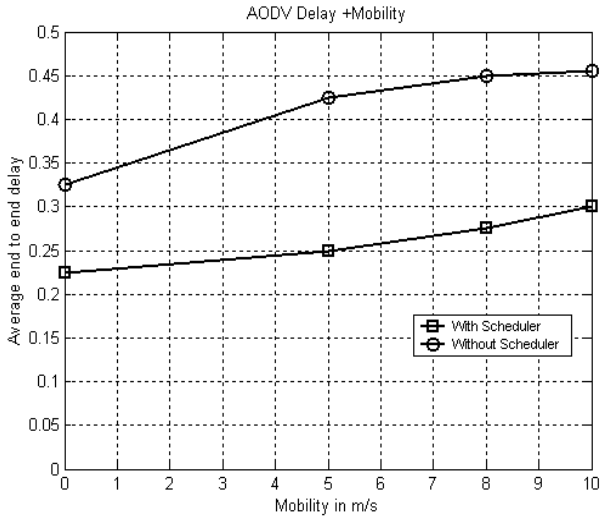


Figure 7. Average end-to-end delay vs. mobility of nodes



improves the performance parameters. As the mobility model is changed to random walk, with uniformly distributed speed, the throughput packet delivery ratio reduces a little and the end-to-end delay increases slightly. This is because the random walk mobility model is characterized by

the abrupt transitions in the direction and speed of nodes.

Scheduler Performance for Multicasting Protocols

Multicasting routing and packets forwarding in ad hoc networks is a fairly unexplored area. In today's network, data transmission between multiple senders and receivers is becoming increasingly important. Three multicast protocols are considered here for testing the proposed scheduler—ODMRP(On-Demand Multicast Routing Protocol), CAMP, and NTPMR (Node Transition Probability-based Multicast Routing) (Gomathy & Shanmugavel, 2005).

Variations in Mobility

In the mobility experiment, 20 nodes are multicast members and five sources transmit packets at the rate of two packets per second each. It is evident from the results that NTPMR provides higher packet delivery ratio as compared to ODMRP and CAMP. This is because NTPMR enables packets to travel distant destinations since a packet is sent to different neighbors during repeated encounters with a node. It is now proposed to include the fuzzy scheduler for these three protocols and test whether there is any improvement in packet delivery ratio. The packet delivery ratio (PDR) increases for all the three protocols. Hence it is verified that even at high mobility speeds, the routing protocols could be used when the fuzzy priority scheduler is included in these routing agents.

Multicast Group Size

The number of multicast members was varied to investigate the scalability of the protocol. The number of senders was fixed at five; the mobility speed at 1 m/s, network traffic rate at 10 packets per second, and the multicast group size was varied from 5 to 20 members. The routing effectiveness of the protocol as a function of multicast group size is now compared. For NTPMR, the packet delivery ratio is found to remain constant with increase in group size. Here the routing of packets does not depend on any forwarding group. CAMP performs better as the number of groups increases. Since the mesh becomes more massive with the growth of members, more redundant routes are formed. In ODMRP, as the number of receivers increases, the number of forwarding group nodes increases; this in turn increases the connectivity. With these results, the fuzzy scheduler is inserted in between the MAC layer and routing agent. The simulation is run and the results are presented. As seen from results, the NTPMR shows an increased performance of 3%. This is again due to the fact that, with already existing best performance, as the data scheduler is added, the packets at the verge of expiry are scheduled immediately. This increases the PDR by 3%. For ODMRP, the scheduler PDR characteristics are closer to the one without scheduler. Again in CAMPamp, the PDR improves by 5% due to the proper selection of a priority index. Thus it is also verified that the proposed scheduler performs well with multicast protocols (Gomathy & Shanmugavel, 2005).

CONCLUSION

This chapter addresses a fuzzy-based priority scheduling scheme, which improves the quality of service parameters in mobile ad hoc networks. The fuzzy scheduler algorithm attaches a priority index to each data packet in the queue of the node. It combines the input parameters such as queue length, data rate, and expiry time to find the priority index. The crisp priority index is calculated by the fuzzy scheduler based on the above inputs, which are derived from the network. The membership functions and rule bases of the fuzzy scheduler are carefully designed. The coding is done in C language and output is verified using MATLAB fuzzy logic toolbox

with FIS editor. Then the inputs are identified in the library of GloMoSim and the fuzzy scheduler is attached.

In this chapter, the performance of the fuzzy scheduler is studied for mobile ad hoc networks using GloMoSim simulator. It is found from the results that priority scheduling helps in effective routing of packets without much loss and with less delay. In a real network environment, where timely reception of each packet plays a crucial role, priority scheduling helps in effective transmission of packets. Based on the studies, we conclude that the proposed fuzzy-based scheduling algorithm performs better compared with the network performance without scheduler. The results are also verified for different unicast and multicast routing protocols under different mobility conditions and network size, with IEEE 802.11 as MAC protocol.

The future extension of the work could be to include the mobility rate, number of nodes in the transmission range, channel state conditions, and fairness among sources as inputs to the fuzzy scheduler, and investigate the effect on the overall performance of the network.

REFERENCES

Bargodia, R., Meyer, R., Takai, M., Chen, Y., Zeng, X., Martin, J., & Song, H. Y. (1999). PARSEC: A parallel simulation environment for complex systems. *IEEE Computers, 31*(10), 77-85.

Bettsetter, C. (2001, July). Smooth is better that sharp: A random mobility model for simulation of wireless networks. *Proceedings of the 4th ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Rome, Italy.

Camp, T., Boleng, J., & Davies, V. (2002). A survey of mobility models for ad hoc network research, wireless communication and mobile computing (WCMC). *Trends and Applications, 2*(5), 483-502.

Chun, B., & Baker, M. (2002). Evaluation of packet scheduling algorithms in mobile ad hoc networks. *ACM Mobile Computing and Communication Review, 6*(3).

Das, S. R., Castaneda, R., Yan, J., & Sengupta, R. (1998). Comparative performance evaluation of routing protocols for mobile ad hoc networks. *Proceedings of the 7th International Conference on Computer Communications and Networks* (pp. 153-161).

Das, S. R., Perkins, C. E., & Royer, E. M. (2001). *IEEE Personal Communications Magazine, 8*(1), 16-29.

Gomathy, C., & Shanmugavel, S. (2004a, February 29-March 4). Fuzzy based priority scheduler for mobile ad hoc networks. *Proceedings of the 3rd International Conference on Networking*, Gosier, Guadeloupe.

Gomathy, C., & Shanmugavel, S. (2004b, March). An efficient fuzzy based priority scheduler for mobile ad hoc networks and performance analysis for various mobility models. *Proceedings of the IEEE Wireless Communication and Networking Conference*, Atlanta, GA.

Gomathy, C., & Shanmugavel, S. (2004c, December 11-14). Effect of packet scheduling and evaluation of fuzzy based priority scheduler on ad hoc network unicast communication. *Proceedings of the IEEE International Conference on Signal Processing and Communication*, Bangalore, India.

Gomathy, C., & Shanmugavel, S. (2004d, December 15-18). Performance evaluation of a novel fuzzy based priority scheduler for mobile ad hoc networks and its effect on MAC protocols. *Proceedings of the 12th International Conference on Advanced Computing and Communication*, Ahmedabad, India.

Gomathy, C., & Shanmugavel, S. (2005a, January 23-25). Design of a priority scheduler using fuzzy logic and the performance analysis with multicast routing protocols. *Proceedings of the IEEE International Conference on Personal Wireless Communication,* New Delhi, India.

Gomathy, C., & Shanmugavel, S. (2005b, January 28-30). Implementation of modified fuzzy priority scheduler for MANET and performance analysis with mixed traffic. *Proceedings of the 11th National Conference on Communication,* Kharagpur, India.

Gomathy, C., & Shanmugavel, S. (2005c). Performance evaluation of a novel fuzzy based priority scheduler for mobile ad hoc networks and its effect on MAC protocols. *International Journal of Information Technology, 4*(1), 78-86.

Jardesh, A., Royer, E. M., Kelvin, C., Almeroth, & Suri, S. (2003). Towards realistic mobility models for mobile ad hoc networks. *Proceedings of MOBICOM 2003,* San Diego, CA.

Kanodia, V., Li, C., Sabharwal, A., Sadeghi, B., & Knightly, E. (2002). Distributed priority scheduling and medium access in ad hoc networks. *ACM Wireless Networks, 8*(1).

Lin, G., Noubir, G., & Rajaraman, R. (2004). Mobility models for ad hoc network simulation. *Proceedings of IEEE INFOCOM 2004.*

Luo, H., Lu, S., & Bhargavan, V. (2000, August). A new model for packet scheduling in multi hop wireless networks. *Proceedings of ACM MobiCom'00,* Boston.

Rea, S., & Pesch, D. (2004, September). Multimetric routing decisions for ad-hoc networks using fuzzy logic. *Proceedings of the 1st IEEE International Symposium on Wireless Communication Systems,* Mauritius.

Royer, E. M., & Toh, C. (1999). A review of current routing protocols for ad hoc networks. *IEEE Personal Communication, 6*(2), 46-55.

UCLA. (n.d.). *Parallel Computing Laboratory and Wireless Adaptive Mobility Laboratory: GloMoSim, a scalable simulation environment for wireless and wired network systems.* Retrieved from http://pcl.cs.ucla.edu/projects/domains/glomosim.html

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 308-321, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.31
Energy-Efficient Cache Invalidation in Wireless Mobile Environment

R. C. Joshi
Indian Institute of Technology Roorkee, India

Manoj Misra
Indian Institute of Technology Roorkee, India

Narottam Chand
Indian Institute of Technology Roorkee, India

ABSTRACT

Caching at the mobile client is a potential technique that can reduce the number of uplink requests, lighten the server load, shorten the query latency and increase the data availability. A cache invalidation strategy ensures that any data item cached at a mobile client has same value as on the origin server. Traditional cache invalidation strategies make use of periodic broadcasting of invalidation reports (IRs) by the server. The IR approach suffers from long query latency, larger tuning time and poor utilization of bandwidth. Using updated invalidation report (UIR) method that replaces a small fraction of the recent updates, the query latency can be reduced. To improve upon the IR and UIR based strategies, this chapter presents a synchronous stateful cache maintenance technique called Update Report (UR). The proposed strategy outperforms the IR and UIR strategies by reducing the query latency, minimizing the disconnection overheads, optimizing the use of wireless channel and conserving the client energy.

INTRODUCTION

The tremendous growth in mobile hardware technology and wireless communication has increased the number of clients that access data remotely. Efficient data access in mobile computing is a

field of increasing importance for a wide range of mobile applications. Users of mobile devices wish to access dynamic data, such as stock quotes, news items, current traffic conditions, weather reports, e-mail, and video clips via wireless networks. However, limited battery power of mobile client and scarce wireless bandwidth hinder the full realization of ubiquitous data access in mobile computing. Caching at the mobile client can relieve bandwidth constraints imposed on wireless mobile computing. Copies of remote data can be kept in the local memory of the mobile client to substantially reduce user requests for retrieval of data from the origin server. This not only reduces the uplink and downlink bandwidth consumption, but also the average query latency. Caching frequently accessed data by a mobile client can also save its power used to retrieve the repeatedly requested data.

Cache invalidation strategy is used to ensure that the data items cached at a mobile client are consistent with those stored on the server. Depending on whether or not the server maintains the state of the mobile client's cache, the invalidation strategies are divided into two categories: the *stateful* server approach and the *stateless* server approach (Barbara & Imielinski, 1994; Tan, Cai, & Ooi, 2001). Barbara and Imielinski (1994) provide a solution where the server periodically broadcasts an invalidation report (IR) in which the changed data items are indicated. Rather than querying the server directly regarding the validation of cached copies, the clients can listen to these IRs over the wireless channel and use them to validate their local cache. The IR-based invalidation may be of two types: *synchronous* and *asynchronous.* In the synchronous method, the invalidation reports are broadcast periodically, whereas in the asynchronous method, the server broadcasts the reports only when some data changes. Because of the nature of periodic broadcast, synchronous methods provide a bound on the waiting time of the next report, whereas in an asynchronous invalidation report, there is no guarantee on how long the client must wait.

Clients use IRs to keep their cache consistent by discarding any obsolete data. If a query cannot be served locally—that is, a cache miss—the client issues an uplink query request for the data items. The IR-based solution is attractive because of its scalability, as the size of IR is independent of the number of clients. It is also energy efficient, as clients can exploit the periodicity of server broadcast to save power, in that mobile devices can operate in doze mode most of the time and only activate during broadcast. However, the solution suffers from the problem of long query latency since a client must listen to the next IR before answering a query. The problem has been tackled with the addition of *updated invalidation report (UIR)* by broadcasting a number of smaller reports (UIRs) between successive IRs (Cao, 2001, 2002a, 2002b, 2003). Each UIR contains information about most recently updated data items since the last IR. In case of cache hit, there is no need to wait for the next IR and hence the query latency is reduced. However, if there is a cache miss, the client still needs to wait for the data to be delivered. Thus, due to cache miss, the UIR strategy has the same query latency as IR strategy.

In IR strategy, if the disconnection time of a client is longer than a fixed period, the client should discard its entire cache even if some of the cached data may still be valid. This issue is addressed in Cao (2002a, 2002b), and Jing, Elmagarmid, Helal, and Alonso (1997). Chand, Joshi, and Misra (2005) have demonstrated more efficient handling of arbitrarily long client disconnection.

To overcome the limitations of existing cache invalidation strategies, we present a synchronous stateful caching strategy where cache consistency is maintained by periodically broadcasting *update reports (URs)* and *request reports (RRs).* The central design of our strategy includes reducing the query latency, improving the cache hit ratio, minimizing the client disconnection overheads, utilizing the wireless channel better, and conserving the client energy. The track of cached

items for each client is maintained at the home mobile support station in the form of *cache state information (CSI*). Use of CSI reduces the size of IR by filtering out non-cached items and handles long disconnection. In various IR-based strategies (Kahol, Khurana, Gupta, & Srimani, 2001; Jing et al., 1997; Barbara & Imielinski, 1994; Chuang & Hsu, 2004), even though many clients cache the same updated data item, all of them have to query the server and get the data separately from the server. It wastes a large amount of wireless bandwidth and client battery energy. To minimize uplink requests and downlink broadcasts, we use a broadcast strategy, called update report (UR) (Chand et al., 2005), where all the recently updated/requested items are broadcast immediately after the invalidation report (IR). To further reduce query latency, the strategy uses *request reports (RRs)*, where all the recently requested items are broadcast after the UIR. Selective tuning is used to conserve the client energy.

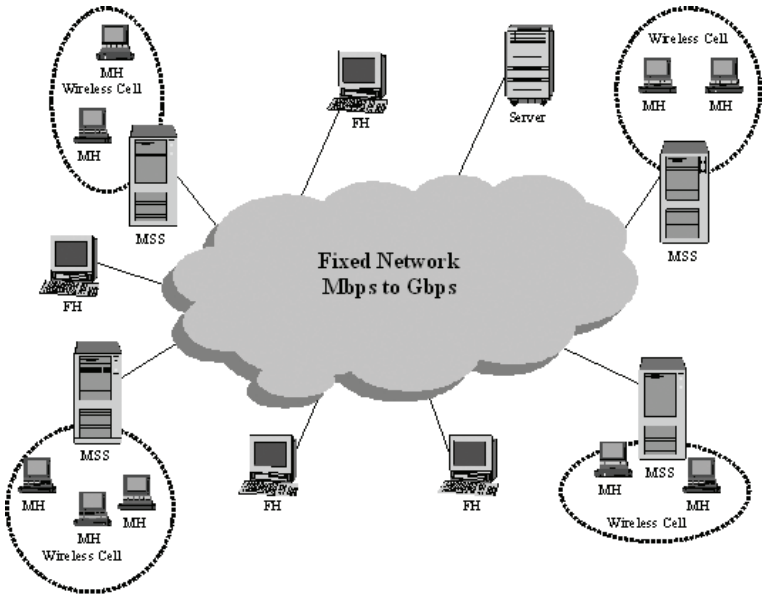
THE PROPOSED CACHE INVALIDATION STRATEGY

In this section, we present our UR-based synchronous stateful caching strategy.

UR Caching Model

As shown in Figure 1, the model consists of two distinct sets of entities: *Mobile Hosts (MHs)* and *Fixed Hosts (FHs).* Some of the fixed hosts, called *Mobile Support Stations (MSSs),* are augmented with a wireless interface in order to communicate with the mobile hosts, which are located within a radio coverage called a cell. Each cell is associated with an id for identification purpose. MSSs are also known as *Base Stations (BSs).* An MSS acts like a gateway between a fixed network and a wireless network. An MH communicates with a fixed host/server via an MSS over a wireless

Figure 1. UR caching model



communication link. The communication is asymmetric (i.e., the uplink bandwidth is much less than that of downlink). The MSSs communicate among themselves over a wired channel and the communication is transparent to a client. A fixed network has a large bandwidth (order of Mbps or Gbps), while the bandwidth of the wireless channel is low (19.2 Kbps-10 Mbps). An MH can move within a cell or between cells while retaining its network connection. When an MH moves from one cell to another (called handoff), its wireless connection is switched to the new cell. An MH either connects to an MSS through a wireless link or disconnects from the MSS by operating in a 'power save' mode (Kahol et al., 2001).

The database D is a collection of N data items with ids: d_1, d_2, …, d_N. A data item is the basic unit for update and query. For each data item d_i, two timestamps t_i and t_i^r are maintained: t_i is the most recent *timestamp* when d_i got updated at the server and t_i^r, called *latest request time*, represents the most recent time when d_i was last requested by any client. MHs only issue simple requests to read the most recent copy of a data item. In order to serve a request sent from a client, the MSS needs to communicate with the database server to retrieve the data items. Caching techniques may also be applied at MSS. Since the communication between the database server and MSSs are through wired link, we assume traditional techniques can be used to maintain cache consistency. Since the communication between the MSS and the database server is transparent to the clients, from the client point of view, the MSS is the same as the database server.

Frequently accessed data items are cached on the client side. We assume that the cache at the mobile client is a nonvolatile memory such as a hard disk so that after a long disconnection, the contents of the cache can still be retrieved. When caching is used, data consistency issues must be addressed. We assume the *latest value* consistency model (Cao, 2002a, 2002b), which is widely used in dissemination-based information systems.

To ensure cache consistency, the server broadcasts UR every L seconds and it also broadcasts (m-1) RRs between two URs. Every active client listens to the report (UR/RR) and invalidates its cache accordingly. To answer a query, the client listens to the IR/UIR part of the next report (UR/RR) and decides its cache validity. If there is a valid cached copy of the requested item, the client returns the item immediately. Otherwise, it sends a query request to the server through the uplink. The simulation architecture of the proposed model is shown in Figure 2.

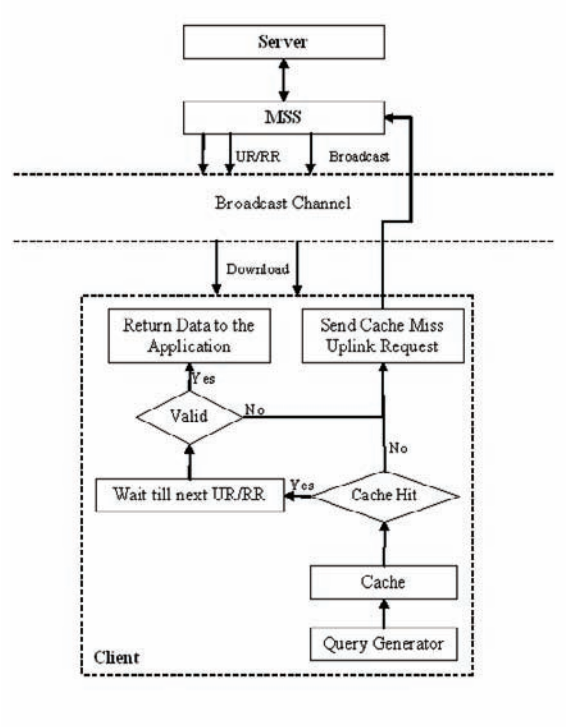
To keep the information about cached items for each MH, a cache state information (CSI) is maintained at the home MSS. The CSI is the list of cached data item ids by the host. For each item d_i a cache count n_i is also maintained at the home MSS. Thus, n_i denotes the number of clients who have cached the item d_i in that particular cell. When a client sends a data request, the MSS updates the relevant counters and the corresponding CSI, and forwards the request to the server.

In order to save energy, an MH may power off most of the times and only turns on during the report broadcast time. Moreover, an MH may be in the power save mode for a long time and it may miss some reports.

The following assumptions are made:

- Database D is a collection of N data items. An item is identified by a unique id d_i ($1 \le i \le N$). D_i denotes the actual data of an item with id d_i. Each item has the same size S_{data} (in bits).
- Each cell has a single MSS such that cell A is managed by MSS_A. Each MSS broadcasts UR every L seconds and RR every L/m seconds.
- A unique host identifier is assigned to each MH in the system. The system has a total of M hosts, and MH_i ($1 \le i \le M$) is a host identifier. Each mobile host moves freely. We use the terms host and client interchangeably.

Figure 2. Simulation architecture



- Each MH has a cache space for C data items.
- CSI stored in the local disk of home MSS maintains the state information for a host. An MH informs its MSS before it stores any data item in its local cache and the MSS updates the CSI accordingly.
- The server is reliable—that is, it handles the failure with some fault tolerance techniques.

Cache State Information to Reduce the Report Size

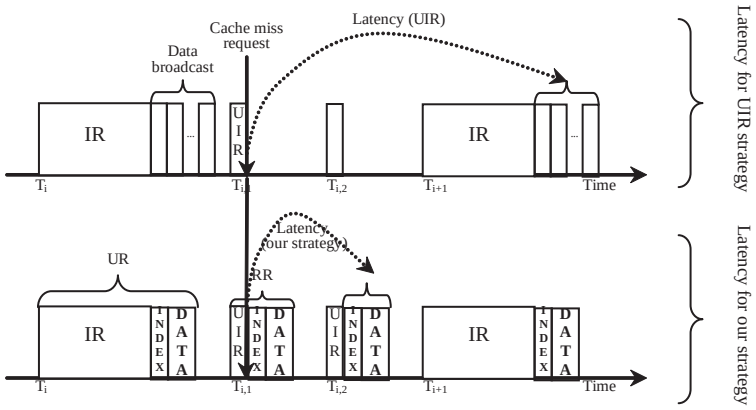
In a stateless strategy when an item updates at the server, its id is broadcast as part of IR irrespective of whether the item has been cached or not. Including an item as part of IR, which has not been cached by any client, makes poor utilization of the available wireless bandwidth. It also increases the client energy consumption since users have to listen to the broadcast channel for a longer duration to download the report. To filter out from a report all those recently updated items that are not cached by any client, we have used a stateful approach in our strategy.

To keep the information about cached items, for each MH a *cache state information (CSI)* is maintained at the home MSS. Consider a cell with H hosts (MH_i, $1 \le i \le H$), at any given time. For any j, CSI_j for MH_j, as maintained on its home MSS, keeps track of what data has been locally cached at MH_j. In general, if $d_k \in CSI_j$, then the client MH_j has cached the item d_k. When an item updates, it will be added as part of a report (UR/RR) at the server. The MSS, upon receiving the report from the server, removes all those items from the IR that are not cached by any client (i.e., with cache count 0) and thus broadcasts reduced report in its cell. When a client moves to a new cell, the copy of its CSI is replicated at the new MSS.

Reducing the Query Latency

In UIR-based caching strategy, the server aggregates data requests from all its clients over the whole invalidation interval (L seconds) and broadcasts the requested data after each IR. This aggregation of requests tremendously reduces the number of data broadcasts and thus makes efficient utilization of the downlink channel. The reduction in the number of broadcasts is at the expense of increased query latency, since a client has to wait longer to download the requested data item. In UIR scheme, the requested data are scheduled for broadcast after the next IR, thus due to cache miss, the expected query latency is L/2 seconds. To reduce the query latency due to cache miss, the UR strategy broadcasts the recently requested data items after the next report

Figure 3. Reducing the query latency



(IR/UIR), whichever arrives earlier, such that the expected query latency is L/(2*m) seconds instead of L/2 seconds. In general, $T_{i,k}$ represents the time of k^{th} RR after the i^{th} UR. When a client receives a cache miss request between $T_{i,1}$ and $T_{i,2}$, it cannot answer the query until T_{i+1} in the UIR approach, but it can answer the query at $T_{i,2}$ in UR approach (see Figure 3). UIR, followed by the broadcast of recently requested data, constitutes *request report (RR).* At interval time $T_{i,k}$, $RR_{i,k}$ is constructed as follows:

$UIR_{i,k}$	$RR\_INDEX_{i,k}$	$RR\_DATA_{i,k}$

$$UIR_{i,k} = \{d_x | (d_x \in D) \wedge (T_{i,0} < t_x \leq T_{i,k}) \wedge (n_x > 0)\} \; (0 < k < m)$$

$$RR\_INDEX_{i,k} = \{d_x | (T_{i,k-1} < t^r_x \leq T_{i,k})\}$$

$$RR\_DATA_{i,k} = \{D_x | d_x \in RR\_INDEX_{i,k}\}$$

This distribution of query replies also reduces the impact of data broadcast on other downlink traffic. To make the selective tuning possible for the clients, the server broadcasts the index information RR_INDEX before the broadcast of actual data. Since the query replies are distributed, the size of the index in our strategy is much smaller than in the UIR strategy.

Improving Wireless Channel Utilization

To reduce the number of uplink requests and downlink broadcasts, we introduce the concept of update report (UR) (Chand et al., 2004). Update reports (URs) are broadcast synchronously with period L.

At interval T_i, the structure of UR_i is as follows:

IR_i	UR_INDEX_i	UR_DATA_i

$$IR_i = \{(d_x, t_x) | (d_x \in D) \wedge (n_x > 0) \wedge (T_i - w*L < t_x \leq T_i)\}$$

$$UR\_INDEX_i = \{d_x | ((T_{i-1} < t_x \leq T_i) \wedge (n_x > 0)) \vee (T_{i-1,m-1} < t^r_x \leq T_i)\}$$

$$UR\_DATA_i = \{D_x | d_x \in UR\_INDEX_i\}$$

UR_INDEX_i defines the order in which data appears in UR_DATA_i. Within UR_INDEX and RR_INDEX, the items are arranged in non-decreasing order of their cache count. This ordering of broadcast items further reduces the query latency (Chand et al., 2004).

IR contains the update history of past w broadcast intervals, whereas UR_DATA contains the actual data value for the items that have been updated during previous UR interval and the items that have been requested during the last RR interval. In our strategy the contents of URs broadcast in different cells depend upon the cache state of the clients lying within a cell, and hence the broadcast URs may be inhomogeneous. In most IR-based algorithms (Kahol et al., 2001; Jing et al., 1997; Barbara & Imielinski, 1994), updating a data item that has been cached may generate many uplink requests and downlink broadcasts, and thus make poor utilization of available wireless bandwidth. This is due to the reason that when an item is updated and IR is broadcast, each client who has cached that item will generate an uplink request for the item and the server responds to each request by broadcasting the item. For example, for an item with id d_x which is cached by n_x clients, there will be n_x uplink requests and n_x downlink broadcasts due to update. We address the problem by asking the server to broadcast all the data items that have been recently requested or updated and are cached by one or more clients. If a client observes that the server is broadcasting an item which is an invalid entry in its local cache, it will download the item. Otherwise, the client may have to send another request to the server, and the server will have to broadcast the data again in the future. So in comparison to n_x uplink requests and downlink broadcasts for an updated item, our strategy makes only single broadcast without any uplink request.

Due to data update at the server, UR strategy has same number of uplink requests and downlink broadcasts as in UIR strategy. Also, during one RR interval, due to cache miss an item may have been requested by many clients, but our scheme broadcasts the item only once. In comparison to UIR strategy, our strategy further improves the wireless channel utilization by using *delayed uplink (DU)* technique as follows.

To improve the efficiency of uplink channel due to cache miss, when an item d_x is requested by a client at time $T_{i,j}$ $(1 \le j \le m\text{-}1)$, in UR strategy the client would download $RR\_INDEX_{i,j}$ to see if the server has planned the broadcast of item d_x. If $d_x \in RR\_INDEX_{i,j}$, the client would download D_x and the item id d_x is piggybacked when a new request is sent to the server; otherwise the client sends an uplink request to the server for d_x. This saving in uplink request also reduces query latency as the client receives the item sooner. Thus, reducing the number of uplink requests and downlink broadcasts due to recent updates or cache misses, UR strategy heavily saves on wireless bandwidth.

Synchronous Broadcasting to Conserve Client Energy

In asynchronous invalidation methodology, there is no guarantee on how long the client must wait for the next report, and hence the clients are in doze mode and may lose some of the reports, thus compromising the cache consistency or further increasing the query latency. By broadcasting UR and RR periodically, we use a synchronous approach where clients may wake up during the UR/RR broadcast time and selectively tune in to the channel to save power. After broadcasting IR/UIR, the server broadcasts UR_INDEX/RR_INDEX followed by the broadcast of actual data UR_DATA/RR_DATA. Every client listens to the report (IR/UIR) if not disconnected. At the end of report, the client downloads index and locates the interesting item that will come, and listens to the channel at that time to download the data. This strategy saves power since the client selectively tunes to the channel and can stay in doze mode most of time.

Handling Client Disconnection

Since a UR broadcasts information about the items that have been updated during past w*L time, our strategy handles the disconnection of clients less than w*L without any additional overhead. When a client reconnects after a disconnection time longer than w*L, it sends an uplink request with the last received UR time stamp T_l (before disconnection) to the home MSS. On receipt of the request, the MSS constructs a binary vector *DIV* called *disconnection information vector*. DIV is of size C bits and contains the validity information about the cached items by the client. For a client MH_i, the MSS constructs DIV_i as follows:

1. Scan the CSI_i for the list of cached items. If $d_j \in CSI_i$, MH_i has cached the item d_j otherwise not ($1 \le j \le N$).
2. For an item d_j which is cached by client MH_i, compare its last update timestamp (t_j) with T_l. If $t_j > T_l$, the item d_j has been updated since MH_i received the last UR before disconnection. In case t_j satisfies the above condition (i.e., $t_j > T_l$), then set $DIV_i[k] = 1$, where MH_i has stored item d_j at k^{th} cache location ($1 \le k \le C$). If $t_j \le T_l$, then set $DIV_i[k] = 0$.

Step 1 gives the list of items that have been cached by the client, and step 2 checks whether the particular cached item has been updated when the client was in disconnection mode. Step 2 is repeated for all the cached items by the client MH_i. The number of bits in DIV_i is C and is equal to the number of items cached by MH_i.

Once the DIV_i has been constructed, the server sends DIV_i to MH_i over the downlink channel. After downloading DIV_i, MH_i finds whether a particular cached item is valid or not. If $DIV_i[k] = 1$, then the k^{th} cached item is invalid, otherwise it is still valid. After checking for each cached item, the client will send an uplink request for all the invalid items, and the server responds by broadcasting the requested items during and following UR/RR.

As compared to UIR strategy (Cao, 2001, 2002a, 2002b, 2003), which handles disconnection by sending the ids for updated items, our strategy uses only one bit for an item, thus reducing the reconnection overheads tremendously. For our strategy, the reconnection overhead is C bits, which is very low as compared to UIR. Because of the smaller size of overheads, our strategy is also very much effective in terms of bandwidth utilization, client tuning time, and energy consumption.

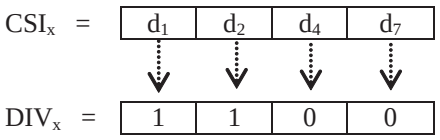
An Example

Consider a database having 10 items with the last update timestamp t_i as follows:

d_i	1	2	3	4	5	6	7	8	9	10
t_i	20	16	17	13	5	6	2	9	23	19

Consider a host MH_x of cache size C = 4 that has cached the items with id d_1, d_2, d_4, and d_7. Let MH_x be disconnected at time 17 such that it has received the last UR at $T_l = 15$ and wakes up at time 30.

Then:



While MH_x receives DIV_x, it is interpreted as: 1st cached item where d_1 is invalid, and 2nd cached item where d_2 is invalid, whereas d_4 and d_7 are still valid. The reconnection overhead for our strategy is 4 bits.

For UIR, the overhead = number of cached items invalidated during disconnection*item id size (S_{id}). Generally S_{id} = 32 bits, therefore the overhead value = 64 bits.

CONCLUSION AND FUTURE DIRECTION

This chapter investigates the cache invalidation issue in a realistic mobile environment, where there exist resource-poor mobile clients, data updates, asymmetric low-quality wireless channel, and client disconnections. UR-based cache invalidation strategy has been proposed that reduces the query latency as compared to existing IR and UIR strategies. The UR strategy employs selective tuning to conserve the client's battery power. Frequent client disconnection is one of the main features in a mobile computing environment. To cater for such an environment, a disconnection information vector (DIV) based algorithm has been proposed that maintains cache consistency at the mobile client with very low overhead as compared to existing strategies. The ad hoc mode of operation, which is now available with new-generation wireless interfaces, makes possible peer-to-peer (P2P) caching in which mobile clients can access data items from the cache in their neighboring peers. Extension of the proposed strategy to peer-enabled caching is a consideration during our future research.

REFERENCES

Barbara, D., & Imielinski, T. (1994, May 24-27). Sleepers and workaholics: Caching strategies in mobile environments. *Proceedings of the ACM SIGMOD Conference on Management of Data,* Minneapolis, MN (pp. 1-12).

Cao, G. (2001, August 6-11). A scalable low-latency cache invalidation strategy for mobile environments. *Proceedings of the ACM International Conference on Computing and Networking* (Mobicom), Massachusetts (pp. 200-209).

Cao, G. (2002a). On improving the performance of cache invalidation in mobile environments. *Mobile Networks and Applications, 7*(4), 291-303.

Cao, G. (2002b). Proactive power-aware cache management for mobile computing systems. *IEEE Transactions on Computers, 51*(6), 608-621.

Cao, G. (2003). A scalable low-latency cache invalidation strategy for mobile environments. *IEEE Transactions on Knowledge and Data Engineering, 15*(5), 1251-1265.

Chand, N., Joshi, R. C., & Misra, M. (2004, December). Broadcast based cache invalidation and prefetching in mobile environment. *Proceedings of the International Conference on High Performance Computing (HiPC)* (pp. 410-419). Berlin: Springer-Verlag (LNCS 3296).

Chand, N., Joshi, R. C., & Misra, M. (2005, January 23-25). Energy efficient cache invalidation in wireless mobile environment. *Proceedings of the IEEE International Conference on Personal Wireless Communications (ICPWC),* New Delhi, India (pp. 244-248).

Chuang, P. J., & Hsu, C. Y. (2004, March 29-31). An efficient cache invalidation strategy in mobile environments. *Proceedings of the IEEE International Conference on Advanced Information Networking and Application (AINA),* Fukuoka, Japan (pp. 260-263).

Jing, J., Elmagarmid, A., Helal, A., & Alonso, R. (1997). Bit-sequences: An adaptive cache invalidation method in mobile client/server environments. *Mobile Networks and Applications,2*(2), 115-127.

Kahol, A., Khurana, S., Gupta, S. K. S., & Srimani, P. K. (2001). A strategy to manage cache consistency in a disconnected distributed environment. *IEEE Transactions on Parallel and Distributed Systems, 12*(7), 686-700.

Tan, K. L., Cai, J., & Ooi, B. C. (2001). An evaluation of cache invalidation strategies in wireless environments. *IEEE Transactions on Parallel and Distributed Systems, 12*(8).

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unhelkar, pp. 132-141, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.32
Ensuring Serializability for Mobile-Client Data Caching

Shin Parker
University of Nebraska at Omaha, USA

Zhengxin Chen
University of Nebraska at Omaha, USA

IMPORTANCE OF ENSURING SERIALIZABILITY IN MOBILE ENVIRONMENTS

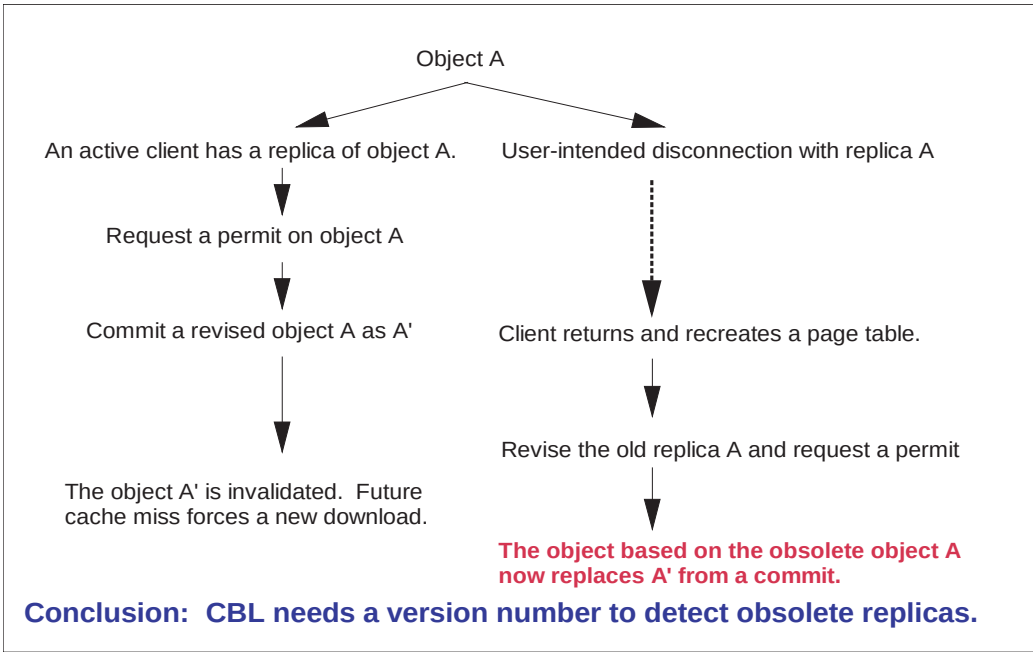
Data management in mobile computing has emerged as a major research area, and it has found many applications. This research has produced interesting results in areas such as data dissemination over limited bandwidth channels, location-dependent querying of data, and advanced interfaces for mobile computers (Barbara, 1999). However, handling multimedia objects in mobile environments faces numerous challenges. Traditional methods developed for transaction processing (Silberschatz, Korth & Sudarshan, 2001) such as concurrency control and recovery mechanisms may no longer work correctly in mobile environments. To illustrate the important aspects that need to be considered and provide a solution for these important yet "tricky" issues in this article, we focus on an important topic of data management in mobile computing, which is concerned with how to ensure serializability for mobile-client data caching. New solutions are needed in dealing with caching multimedia data for mobile clients, for example, a cooperative cache architecture was proposed in Lau, Kumar, and Vankatesh (2002). The particular aspect considered in this article is that when managing a large number of multimedia objects within mobile client-server computing environments, there may be multiple physical copies of the same data object in client caches with the server as the primary owner of all data objects. Invalid-access prevention policy protocols developed in traditional DBMS environment will not work correctly in the new environment, thus, have to be extended to ensure that the serializability involving data updates is achieved in mobile environments. The research by Parker and Chen (2004) performed the analysis, proposed three extended protocols, and conducted experimental studies under the invalid-access prevention policy

in mobile environments to meet the serializability requirement in a mobile client/server environment that deals with multimedia objects. These three protocols, referred to as extended server-based two-phase locking (ES2PL), extended call back locking (ECBL), and extended optimistic two-phase locking (EO2PL) protocols, have included additional attributes to ensure multimedia object serializability in mobile client/server computing environments. In this article, we examine this issue, present key ideas behind the solution, and discuss related issues in a broader context.

BACKGROUND

In a typical client-server computing architecture, there may exist multiple physical copies of the same data object at the same time in the network with the server as the primary owner of all data objects. The existence of multiple copies of the same multimedia object in client caches is possible when there is no data conflict in the network. In managing multiple clients' concurrent read/write operations on a multimedia object, no transactions that accessed the old version should be allowed to commit. This is the basis of the invalid-access prevention policy, from which several protocols have been proposed. The purpose of these protocols is to create an illusion of a single, logical, multimedia data object in the face of multiple physical copies in the client/server network when a data conflict situation arises. When the server becomes aware of a network-wide data conflict, it initiates a cache consistency request to remote clients on behalf of the transaction that caused the data conflict. Three well-known invalid-access prevention protocols are Server-based Two-Phase

Figure 1. CBL failure analysis tree in a mobile environment



Locking (S2PL), Call-Back Locking (CBL), and Optimistic Two-Phase Locking (O2PL).

In order to extend these policies to the mobile environment, we should understand that there are four key constraints of mobility which forced the development of specialized techniques, namely, unpredictable variation in network quality, lowered trust and robustness of mobile elements, limitations on local resources imposed by weight and size constraints, and concern for battery power consumption (Satyanarayanan, 1996). The inherent limitations of mobile computing systems present a challenge to the traditional problems of database management, especially when the client/server communication is unexpectedly severed from the client site. The standard policy does not enforce the serializability to the mobile computing environment. Transactions executing under an avoidance-based scheme must obey the Read-Once Write-All (ROWA) principle, which guarantees the correctness of the data from the client cache under the CBL or the O2PL protocol. The standard CBL and O2PL protocols cannot guarantee the currency of the mobile clients' cache copies or prevent serializability violations when they reconnect to the network. Figures 1 illustrates how error conditions (appearing toward the end of the figure) arise after mobile clients properly exit the client application when the traditional CBL protocol is used.

FUNDAMENTAL ISSUES AND APPROACHES TO DEALING WITH THESE ISSUES

In order to extend invalid-access prevention policy protocols to mobile environments, there are three fundamental issues that need to be addressed for mobile-client multimedia data caching, namely:

- to transform multimedia objects from databases' persistent data type to the clients' persistent data type;
- to handle client-server communication for multimedia objects; and
- to deal with the impact of mobility, particularly to deal with the case when the client-server communication is unexpectedly severed from the client site.

Research work from various authors (Breitbart et al., 1999; Franklin, Carey & Livny, 1997; Jensen & Lomer, 2001; Pacitti, Minet & Simon, 1999; Shanmugasundaram et al., 1999; Schuldt, 2001) have contributed to the investigation of aspects related to ensuring serializability of data management. Based on these studies, Parker and Chen (2004) have conducted a more recent research to deal with the three issues mentioned above and developed algorithms to achieve extended invalid-access prevention protocols. The basic ideas of this research are summarized below.

First, in order to prevent the serializability failure scenario described above, we summarize important features of the extended invalid-access prevention policy protocols for the mobile client/server environments that guarantee the serializability. As shown in Table 1, an X denotes an attribute under the standard invalid-access prevention policy, while a bold-face **X** as an additional attribute under the extended invalid-access prevention policy. The revised algorithms for extended invalid-access prevention policy protocols are developed based on these considerations.

As an example of these attributes, here we take a brief look at the important role of the page table. To detect or avoid invalid-accesses from all transactions, all clients and the server each need to keep a separate table to detect or avoid data conflict situations. For clients, page tables are the current inventories of their cached multimedia objects. For the server, a page table is the information about their active constituent clients to detect or avoid data conflicts in the network. Figure 2 depicts a proper page table procedure for logical invalidations to deal with serializability problems through page table consistency check.

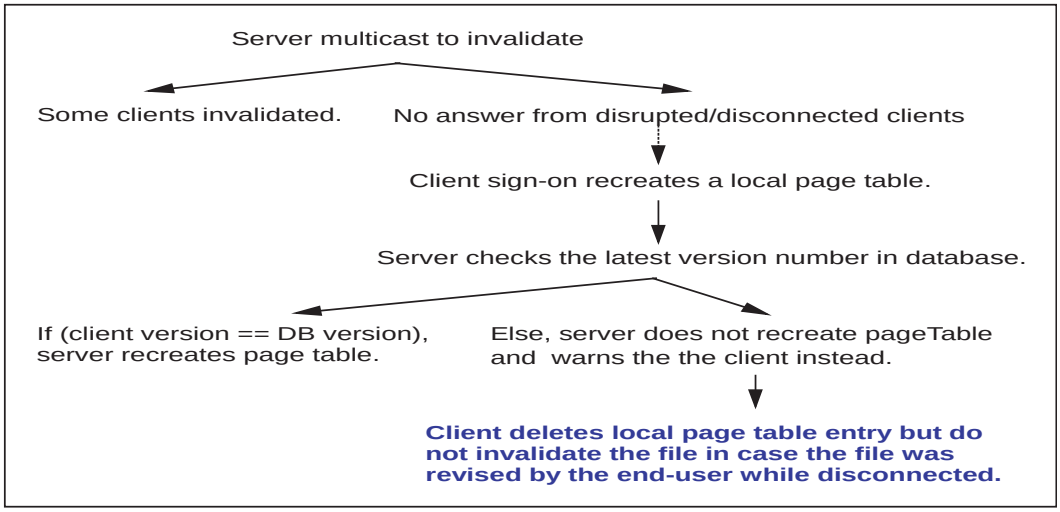
Table 1. Extended invalid-access prevention policy

ATTRIBUTE	S2PL	O2PL	CBL
Version Numbers	X	**X**	**X**
Recreate/release page table rows		**X**	**X**
Permit before Commit			X
Lock before Commit	X		
Commit before Lock		X	
Invalidation		X	X
Dynamic Replication	X	X	
Broadcast		X	X
Read-write conflict		X	X
Write-read conflict	X	X	X
Write-write conflict	X	X	
Relinquish unused locks at sign-off	**X**	**X**	**X**
Maximum lock duration	**X**	**X**	**X**
Server knows who has locks	X	X	X
Server knows who has what objects		X	X

EXTENDING TRADITIONAL PROTOCOLS: BASIC IDEA AND RESULTS OF EXPERIMENTS

To illustrate the basic ideas involved in extending standard protocols, let us take a look at the case of extended O2PL algorithm with dynamic replication. In its original form, the O2PL protocol defers the write declaration until the end of a transaction's execution phase. Dynamic replication is a natural choice for the O2PL because when the server issues a write lock after multicasting

Figure 2. Consistency check of the page table



to all cached remote clients of the same object, the server already has an updated object at hand from the committing client. Just before the server commits the new object to the database with an exclusive lock, there are two copies of the new version object in the network, the server's binary array variable and the local client's new-version cache copy and only one primary copy of the old version object in the database.

The correctness criterion in the replicated database is one-copy serializability (Holiday, Agrawal & Abbadi, 2002) under the ROWA principle where write operations must write to all copies before the transaction can complete. This is accomplished via the server's multicast transmission to a subset of active clients after the primary new-version copy is safely stored in the database.

The primary copy of an object is with the server, and the replicas are at client caches for read transactions. For write transactions, the primary copy is at the transaction's originating site temporarily. After commit operations, however, the server becomes the primary site, and the originating client becomes one of the replicated sites. The server then multicasts the replica to remote clients with the previous version object. When a client downloads an object explicitly, a local lock is given automatically, but the end user can release the local lock manually. Local locks will not be automatically given after dynamic replications.

To enforce the network-wide unique object ID, the server application will verify the uniqueness of the file name at the insert transaction, and the client application will verify that the file name is not altered at the commit transaction as an early abort step.

RESULT OF EXPERIMENTS ON EXTENDED PROTOCOLS

The three extended protocols have been implemented, and comparative studies through experiments have been conducted. Below is the result

of an experiment where identical transactions of four clients are used, each with two multimedia objects (one small size and the other 15 times as large). Table 2 summarizes the number of messages clients sent to the server, total kilobytes of the messages, the number of the server aborts, and the abort rate which is the percentage of aborts from the entire number of messages clients sent to the server. Any dynamic replications do not count toward the messages sent since clients do not send any messages to the server for them. All client-initiated abort messages, such as the permit abort in the ECBL or the lock abort in the ES2PL, are counted toward the MESSAGE, not the ABORT.

Experiments have shown that extended invalid-access prevention policy algorithms enforce a guaranteed serializability of multimedia objects in RDBMS applications under a mobile client/server environment. As for the pros and cons of each extended algorithm, we have the following general observations. Extended S2PL protocol brings the lowest number of client messages to the server but at the highest server abort rate leaving the network with multiple versions. Extended CBL protocol with invalidation carries the highest number of client messages sent to the server and a moderate server abort rate in the expense of reliability. Extended O2PL protocol with replication offers a moderate number of client messages sent to the server with the lowest server abort rate that may make it desirable for most applications.

FUTURE TRENDS

Due to its importance for data management in a mobile environment, techniques for ensuring serializability in dealing with multiple copies of multimedia objects should be further explored. New solutions are needed for existing problems in new environments, and new problems emerge, demanding solutions, as well. In this article, we have deliberately focused on a well-selected specific topic to show the need for an in-depth study of dealing with mobility in databases involving multimedia objects. However, the methodology used here can be extended to many other topics in regard to data management in a mobile computing environment. The key to success in such studies lies in a good understanding of important features of mobile environments, as well as in-

Table 2. Comparison of extended invalid-access prevention policy protocols

PROTOCOL	MSSG Nr	KB*	ABORT	ABORT RATE
ES2PL-Replication	18	70	3	17%
ECBL-Invalidation	34	44	2	6%
EO2PL-Repl+Inval	30	42	0	0%
EO2PL-Replication	24	41	0	0%

*** Total size of client messages**

herent limitations of involved resources in such environments.

In addition, ensuring serializability has implications beyond guaranteeing correctness of transaction execution in a mobile environment. For example, based on the work reported above, Parker, Chen, and Sheng (2004) further identified four core areas of issues to be studied in database-centered mobile data mining, with an emphasis on issues related to DBMS implementation such as query processing and transaction processing. Other aspects related to data mining techniques and distributed or mobile (or even pervasive) computing environments have also been explored (e.g., Kargupta & Joshi, 2001; Lim et al., 2003; Liu, Kargupta & Ryan, 2004; Saygin & Ulusoy, 2002).

There are many other database issues related to mobile computing, such as location-dependent queries (Dunham, Helal & Balakrishnan, 1997; Seydim, Dunham & Kumar, 2001; Yan, Chen & Zhu, 2001). In addition, many issues related to mobile computing can be examined in a more general context of pervasive computing (or ubiquitous computing). Satyanarayanan (1996, 2001) discussed several important issues and future directions on pervasive computing. A wide range of data management aspects should be explored in the future, with the following as sample topics:

- Infrastructure for mobile computing research
- User interface management for pervasive devices
- Data models for mobile information systems
- Mobile database management and mobility-aware data servers
- Mobile transaction and workflow management and models
- Data and process migration, replication/ caching and recovery
- Moving objects and location-aware data management
- Adaptability and stability of pervasive systems in ever-changing wireless environments
- Quality of service (QOS) mechanism for mobile data management

CONCLUSION

As noted earlier, handling multimedia objects in mobile environments faces numerous challenges. Traditional methods developed for transaction processing such as concurrency control and recovery mechanisms may no longer work correctly in mobile environments. In this article, we have focused on a specific issue to ensure serializability for mobile-client data caching. We have explained why the traditional approaches need to be revised and demonstrated the basic idea of extended approaches. Extending from this particular study, we have also discussed related issues in a more general perspective. As indicated in the Future Trends section , there are numerous challenging issues to be resolved in the near future.

REFERENCES

Barbara, D. (1999). Mobile computing and databases: A survey. *IEEE Transactions on Knowledge and Data Engineering, 11*(1), 108-117.

Breitbart, Y., Komondoor, R., Rastogi, R., Seshadri, S., & Silberschatz., A. (1999). Update propagation protocols for replicated databases. *Proceedings of the 1999 ACM SIGMOD Conference* (pp. 97-108).

Dunham, M.H., Helal, A., & Balakrishnan, T. (1997). Mobile transaction model that captures both the data and movement behavior. *Mobile Networks and Applications, 2*(2), 149-162.

Franklin, M.J., Carey, M.J., & Livny, M. (1997). Transactional client-server cache consistency: Alternatives and performance. *ACM Transactions on Database Systems, 22*(3), 315-363.

Holiday, J., Agrawal, D., & Abbadi, A. (2002). Disconnection modes for mobile databases. *Wireless Networks, 8*(4), 391-402.

Jensen, C.S., & Lomer, D.B. (2001). Transaction timestamping in temporal databases. *Proceedings of the 27th International Conference on Very Large Data Bases* (pp. 441-450).

Kargupta, H., & Joshi, A. (2001). Data mining "to go": Ubiquitous KDD for mobile and distributed environments. *Tutorial Notes of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 4.1-4.78).

Lau, W.H.O., Kumar, M., & Vankatesh, S. (2002). A cooperative cache architecture in support of caching multimedia objects in MANETs. *Proceedings of the WoMMoM 02* (pp. 56-63).

Lim, E.-P., Wang, Y., Ong, K.-L., & Hwang, S.-Y. (2003, July). In search of knowledge about mobile users. Center for Advanced Information Systems at Nanyang Technological University, Singapore. Retrieved February 5, 2005, from *http://www.ercim.org/publication/Ercim_News/enw54/lim.html*

Liu, K., Kargupta, H., & Ryan, J. (2004, January). Distributed data mining bibliography. University of Maryland Baltimore County. Retrieved February 5, 2005, from *http://www.cs.umbc.edu/~hillol/DDMBIB/ddmbib.pdf*

Pacitti, E., Minet, P., & Simon, E. (1999). Fast algorithms for maintaining replica consistency in lazy master replicated databases. *Proceedings of the 25th International Conference on Very Large Data Bases* (pp. 126-137).

Parker, S., Chen, Z., & Sheng, E. (2004). Ensuring serializability for mobile data mining on multimedia objects. *Proceedings of the CASDMKM 2004* (pp. 90-98).

Parker, S., & Chen, Z. (2004). Extending invalid-access prevention policy protocols for mobile-client data caching. *Proceedings of the ACM SAC 2004* (pp. 1171-1176).

Satyanarayanan, M. (1996). Mobile information access. *IEEE Personal Communications, 3*(1), 26-33.

Satyanarayanan, M. (2001). Pervasive computing: Vision and challenges. *IEEE Personal Communications, 8*, 10-17.

Saygin, Y., & Ulusoy, Ö. (2002). Exploiting data mining techniques for broadcasting data in mobile computing environments. *IEEE Transactions on Knowledge Data Engineering, 14*(6), 1387-1399.

Schuldt, H. (2001). Process locking: A protocol based on ordered shared locks for the execution of transactional processes. *Proceedings of the 20th ACM SIGMOD SIGACT SIGART Symposium on Principles of Database Systems* (pp. 289-300).

Seydim, A.Y., Dunham, M.H., & Kumar, V. (2001). Location dependent query processing. *Proceedings of the 2nd ACM International Workshop on Data Engineering for Wireless and Mobile Access* (pp. 47-53).

Shanmugasundaram, S., Nithrakashyap, A., Sivasankaran, R., & Ramamritham, K. (1999). Efficient concurrency control for broadcast environments. *Proceedings of the 1999 ACM SIGMOD International Conference in Management of Data* (pp. 85-96).

Silberschatz, A., Korth, H.F., & Sudarshan, S. (2001). *Database system concepts* (4th ed.). New York: WCB McGraw-Hill.

Yan, J., Chen, Z., & Zhu, Q. (2001). An approach for query optimizing in a mobile environment.

Proceedings of the Joint Conference on Information Systems (JCIS 2001) (pp. 507-510).

KEY TERMS

Call-Back Locking (CBL): CBL is an avoidance-based protocol that supports inter-transactional page caching. Transactions executing under an avoidance-based scheme must obey the read-once write-all (ROWA) replica management approach, which guarantees the correctness of data from the client cache by enforcing that all existing copies of an updated object have the same value when an updating transaction commits.

Data Management for Mobile Computing: Numerous database management issues exist in mobile computing environments, such as resource management and system support, representation/dissemination/management of information, location management, as well as others. Various new techniques for cache management, data replication, data broadcasting, transaction processing, failure recovery, as well as database security, have been developed. Applications of these techniques have been found distributed mobile database systems; mobile information systems; advanced mobile computing applications; and the Internet. Yet there are still many other issues need to be dealt with, such as the problem described in this article.

Invalid Access Prevention Policy: The invalid-access prevention policy requires that in order to manage multiple clients' concurrent read/write operations in the client/server architecture, no transactions that access stale multimedia data should be allowed to commit. In general, there are two different approaches to achieve this policy. The detection-based (lazy) policy ensures the validity of accessed multimedia data, and the avoidance-based (eager) policy ensures that invalid multimedia data is preemptively removed from the client caches.

Multimedia Database: A particular challenge for a multimedia database is the ability of dealing with multimedia data types. Retrieval of structured data from databases is typically handled by a database management system (DBMS), while retrieval of unstructured data from databases requires techniques developed for information retrieval (IR). (A survey on content-based retrieval for multimedia databases can be found in Yoshitaka and Ichikawa, A survey on content-based retrieval for multimedia databases, *IEEE Transactions of Knowledge and Data Engineering, 11*(1), pp. 81-93, 1999.) Yet the rigid resource requirement demands more advanced techniques in dealing with multimedia objects in a mobile computing environment.

Optimistic Two-Phase Locking (O2PL): This is avoidance-based and is more optimistic about the existence of data contention in the network than CBL. It defers the write intention declaration until the end of a transaction's execution phase. Under the ROWA protocol, an interaction with the server is required only at client cache-miss or for committing its cache copy under the O2PL. As in CBL, all clients must inform the server when they erase a page from their buffer so that the server can update its page list.

Pervasive Computing (or Ubiquitous Computing): Pervasive computing "has as its goal the enhancing of computer use by making many computers available throughout the physical environment, but making them effectively invisible to the user" (Mark Weiser, Hot topics: Ubiquitous computing, *IEEE Computer*, October 1993, p. 000). Pervasive computing is the trend towards increasingly ubiquitous, connected computing devices in the environment. As the result of a convergence of advanced electronic (particularly, mobile wireless) technologies and the Internet,

pervasive computing is becoming a new trend of contemporary technology.

Serializability: Serializability requires that a schedule for executing concurrent transactions in a DBMS is equivalent to one that executes the transactions serially in a certain order.

Server-Based Two-Phase Locking (S2PL): The S2PL uses a detection-based algorithm and supports inter-transaction caching. It validates cached pages synchronously on a transaction's initial access to the page. Before a transaction is allowed to commit, it must first access the primary copies from the server on each data item that it has read at the client. The new value must be installed at the client if the client's cache version is outdated. The server is aware of a list of clients who requested locks only, and no broadcast is used by the server to communicate with clients.

This work was previously published in Encyclopedia of Database Technologies and Applications, edited by L. Rivero, J. Doorn, and V. Ferraggine, pp. 223-228, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.33
Mobile Caching for Location-Based Services

Jianliang Xu
Hong Kong Baptist University, Hong Kong

INTRODUCTION

Location-based services (LBS) are services that answer queries based on the locations with which the queries are associate; normally the locations where the queries are issued. With a variety of promising applications, such as local information access (e.g., traffic reports, news, and navigation maps) and nearest neighbor queries (e.g., finding the nearest restaurants) (Barbara, 1999; Ren & Dunham, 2000; D. L. Lee, Lee, Xu, & Zheng, 2002; W. C. Lee, Xu, & Zheng, 2004), LBS is emerging as an integral part of daily life.

The greatest potential of LBS is met in a mobile computing environment, where users enjoy unrestricted mobility and ubiquitous information access. For example, a traveler could issue a query like "Find the nearest hotel with a room rate below $100" from a wireless portable device in the middle of a journey. To answer such a query, however, three major challenges have to be overcome:

- **Constrained Mobile Environments:** Users in a mobile environment suffer from various constraints, such as scarce bandwidth, low-quality communication, frequent network disconnections, and limited local resources. These constraints pose a great challenge for the provision of LBS to mobile users.
- **Spatial Data:** In LBS, the answers to a query associated with different locations may be different. That is, query results are dependent on spatial properties of queries. For a query bound with a certain query location, the query result should be relevant to the query as well as valid for the bound location. This requirement adds additional complexity to traditional data management techniques such as data placement, indexing, and query processing (D. L. Lee, 2002).
- **User Movement:** The fact that a mobile user may change its location makes some tasks in LBS, such as query scheduling and cache management, particularly tough. For example, suppose that a mobile user issues a query "Find the nearest restaurant" at loca-

tion *A*. If the query is not scheduled timely enough on the server, the user has moved to location *B* when he or she gets the answer *R*. However, *R* is no longer the nearest restaurant at location *B*.

Caching has been a commonly used technique for improving data access performance in a mobile computing environment (Acharya, Alonso, Franklin, & Zdonik, 1995). There are several advantages for caching data on mobile clients:

- It improves data access latency since a portion of queries, if not all, can be satisfied locally.
- It helps save energy since wireless communication is required only for cache-miss queries.
- It reduces contention on the narrow-bandwidth wireless channel and off-loads workload from the server; as such, the system throughput is improved.
- It improves data availability in circumstances where clients are disconnected or weakly connected because cached data can be used to answer queries.

However, as discussed above, the *constrains* of mobile computing environments, the *spatial* property of location-dependent data, and the *mobility* of mobile users have opened up many new research problems in client caching for LBS. This chapter discusses the research issues arising from caching of location-dependent data in a mobile environment and briefly describes several state-of-the-art solutions.

BACKGROUND

Location Model

Location plays a central role in LBS. A location needs to be specified explicitly or implicitly for any information access. The available mechanisms for identifying locations of mobile users are based on two models:

- **Geometric Model:** A location is specified as an *n*-dimensional coordinate (typically, $n = 2$ or 3); for example , the latitude/longitude pair returned by the global positioning system (GPS). The main advantage of the geometric model is its compatibility across heterogeneous systems. However, providing such fine-grained location information may involve considerable cost and complexity.
- **Symbolic Model:** The location space is divided into disjointed zones, each of which is identified by a unique name. Examples are the Cricket system (Priyantha, Chakraborty, & Balakrishnan, 2000) and the cellular infrastructure. The symbolic model is in general cheaper to deploy than the geometric model because of the lower cost of employing a coarser location granularity. Also, being discrete and well-structured, location information based on the symbolic model is easier to manage.

For ease of illustration, two notions are defined: *valid scope* and *valid scope distribution*. A dataset is a collection of data instances. The *valid scope* of a data instance is defined as the area within which this instance is the only answer with respect to a location-dependent query. With the symbolic location model, a valid scope is represented by a set of logical zone ids. With the geometric location model, a valid scope often takes the shape of a polygon in a two-dimensional space. Since a query may return different instances at different locations, it is associated with a set of valid scopes, which collectively is called the *scope distribution* of the query. To illustrate, consider a four-cell system with a wireless-cell-based location model. Suppose that the nearby restaurant for cell 1 and cell 2 is instance *X*, and the nearby restaurant for cell 3 and cell 4 is instance *Y*. Then,

the valid scope of X is {1, 2}, the valid scope of Y is {3, 4}, and the scope distribution of the nearby restaurant query is {{1, 2}, {3, 4}}.

Client Caching Model

There is a cache management module in the client. Whenever an application issues a query, the local cache manager first checks whether the desired data item is in the cache. If it is a cache hit, the cache manager still needs to validate the consistency of the cached item with the master copy at the server. This process is called *cache validation*. In general, data inconsistency is incurred by data updates at the server (called *temporal-dependent invalidation*). For location-dependent information in a mobile environment, cache inconsistency can also be caused by location change of a client (called *location-dependent invalidation*). If it is a cache hit but the cached content is obsolete or invalid, or it is a cache miss, the cache manager requests the data from the server via on-demand access. When the requested data item arrives, the cache manager returns it to the user and retains a copy in the cache. The issue of *cache replacement* arises when the free cache space is not enough to accommodate a data item to be cached. It determines the victim data item(s) to be dropped from the cache in order to allocate sufficient cache space for the incoming data item.

Survey of Related Work

This section reviews the existing studies on cache invalidation and replacement strategies for mobile clients. Most of them were designed for general data services and only a few addressed the caching issues for location-dependent data. Temporal-dependent invalidation has been studied for many years (Barbara & Imielinski, 1994; Cao, 2000; Wu, Yu, & Chen, 1996). To carry out temporal-dependent invalidation, the server keeps track of the update history (for a reasonable length of time) and sends it, in the form of an invalidation report (IR), to the clients, either by periodic/aperiodic broadcasting or upon individual requests from the clients. In the basic IR approach, the server broadcasts a list of IDs for the items that have been changed within a history window. The mobile client, if active, listens to the IRs and updates its cache accordingly. Most existing temporal-dependent invalidation schemes are variations of the basic IR approach. They differ from one another mainly in the organization of IR contents and the mechanism of uplink checking. A good survey can be found in Tan et al. (2001).

Semantic data caching has been suggested for managing location-dependent query results (Dar, Franklin, Jonsson, Srivatava, & Tan, 1996; Lee, Leong, & Si, 1999), where a cached result is described with the location associated with the query. Unfortunately, the possibility was not explored that a cached data value may be valid for queries issued from locations different from that associated with the original query. As demonstrated in Zheng, Xu, and Lee (2002), the exploration of this possibility can significantly enhance the performance of location-dependent data caching. As a matter of fact, the invalidation information in the proposed methods (to be discussed later in this chapter) can be considered a kind of semantic description, which could improve cache hit rates.

Cache replacement policies for wireless environments were first studied in the *broadcast disk* project (Acharya et al., 1995; Acharya, Franklin, & Zdonik, 1996). In Acharya et al. (1995), the PIX policy takes into consideration both data access probability and broadcast frequency during replacement. In Khanna and Liberatore (2000), the Gray scheme makes replacement decisions based on both data access history and retrieval delay. Motivated by a realistic broadcast environment, an optimal cache replacement policy, called Min-SAUD, was investigated in Xu, Hu, Lee, and Lee (2004). The Min-SAUD policy incorporates

various factors that affect cache performance, that is, access probability, retrieval delay, item size, update frequency, and cache validation delay.

In the studies on location-dependent data caching, data-distance based cache replacement policies, Manhattan distance (Dar et al., 1996) and FAR (Ren & Dunham, 2000), have been proposed. Under these two policies, the data that is farthest away from the client's current location is removed during replacement. However, data distance was considered alone and not integrated with other factors such as access probability. Moreover, they did not consider the factor of valid scope area.

CACHING FOR LOCATION-BASED SERVICES

Location-Dependent Cache Invalidation

When the client moves around, location-dependent data cached at a mobile client may become invalid with respect to the new location. The procedure of verifying the validity of location-dependent data with respect to the current location is referred to as *location-dependent cache invalidation*. To perform location-dependent invalidation efficiently, the idea is to make use of validity information of data instances. Specifically, the server delivers the valid scope along with a data instance to a mobile client and the client caches the data as well as its valid scope for later validity checking. The strategy involves two issues, namely validity checking time and validity information organization. Since a query result depends on the location specified with the query only, it is suggested to perform validity checking for a cached data instance until it is queried. For validity information organization, a number of schemes have been proposed (Zheng et al., 2002; Xu, Tang, & Lee, 2003). The proposed schemes can be classified into two categories according to the underlying location model employed. This section introduces two methods, that is, implicit scope information (ISI) and caching-efficiency-based method (CEB), for a symbolic and geometric location model respectively.

Implicit Scope Information (ISI)

Assume a wireless-cell-ID-based symbolic location model. Under the ISI scheme, the server enumerates the scope distributions of all items and numbers them sequentially. The valid scopes within a scope distribution are also numbered sequentially. For any instance of data item *i*, its valid scope is specified by a 2-tuple (SDN_i, SN_i), where SDN_i is the scope distribution number and SN_i denotes the scope number within this distribution. The 2-tuple is attached to a data instance as its valid scope. For example, suppose there are three different scope distributions (see Table 1) and data item 4 follows distribution 3. If item 4 is cached from cell 6 (i.e., CID = 6), then SDN_4 = 3 and SN_4 = 3. This implies that item 4's instance is valid in cells 6 and 7 only.

It can be observed that the size of the validity information for a data instance is small and independent of the actual number of cells in which the instance is valid. Another observation is that a set of data items may share the same scope distribution. As such, the number of scope distributions could be much smaller than the number of items in the database.

At the server-side, a *location-dependent IR* is periodically broadcast in each cell. It consists of the ordered valid scope numbers (SN) for each scope distribution in the cell. For example, in cell 8, the server broadcasts {8, 3, 4} to mobile clients, where the three numbers are the SN values in cell 8 for scope distributions 1, 2, and 3, respectively (see Table 1).

The validity checking algorithm for item *i* works as follows. After retrieving a location-dependent IR, the client compares the cached SN_i with the SDN_i-th SN in the location-dependent IR received. If they are the same, the cached data

Table 1. An example of data items with different distributions

Cell ID	1	2	3	4	5	6	7	8	9	10	11	12
Scope Distribution (SDN) #1	1	2	3	4	5	6	7	8	9	10	11	12
Scope Distribution (SDN) #2	1			2			3			4		
Scope Distribution (SDN) #3	1			2		3		4			5	

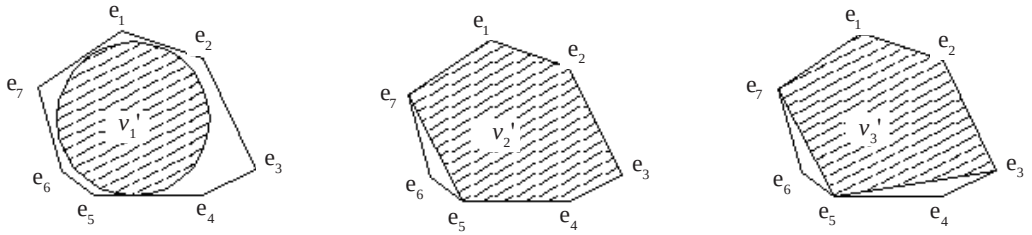
instance is valid. Otherwise, the data instance is invalid. For example, in cell 8, the client checks for the cached instance of data item 4 whose $SDN_4 = 3$ and $SN_4 = 3$. In the broadcast report, the SDN_4-th (i.e., third) *SN* equals to 4. Therefore, the client knows that the cached instance is invalid. The performance analysis conducted in Xu et al. (2003) shows that the ISI method performs close to an optimal strategy which assumes perfect location information is available on mobile clients.

Caching-Efficiency-Based Method (CEB)

This section discusses location-dependent cache invalidation strategies for a geometric location model. Under this model, there are two basic schemes for representing valid scopes, that is, *polygonal endpoints* and *approximate circle* (Zheng et al., 2002). However, these two schemes perform poorly due to either high overhead or imprecision of the invalidation information. To enhance performance, a generic *caching-efficiency-based* (CEB) method for balancing the overhead and the knowledge of valid scopes was proposed in Zheng et al. (2002).

In the CEB method, a new metric *caching efficiency* was introduced. Suppose that the valid scope of a data instance is *v*, and v_i' is a subregion contained in *v* (see Figure 1). Let *s* be the data size, $A(v_i')$ the area of any scope of v_i', and $O(v_i')$ the storage overhead needed to record the scope v_i'. The caching efficiency of the data instance with respect to a scope v_i' is defined as follows:

Figure 1. An example of possible candidate valid scopes ($v = p(e_1,e_2,...,e_7)$)



(a)v_1'=inscribed circle of *v* (b)$v_2' = p(e_1, e_2, e_3, e_4, e_5, e_7)$ (c)$v_3' = p(e_1, e_2, e_3, e_5, e_7)$

$$E(v_i') = \frac{A(v_i')/A(v)}{(s + O(v_i')/s} = \frac{A(v_i')s}{A(v)(s + O(v_i')} \quad (1)$$

Let v_i' be the approximated scope information stored in the client cache. Assuming that the cache size is infinite and the probabilities of a client issuing queries at different locations are uniform, $A(v_i')$ / $A(v)$ is the data instance's cache hit ratio when the client issues the query within the valid scope v. In contrast, $(s + O(v_i'))$ / s is the normalized overhead for achieving such a hit ratio. The rationale behind this definition is as follows. When none of the invalidation information is cached, $E(v_i')$ is 0 because the cached data is completely useless; $E(v_i')$ increases with more invalidation information attached. However, if too much overhead is therefore introduced, $E(v_i')$ would decrease again. Thus, a generic method for balancing the overhead and the precision of invalidation information works as follows:

- For a data instance with a valid scope of v, given a candidate valid scope set $V' = \{v_1', v_2', ..., v_k'\}$, $v_i' \subseteq v$, $1 \le i \le k$, the CEB method chooses the scope v_i' that maximizes caching efficiency $E(v_i')$ as the valid scope to be attached to the instance.

Figure 1 illustrates an example where the valid scope of the data instance is $v = p(e_1,e_2,...,e_7)$, and v_1', v_2', v_3' are three different subregions of v, $A(v_1')$ / $A(v)$ = 0.788, $A(v_2')$ / $A(v)$ = 0.970, and $A(v_3')$ / $A(v)$ = 0.910 . Assume that the data size s is 128 bytes, 8 bytes are needed to represent an endpoint, and 4 bytes for the radius of an inscribed circle; hence $O(v)$ = 56, $O(v_1')$ = 12, $O(v_2')$ = 48, and $O(v_3')$= 40. Thus, $E(v)$ = 0.696, $E(v_1')$ = 0.721, $E(v_2')$ = 0.706, and $E(v_3')$= 0.694. As a result, v_1' is chosen as the valid scope to be attached to the data instance. The simulation based evaluation demonstrates that the CEB method is very effective and outperforms other invalidation methods (Zheng et al., 2002).

Cache Replacement Policies

Because a mobile client has only limited cache space, cache replacement is another important issue to be tackled in client cache management. In traditional cache replacement policies, access probability is considered the most important factor that affects cache performance. A probability-based policy is to replace the data with the least access probability. However, in LBS, besides access probability, there are two other factors, namely *data distance* and *valid scope area*, which have to be considered in cache replacement strategies.

Generally, a promising cache replacement policy should choose as its victim the data item with a low access probability, a small valid scope area, and a long distance if data distance is also an influential factor. This section presents two cost-based cache replacement policies, PA and PAID, which integrate the three factors that are supposed to affect cache performance. The discussions are based on a geometric location model.

- **Probability Area (PA):** As the name suggests, the cost of a data instance under this policy is defined as the product of the access probability of the data item and the area of the attached valid scope. That is, the cost function for data instance j of item i is as follows:

$$c_{i,j} = p_i \cdot A(v'_{i,j}), \quad (2)$$

  where p_i is the access probability of item i and $A(v'_{i,j})$ is the area of the attached valid scope $v'_{i,j}$ for data instance j. The PA policy chooses the data with the least cost as its victim for cache replacement.
- **Probability Area Inverse Distance (PAID):** Compared with PA, this scheme further integrates the data distance factor. For the PAID policy, the cost function for

data instance j of item i is defined as follows:

$$c_{i,j} = \frac{p_i \cdot A(v'_{i,j})}{D(v'_{i,j})}, \qquad (3)$$

where p_i and $A(v'_{i,j})$ are defined the same as above, and $D(v'_{i,j})$ is the distance between the current location and the valid scope $v'_{i,j}$. Similar to PA, PAID ejects the data with the least cost during each replacement.

Zheng et al. (2002) have evaluated the performance of PA and PAID and demonstrated that PA and PAID substantially outperform the existing policies including LRU and FAR. In particular, consideration of the valid scope area improves performance in all settings, and consideration of the moving direction in calculating data distance is effective only for short query intervals and short moving intervals.

FUTURE TRENDS

Caching of location-dependent data opens up a new dimension of research in mobile computing. As for future work, per user based adaptive techniques can be developed since mobile clients may have different movement patterns. Besides cache invalidation and replacement schemes, it is interesting to investigate *cache prefetching* which preloads data onto the mobile client cache by taking advantage of user mobility. Furthermore, how to incorporate location-dependent data invalidation schemes and semantic caching would be an interesting topic. In addition, battery power is a scarce resource in a mobile computing environment; it is believed that power-aware cache management deserves further in-depth study.

CONCLUSION

LBS has been emerging as the result of technological advances in high-speed wireless networks, personal portable devices, and location positioning techniques. This chapter discussed client cache management issues for LBS. Two location-dependent cache invalidation methods, that is, ISI and CEB, are introduced. The cache replacement issue for location-dependent data was also investigated. Two cache replacement policies, that is, PA and PAID that consider the factors of valid scope area (for both methods) and data distance (for PAID only) and combine these factors with access probability, were presented. With an increasing popularity of LBS, caching of location-dependent data remains a fertile research area that aims to overcome inherent constraints (including power, bandwidth, storage, etc.) in a mobile environment.

REFERENCES

Acharya, S., Alonso, R., Franklin, M., & Zdonik, S. (1995). Broadcast disks: Data management for asymmetric communications environments. *Proceedings of ACM SIGMOD Conference on Management of Data* (pp. 199-210).

Acharya, S., Franklin, M., & Zdonik, S. (1996). Prefetching from a broadcast disk. *Proceedings of the 12th International Conference on Data Engineering* (pp. 276-285).

Barbara, D. (1999). Mobile computing and databases—A survey. *IEEE Transactions on Knowledge and Data Engineering, 11*(1), 108-117.

Barbara, D., & Imielinski, T. (1994). Sleepers and workaholics: Caching strategies for mobile environments. *Proceedings of ACM SIGMOD Conference on Management of Data* (pp. 1-12).

Cao, G.. (2000). A scalable low-latency cache invalidation strategy for mobile environments. *Proceedings of the Sixth ACM International Conference on Mobile Computing and Networking* (pp. 200-209).

Dar, S., Franklin, M. J., Jonsson, B. T., Srivatava, D., & Tan, M. (1996). Semantic data caching and replacement. *Proceedings of the 22nd International Conference on Very Large Data Bases* (pp. 330-341).

Khanna, S., & Liberatore, V. (2000). On broadcast disk paging. *SIAM Journal on Computing, 29*(5), 1683-1702.

Lee, D. L., Lee, W.-C., Xu, J., & Zheng, B. (2002). Data management in location-dependent information services. *IEEE Pervasive Computing, 1*(3), 65-72.

Lee, K. C. K., Leong, H. V., & Si, A. (1999). Semantic Query caching in a mobile environment. *Mobile Computing and Communication Review, 3*(2), 28-36.

Lee, W. C., Xu, J., & Zheng, B. (2004). Data management in location-dependent information services. *Tutorial at the 20th IEEE International Conference on Data Engineering* (pp. 871).

Priyantha, N. B., Chakraborty, A., & Balakrishnan, H. (2000). The cricket location-support system. *Proceedings of the Sixth ACM International Conference on Mobile Computing and Networking* (pp. 32-43).

Ren, Q., & Dunham, M. H. (2000). Using semantic caching to manage location dependent data in mobile computing. *Proceedings of the Sixth ACM International Conference on Mobile Computing and Networking* (pp. 210-221).

Tan, K. L., Cai, J., & Ooi, B. C. (2001). An evaluation of cache invalidation strategies in wireless environments. *IEEE Transactions on Parallel and Distributed Systems (TPDS), 12*(8), 789-807.

Wu, K.-L., Yu, P. S., & Chen, M.-S. (1996). Energy-efficient caching for wireless mobile computing. *Proceedings of the 12th International Conference on Data Engineering* (pp. 336-343).

Xu, J., Hu, Q., Lee, W.-C., & Lee, D. L. (2004). Performance evaluation of an optimal cache replacement policy for wireless data dissemination. *IEEE Transactions on Knowledge and Data Engineering, 16*(1), 125-139.

Xu, J., Tang, X., & Lee, D. L. (2003). Performance analysis of location-dependent cache invalidation schemes for mobile environments. *IEEE Transactions on Knowledge and Data Engineering, 15*(2), 474-488.

Zheng, B., Xu, J., & Lee, D. L. (2002). Cache invalidation and replacement policies for location-dependent data in mobile environments. *IEEE Transactions on Computers (TC), 51*(10), 1141-1153.

KEY TERMS

Cache Invalidation: The procedure of validating whether the cached data is consistent with the master copy at the server.

Cache Replacement: The procedure of finding the victim data item(s) to be dropped from the cache in order to allocate sufficient cache space for an incoming data item.

Location-Based Services (LBS): The services that answer queries based on the locations with which the queries are associate.

Location-Dependent Cache Invalidation: The procedure of verifying the validity of cached location-dependent data with respect to the current location.

Mobile Client: A portable device that is augmented with a wireless communication interface.

Valid Scope: The area within which the data instance is the only answer with respect to a location-dependent query.

Valid Scope Distribution: The collective set of valid scopes for a data item.

Wireless Cell: The radio coverage area in which a mobile client can communicate with the wireless infrastructure.

This work was previously published in Encyclopedia of E-Commerce, E-Government, and Mobile Commerce, edited by M. Khosrow-Pour, pp. 760-765, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.34
Location-Aware Query Resolution for Location-Based Mobile Commerce:
Performance Evaluation and Optimization

James E. Wyse
Memorial University of Newfoundland, Canada

ABSTRACT

Location-based mobile commerce incorporates location-aware technologies, wire-free connectivity, and locationalized Web-based services to support the processing of location-referent transactions. In order to provide usable transaction processing services to mobile consumers, location-referent transactions require timely resolution of queries bearing transaction-related locational criteria. This research evaluates Wyse's location-aware method of resolving these queries. Results obtained in simulated mobile commerce circumstances (1) reveal the query resolution behavior of the location-aware method, (2) confirm the method's potential to improve the timeliness of transactional support provided to mobile consumers, and (3) identify the method-related adjustments required to maintain optimal levels of query resolution performance. The article also proposes and provides a preliminary evaluation of a heuristic that may be used in efficiently determining the method-related adjustments needed in order to maximize query resolution performance.

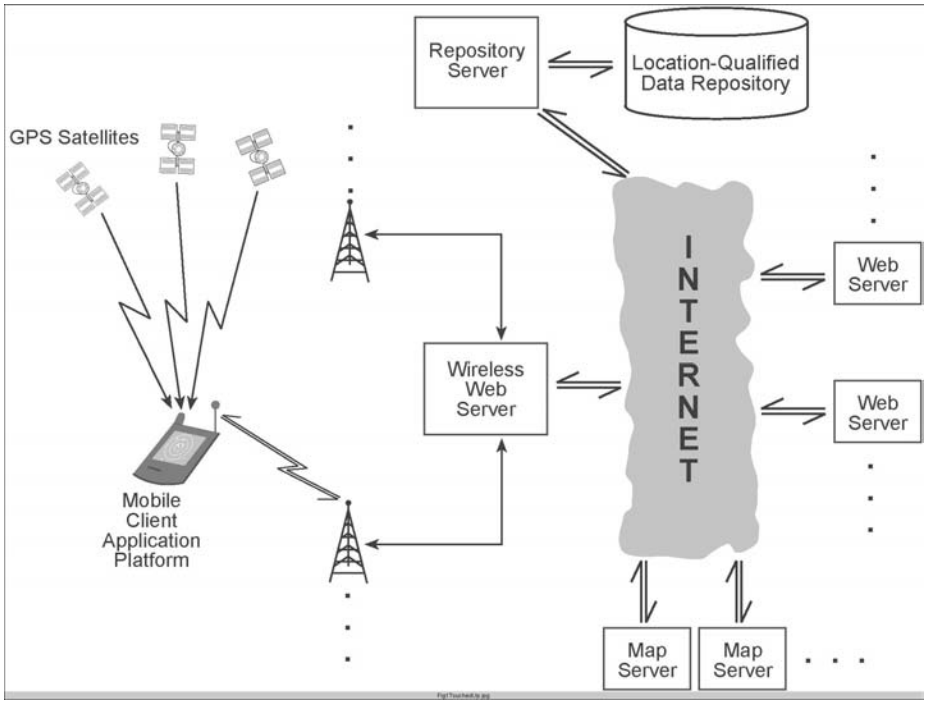
INTRODUCTION

Recent years have witnessed the emergence of transaction-supporting devices directed toward the mobile consumer. Devices range from simple handsets in mobile/cellular phone systems to those involving the convergence of palm-top computing, location-determining technology, and wireless Internet connectivity. Minimally,

devices utilized by mobile consumers must incorporate wireless communication capabilities that permit a significant degree of mobility (Leung & Atypas, 2001; Santami, Leow, Lim, & Goh, 2003). Yuan and Zhang (2003) assert that mobile devices with capabilities extending beyond wireless communication to include those that support location awareness add a "much emphasised … new dimension for value creation" (p. 41) to mobile commerce. Location awareness refers to the capability of a device to obtain data about geographical position and then to use the data to retrieve, select, and report information with respect to that position (Butz, Bauss, & Kruger, 2000). Figure 1 illustrates a location-aware mobile commerce (mcommerce) context in which location-aware applications operating on mobile, GPS-enabled, handheld computing devices avail of wireless connectivity to access a variety of Internet-based servers providing information and functionality to support the transactional activities of mobile consumers.

An essential component in large-scale, location-aware, mobility-supporting applications is a specialized database of transaction-supporting information (Location-Qualified Data Repository, Figure 1). Locational content from the repository is required for the resolution of queries arising from location-referent transactions, transactions in which the relative geographical locations of the prospective transactional parties is a material transactional concern. Siau, Lim, and Shen (2001) and, later, Siau and Shen (2003) call for research on improving the processing of transactional queries in circumstances "where users are constantly on the move and few [end user device] computing resources are available" (p. 13). The research reported here responds to this call; it is

Figure 1. Illustrative configuration of m-commerce components



concerned with the timely processing of queries initiated by location-variant (on-the-move) consumers operating resource-limited devices that must rely on centralized repositories of location-qualified information.

The emergence of m-commerce has spawned several streams of research in areas related to the components shown in Figure 1. Some recent studies with respect to these components have been conducted in such areas as mobile user location determination (McGuire, Plataniotis, & Venetsanopoulos, 2005; Quintero, 2005; Samaan & Karmouch, 2005), mobile device interface design (Lee & Benbasat, 2004), mobile business application design (Gebauer & Shaw, 2004; Khungar & Reikki, 2005), and mobility-related wireless connectivity (Chao, Tseng, & Wang, 2005; Chou & Shin, 2005; Cinque, Cotroneo, & Russo, 2005; Lin, Juang, & Lin, 2005; Xu, Shen, & Mark, 2005; Yeung & Kwok, 2005). Research in areas related to mobile commerce has also addressed in various ways the issue of query resolution in mobile computing environments. Kottkamp and Zukunft (1998) developed and evaluated a mobility aware cost model for location-aware query optimization in the context of mobile user location management; Choy, Kwan, and Hong (2000) proposed a distributed database system architecture to support query processing in mobile geographical applications; Lee and Ke (2001) conducted a cost analysis of strategies for query processing in a mobile commerce environment; Lee, Xu, Zheng, and Lee (2002) and Huang, Lin, and Deng (2005) dealt with the validity of query results and the efficiency of query processing through improved mobile device cache management; while Wyse (2003) proposed a location-aware method of locations repository management to support m-commerce transactions. It is the latter area that is addressed here; specifically, this article examines the extent to which query resolution time is affected by implementing Wyse's (2003) location-aware method (LAM) of managing a server-based locational repository.

A synopsis of LAM is provided in Appendix A. The method employs the linkcell construct as a means of transforming locational coordinates in geographical space to spatially-oriented table names in relational space. A specialized search method operates on the transformation to resolve location-referent queries. Results from Wyse's (2003) work suggest that the method significantly improves query resolution performance over that realized from the use of naïve enumerative methods. However, the work notes that the location-aware method's performance was evaluated in limited circumstances (small repository sizes, fixed geographical coverage, limited business categorization) and also points out that the effect of variations in linkcell size on resolution performance remains unexamined. Wyse (2003) also contemplated the existence of a linkcell size that would optimize query resolution performance but offered no approach that would result in its determination.

These contemplations and limitations give rise to four questions to be addressed by the research reported here: (1) Will the location-aware method yield resolution performance profiles consistent with those previously observed when greater repository sizes, larger variations in geographical coverage, and differing business category sets are used? (2) How is linkcell size related to query resolution performance? (3) Is there a specific linkcell size that will optimize resolution performance? and (4) How might an optimal linkcell size be determined? Before providing results that address these questions, some discussion is warranted on the nature of the problem for which the location-aware method is proposed as a solution.

THE REPOSITORY MANAGEMENT PROBLEM

Mobile consumers frequently require information presented in some consumer-centric proximity

pattern on the locations of businesses offering products and services in a specified business category. Consumer-centric information may be requested in relation to questions such as *Where is the nearest health food outlet? How far away am I from a golf course? Where am I situated in relation to a medical facility?* The queries arising from such questions must incorporate both a product/service criterion (e.g., medical facility) and a consumer-centric, distance-related criterion (e.g., nearest). Two distinctions between product/service-related criteria and consumer-centric, distance-related criteria have implications for the management of locational repositories. First, product/service-related criteria are invariant with respect to a mobile consumer's location, while consumer-centric, distance-related criteria are not. Nievergelt and Widmayer (1997) recognize the distinction between the two types of criteria and point out its efficiency-related implication: "Spatial data differs from all other types of data in important respects. Objects are embedded in an Euclidean space … and most queries involve proximity rather than intrinsic properties of the objects to be retrieved. Thus, data structures developed for conventional database systems are unlikely to be efficient" (p. 186). The issue of efficiency is readily seen in the second distinction: product-service attribute values are patently resident in a repository, while consumer-centric, distance-related attribute values must be derived from the locational attributes of both the consumer and the business location offering a consumer-targeted product or service. Thus, each change in a consumer's geographical position in general will necessitate a redetermination of values for an appropriate consumer-centric, distance-related attribute.

The requirement to continually requery a repository and redetermine a consumer-centric proximity pattern places an extensive burden on server-side repository functionality. For a given level of computational capability, continual requerying and redetermination eventually results in service time degradation as repository sizes increase and/or as the number of consumers increases and/or as consumers more frequently change geographic positions. Increased repository size (i.e., a richer set of locations from which the mobile consumer may obtain information on targeted products or services) would likely attract greater numbers of mobile consumers. In turn, greater numbers of consumers would likely motivate the construction of larger, more richly populated repositories, which then would attract even more consumers (Lee, Zhu, & Hu, 2005). Thus, a cycle is created wherein repository sizes will increase and, in the absence of mitigating investments in computational capability, result in a degradation of the service times experienced by consumers accessing the repository. Thus, an important challenge facing those who are tasked with managing large-scale locational repositories is one of minimizing the increase (i.e., degradation) in the service times realized by mobile consumers while at same time enriching (i.e., enlarging) the location-qualified data repository available to mobile consumers.

SOLUTION APPROACHES

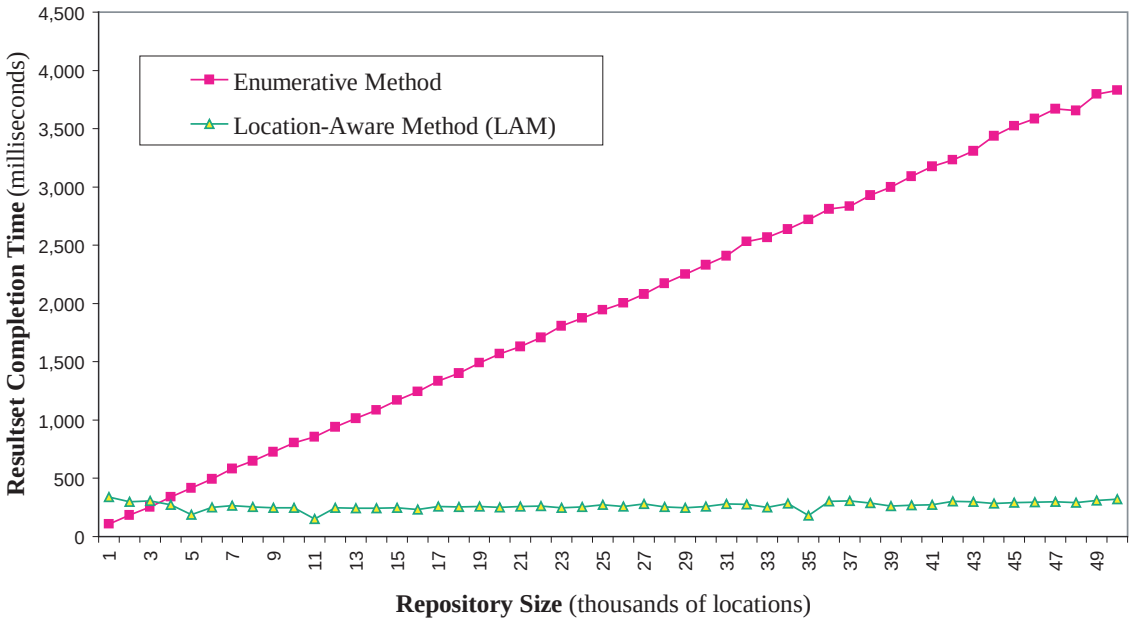
The nature of queries initiated by mobile consumers (e.g., Where's the nearest health food outlet?) suggests that the problem of query resolution is conceptually similar to the nearest neighbor (NN) problem, a problem that has received considerable attention in computational geometry. Formally, solutions to the NN problem incorporate constructs and procedures that, when given a set P of n points and a query point q, result in $p \in P$ such that for all $p' \in P$ we have $d(p, q) \leq d(p', q)$ where $d(p', q)$ is the distance between p' and q (Cary, 2001). Several works have developed NN solution algorithms. Arnon, Efrat, Indyk, and Samet (1999), Lee (1999), and Cary (2001) propose solutions from computational geometry, while Kuznetsov (2000) proposes a solution based on

the space-filling curves developed by Sierpinski (1912) and Hilbert (Butz, 1969, 1971). These algorithms yield solution times that improve upon those realized from naïve enumerative methods; however, the solution time derived for each algorithm's execution is positively related to *n*, the number of points in the set *P*.

A mapping of the terms of the NN problem to aspects of the problem of managing location-qualified data repositories gives *n* as the repository size, *P* as the repository, *q* as the mobile consumer's location, *p* as the nearest location, and $d(p', q)$ as the distance-related attribute needed to resolve the consumer's query. This mapping formalizes the dependency of this attribute on both the consumer's location (q) and the locational attributes contained in the repository for each location, p'. Furthermore, the condition that $d(p, q) \leq d(p', q)$, for all $p' \in P$, implies that new distance-related attribute values are required for all records whenever there is any change in the consumer's location (q). This condition in combination with algorithm solution times that are related positively to repository size (n) corroborates the assessment reached in the previous section that the service times associated with mobile consumer access to location-qualified data repositories will degrade as repository size is increased.

An important aspect of mitigating service time degradation is the use of a retrieval algorithm that does not require a determination of $d(p', q)$, for all $p' \in P$. In other words, new distance-related attribute values need not be calculated for all locations in the repository whenever a new q is encountered (i.e., whenever the mobile consumer changes location). The solution algorithms developed in computational geometry generally take this approach; however, these algorithms have

Figure 2. Resultset completion times for enumerative and location-aware methods by repository size



Source: Wyse (2003), p. 135. Reproduced with the permission of Inderscience Publishers.

solution times that increase, albeit in varying ways, as the size of the set P increases. In this report, the location-aware solution approach to be examined also does not require a determination of $d(p', q)$, for all $p' \in P$. As shown by the LAM curve in Figure 2, the method appears capable of producing resultset completion times (RCTs) that, for practical purposes, are invariant with respect to repository size.[1] Thus, LAM implementations would appear to be useful in mitigating the service time degradation attributable to increases in repository size that otherwise would be realized from implementing an enumerative method of query resolution. The Enumerative Method curve shown in Figure 2 illustrates the degradation that otherwise would occur.[2]

Although the goal of the location-aware method's solution algorithm essentially is the same as that for the NN solution algorithms of computational geometry, the method does not draw upon that discipline's constructs and procedures. Instead, as seen in Appendix A, the method relies upon relational database constructs and functionality combined with the structure of a commonly used convention for designating geographical position (latitude and longitude) to construct and manipulate specialized relational tables (linkcells). Previous work used a simulation-based methodology to demonstrate that the location-aware method is a potential solution (illustrated by the LAM curve in Figure 2) to an important problem (illustrated by the Enumerative curve in Figure 2) associated with managing locational repositories. In what follows, the location-aware method is evaluated in simulated circumstances that extend beyond those of previous work. Also, in contrast to previous work, an important task here is determining the existence of a specific (optimal) linkcell construction that would maximize the resolution performance of queries associated with location-referent transactions.

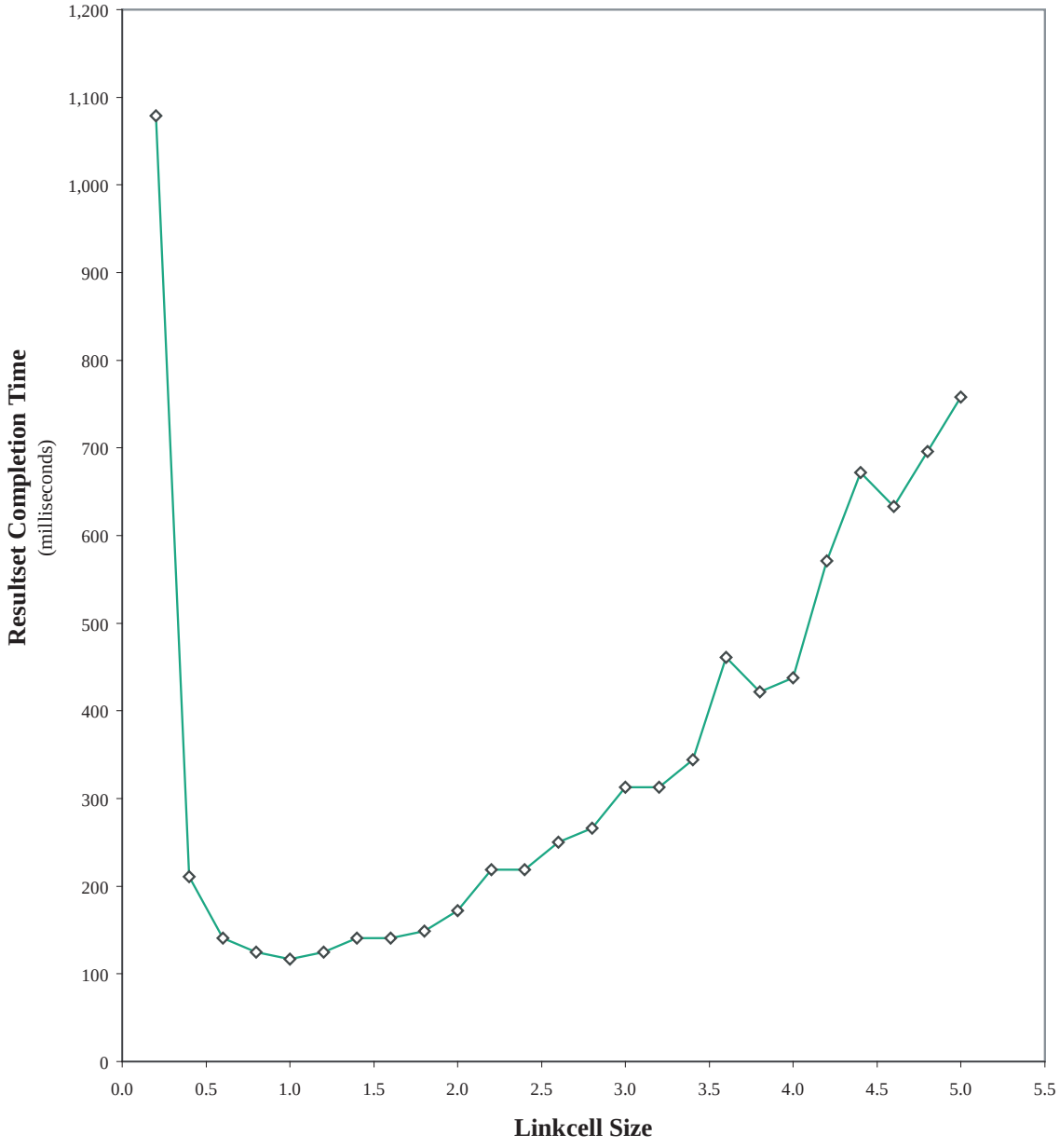
OPTIMAL LINKCELL SIZE DETERMINATION: METHODS AND MEASURES

Appendix A discusses the formulation of the linkcell construct and defines linkcell size. The appendix also notes that changes in linkcell size result in redistributions of repository locations among linkcells and that these redistributions may have a substantial impact on query resolution performance. Wyse's (2003) work on simulated locational repositories used a single linkcell size of 1.0 for all of the various analyses that were conducted, and beyond noting that other linkcell sizes could be used, the work provided little guidance on searching for an optimal linkcell size. Thus, a brute force search strategy was employed in preliminary work here wherein linkcell size (arbitrarily) took an initial value of 0.2 and then was incremented successively (also arbitrarily) by 0.2. This sequence of linkcell sizes was imposed on a series of simulated repository scenarios, each of which consisted of a selected number of randomly generated locations, a selected area of geographical coverage, and a selected number of product-service categories. Resultset completion times (RCTs) were determined at each linkcell size based on queries initiated from 100 randomly chosen locations within the selected geographical area. As previously noted, RCT values indicate the time taken to extract a set of repository locations that represent a resolution of a consumer-initiated query.

Figures 3 and 4 show results from two of the many scenarios on which a brute force search was carried out. All scenarios revealed what appeared to be an optimal linkcell size; however, the optimal value was not generally the same across the scenarios. An optimal linkcell size of 1.0 was revealed for the scenario in Figure 3, while an optimal linkcell size of 0.8 was revealed for the scenario in Figure 4. In its comparison of RCT

performance profiles for the Enumerative and LAM methods, Figure 4 illustrates that a poorly chosen linkcell size (e.g., 5.0) could result not just in RCTs that are far removed from optimal values but also in the complete loss of any query resolution performance advantage attributable to LAM. Table 1 provides the results of an analysis of the scenario whose RCT profile appears in Figure 3.

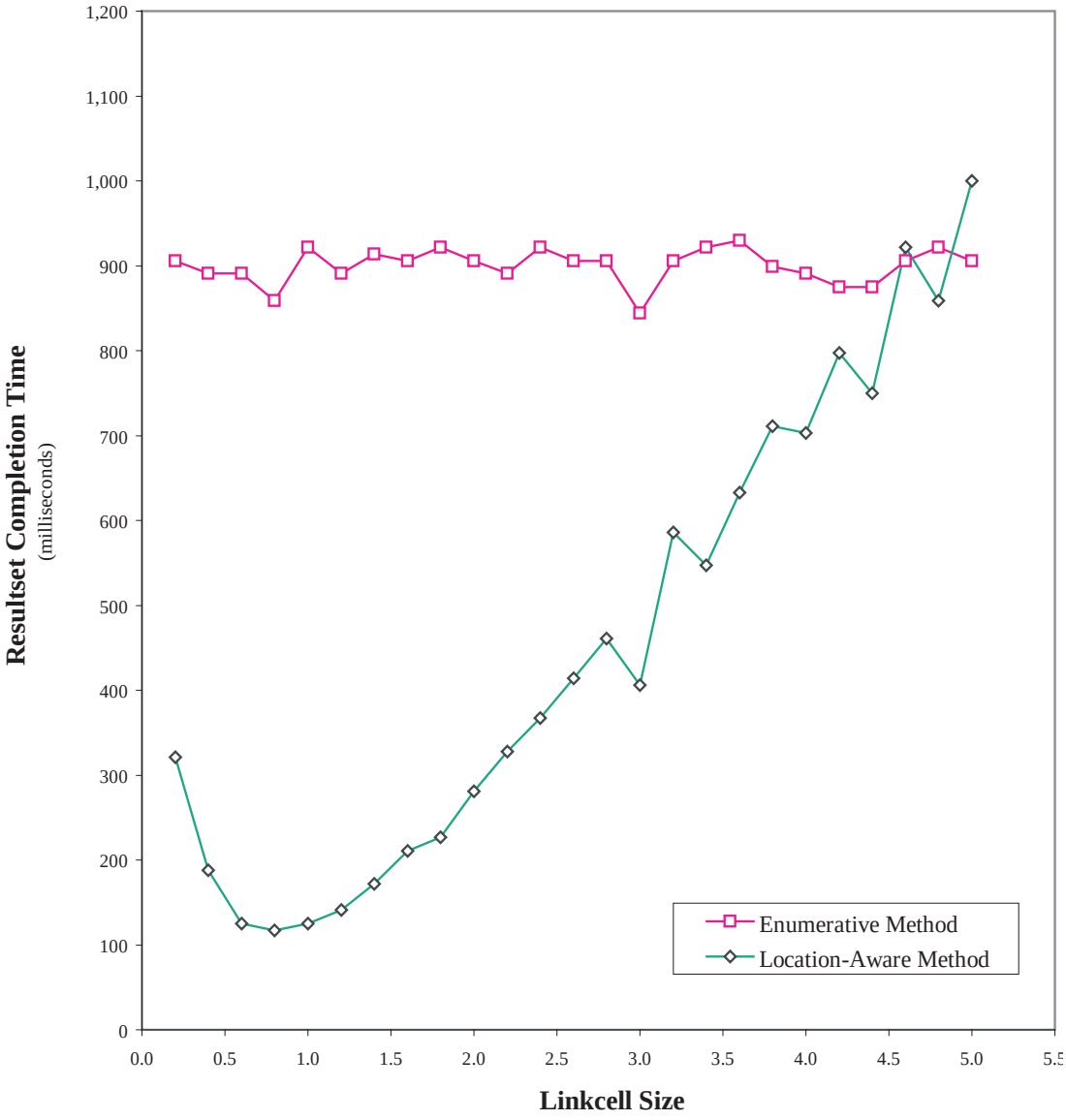
Figure 3. LAM resultset completion times (RCT_L) by linkcell size (Repository: 100,000 locations, 100 product-service categories, area N30° to N50° and W070° to W130°)



The following discussion of the table's content will provide more detail on the simulation-based methodology used to (1) obtain resultset completion times, (2) explain the mechanism underlying the existence of an optimal linkcell size, and (3) illustrate the trade-offs in repository space consumption required to realize LAM-related gains in query resolution performance.

The methodology used to generate the RCT curves in Figures 3 and 4 is similar to the simula-

Figure 4. Resultset completion times for location-aware and enumerative methods by linkcell size (Repository: 100,000 locations, 200 product-service categories, area N35° to N45° and W080° to W110°)



tion-based approach used in Wyse's (2003) evaluative work on the location-aware method. Software called the Linkcell Performance Analyzer (LPA) was developed that (1) generates locational repositories with varying numbers of locations ranging over various geographical areas and referencing different product-service category lists; (2) creates linkcell sets based on repository locations with respect to a specified linkcell size; (3) assembles and processes simulated queries bearing randomly assigned product-service criteria for randomly located consumers; (4) resolves the location-referent queries using both enumerative (E) and location-aware (L) methods; and (5) determines resultset completion times (RCT_E and RCT_L) for each method. Table 1 provides LPA-generated results associated with the RCT_L values plotted in Figure 3. The first column shows the series of linkcell sizes (from 0.2 to 5.0) for the plot's horizontal axis. Columns 6 through 10 report RCT_L statistics derived from simulated queries originating from 100 randomly chosen mobile consumer locations. All five statistics reveal their lowest resultset completion time values at a linkcell size

Table 1. Selected repository scenario (100,000 locations, 100 product-service categories, area N30° to N50° and W070° to W130°)

(1) Selected Linkcell Size	(2) Maximum Number of Linkcells	(3) Linkcells Actually Generated	(4) Mean Linkcell Entries	(5) Probability of Linkcell with Targeted Category	RCT_L					(11) Compacted Disk Space (MB)	(12) Repository xLinkcells Multiple
					(6) Mean	(7) Min	(8) Max	(9) 50th	(10) 90th		
0.2	30,000	28,923	3.5	0.03	1,580	63	8,094	1,079	3,453	358.0	46.6
0.4	7,500	7,500	13.3	0.13	268	16	984	211	516	98.7	12.9
0.6	3,434	3,434	29.1	0.25	182	16	1,047	141	344	49.5	6.4
0.8	1,976	1,976	50.6	0.40	176	16	969	125	328	31.7	4.1
1.0	1,200	1,200	83.3	0.57	144	16	391	117	281	22.4	2.9
1.2	867	867	115.3	0.69	185	16	922	125	373	18.4	2.4
1.4	645	645	155.0	0.79	164	31	531	141	313	17.6	2.3
1.6	546	546	183.2	0.84	165	16	496	141	313	16.2	2.1
1.8	420	420	238.1	0.91	167	16	750	149	281	14.6	1.9
2.0	300	300	333.3	0.96	202	31	969	172	391	13.7	1.8
2.2	290	290	344.8	0.97	231	31	875	219	406	13.4	1.7
2.4	234	234	427.4	0.99	236	31	906	219	422	12.9	1.7
2.6	216	216	463.0	0.99	267	31	656	250	453	12.6	1.6
2.8	176	176	568.2	1.00	296	31	750	266	547	12.2	1.6
3.0	147	147	680.3	1.00	331	31	1,172	313	641	12.1	1.6
3.2	140	140	714.3	1.00	337	31	734	313	594	12.0	1.6
3.4	133	133	751.9	1.00	367	31	1,328	344	578	11.9	1.5
3.6	108	108	925.9	1.00	461	63	1,500	461	797	11.6	1.5
3.8	119	119	840.3	1.00	476	78	1,375	422	895	11.7	1.5
4.0	96	96	1,041.7	1.00	473	31	1,984	438	813	11.5	1.5
4.2	75	75	1,333.3	1.00	585	47	1,703	571	100	11.3	1.5
4.4	90	90	1,111.1	1.00	653	47	2,000	672	1,156	11.5	1.5
4.6	70	70	1,428.6	1.00	696	109	2,500	633	1,156	11.3	1.5
4.8	70	70	1,428.6	1.00	751	78	2,281	696	1,281	11.3	1.5
5.0	48	48	2,083.3	1.00	758	141	1,859	758	1,344	11.0	1.4

of 1.0, a result reflected in the plot of RCT_L's 50th percentile values, as shown in Figure 3.[3]

The results shown in Columns 2, 3, and 4 of Table 1 help to explain why an optimal linkcell size can be expected to exist. Column 2 indicates the maximum number of linkcells that may be created for each linkcell size, a quantity that varies from 30,000 linkcells for a linkcell size of 0.2 down to 48 linkcells for a linkcell size of 5.0. The numbers in Column 2 are the result of allocating the linkcell size along the (arbitrarily chosen) horizontal and vertical extents that encompass a repository's locations. For a linkcell size of 0.2, the chosen extents potentially result in 30,000 relational tables (linkcells), a number that is the product of 300 linkcell size intervals along the repository's horizontal extent (W070° to W130°) and 100 linkcell size intervals along its vertical extent (N30° to N50°).[4] Not all of the maximum number of relational tables may be actually created. Appendix A's discussion on linkcell creation indicates that a relational table corresponding to a linkcell only comes into existence when its name is derived from a repository location. Thus, the number of linkcells actually generated from a repository's locations may be less than the maximum potential number of linkcells. Such an outcome is seen in Column 3 of Table 1, where, for a linkcell size of 0.2, only 28,923 of the 30,000 possible linkcells actually were created.

Column 4 of Table 1 reports the mean number of linkcell entries for each linkcell size and is obtained by dividing the total number of repository locations (100,000, in this case) by the number of linkcells (Column 3) created from those locations. Comparing Columns 3 and 4 will help to reveal why an optimal linkcell size exists. Note that small linkcell sizes result in the generation of large numbers of small relational tables, while large linkcell sizes result in the generation of small numbers of large relational tables. Consequently, as linkcells initially increase in size, query resolution times will improve, because fewer relational tables have to be examined in order to find a location in the targeted product-service category; however, with each increase in linkcell size, query resolution times also will degrade, because more relational table entries have to be examined in order to find a location in the targeted product-service category. The optimal linkcell size corresponds to the size at which the RCT gains from processing fewer linkcells begin to be overwhelmed by the RCT losses incurred from processing linkcells with greater numbers of linkcell entries. The RCT statistics in Columns 6 through 10 of Table 1 indicate that such gains and losses combine to reveal an optimal linkcell size of 1.0.

The generation of a relational table for each linkcell results in the database containing a location-qualified data repository that is larger than the database for the same repository without linkcells. Column 11 of Table 1 shows the disk space consumed by the location-qualified repository for each linkcell size. A linkcell size of 0.2 results in the generation of 28,923 linkcells and requires disk storage of approximately 358 MB, an amount of storage that is almost 47 times the 7.7 MB storage amount consumed by the repository without linkcells (referred to as Repository x-Linkcells in Column 12 of Table 1). At the other end of the linkcell range, a linkcell size of 5.0 results in the generation of 48 linkcells and requires 11.0 MB of storage, or 1.4 times as much storage as the repository x-linkcells. At the observed optimal linkcell size of 1.0, the repository requires 2.9 times the storage required by the repository x-linkcells. In general, disk storage consumption (manifested here by repository xlinkcell multiples) varies across the scenarios investigated. With respect to the results shown in Figure 4, the optimal linkcell size requires a repository that is 2.0 times the repository x-linkcells. Thus, service-level performance gains from using the location-aware method come at a repository space cost that may be several multiples of that required when using the enumerative method.

The fifth column of Table 1 reports the probability $P_{TC}(S)$ that a linkcell of size S contains an

entry for a location in the product-service category that is targeted by the mobile consumer. For example, if a mobile consumer initiates a query about the nearest medical facility, which has been assigned, for instance, a product-service code of *C016*, then the targeted category TC is *C016*, and with reference to the results in Table 1 for a linkcell size of 0.2, we see that $P_{C016}(0.2)$ is 0.03. Note that *C016* is one of the 100 product-service categories used to qualify the 100,000 locations in Table 1's repository scenario. As will be explained next, values for $P_{TC}(S)$ were generated to facilitate the search for optimal linkcell sizes in a way that is more computationally efficient and managerially usable than searches using brute force methods.

Formally, the probability that a linkcell contains a location in the targeted product-service category TC is given by:

$$P_{TC}(S) = 1 - (1 - n_{TC}/N)^{N/CS} \qquad (I)$$

where n_{TC} is the number of locations in the repository with product-service code TC,
N is the total number of locations contained in the repository,
C_S is the number of linkcells of size S created from the repository's N locations, and
N/C_S is the mean number of entries per linkcell.

Equation (I) was formulated on the following basis:

(1) As noted in Appendix A, the manner in which linkcells are created, populated, and destroyed results in one and only one linkcell entry for each repository location. Thus, the probability that any linkcell entry bears the targeted product-service code is the ratio of the number of locations in the repository in the targeted product-service category to the total number of locations in the repository in all product-service categories, or n_{TC} /N.

(2) As also noted in Appendix A during its discussions on repository structure, each location in the repository is qualified by one and only one product-service code. Thus, each linkcell entry either bears the targeted product-service code or it does not. Consequently, if the probability that a linkcell entry bears TC is n_{TC} /N, then the probability that a linkcell entry does not bear TC is (1 - n_{TC} /N).

(3) The probability that none of a linkcell's entries bears the targeted code is given by the product of the probabilities that each linkcell entry does not bear the targeted code, or in other terms, (1 – n_{TC} /N) x (1 – n_{TC} /N) x …. x (1 – n_{TC} /N), which may be estimated by $(1 - n_{TC}/N)^{N/CS}$.

(4) Hence, the probability that at least one of the linkcell entries bears the targeted code is $(1 - (1 - n_{TC}/N)^{N/CS})$, which is the right-hand side of Equation (I).

An important assumption underlining Equation (I) is the independence of the probabilities of occurrence of repository entries in the same product-service category with respect to geographical location. Since the repositories used here are generated based on a uniform distribution of locations within specified geographical boundaries, the assumption of independent probabilities in this respect is not an unreasonable one, given the simulated circumstances employed here. However, as discussed later in this article, this assumed distribution of locations may not hold in many practical mcommerce circumstances.

An examination of various repository scenarios indicated the potential usefulness of Equation (I) in identifying optimal linkcell sizes more efficiently and conveniently than doing so using a brute force identification approach. Instead of starting the search for an optimal linkcell size at some arbitrary point, the search was started at the linkcell size S that results in a value of $P_{TC}(S)$ that

is close to 0.5.[5] Letting $S_{0.5}$ denote the linkcell size such that $P_{TC}(S_{0.5}) \cong 0.5$, then for smaller linkcell sizes ($S < S_{0.5}$), a linkcell examined during a LAM search probably does not contain a location in the consumer-targeted category; however, for larger linkcell sizes ($S > S_{0.5}$), an examined linkcell probably does contain a consumer-targeted location. Thus, somewhere in the vicinity of $S_{0.5}$, it starts to become likely that a linkcell contains a consumer-targeted location, and consequently, searches for RCT minima that begin at $S_{0.5}$ are likely to more quickly identify an optimal linkcell size than would brute search methods.

An analysis of selected repository scenarios suggested that searches initiated at a linkcell size of $S_{0.5}$ were effective in quickly identifying optimal linkcell sizes. In the case of the repository scenario associated with Table 1 and Figure 3, Equation (I) yields $S_{0.5} \cong 0.9$ as the starting linkcell size. The optimal linkcell size revealed by brute force is close by at 1.0. With respect to repository scenario associated with Figure 4, Equation (I) yields $S_{0.5} \cong 0.7$, a linkcell size that is close to the 0.8 linkcell size revealed by brute force. Application of Equation (I) in the context of these two and various other repository scenarios indicated its usefulness in improving the efficiency of optimal linkcell size identification. However, the repository scenarios were chosen arbitrarily, and the results of their analysis provide only a rough indication of both the existence of linkcell size optima and Equation (I)'s managerial usefulness in optima identification.

In the following sections, more comprehensive and rigorous assessments of the existence of linkcell size optima are reported with respect to (1) changes in repository size, (2) variations in a repository's geographical coverage, and (3) differences in the product-service code sets used to qualify a repository's locations. A simulation-based methodology is used wherein RCT curves like those seen in Figures 3 and 4 are constructed. The methodology is similar to that used previously but with three differences: (1) RCT curves are based on finer linkcell size increments (0.1 vs. 0.2); (2) RCT values are obtained using a greater number of randomized consumer-initiated queries (200 vs. 100); and (3) the $S_{0.5}$-method is used to initiate a brute force search for linkcell size optima.

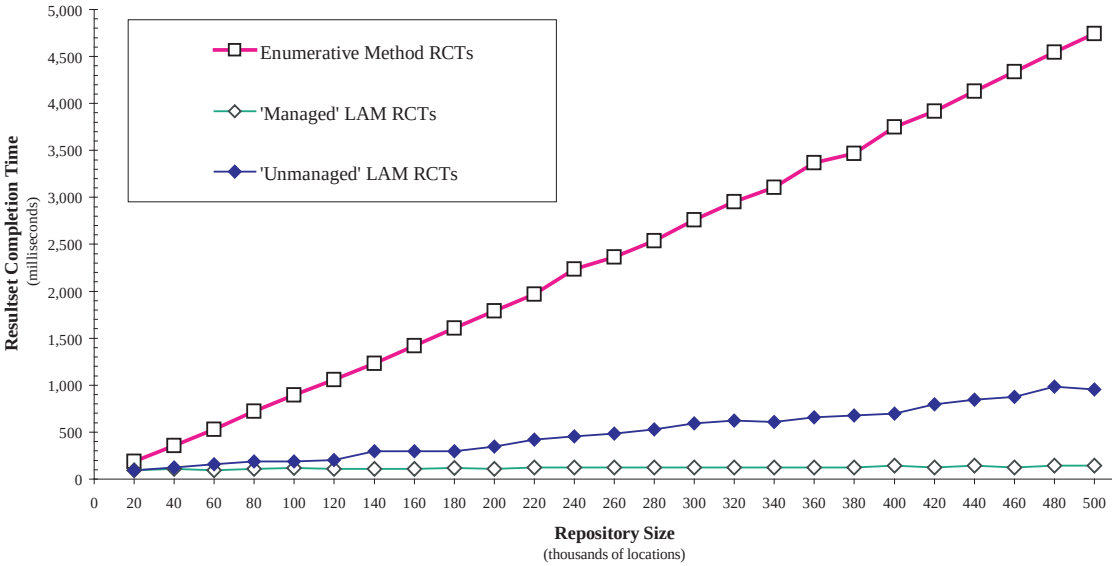
LINKCELL SIZE OPTIMA AND REPOSITORY SIZE VARIABILITY

Figure 5 plots resultset completion times (RCTs) for repositories that range in size from 20,000 to 500,000 locations. Each point is the 50th percentile of RCTs for 200 queries issued by mobile consumers from randomly selected geographical locations. Table 2 presents linkcell sizes, P_{TC} values, and RCTs for the repositories associated with Figure 5. The geographical coverage area and the number of product-service categories remained fixed for all repository sizes. Three RCT curves are shown in Figure 5(a): (1) the RCT_E values plotted in the topmost curve (and shown in Column 5 of Table 2) are the result of using enumerative query resolution; (2) the RCT_L values plotted in the bottom curve (and shown in Column 6 of Table 2) are the result of using location-aware query resolution and doing so at each repository's observed optimal linkcell size (seen in Column 3 of Table 2); and, (3) the RCT_L values plotted in the middle curve are the result of using location-aware query resolution, but here, the linkcell size is the same for all repository sizes and is set to the observed optimal linkcell size (3.3) for a repository size of 20,000 locations (note the first linkcell size shown in Column 3 of Table 2).

Managerially, the middle curve in Figure 5 reflects a circumstance in which linkcell size is set to its optimal value with respect to some initial repository size and then remains unchanged (unmanaged) as repository growth occurs. The results indicate that an unmanaged linkcell size results in query resolution time deterioration as a repository grows in size. As seen from the point

Figure 5. Resultset completion times by repository size

(a) RCT_E, "Unmanaged" RCT_L, and "Managed" RCT_L



(b) RCT_E and RCT_L at Mean $P_{TC}(S_{OPT})$

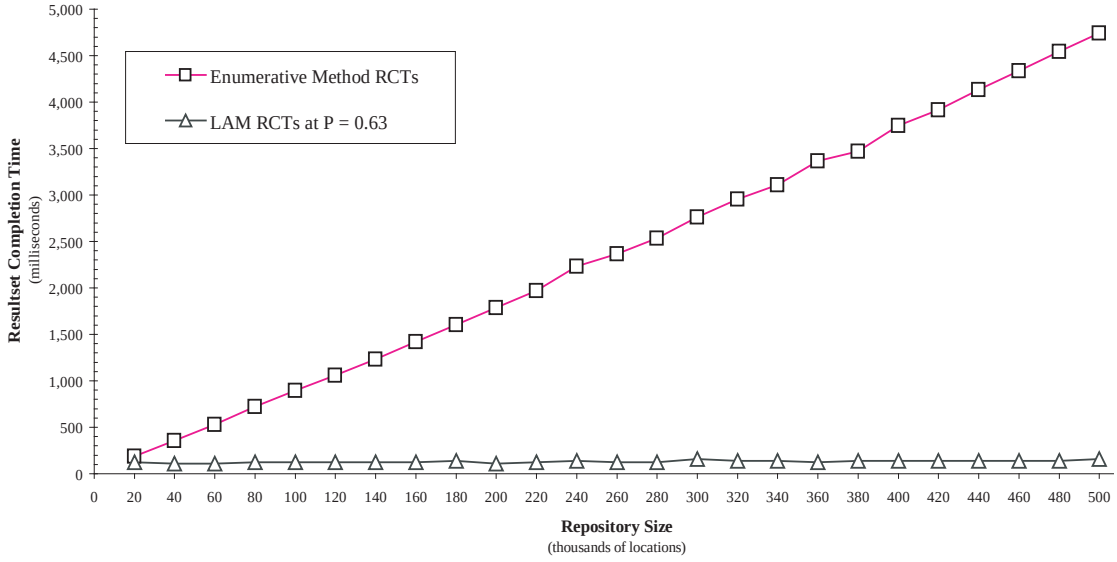


Table 2. Linkcell sizes and resultset completion times by repository size (200 product-service categories, area N30° to N50° and W070° to W130°)

(1) Repository Size (Locations)	(2) $S_{0.5}$ Linkcell Size	(3) Observed S_{OPT} Linkcell Size	(4) $P_{TC}(S_{OPT})$ Prob. linkcell has Targeted Category	(5) RCT_E 50th Percentile	(6) RCT_L 50th Percentile	(7) Linkcell Size at $P_{TC}(S_{OPT})$'s Mean Value	(8) RCT_L for Mean $P_{TC}(S_{OPT})$ Linkcell Size
20,000	3.1	3.3	0.53	187	94	3.9	125
40,000	2.2	2.5	0.65	358	109	2.5	109
60,000	1.7	2.4	0.72	531	94	2.0	109
80,000	1.5	1.4	0.46	723	109	1.9	125
100,000	1.3	1.4	0.54	897	117	1.7	125
120,000	1.3	1.7	0.72	1057	109	1.5	125
140,000	1.1	1.2	0.55	1232	109	1.3	125
160,000	1.1	1.4	0.71	1420	109	1.3	125
180,000	1.0	1.3	0.70	1605	117	1.2	141
200,000	0.9	1.1	0.61	1788	109	1.1	109
220,000	0.9	1.0	0.60	1970	125	1.1	125
240,000	0.9	0.9	0.54	2234	125	1.0	141
260,000	0.8	1.0	0.66	2365	125	1.0	125
280,000	0.8	0.9	0.59	2537	125	0.9	125
300,000	0.8	1.1	0.76	2760	125	0.9	156
320,000	0.7	1.0	0.74	2954	125	0.9	141
340,000	0.7	0.8	0.58	3107	125	0.9	141
360,000	0.7	0.8	0.60	3366	125	0.8	125
380,000	0.7	0.9	0.70	3469	125	0.8	141
400,000	0.7	0.7	0.54	3747	141	0.8	141
420,000	0.6	1.0	0.83	3917	125	0.8	141
440,000	0.6	0.6	0.47	4132	141	0.8	141
460,000	0.6	0.8	0.69	4337	125	0.7	141
480,000	0.6	0.6	0.50	4543	141	0.7	141
500,000	0.6	0.9	0.80	4741	141	0.7	156

on the curve for a 500,000-location repository, RCT_L eventually reaches 953 ms when linkcell size remains unmanaged (i.e., unoptimized), a query resolution time that is almost seven times the RCT_L of 141 ms for a managed (i.e., optimized) linkcell size. This outcome suggests that in order to continually realize optimal query resolution performance, linkcell size must be adjusted as repository size changes.

The observed optimal linkcell size S_{OPT} was identified for each repository size as the linkcell size corresponding to the minimum observed RCT. The search for S_{OPT} began at a linkcell size of $S_{0.5}$ (determined from Equation (I)) and then was expanded above and below $S_{0.5}$ in increments of 0.1 until an RCT minimum was discernable. Table 2 shows $S_{0.5}$ (Column 2) and the observed optimal linkcell size S_{OPT} (Column 3) for each repository size. Comparisons of the values of $S_{0.5}$ and S_{OPT} provide an indication of the usefulness of $S_{0.5}$ in the identification of S_{OPT}. Of the 25 repository sizes, $S_{0.5}$ is within 0.3 for 22 of them and never exceeds 0.7 for any of them. Furthermore, for 24 of the 25 repository sizes, S_{OPT} >=

$S_{0.5}$, a result consistent with the previously noted implication of Equation (I) that when $S > S_{0.5}$, a linkcell examined in the course of a location-aware search probably contains a location in the consumer-targeted category. Column 4 of Table 2 shows values for $P_{TC}(S_{OPT})$, the probability that a linkcell contains a consumer-targeted location at the observed optimal linkcell size. These probability figures are consistent with the expectation that optimal linkcell sizes would be rarely observed at a $P_{TC}(S_{OPT})$ value that is substantially below 0.50.

A comparison of the RCT curves in Figure 5(a) indicates that the methodology used to identify S_{OPT} yields linkcell sizes that result in query resolution performance that is not only superior to the conventional enumerative methodology but also substantively independent of repository size. Although these two outcomes are managerially important, the methodology by which they are realized is likely to be regarded as cumbersome and inconvenient by those tasked with repository management. Thus, a simpler method was sought that would be more readily applicable in practical circumstances. In the course of the investigation, it was observed that the use of linkcell sizes derived from Equation (I) with $P_{TC}(S)$ set to the mean value of $P_{TC}(S_{OPT})$ results in RCT values that, for practical purposes, very closely approximate the minimal RCT values associated with optimal linkcell sizes. With respect to the probability values in Column 4 of Table 2, the mean value of $P_{TC}(S_{OPT})$ is 0.63. Column 7 of Table 2 shows the linkcell sizes that result from setting Equation (I) = 0.63 and solving for S. Column 8 presents the RCT values that result from using the linkcell sizes in Column 7. The bottom curve in Figure 5(b) is a plot of Column 8's RCT values. A comparison of Columns 8 and 6 in Table 2 or, equivalently, a comparison of the bottom curves in Figures 5(a) and 5(b) reveals minimal differences in query resolution performance across the examined range of repository sizes.

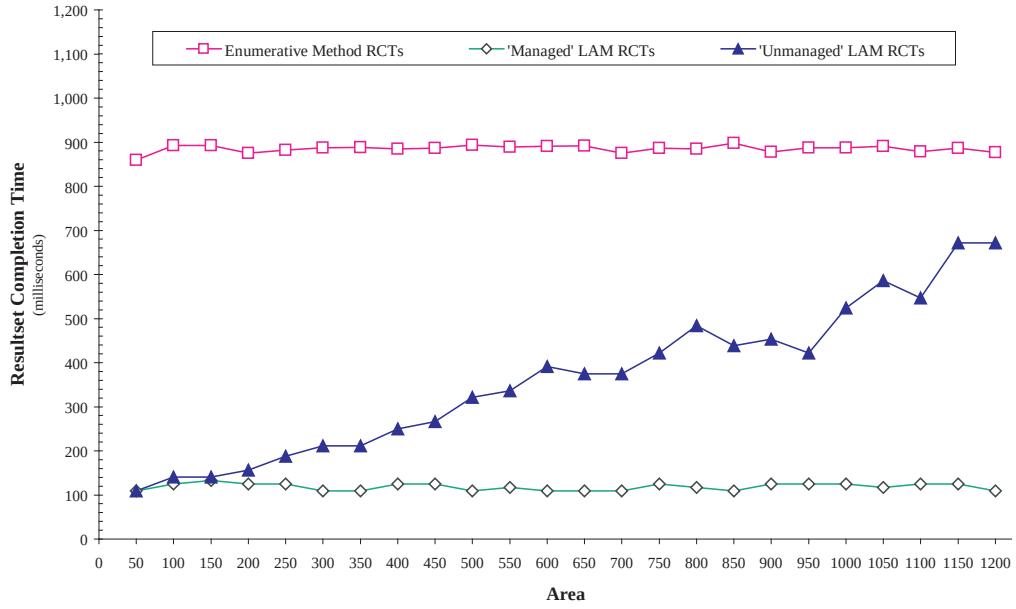
The outcomes associated with $P_{TC}(S) = 0.63$ suggest that it may be the basis for a practical method of directly identifying performance-optimizing linkcell sizes and one with considerable potential to simplify the repository manager's task of linkcell size determination. Furthermore, the method is structured to an extent that it may be captured in a software module in a straightforward manner. Although this method of linkcell size identification is more convenient and considerably less cumbersome than both the brute force method and the $S_{0.5}$method, its applicability relies heavily on the validity of setting $P_{TC}(S)$ = mean value of $P_{TC}(S_{OPT})$ as a basis for estimating optimal linkcell sizes. The next two sections provide further assessments of validity in this respect through examinations of circumstances in which (1) geographical area is varied and (2) different product-service code sets are used to qualify the repository's locations. Consistent with the methodological approach used previously, optimal linkcell sizes are identified in both cases, first by the method of constructing RCT curves at successive incremental linkcell sizes in the region of $S_{0.5}$ (i.e., the $S_{0.5}$method) and, second, by the method of determining linkcell sizes with reference to the mean value of $P_{TC}(S_{OPT})$.

LINKCELL SIZE OPTIMA AND AREA VARIABILITY

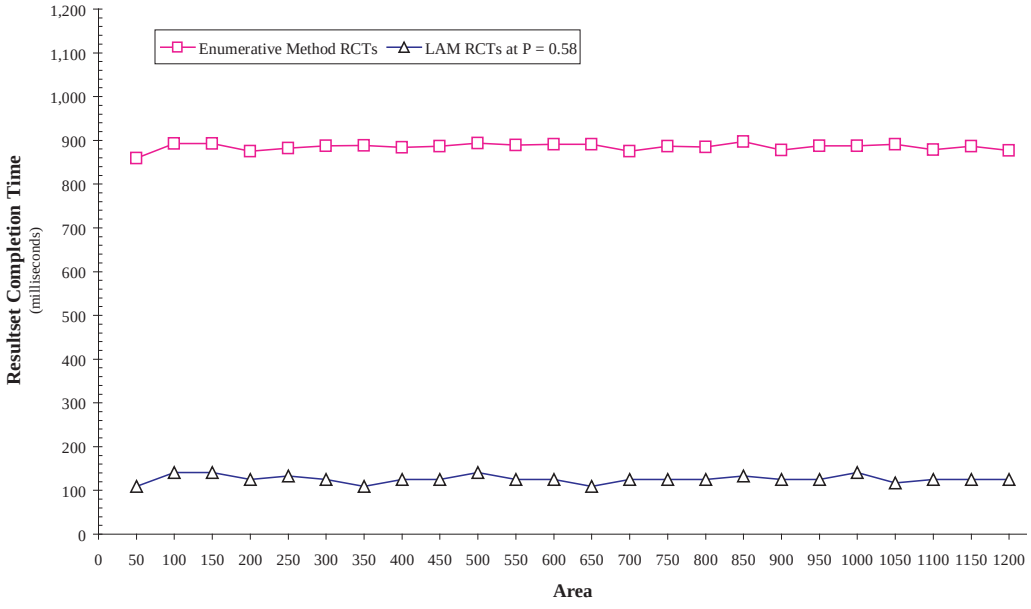
Figure 6 plots resultset completion times for a 100,000-location repository whose area of geographical coverage varies over a sequence of 24 areas of increasing size beginning (arbitrarily) with an area bounded by N35° to N40° and W095° to W105° (or 5 degrees of latitude by 10 degrees of longitude) and ending (arbitrarily) with area bounded by N30° to N50° and W070° to W130° (or 20 degrees of latitude by 60 degrees of longitude). For convenience, the areas of increasing size are shown in Figure 6 by the product of their latitu-

Figure 6. Resultset completion times by geographical area

(a) RCT_E, "Unmanaged" RCT_L, and "Managed" RCT_L



(b) RCT_E and RCT_L at Mean $P_{TC}(S_{OPT})$



dinal and longitudinal extents (50, 100, 150, …, 1200). Note that in contrast to the results seen in Figure 5, RCT_E values in this case are essentially invariant with respect to geographical area. A consideration of the enumerative method's procedural details will reveal that although changes in either the number of locations or the number of product-service codes will affect query resolution time, changes in geographical area will not. Changes in geographical area affect the range of values over which the coordinates of locations will vary; however, there is no additional computational burden placed on the enumerative method when different coordinate values are assigned to the same repository locations. Thus, RCT_E values are substantively invariant with respect to geographical area.

Three curves are shown in Figure 6(a): (1) the topmost curve is the result of using the enumerative method; (2) the bottom curve is the result of using the location-aware method and doing so for each area's observed optimal linkcell size; and (3) the middle curve is the result of using the location-aware method with the linkcell size unchanged for all areas from the optimal linkcell size (0.3) for the smallest area. The third curve reflects a circumstance in which linkcell size is set at its optimal value with respect to some arbitrary area and then remains unchanged (unmanaged) as the repository's area of geographical coverage is enlarged. Here, as before, unmanaged linkcell sizes result in query resolution time degradation. As seen from the point on the RCT curve for the largest area, RCT_L eventually reaches 672 ms when linkcell sizes remain unoptimized, a query resolution time that is more than six times the RCT_L of 109 ms for an optimized linkcell size.

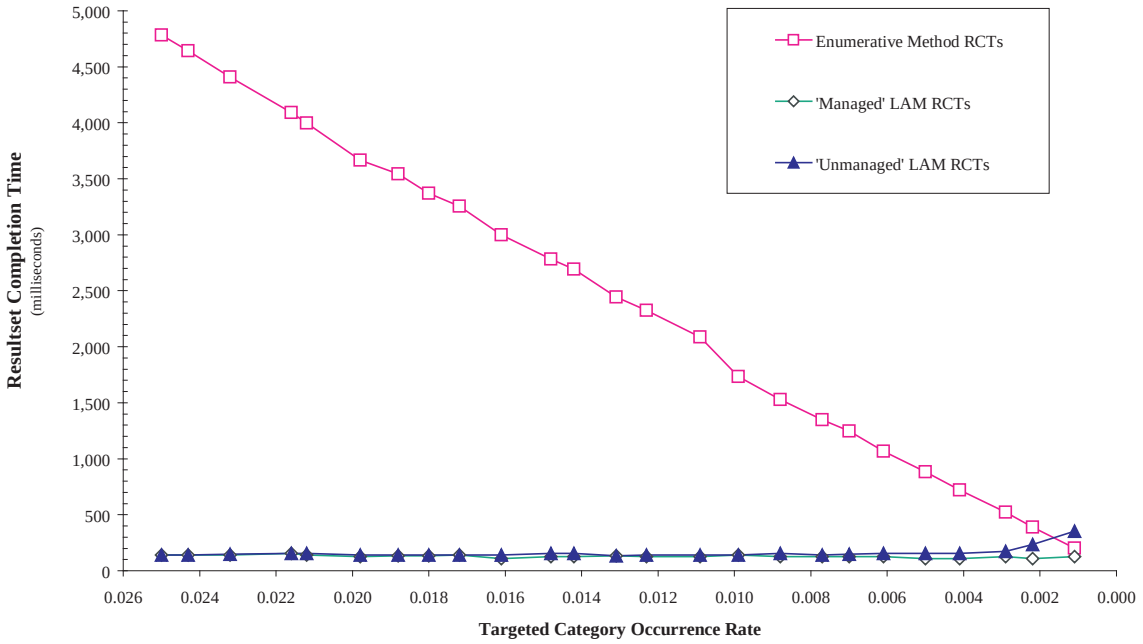
The results presented in Figure 6 (a) indicate that in the observed range of geographical coverage, the location-aware method yields query resolution performance that is superior to the enumerative method. However, the results also indicate that linkcell size must be adjusted appropriately in order to maintain this performance as the area of geographical coverage changes. As previously discussed with respect to repository size, this leads to a consideration of how the burden associated with determining linkcell size optima may be lightened by using the method of assigning sizes with reference to the mean value of $P_{TC}(S_{OPT})$. For the 24 areas associated with the results shown in Figure 6, the mean of $P_{TC}(S_{OPT})$ is 0.58. The bottom curve in Figure 6(b) shows the RCT_L values that result for each of the 24 areas when a linkcell size S is determined by setting Equation (I) =0.58. The curve suggests that as was seen previously for repository size, the application of this method results in linkcell sizes giving query resolution performance that is approximately the same as the performance realized when linkcell sizes are determined by the more cumbersome method of constructing RCT curves at successive incremental linkcell sizes in the region of $S_{0.5}$. Thus, the method of assigning linkcell sizes with reference to the mean value of $P_{TC}(S_{OPT})$ appears to be as useful in the context of variations in geographical area as it is for variations in repository size. Next, the method is assessed with respect to variations in the rate of occurrence of a specific product-service code or, in other words, with respect to using different product-service category sets to qualify repository locations.

LINKCELL SIZE OPTIMA AND TARGETED CATEGORY OCCURRENCE RATE VARIABILITY

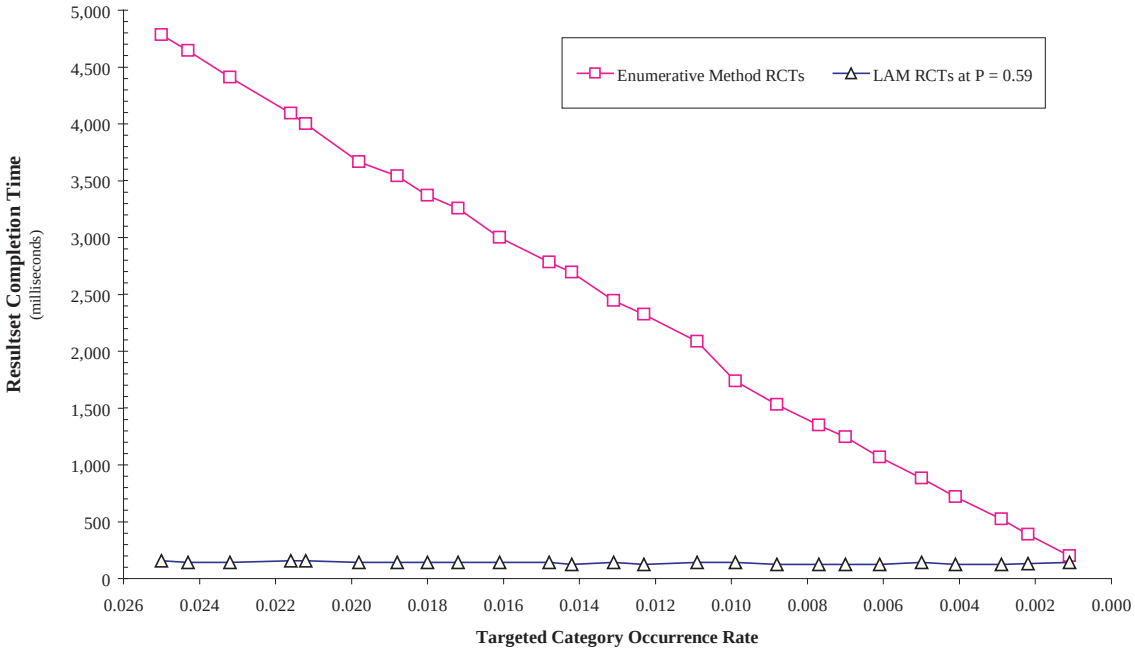
Figure 7 plots RCTs for variations in targeted category occurrence rate (TCOR) for a 100,000-location repository whose locations are distributed over a fixed geographical area (the largest of those in Figure 6). TCOR refers to the portion of a repository's locations falling into the product-service category that is targeted by a mobile consumer's query. The results seen previously in Figures 5 and 6 are based on repositories in which

Figure 7. Resultset completion times by targeted category occurrence rate

(a) RCT_E, "Unmanaged" RCT_L, and "Managed" RCT_L



(b) RCT_E and RCT_L at Mean $P_{TC}(S_{OPT})$



the product-service category attribute for each location was assigned randomly from a set of 200 product-service category codes {C001, C002, …, C200}. This, in effect, resulted in the rate with which each product-service code occurs across a repository's locations of 1/200 or, equivalently, a fixed TCOR of 0.005. The RCT values shown in Figure 7 correspond to TCOR rates that vary from 0.0250 to 0.0011. The first TCOR value indicates that 0.250 of a repository's locations bear a product-service category code matching a specific consumer-targeted category and corresponds to using a set of 40 product-service codes wherein each code occurs with equal frequency across the repository's locations. The latter TCOR value indicates that 0.0011 of a repository's locations bear a product-service category code matching a specific consumer-targeted category and corresponds to using a set of 909 equally occurring product-service codes.[6]

Figure 7(a) presents the usual three curves: (1) the topmost curve shows enumerative results; (2) the bottom curve shows optimized LAM results; and (3) the middle curve shows unmanaged LAM results. As before, the third curve reflects a circumstance in which linkcell size is set to an optimal value with respect to some initial TCOR value (0.0250, in this case) and then remains unadjusted as TCOR is changed. It is readily seen that in this case, unmanaged linkcell sizes have little appreciable effect on query resolution performance except at the smallest TCOR levels. Not until TCOR reaches 0.002 (500 product-service categories) is there a substantive separation of the two curves. The separation attains a managerially significant level in the vicinity of 0.001 (1,000 product-service categories), in which RCT_L eventually reaches 352 ms, a query resolution time that not only exceeds RCT_E but is also 2.5 times the optimized RCT_L value of 141ms.

As with the analyses respecting repository size and geographical area, TCOR-related analysis also leads to a consideration of how the repository manager's burden associated with determining linkcell size optima may be lightened by using the method of assigning sizes with reference to the mean value of $P_{TC}(S_{OPT})$. For the 24 TCOR values associated with the results shown in Figure 7, the mean of $P_{TC}(S_{OPT})$ is 0.59. The bottom curve in Figure 7(b) shows the RCT_L values that result for each of the 24 TCOR values when a linkcell size S is derived from Equation (I) = 0.59. The curve suggests, as was seen previously for repository size and geographical area, that the application of this method produces linkcell sizes resulting in query resolution performance that is approximately the same as the performance realized when linkcell sizes are determined by the $S_{0.5}$-method. Thus, the method of assigning linkcell sizes with reference to the mean value of $P_{TC}(S_{OPT})$ appears to be as valid in the TCOR context as it is in the previous two contexts.

LINKCELL SIZE DETERMINATION IN PRACTICAL M-COMMERCE OPERATIONAL CIRCUMSTANCES

The results obtained from the mean value approach to estimating optimal linkcell size in all three contexts, along with the observation that the three mean values (0.63, 0.58, and 0.59) are close to their average value of 0.60, suggest that reasonably valid estimates for optimal linkcell sizes may be obtained for practical purposes in a wide range of circumstances on the basis that:

$$P_{TC}(S) = 1 - (1 - n_{TC}/N)^{N/CS} = 0.6 \qquad \text{(II)}$$

The validity of Equation (II)'s use in linkcell size determination was assessed further by revisiting the RCT_L curves with linkcell sizes determined using Equation (II). Doing so yields query resolution performance profiles that are essentially the same as those shown in Figures 5(b), 6(b), and 7(b). Although the linkcell sizes identified through Equation (II) generally differed from those identified by the $S_{0.5}$-method, differ-

ences were minimal, and the resulting values for S always fell in a range of linkcell sizes associated with a region of minimal RCT values. Regions in this respect may be discerned in Figures 3 and 4; minimal RCT values are seen in the region in Figure 3 where linkcells vary in size from about 0.8 to 1.2 and in Figure 4 for sizes from about 0.5 to 1.0. These results, along with those seen previously, form the basis for proposing that Equation (II) represents, to this point, a heuristic with some potential to assist repository managers in realizing close-to-optimal query resolution performance and, ultimately, improved support to the location-referent transactions initiated by mobile consumers.

FURTHER WORK

Although the query resolution optimization methods used here appear to be potentially useful in realizing m-commerce service-level improvements, further work is needed in several respects. Preliminary results obtained when the location-aware method is implemented on a different computing platform reveal essentially the same RCT-linkcell size relationships that are reported here but with different RCT values. Such differences are expected and are largely attributable to differences in computational speed; however, further work is needed in order to confirm the validity of the results obtained across a greater variety of computing platforms and operational environments.

Although the major dimensions (repository size, geographical area, product-service category) associated with a location-qualified data repository in an m-commerce operational setting were addressed here, LAM's performance and the methods by which it may be optimized should be assessed in a wider variety of dimensional circumstances. For example, preliminary results from situations in which mobile consumers initiate location-referent queries from locations that are well beyond a repository's geographical boundaries suggest that values greater than 0.6 should be used to identify optimal linkcell sizes. However, these results also suggest that $P_{TC}(S) \rightarrow 0.6$ as a consumer's geographical position approaches the repository's boundaries. Further work in this respect will validate the method's applicability and assess its performance when repositories with highly localized information are used to support the location-referent queries of remotely located consumers.

Finally, the robustness of $P_{TC}(S)$'s application in circumstances that relax the assumption of a uniform distribution of locations in the same product-service category requires further examination. Product and service providers of similar type often choose locations in a non-independent, proximal fashion (e.g., law firms in legal districts, fast-food services in shopping mall food courts, retail petroleum outlets at highway intersections, etc.). Consequently, pending the outcome of further research in this respect, the use of uniform distributions of locations should be considered an important limitation on the applicability of this study's results in practical mcommerce operational settings.

CONCLUSION

Although any application of the research reported here must be done with an appreciation of its limitations and/or await the outcome of further work, the results obtained address to varying extents the four questions posed at the beginning of this article. With respect to question (1), the query resolution performance profiles observed in previous work for an invariant linkcell size are not inconsistent with those observed here. However, the present study's variant linkcell size combined with its examination of larger repository sizes, variability in geographical area, and differing product-service code sets permitted the observation of appreciable performance

degradation in unmanaged circumstances. With respect to question (2), the relationship of query resolution time to linkcell size reflects the varying dominance of two types of retrieval tasks: (1) the processing of relatively large numbers of generally smaller relational tables when linkcell sizes are small and (2) the processing of relatively small numbers of generally larger relational tables when linkcell sizes are large. The interplay of the two retrieval task types consistently produces U-shaped performance curves similar to those presented earlier in this article.

With respect to question (3), the U-shaped relationship between query resolution time and linkcell size always revealed a distinct minima or narrow region of minima indicative of the existence of a specific linkcell size that could be associated with maximum query resolution performance. Finally, with respect to question (4), the optimal linkcell size may be determined in three ways: (1) by brute force, (2) by the $S_{0.5}$-method, and (3) by solving for S such that $P_{TC}(S) = 0.6$. While the first two linkcell size determination methods were effective in revealing optimal linkcell sizes, the logistics associated with their application limits the feasibility of their deployment in practical mcommerce settings. The third method is considerably less burdensome to deploy, and results suggest that it is a useful linkcell size determination heuristic; however, further work is needed in order to assess its robustness in the face of departures from underlying assumptions and its predictive ability in a wider range of mcommerce circumstances.

REFERENCES

Arnon, A., Efrat, A., Indyk, P., & Samet, H. (1999, October 17-19). Efficient regular data structures and algorithms for location and proximity problems. *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, New York (pp. 160-170).

Butz, A.R. (1969, May). Convergence with Hilbert's space-filling curve. *Journal of Computer and System Sciences, 3*, 128-146.

Butz, A.R. (1971, April). Alternative algorithm for Hilbert's space-filling curve. *IEEE Transactions on Computers, C-20*, 424-426.

Butz, A., Bauss, J., & Kruger, A. (2000). *Different views on location awareness.* Retrieved September 16, 2005, from http://www.coli.uni-sb.de/sfb378/1999-2001/publications/butzetal2000d-de.html

Cary, M. (2001). Towards optimal ε-approximate nearest neighbor algorithms. *Journal of Algorithms, 41*(2), 417-428.

Chao, C.-M., Tseng, Y.-C., & Wang, L.-C. (2005). Dynamic bandwidth allocation for multimedia traffic with rate guarantee and fair access in WCDMA systems. *IEEE Transactions on Mobile Computing, 4*(5), 420-429.

Chou, C.-T., & Shin, K. (2005). An enhanced inter-access point protocol for uniform intra and intersubnet handoffs. *IEEE Transactions on Mobile Computing, 4*(4), 321-334.

Choy, M., Kwan, M.-P., & Hong, V. (2000). Distributed database design for mobile geographical applications. *Journal of Database Management, 11*(1), 3-15.

Cinque, M., Cotroneo, D., & Russo, S. (2005). Achieving all the time, everywhere access in next-generation mobile networks. *ACM SIGMOBILE Mobile Computing and Communications Review, 9*(2), 29-39.

Gebauer, J., & Shaw, M. (2004). Success factors and impacts of mobile business applications: Results from a mobile e-procurement study. *International Journal of Electronic Commerce, 8*(3), 19-41.

Huang, S.-M., Lin, B., & Deng, Q.-S. (2005). Intelligent cache management for mobile data

warehouse systems. *Journal of Database Management, 16*(2), 46-65.

Khungar, S., & Reikki, J. (2005). A context based storage system for mobile computing applications. *ACM SIGMOBILE Mobile Computing and Communications Review, 9*(1), 64-68.

Kottkamp, H.-E., & Zukunft, O. (1998, February 27-March 1). Location-aware query processing in mobile database systems. *Proceedings of the 1998 ACM Symposium on Applied Computing,* Atlanta, Georgia (pp. 416-423).

Kuznetsov, V.E. (2000). Method for storing map data in a database using space filling curves and a method of searching the database to find objects in a given area and to find objects nearest to a location. United States Patent Number 6,021,406, issued February 1, 2000.

Lee, C., & Ke, C.-H. (2001). A prediction-based query processing strategy in mobile commerce systems. *Journal of Database Management, 12*(3), 14-26.

Lee, D. (1999). Computational geometry II. In M. Atallah (Ed.), *Algorithms and theory of computation handbook* (pp. 20-1–20-31). Boca Raton, FL: CRC Press.

Lee, D., Xu, J., Zheng, B., & Lee, W.-C. (2002, July-September). Data management in location-dependent information services. *IEEE Pervasive Computing, 1*(3), 65-72.

Lee, D., Zhu, M., & Hu, H. (2005). When location-based services meet databases. *Mobile Information Systems, 1*(2), 81-90.

Lee, E., & Benbasat, I. (2004). A framework for the study of customer interface design for mobile commerce. *International Journal of Electronic Commerce, 8*(3), 79-102.

Leung, K., & Atypas, J. (2001). Improving returns on m-commerce investments. *The Journal of Business Strategy, 22*(5), 12-13.

Lin, H.-P., Juang, R.-T., & Lin, D.-B. (2005). Validation of an improved location-based handover algorithm using GSM measurement data. *IEEE Transactions on Mobile Computing, 4*(5), 530-536.

McGuire, M., Plataniotis, K., & Venetsanopoulos, A. (2005). Data fusion of power and time measurements for mobile terminal location. *IEEE Transactions on Mobile Computing, 4*(2), 142-153.

Nievergelt, J., & Widmayer, P. (1997). Spatial data structures: Concepts and design choices. In M. van Kreveld, J. Nievergelt, T. Roos, & P. Widmayer (Eds.), *Algorithmic foundations of geographic information systems* (pp. 153-197). Berlin: Springer Verlag.

Quintero, A. (2005). A user pattern learning strategy for managing users' mobility in UMTS networks. *IEEE Transactions on Mobile Computing, 4*(6), 552-566.

Samaan, N., & Karmouch, A. (2005). A mobility prediction architecture based on contextual knowledge and spatial conceptual maps. *IEEE Transactions on Mobile Computing, 4*(6), 537-551.

Santami, A., Leow, T., Lim, H., & Goh, P. (2003). Overcoming barriers to the successful adoption of mobile commerce in Singapore. *International Journal of Mobile Communications, 1*(1/2), 194-231.

Siau, K., Lim, E., & Shen, Z. (2001). Mobile commerce: Promises, challenges, and research agenda. *Journal of Database Management, 12*(3), 4-13.

Siau, K., & Shen, Z. (2003). Mobile communications and mobile services. *International Journal of Mobile Communications, 1*(1/2), 3-14.

Sierpinski, W. (1912). Sur une novelle courbe continue qui remplit tout une aire plaine. *Bulletin International De L'Academie Des Sciences de Cracovie, A,* 462-478.

Wyse, J. (2003). Supporting m-commerce transactions incorporating locational attributes: An evaluation of the comparative performance of a location-aware method of locations repository management. *International Journal of Mobile Communications, 1*(1/2), 119-147.

Xu, L., Shen, X., & Mark, J. (2005). Fair resource allocation with guaranteed statistical QoS for multimedia traffic in a wideband CDMA cellular network. *IEEE Transactions on Mobile Computing, 4*(2), 166-177.

Yeung, M., & Kwok, Y.-K. (2005). Wireless cache invalidation schemes with link adaptation and downlink traffic. *IEEE Transactions on Mobile Computing, 4*(1), 68-83.

Yuan, Y., & Zhang, J. (2003). Towards an appropriate business model for m-commerce. *International Journal of Mobile Communications, 1*(1/2), 35-56.

ENDNOTES

1 Resultset Completion (RCT) is the time required to extract a set of repository locations that represents a resolution of a consumer-initiated query.

2 The Enumerative Method of query resolution used both in Wyse's (2003) work and in the work here is a method that (1) selects repository locations in the consumer-targeted, product-service category, (2) calculates consumer-relative distances for each of the selected locations, (3) orders the selected locations in ascending order by consumer-relative distance, and (4) presents the first N ordered locations as the resultset that resolves the consumer's query about the nearest N locations (N = 3, in the case of Figure 2).

3 Wyse (2003) used mean RCT values as the primary statistic to measure query resolution performance. The work here has chosen to use 50th percentile RCT values as the primary performance measurement statistic, a choice that (1) minimizes the disproportionate impact of the infrequent occurrence of very large query resolution times and (2) is consistent with widely used approaches to measuring and monitoring the response time performance for computer-based transaction processing.

4 Formally, the number of linkcells C_S for linkcell size S is given by:

$$C_S = ([UVL/S] - [LVL/S] + 1)([LHL/S] - [RHL/S] + 1)$$

where UVL and LVL represent the upper and lower limits, respectively, of the vertical extent of the geographical area covered by the repository's locations, and LHL and RHL represent the left and right limits, respectively, of the area's horizontal extent. Note that [ ] denotes the greatest integer function.

5 The phrase *close to* is used deliberately, since the integrally valued components of $P_{TC}(S)$, primarily C_S, result in values for $P_{TC}(S)$ that rarely will equal 0.5.

6 Two observations with respect to Figure 7 are worthy of note: (1) query resolution times are measured at unequal TCOR intervals, and (2) RCT_E values decline as TCORs become smaller (or, equivalently, product-service code sets become larger). With respect to the first observation, the TCOR value of 0.0011 (or 909 equally occurring product-service codes) resulted when LPA was supplied with a TCOR of 0.0010 (or 1,000 equally occurring product-service codes) and then asked to generate a repository in which locations are randomly assigned a product-service code. This randomized assignment results in realized (or output) TCORs that are close to, but generally different from,

supplied (or input) TCORs. Thus, unlike the RCT values seen in Figures 5 and 6, those in Figure 7 generally do not occur at equal intervals. With respect to the second observation, note that increases in the number of product-service codes for a given repository size will result in fewer repository locations in each product-service category, including the category matching the category criteria on a consumer's query. A consideration of the enumerative method's procedural details will reveal that this circumstance results in fewer instances in which consumer-relative distances must be determined as well as in smaller resultsets that must be sorted. Thus, as TCORs become smaller, the enumerative method completes its work faster, an outcome reflected in the downward sloping curve for RCT_E.

APPENDIX A: THE LOCATION-AWARE METHOD (LAM)

A synopsis of LAM's fundamental components is given in the following with respect to (1) the requisite structure of a location-qualified data repository, (2) a formulation of the linkcell construct, (3) the general tasks of linkcell management (creation, modification, and destruction), and (4) the method's linkcell-based retrieval process. Although LAM's essentials are disclosed here, many of the method's details are not presented. A more comprehensive description may be found in Wyse (2003).

Repository Structure: The solution approach assumes that the location-qualified data repository is a relational database table containing a tuple for each repository location minimally consisting of four attributes: (1) a unique identifier for the location; (2) a horizontal coordinate (e.g., the location's longitude); (3) a vertical coordinate (e.g., the location's latitude); and (4) a code that qualifies each location in terms of its product or service offering. Table A1 provides a sample repository segment.

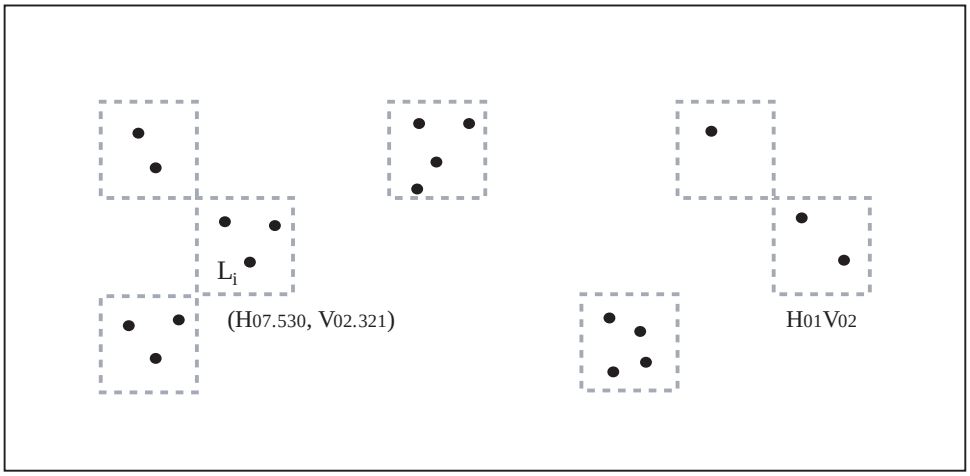
The Linkcell Construct: LAM relies on a set of auxiliary relational tables referred to as linkcells, which contain subsets of repository content and take relational table names derived from the coordinates of repository locations. Figure A1 illustrates the relationship between a repository's locations and its linkcells. Linkcells are generated based on the existence of repository locations within the area covered by the linkcell. A linkcell's name is derived from the coordinates of any location situated within the linkcell. In order to illustrate how this is accomplished, note L_i's coordinates in Figure A1. Truncating the fractional part of each coordinate yields the linkcell name. Thus, the name for the linkcell containing location L_i is H07V02. The same linkcell name also would be derived from the other two locations contained in the linkcell with L_i.

Formally, a linkcell with the relational table name HNNVMM will contain all repository locations with horizontal coordinate values HNN.0 through HNN.999··· and vertical coordinate values VMM.0 through VMM.999···. Each linkcell in its relational table form contains a tuple for each of the repository locations encompassed by the linkcell's

Table A1. Repository segment

Location Identifier	Horizontal Coordinate	Vertical Coordinate	Category Code
•	•	•	•
•	•	•	•
•	•	•	•
L0340	W112.91761	N40.71098	C001
L0341	W089.45995	N49.70451	C007
L0342	W097.81718	N47.78187	C014
L0343	W076.55539	N45.00473	C013
•	•	•	•
•	•	•	•

Figure A1. Locations and linkcells



Source: Wyse (2003), p. 125. Reproduced with the permission of Inderscience Publishers.

boundaries. Linkcells, manifested as relational tables, could appear as shown in Table A2 (with longitudes treated as horizontal coordinates and latitudes treated as vertical coordinates).

Linkcells may be varied in size through coordinate scaling. For example, if L_i's positional coordinates are scaled by 10, for example, from (H07.530, V02.321) to (H075.30, V023.21), then L_i's linkcell name becomes H075V023, and the linkcell now relates to a smaller geographical area. The smaller area in this instance consists of vertical and horizontal extents that are $1/10^{th}$ of the respective extents of the original linkcell. (It also should be noted that the coordinate scaling factor for a linkcell's horizontal component may differ from the factor used for its vertical component.) The use of a different linkcell size does not affect the relative positions of a repository's locations; however, using a different linkcell size results in a redistribution of a repository's locations among the linkcells. As seen in the results reported here, redistributions attributable to changes in linkcell size often have a substantial impact on query resolution performance.

Linkcell Creation, Modification, and Destruction: Whenever a location is added to the repository, a linkcell name is derived from the location's horizontal and vertical coordinates. The derived name is used in order to query the repository about the existence of a corresponding linkcell (relational table). If the linkcell exists, the location's identifier and category code are placed in the linkcell. If the linkcell does not exist, it is created using the name derived from the location's coordinates, and then, the location's identifier and category code are placed in the newly created linkcell. Whenever a location is removed from the repository, a linkcell name is derived from the location's coordinates. Since the location had been previously added to the repository, it is assumed that a linkcell with the derived name already exists. If the location to be removed is the only remaining location in the linkcell, then the linkcell is destroyed. If the linkcell contains

Table A2. Linkcells as relational tables

Linkcell: W112N40	
Location Id	Category Code
L0340	C001
L0736	C016
L2043	C010
L2063	C010

Linkcell: W089N49	
Location Id	Category Code
L0341	C007
L4028	C011

Linkcell: W097N47	
Location Id	Category Code
L0342	C014
L0856	C006
L1021	C001
L1326	C004
L1593	C006
L2148	C016

other location identifiers, then only the attribute tuple for the location to be removed is deleted, and the linkcell is not destroyed. Thus, linkcells (manifested as relational tables) are created, destroyed, and modified dynamically, based on repository changes.

Retrieval Procedure: The procedure relies on two types of linkcells: (1) the Core Linkcell and (2) the Cursor Linkcell. The Core Linkcell obtains its name using the method described previously but not from the coordinates of any repository location; instead, from the coordinates of the

Figure A2. LAM retrieval procedure — Cursor linkcell's naming sequence

(9) H09V03	(2) H08V03	(3) H07V03
(8) H09V02	(1) Core Linkcell H08V02	(4) H07V02
(7) H09V01	(6) H08V01	(5) H07V01

Source: Wyse (2003), p. 128. Reproduced with the permission of Inderscience Publishers.

consumer's location. The structure of the linkcell construct and its manifestation as a relational table implies that once the Core Linkcell's name is obtained, the search procedure is immediately aware of the existence or non-existence of any repository locations in the immediate vicinity of the consumer. Once derived, the Core Linkcell name remains unchanged; however, the Cursor Linkcell takes on a sequence of linkcell names that effectively moves it in search of other locations in the vicinity of the consumer.

The procedure begins by setting the Cursor Linkcell name to the Core Linkcell name and then checking for the existence of a linkcell with the same name as the Cursor Linkcell. If the linkcell exists, its contents are examined for locations with a category code equal to that sought by the query. When a sought-after location is found, its attributes are placed in the query's resultset. The procedure then expands the search area by generating a sequence of Cursor Linkcell names. This is done by systematically changing the numeric sections of the Cursor Linkcell's name using a sequence that "moves" the Cursor Linkcell around the Core Linkcell in a clockwise pattern. Whenever the Cursor Linkcell is assigned a new name, it checks for the existence of a linkcell with its currently assigned name. If the linkcell exists, then its contents are examined, and the actions outlined previously are performed, resulting in further locations being accumulated in the query resultset. The numbers in parenthesis in Figure A2 indicate the sequence in which relational table names are generated and examined in the course of a clockwise movement of the Cursor Linkcell. The search area may be expanded further by moving the Cursor Linkcell through a layer of linkcells on the outer periphery of the linkcells previously examined. This outward-spiraling, clockwise-moving process continues until the sought-after number of locations is found.

This work was previously published in the Journal of Database Management, edited by K. Siau, Volume 17, Issue 3, pp. 41-65, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.35
Data Dissemination in Mobile Environments

Panayotis Fouliras
University of Macedonia, Greece

ABSTRACT

Data dissemination today represents one of the cornerstones of network-based services and even more so for mobile environments. This becomes more important for large volumes of multimedia data such as video, which have the additional constraints of speedy, accurate, and isochronous delivery often to thousands of clients. In this chapter, we focus on video streaming with emphasis on the mobile environment, first outlining the related issues and then the most important of the existing proposals employing a simple but concise classification. New trends are included such as overlay and p2p network-based methods. The advantages and disadvantages for each proposal are also presented so that the reader can better appreciate their relative value.

INTRODUCTION

A well-established fact throughout history is that many social endeavors require dissemination of information to a large audience in a fast, reliable, and cost-effective way. For example, mass education could not have been possible without paper and typography. Therefore, the main factors for the success of any data dissemination effort are supporting technology and low cost.

The rapid evolution of computers and networks has allowed the creation of the Internet with a myriad of services, all based on rapid and low cost data dissemination. During recent years, we have witnessed a similar revolution in mobile devices, both in relation to their processing power as well as their respective network infrastructure. Typical representatives of such networks are the 802.11x for LANs and GSM for WANs.

In this context, it is not surprising that the main effort has been focusing on the dissemination of multimedia content–especially audio and video, since the popularity of such services is high, with RTP the de-facto protocol for multimedia data transfer on the Internet. Although both audio and video have strict requirements in terms of packet jitter (the variability of packet delays within the same packet stream), video additionally requires

significant amount of bandwidth due to its data size. Moreover, a typical user requires multimedia to be played in real-time, (i.e., shortly after his request, instead of waiting for the complete file to be downloaded; this is commonly referred to as *multimedia streaming.*

In most cases, it is assumed that the item in demand is already stored at some server(s) from where the clients may request it. Nevertheless, if the item is popular and the client population very large, additional methods must be devised in order to avoid a possible drain of available resources. Simple additional services such as fast forward (FF) and rewind (RW) are difficult to support, let alone interactive video. Moreover, the case of asymmetric links (different upstream and downstream bandwidth) can introduce more problems. Also, if the item on demand is not previously stored but represents an ongoing event, many of the proposed techniques are not feasible.

In the case of mobile networks, the situation is further aggravated, since the probability of packet loss is higher and the variation in device capabilities is larger than in the case of desktop computers. Furthermore, ad-hoc networks are introduced, where it is straightforward to follow the bazaar model, under which a client may enter a wall mart and receive or even exchange videos in real time from other clients, such as specially targeted promotions, based on its profile. Such a model complicates the problem even further.

In this chapter, we are focusing on video streaming, since video is the most popular and demanding multimedia data type (Sripanidkulchai, Ganjam, Maggs, & Zhang, 2004). In the following sections, we are identifying the key issues, present metrics to measure the efficiency of some of the most important proposals and perform a comparative evaluation in order to provide an adequate guide to the appropriate solutions.

ISSUES

As stated earlier, streaming popular multimedia content with large size such as video has been a challenging problem, since a large client population demands the same item to be delivered and played out within a short period of time. This period should be smaller that the time t_w a client would be willing to wait after it made its request. Typically there are on average a constant number of requests over a long time period, which suggests that a single broadcast should suffice for each batch of requests. However, the capabilities of all entities involved (server, clients, and network) are finite and often of varying degree (e.g., effective available network and client bandwidth). Hence the issues and challenges involved can be summarized as follows:

- What should the broadcasting schedule of the server be so that the maximum number of clients' requests is satisfied without having them wait more than t_w?
- How can overall network bandwidth be minimized?
- How can the network infrastructure be minimally affected?
- How can the clients assist if at all?
- What are the security considerations?

In the case of mobile networks, the mobile devices are the clients; the rest of the network typically is static, leading to a mixed, hybrid result. Nevertheless, there are exceptions to this rule, such as the ad hoc networks. Hence, for mobile clients there are some additional issues:

- Mobile clients may leave or appear to leave a session due to higher probability of packet loss. How does such a system recover from this situation?

- How can redirection (or *handoff*) take place without any disruption in play out quality?
- How can the bazaar model be accommodated?

BACKGROUND

In general, without prior knowledge on how the data is provided by the server, a client has to send a request to the server. The server then either directly delivers the data (on demand service) or replies with the broadcast channel access information (e.g., channel identifier, estimated access time, etc.). In the latter case, if the mobile client decides so, it monitors the broadcast channels (Hu, Lee, & Lee, 1998). In both cases, there have been many proposals, many of which are also suitable for mobile clients. Nevertheless, many proposals regarding mobile networks are not suitable for the multimedia dissemination. For example, Coda is a file replication system, Bayou a database replication system and Roam a slightly more scalable general file replication system (Ratner, Reiher, & Popek, 2004), all of which do not assume strict temporal requirements.

The basic elements which comprise a dissemination system are the server(s), the clients, and the intermittent network. Depending on which of these is the focus, the various proposals can be classified into two broad categories: Proposals regarding the server organization and its broadcast schedule, and those regarding modifications in the intermittent network or client model of computation and communication.

Proposals According to Server Organization and Broadcasting Schedule

Let us first examine the various proposals in terms of the server(s) organization and broadcasting schedule. These can be classified in two broad classes, namely *push*-based scheduling (or *proactive*) and *pull*-based scheduling (or *reactive*). Under the first class, the clients continuously monitor the broadcast process from the server and retrieve the required data without explicit requests, whereas under the second class the clients make explicit requests which are used by the server to make a schedule which satisfies them. Typically, a hybrid combination of the two is employed with push-based scheduling for popular and pull-based scheduling for less popular items (Guo, Das, & Pinotti, 2001).

Proposals for Popular Videos

For the case of pushed-based scheduling broadcasting schedules of the so-called *periodic broadcasting* type are usually employed: The server organizes each item in segments of appropriate size, which it broadcasts periodically. Interested clients simply start downloading from the beginning of the first segment and play it out immediately. The clients must be able to preload some segments of the item and be capable of downlink bandwidth higher than that for a single video stream. Obviously this scheme works for popular videos, assuming there is adequate bandwidth at the server in relation to the amount and size of items broadcasted.

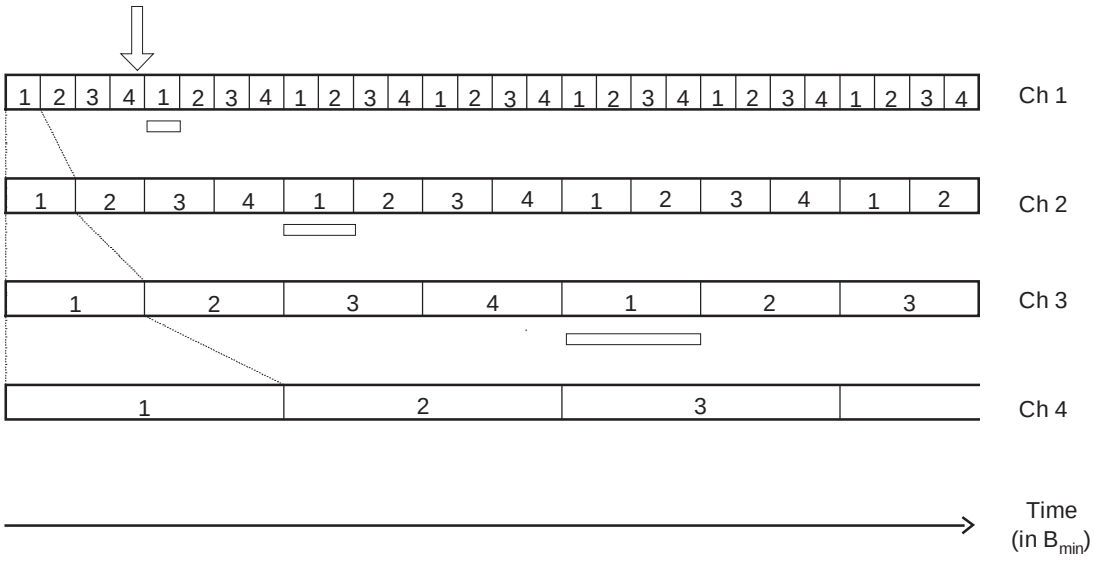
Pyramid broadcasting (PB) (Viswanathan & Imielinski, 1995) has been the first proposal in this category. Here, each client is capable of downloading from up to two channels simultaneously. The video is segmented in s segments of increasing size, so that $s_{i+1}=\alpha \cdot s_i$, where $\alpha = \frac{B}{MK}$ and B is the total server bandwidth expressed in terms of the minimum bandwidth b_{min} required to play out a single item, M the total number of videos and K the total number of virtual server channels. Each channel broadcasts a separate segment of the same video periodically, at a speed higher than b_{min}. Thus, with M=4, K=4 and B=32,

we have α=2, which means that each successive segment is twice the size of the previous one. Each segment is broadcasted continuously from a dedicated channel as depicted in Figure 1. In our example, each server channel has bandwidth $B'=B/K=8 \cdot b_{min}$, which means that the clients must have a download bandwidth of $16 \cdot b_{min}$.

If D is the duration of the video, then the waiting time of a client is at most $M \cdot s_1/B'$. With D = 120 and $K = M = 4$, we have $M \cdot s_1/B' = 4 \cdot 8/8 = 4$ time units. Each segment from the first channel requires 1 time unit to be downloaded, but has a play out time of 8 units. Consider the case that a client requests video 1 at the time indicated by the thick vertical arrow. Here the first three segments to be downloaded are indicated by small grey rectangles. By the time the client has played out half of the first segment from channel 1 it will start downloading the second segment from channel 2 and so on. The obvious drawback of this scheme is that it requires a very large download bandwidth at the client as well as a large buffer to store the preloaded segments (as high as 70% of the video).

In order to address these problems, other methods have been proposed, such as *permutation-based pyramid broadcasting* (PPB) (Aggarwal, Wolf, & Yu, 1996) and *skyscraper broadcasting* (SB) (Hua & Sheu, 1997). Under PPB each of the K channels is multiplexed into P subchannels with P times lower rate, where the client may alternate the selection of subchannel during download. However, the buffer requirements are still high (about 50% of the video) and synchronization is difficult. Under SB, two channels are used for downloading, but with a rate equal to the playing rate B_{min}. Relative segment sizes are 1, 2, 2, 5, 5, 12, 12, 25, 25,…W, where W the width of the skyscraper. This leads to much lower demand on the client, but is inefficient in terms of server bandwidth. The latter goal is achieved by *fast broadcasting* (FB) (Juhn & Tseng, 1998) which

Figure 1. Example of pyramid broadcasting with 4 videos and 4 channels



divides the video into segments of geometric series, with K channels of B_{min} bandwidth, but where the clients download from *all K* channels.

Yet another important variation is *harmonic broadcasting* (HB) (Juhn & Tseng, 1997) which divides the video in segments of equal size and broadcasts them on K successive channels of bandwidth B_{min}/i, where $i = 1,\ldots K$. The client downloads from all channels as soon as the first segment has started downloading. The client download bandwidth is thus equal to the server's and the buffer requirements low (about 37% of the total video). However, the timing requirements may not be met, which is a serious drawback. Other variations exist that solve this problem with the same requirements (Paris, Carter, & Long, 1998) or are hybrid versions of the schemes discussed so far, with approximately the same cost in resources as well as efficiency.

Proposals for Less Popular Videos or Varying Request Pattern

In the case of less popular videos or of a varying request pattern pulled-based or reactive methods are more appropriate. More specifically, the server gathers clients' requests within a specific time interval $t_{in} < t_w$. In the simplest case all requests are for the beginning of the same video, although they may be for different videos or for different parts of the same video (e.g., after a FF or RW). For each group (*batch*) of similar requests a new broadcast is scheduled by reserving a separate server channel, (*batching*). With a video duration t_D a maximum of $\lceil t_D/t_{in} \rceil$ server channels are required for a single video assuming multicast.

The most important proposals for *static* multicast batching are: *first-come-first-served* (FCFS) where the oldest batch is served first, *maximum-queue-length-first* (MQLF) where the batch containing the largest amount of requests is served first, reducing average system throughput by being unfair and *maximum-factor-queue-length* (MFQL) where the batch containing the largest amount of requests for some video weighted by the factor $1/\sqrt{f_i}$ is selected, where f_i is the access frequency of the particular video. In this way the popular videos are not always favored (Hua, Tantaoui, & Tavanapong, 2004).

A common drawback of the proposals above is that client requests which miss a particular video broadcasting schedule cannot hope for a reasonably quick service time, in a relatively busy server. Hence, *dynamic* multicast proposals have emerged, which allow the existing multicast tree for the same video to be extended in order to include late requests. The most notable proposals are *patching*, *bandwidth skimming*, and *chaining*.

Patching (Hua, Cai, & Sheu, 1998) and its variations allow a late client to join an existing multicast stream and buffer it, while simultaneously the missing portion is delivered by the server via a separate patching stream. The latter is of short duration, thus quickly releasing the bandwidth used by the server. Should the clients arrive towards the end of the normal stream broadcast, a new normal broadcast is scheduled instead of a patch one. In more recent variations it is also possible to have double patching, where a patching stream is created on top of a previous patching stream, but requires more bandwidth on both the client(s) and the server and synchronization is more difficult to achieve.

The main idea in Bandwidth Skimming (Eager, Vernon, & Zahorjan, 2000) is for clients to download a multicast stream, while reserving a small portion of their download bandwidth (*skim*) in order to listen to the closest active stream other than theirs. In this way, hierarchical merging of the various streams is possible to achieve. It has been shown that it is better than patching in terms of server bandwidth utilization, though more complex to implement.

Chaining (Sheu, Hua, & Tavanapong, 1997) on the other hand is essentially a pipeline of clients, operating in a peer-to-peer scheme, where the server is at the root of the pipeline. New clients are added at the bottom of the tree,

receiving the first portion of the requested video. If an appropriate pipeline does not exist, a new one is created by having the server feed the new clients directly. This scheme reduces the server bandwidth and is scalable, but it requires a collaborative environment and implementation is a challenge, especially for clients who are in the middle of a pipeline and suddenly lose network connection or simply decide to withdraw. It also requires substantial upload bandwidth to exist at the clients, so it is not generally suitable for asymmetric connections.

Proposals According to Network and Client Organization

Proxies and Content Distribution Networks

Proxies have been used for decades for delivering all sorts of data and especially on the Web, with considerable success. Hence there have been proposals for their use for multimedia dissemination. Actually, some of the p2p proposals discussed later represent a form of proxies, since they cache part of the data they receive for use by their peers. A more general form of this approach, however, involves dedicated proxies strategically placed so that they are more effective.

Wang, Sen, Adler, and Towsley, (2004) base their proposal on the principle of prefix proxy cache allocation in order to reduce the aggregate network bandwidth cost and startup delays at the clients. Although they report substantial savings in transmission cost, this is based on the assumption that all clients request a video from its beginning.

A more comprehensive study based on Akamai's streaming network appears in (Sripanidkulchai, Ganjam, Maggs, & Zhang, 2004). The latter is a static overlay composed of edge nodes located close to the clients and intermediate nodes that take streams from the original content publisher and split and replicate them to the edge nodes. This scheme effectively constitutes a content distribution network (CDN), used not only for multimedia, but other traffic as well. It is reported that under several techniques and assumptions tested, application end-point architectures have enough resources, inherent stability and can support large-scale groups. Hence, such proposals (including p2p) are promising for real-world applications. Client buffers and uplink bandwidth can contribute significantly if it is possible to use them.

Multicast Overlay Networks

Most of the proposals so far work for multicast broadcasts. This suggests that the network infrastructure supports IP multicasting completely. Unfortunately, most routers in the Internet do not support multicast routing. As the experience from MBone (multicast backbone) (Kurose, & Ross, 2004) shows, an overlay virtual network interconnecting "islands" of multicasting-capable routers must be established over the existing Internet using the rest of the routers as end-points of "tunnels." Nevertheless, since IP multicasting is still a best effort service and therefore unsuitable for multimedia streaming, appropriate reservation of resources at the participating routers is necessary. The signaling protocol of choice is RSVP under which potential receivers signal their intention to join the multicast tree. This is a defacto part of the Intserv mechanism proposed by IETF. However, this solution does not scale well. A similar proposal but with better scaling is DiffServ which has still to be deployed in numbers (Kurose, & Ross, 2004).

A more recent trend is to create an overlay multicast network at the application layer, using unicast transmissions. Although worse than pure multicast in theory, it has been an active area of research due to its relative simplicity, scalability and the complete absence of necessity for modifications at the network level. Thus, the complexity is now placed at the end points, (i.e.,

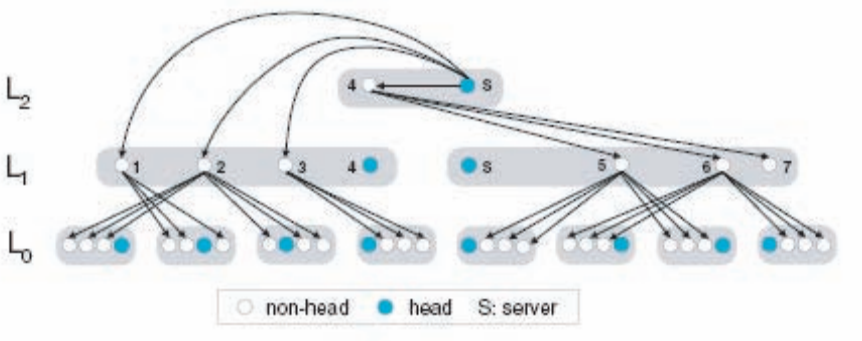
the participating clients and server(s)) and the popular point-to-point (p2p) computation model can be employed in most cases. Asymmetric connections must still include uplink connections of adequate bandwidth in order to support the p2p principle.

Variations include P2Cast (Guo, Suh, Kurose, & Towsley, 2003) which essentially is patching in the p2p environment: Late clients receive the patch stream(s) from old clients, by having two download streams, namely the normal and the patch stream. Any failure of the parent involves the source (the initial server), which makes the whole mechanism vulnerable and prone to bottlenecks.

ZigZag (Tran, Hua, & Do, 2003) creates a logical hierarchy of clusters of peers, with each member at a bounded distance from each other and one of them the cluster leader. The name of this technique emanates from the fact that the leader of each cluster forwards data only to peers in different clusters from its own. An example is shown in Figure 2, where there are 16 peers, organized in clusters of four at level 0. One peer from each cluster is the cluster leader or *head* (additionally depicted for clarity) at level 1. The main advantages of ZigZag are the small height of the multicast tree and the amount of data and control traffic at the server. However, leader failures can cause significant disruption, since both data and control traffic pass through a leader.

LEMP (Fouliras, Xanthos, Tsantalis., & Manitsaris, 2004) is a another variation which forms a simple overlay multicast tree with an upper bound on the number of peers receiving data from their parent. However, each level of the multicast tree forms a virtual cluster where one peer is the local representative (LR) and another peer is its backup, both initially selected by the server. Most of the control traffic remains at the same level between the LR and the rest of the peers. Should the LR fail, the backup takes its place, selecting a new backup. All new clients are assigned by the server to an additional level under the most recent or form a new level under the server with a separate broadcast. Furthermore, special care has been made for the case of frequent disconnections and re-connections, typical for mobile environments; peers require a single downlink channel at play rate and varying, but bounded uplink channels. This scheme has better response to failures and shorter trees than ZigZag, but for very populous levels there can be some bottleneck for the light control traffic at the LR.

Figure 2. ZigZag: Example multicast tree of peers (3 layers, 4 peers per cluster)



Other Proposals

Most of the existing proposals have been designed without taking into consideration the issues specific to mobile networks. Therefore, there has recently been considerable interest for research in this area. Most of the proposed solutions, however, are simple variations of the proposals presented already. This is natural, since the network infrastructure is typically static and only clients are mobile. The main exception to this rule comes from ad hoc networks.

Add hoc networks are more likely to show packet loss, due to the unpredictable behavior of all or most of the participant nodes. For this reason there has been considerable research effort to address this particular problem, mostly by resorting to multipath routing, since connectivity is less likely to be broken along multiple paths. For example, (Zhu, Han, & Girod, 2004) elaborate on this scheme, by proposing a suitable objective function which determines the appropriate rate allocation among multiple routes. In this way congestion is also avoided considerably, providing better results at the receiver. Also (Wei, & Zakhor, 2004) propose a multipath extension to an existing on-demand source routing protocol (DSR), where the packet carries the end-to-end information in its header and a route discovery process is initiated in case of problems and (Wu, & Huang, 2004) for the case of heterogeneous wireless networks.

All these schemes work reasonably well for small networks, but their scalability is questionable, since they have been tested for small size networks.

COMPARATIVE EVALUATION

We assume that the play out duration t_D of the item on demand is in general longer than at least an order of magnitude compared to t_w. Furthermore, we assume that the arrival of client requests is a Poisson distribution and that the popularity of items stored at the server follows the Zipf distribution. These assumptions are in line with those appearing in most of the proposals.

In order to evaluate the various proposals we need to define appropriate metrics. More specifically:

- Item access time: this should be smaller than t_w as detailed previously
- The bandwidth required at the server as a function of client requests
- The download and upload bandwidth required at a client expressed in units of the minimum bandwidth b_{min} for playing out a single item
- The minimum buffer size required at a client
- The maximum delay during redirection, if at all; obviously this should not exceed the remainder in the client's buffer
- The overall network bandwidth requirements
- Network infrastructure modification; obviously minimal modification is preferable
- Interactive capabilities

Examining the proposals for popular videos presented earlier, we note that they are unsuitable for mobile environments, either because they require a large client buffer, large bandwidth for downloads or very strict and complex synchronization. Furthermore, they were designed for popular videos with a static request pattern, where clients always request videos from their beginning.

On the other hand, patching, bandwidth skimming are better equipped to address these problems, but unless multicasting is supported, may overwhelm the server. Chaining was designed for multicasting, but uses the p2p computation model, lowering server load and bandwidth.

Nevertheless, unicast-based schemes are better in practice for both wired and mobile networks as

stated earlier. Although several proposals exist, Zigzag and LEMP are better suited for mobile environments, since they have the advantages of chaining, but are designed having taken into consideration the existence of a significant probability of peer failures, as well as the case of ad hoc networks and are scalable. Their main disadvantage is that they require a collaborative environment and considerable client upload bandwidth capability, which is not always the case for asymmetric mobile networks. Furthermore, they reduce server bandwidth load, but not the load of the overall network.

The remaining proposals either assume a radical reorganization of the network infrastructure (CDN) or are not proven to be scalable.

CONCLUSION AND FUTURE TRENDS

The research conducted by IETF for quality of service (QoS) in IP-based mobile networks and QoS policy control is of particular importance. Such research is directly applicable to the dissemination of multimedia data, since the temporal requirement may lead to an early decision for packet control, providing better network bandwidth utilization. The new requirements of policy control in mobile networks are set by the user's home network operator, depending upon a profile created for the user. Thus, certain sessions may not be allowed to be initiated under certain circumstances (Zheng, & Greis, 2004).

In this sense, most mobile networks will continue being hybrid in nature for the foreseeable future, since this scheme offers better control for administrative and charging reasons, as well as higher effective throughput and connectivity to the Internet. Therefore, proposals based on some form of CDN are better suited for commercial providers. Nevertheless, from a purely technical point of view, the p2p computation model is better suited for the mobile environment, with low server bandwidth requirements, providing failure tolerance and, most important, inherently supporting ad hoc networks and interactive multimedia.

REFERENCES

Aggarwal, C., Wolf, J., & Yu, P. (1996). A permutation based pyramid broadcasting scheme for video on-demand systems. *IEEE International Conference on Multimedia Computing and Systems (ICMCS '96),* (pp. 118-126), Hiroshima, Japan.

Eager, D., Vernon, M., & Zahorjan, J. (2000). Bandwidth skimming: A technique for cost-effective video-on-demand. *Proceedings of IS&T/SPIE Conference on Multimedia Computing and Networking (MMCN 2000)* (pp. 206-215).

Fouliras, P., Xanthos, S., Tsantalis, N., & Manitsaris, A. (2004). LEMP: Lightweight efficient multicast protocol for video on demand. *ACM Symposium on Applied Computing (SAC'04)* (pp. 1226-1231), Nicosia, Cyprus.

Guo, Y., Das, S., & Pinotti, M. (2001). A new hybrid broadcast scheduling algorithm for asymmetric communication systems: Push and pull data based on optimal cut-off point. *Mobile Computing and Communications Review (MC2R), 5*(3), 39-54. ACM.

Guo, Y., Suh, K., Kurose, J., & Towsley, D. (2003). A peer-to-peer on-demand streaming service and its performance evaluation. *IEEE International Conference on Multimedia Expo (ICME '03)* (pp. 649-652).

Hu, Q., Lee, D., & Lee, W. (1998). Optimal channel allocation for data dissemination in mobile computing environments. *International Conference on Distributed Computing Systems* (pp. 480-487).

Hua, K., Tantaoui, M., & Tavanapong, W. (2004). Video delivery technologies for large-scale de-

ployment of multimedia applications. *Proceedings of the IEEE, 92*(9), 1439-1451.

Hua, K., & Sheu, S. (1997). Skyscraper broadcasting: A new broadcasting scheme for metropolitan video-on-demand systems. *ACM Special Interest Group on Data Communication (SIGCOMM '97)* (pp. 89-100), Sophia, Antipolis, France.

Hua, K., Cai, Y. & Sheu, S. (1998). Patching: A multicast technique for true video-on-demand services. *ACM Multimedia '98* (pp. 191-200), Bristol, UK.

Juhn, L., & Tseng, L. (1997). Harmonic broadcasting for video-on-demand service. *IEEE Transactions on Broadcasting, 43*(3), 268-271.

Juhn, L., & Tseng, L. (1998). Fast data broadcasting and receiving scheme for popular video service. *IEEE Transactions on Broadcasting, 44*(1), 100-105.

Kurose, J., & Ross, K. (2004). *Computer networking: A top-down approach featuring the Internet* (3rd ed.). Salford, UK: Addison Wesley; Pearson Education.

Paris, J., Carter, S., & Long, D. (1998). A low bandwidth broadcasting protocol for video on demand. *IEEE International Conference on Computer Communications and Networks (IC3N'98)* (pp. 690-697).

Ratner, D., Reiher, P., & Popek, G. (2004). Roam: A scalable replication system for mobility. *Mobile Networks and Applications, 9,* 537-544). Kluwer Academic Publishers.

Sheu, S., Hua, K., & Tavanapong, W. (1997). Chaining: A generalized batching technique for video-on-demand systems. *Proceedings of the IEEE ICMCS'97* (pp. 110-117).

Sripanidkulchai, K., Ganjam, A., Maggs, B., & Zhang, H. (2004). The feasibility of supporting large-scale live streaming applications with dynamic application end-points. *ACM Special Interest Group on Data Communication (SIGCOMM'04)* (pp. 107-120), Portland, OR.

Tran, D., Hua, K., & Do, T. (2003). Zigzag: An efficient peer-to-peer scheme for media streaming. *Proceedings of IEEE Infocom* (pp. 1283-1293).

Viswanathan, S., & Imielinski, T. (1995). Pyramid broadcasting for video-on-demand service. *Proceedings of the SPIE Multimedia Computing and Networking Conference* (pp. 66-77).

Wang, B., Sen, S., Adler, M., & Towsley, D. (2004). Optimal proxy cache allocation for efficient streaming media distribution. *IEEE Transaction on Multimedia, 6*(2), 366-374.

Wei, W., & Zakhor, A. (2004). Robust multipath source routing protocol (RMPSR) for video communication over wireless ad hoc networks. *International Conference on Multimedia and Expo (ICME)* (pp. 27-30).

Wu, E., & Huang, Y. (2004). Dynamic adaptive routing for a heterogeneous wireless network. *Mobile Networks and Applications, 9,* 219-233.

Zheng, H., & Greis, M. (2004). Ongoing research on QoS policy control schemes in mobile networks. *Mobile Networks and Applications, 9,* 235-241. Kluwer Academic Publishers.

Zhu, X., Han, S., & Girod, B. (2004). Congestion-aware rate allocation for multipath video streaming over ad hoc wireless networks. *IEEE International Conference on Image Processing (ICIP-04).*

KEY TERMS

CDN: Content distribution network is a network where the ISP has placed proxies in strategically selected points, so that the bandwidth used and response time to clients' requests is minimized.

Overlay Network: A virtual network built over a physical network, where the participants communicate with a special protocol, transparent to the non-participants.

QoS: A notion stating that transmission quality and service availability can be measured, improved, and, to some extent, guaranteed in advance. QoS is of particular concern for the continuous transmission of multimedia information and declares the ability of a network to deliver traffic with minimum delay and maximum availability.

Streaming: The scheme under which clients start playing out the multimedia immediately or shortly after they have received the first portion without waiting for the transmission to be completed.

This work was previously published in Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 38-48, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.36
Data Broadcasting in a Mobile Environment

A. R. Hurson
The Pennsylvania State University, USA

Y. Jiao
The Pennsylvania State University, USA

ABSTRACT

The advances in mobile devices and wireless communication techniques have enabled anywhere, anytime data access. Data being accessed can be categorized into three classes: private data, shared data, and public data. Private and shared data are usually accessed through on-demand-based approaches, while public data can be most effectively disseminated using broadcasting. In the mobile computing environment, the characteristics of mobile devices and limitations of wireless communication technology pose challenges on broadcasting strategy as well as data-retrieval method designs. Major research issues include indexing scheme, broadcasting over single and parallel channels, data distribution and replication strategy, conflict resolution, and data retrieval method. In this chapter, we investigate solutions proposed for these issues. High performance and low power consumption are the two main objectives of the proposed schemes. Comprehensive simulation results are used to demonstrate the effectiveness of each solution and compare different approaches.

INTRODUCTION

The increasing development and spread of wireless networks and the need for information sharing has created a considerable demand for cooperation among existing, distributed, heterogeneous, and autonomous information sources. The growing diversity in the range of information that is accessible to a user and rapidly expanding technology have changed the traditional notion of timely and reliable access to global information in a distributed system. Remote access to data refers to both mobile nodes and fixed nodes accessing data within a platform characterized by the following:

- low bandwidth,
- frequent disconnection,
- high error rates,
- limited processing resources, and
- limited power sources.

Regardless of the hardware device, connection medium, and type of data accessed, users require timely and reliable access to various types of data that are classified as follows:

- Private data, that is, personal daily schedules, phone numbers, and so forth. The reader of this type of data is the sole owner or user of the data.
- Public data, that is, news, weather information, traffic information, flight information, and so forth. This type of data is maintained by one source and shared by many—a user mainly queries the information source(s).
- Shared data, that is, traditional, replicated, or fragmented databases. Users usually send transactions as well as queries to the information source(s).

  Access requests to these data sources can be on-demand-based or broadcast-based.

On-Demand-Based Requests

In this case users normally obtain information through a dialogue (two-way communication) with the database server—the request is pushed to the system, data sources are accessed, operations are performed, partial results are collected and integrated, and the final result is communicated back to the user. This access scenario requires a solution that addresses the following issues.

- **Security and access control.** Methods that guarantee authorized access to the resources.
- **Isolation.** Means that support operations off-line if an intentional or unintentional disconnection has occurred.
- **Semantic heterogeneity.** Methods that can handle differences in data representation, format, structure, and meaning among information sources and hence establish interoperability.
- **Local autonomy.** Methods that allow different information sources to join and depart the global information-sharing environment at will.
- **Query processing and query optimization.** Methods that can efficiently partition global queries into subqueries and perform optimization techniques.
- **Transaction processing and concurrency control.** Methods that allow simultaneous execution of independent transactions and interleave interrelated transactions in the face of both global and local conflicts.
- **Data integration.** Methods that fuse partial results to draw a global result.
- **Browsing.** Methods that allow the user to search and view the available information without any information processing overhead.
- **Distribution transparency.** Methods to hide the network topology and the placement of the data while maximizing the performance for the overall system.
- **Location transparency.** Methods that allow heterogeneous remote access (HRA) to data sources. Higher degrees of mobility argue for higher degrees of heterogeneous data access.
- **Limited resources.** Methods that accommodate computing devices with limited capabilities.

The literature is abounded with solutions to these issues (Badrinath, 1996; Bright, Hurson, & Pakzad, 1992, 1994; Joseph, Tauber, & Kaashoek, 1997; Satyanarayanan, 1996). Moreover, there are existing mobile applications that address the limited bandwidth issues involved in mobility (Demers, Pertersen, Spreitzer, Terry, Theier,

& Welch, 1994; Fox, Gribble, Brewer, & Amir, 1996; Honeyman, Huston, Rees, & Bachmann, 1992; Joseph et al., 1997; Kaashoek, Pinckney, & Tauber, 1995; Lai, Zaslavsky, Martin, & Yeo, 1995; Le, Burghardt, Seshan, & Rabaey, 1995; Satyanarayanan, 1994, 1996).

Broadcast-Based Requests

Public information applications can be characterized by (a) massive numbers of users and (b) the similarity and simplicity in the requests solicited by the users. The reduced bandwidth attributed to the wireless environment places limitations on the rate of the requests. Broadcasting (one-way communication) has been suggested as a possible solution to this limitation. In broadcasting, information is provided to all users of the air channels. Mobile users are capable of searching the air channels and pulling the desired data. The main advantage of broadcasting is that it scales up as the number of users increases and, thus, eliminates the need to multiplex the bandwidth among users accessing the air channel. Furthermore, broadcasting can be considered as an additional storage available over the air for mobile clients. Within the scope of broadcasting one needs to address three issues:

- effective data organization on the broadcast channel,
- efficient data retrieval from the broadcast channel, and
- data selection.

The goal is to achieve high performance (response time) while minimizing energy consumption. Note that the response time is a major source of power consumption at the mobile unit (Imielinski & Badrinath, 1994; Imielinski & Korth, 1996; Imielinski, Viswanathan, & Badrinath, 1997; Weiser, 1993). As a result, the reduction in response time translates into reducing the amount of time a mobile unit spends accessing the channel(s) and thus has its main influence on conserving energy at the mobile unit.

Chapter Organization

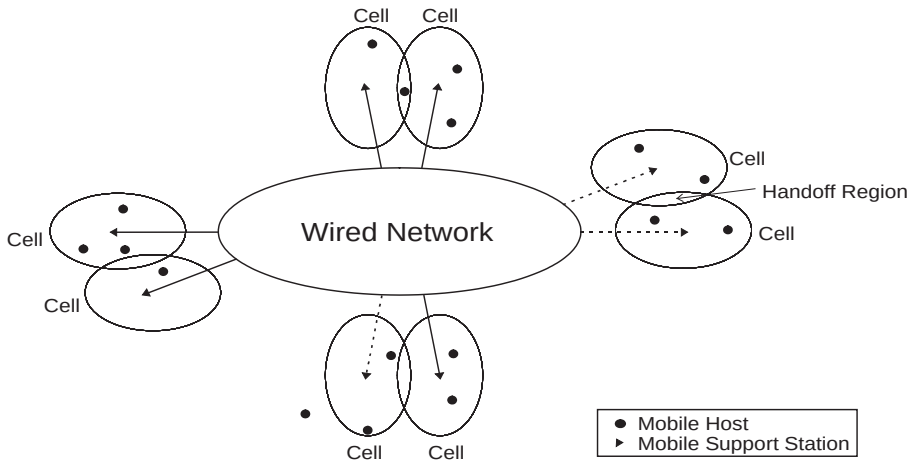
In this chapter, we first introduce the necessary background material. Technological limitations are outlined and their effects on the global information-sharing environment are discussed. Issues such as tree-based indexing, signature-based indexing, data replication, broadcasting over single and parallel channels, data distribution, conflict, and data access are enumerated and analyzed next. Then we present solutions to these issues with respect to the network latency, access latency, and power management. Finally, we conclude the chapter and point out some future research directions.

MOBILE COMPUTING

The mobile computing environment is composed of a number of network servers enhanced with wireless transceivers—mobile support stations (MSSs) and a varying number of mobile hosts (MHs) free to move at will (Figure 1).

The role of the MSS is to provide a link between the wireless network and the wired network. The link between an MSS and the wired network could be either wireless (shown as a dashed line) or wire based. The area covered by the individual transceiver is referred to as a cell. To satisfy a request, an MH accesses the MSS responsible for the cell where the MH is currently located. It is the duty of the MSS to resolve the request and deliver the result back to the client. Once an MH moves across the boundaries of two cells, a handoff process takes place between the MSSs of the corresponding cells. The MH is normally small, lightweight, and portable. It is designed to be compact with limited resources relying on temporary power supplies (such as batteries) as its main power source.

Figure 1. Architecture of the mobile-computing environment



Characteristics of the Mobile Environment

Wireless communication is accomplished via modulating radio waves or pulsing infrared light. Table 1 summarizes a variety of mobile network architectures. Mainly, three characteristics distinguish the mobile computing environment from traditional wired computing platforms, namely, wireless medium, mobility, and portability.

Wireless Medium

The common ground among all wireless systems is the fact that communication is done via the air (and not via cables). This fact changes a major underlying assumption behind the conventional distributed algorithms. The physical layer of the connection is no longer the reliable coaxial or optic cable. Communication over the air is identified by frequent disconnections, low data-rate, high cost, and lack of security (Alonso & Ganguly, 1992; Alonso & Korth, 1993; Chlamtac & Lin, 1997; Imielinski & Badrinath, 1994; Imielinski & Korth, 1996; Imielinski et al., 1997; Weiser, 1993).

Mobility

Mobility introduces new challenges beyond the scope of the traditional environment. Mobile devices can be used at multiple locations and in transition between these locations. Mobility results in several issues including disconnections due to handoff processes, motion management, location-dependent information, heterogeneous and fragmented networks, security, and privacy.

Portability

There are many variations of portable computer systems with different physical capabilities. However, they share many common characteristics such as limited memory, processing power, and power source. The ideal goal would be to develop a device that is compact, durable, lightweight, and that consumes a minimum amount of power.

Table 1. Mobile network architectures

Architecture	Description
Cellular Networks	• Provides voice and data services to users with handheld phones • Continuous coverage is restricted to metropolitan regions • Movement over a wide area may need user to inform the network of the new location • Low bandwidth for data-intensive applications • Could be based on either analog technology or digital technology
Wireless LANs	• A traditional LAN extended with a wireless interface • Serves small, low-powered, portable terminals capable of wireless access • Connected to a more extensive backbone network, such as a LAN or WAN
Wide Area Wireless Networks	• Special mobile radio networks provided by private service providers (RAM, ARDIS) • Provides nationwide wireless coverage for low-bandwidth data services, including e-mail or access to applications running on a fixed host
Paging Networks	• Receive-only network • No coverage problems • Low bandwidth • Unreliable

Satellite Networks	• Unlike the static, grounded MSSs, satellites are not fixed
	• Normally classified based on their altitudes (from earth) into three classes:
	  • Low Earth Orbit Satellites (LEOS)
	  • Medium Earth Orbit Satellites (MEOS)
	  • Geostationary Satellites (GEOS)

Table 2. Limitations of the mobile environment

Limitations	Concerns/Side Effects
Frequent Disconnections	• Handoff blank out in cellular networks • Long down time of the mobile unit due to limited battery power • Voluntary disconnection by the user • Disconnection due to hostile events (e.g., theft, destruction) • Roaming off outside the geographical coverage area of the window service
Limited Communication Bandwidth	• Quality of service (QoS) and performance guarantees • Throughput and response time and their variances • Efficient battery use during long communication delays
Heterogeneous and Fragmented Wireless Network Infrastructure	• Rapid and large fluctuations in network QoS • Mobility transparent applications perform poorly without mobility middleware or proxy • Poor end-to-end performance of different transport protocols across network of different parameters and transmission characteristics

Table 2 highlights some limitations of the mobile environment.

Broadcasting

The cost of communication is normally asymmetric: Sending information requires 2 to 10 times more energy than receiving the information (Imielinski, Viswanathan, & Badrinath, 1994). In the case of accessing public information, instead of the two-way, on-demand, traditional communication pattern, popular public information can be generated and disseminated over the air channel. The MH requiring the information can tune to the broadcast and access the desired information from the air channel.

In general, data can be broadcast either on one or several channels. Broadcasting has been used extensively in multiple disciplines, that is, management of communication systems (Comer, 1991) and distributed database environments (Bowen, 1992). In this chapter, the term *broadcast* is referred to as the set of all broadcast data elements (the stream of data across all channels). A broadcast is performed in a cyclic manner. The MH can only read from the broadcast, whereas the database server is the only entity that can write to the broadcast.

In the data-broadcasting application domain, power consumption and network latency are proven constraints that limit "timely and reliable" access to information. The necessity of minimizing power consumption and network latency lies in the limitation of current technology. The hardware of the mobile units have been designed to mitigate this limitation by operating in various operational modes such as active, doze, sleep, nap, and so forth to conserve energy. A mobile unit can be in active mode (maximum power consumption) while it is searching or accessing data; otherwise, it can be in doze mode (reduced power consumption) when the unit is not performing any computation. Along with the architectural and hardware enhancements, efficient power management and energy-aware algorithms can be devised to manage power resources more effectively. In addition, appropriate retrieval protocols can be developed to remedy network latency and hence to allow faster access to the information sources. In general, two issues need to be considered.

- The MH should not waste its energy in continuously monitoring the broadcast to search for information. As a result, the information on the broadcast should be organized based on a disciplined order. Techniques should be developed to (a) instruct the MH of the availability of the data element on the broadcast and (b) if the data element is available, instruct the MH of the location of the data element on the broadcast.
- An attempt should be made to minimize the response time. As will be seen later, this is achieved by shortening the broadcast length and/or reducing the number of passes over the air channel(s).

Data Organization on the Air Channel

Unlike the conventional wired environment, where a disk is assumed to be the underlying storage, data in the mobile environment are stored on air channel(s). A disk and an air channel have major structural and functional differences. The disk has a three-dimensional structure (disks can have a four-dimensional structure if multiple disks are used — redundant arrays of independent disks [RAID]). An air channel, on the other hand, is a one-dimensional structure. The disk has a random-access feature and the air channel is sequential in nature. Finally, the current raw data rate of a disk is generally much higher than that of the air channel.

Zdonik, Alonso, Franklin, and Acharya (1994) and Acharya, Alonso, Franklin, and Zdonik (1995) investigated the mapping of disk pages onto a

broadcast channel and the effects of that mapping on the management of cache at the MH. In order to place disk pages onto the data channel, the notion of multiple disks with different sizes spinning at multiple speeds was used. Pages available on faster spinning disks get mapped more frequently than those available on slower disks. In cache management, a nonconventional replacement strategy was suggested. Such a policy assumed that the page to be replaced might not be the least-recently used page in the cache. This is justifiable since the set of pages that are most frequently in demand are also the most frequently broadcast. This work was also extended to study the effect of prefetching from the air channel into the cache of the MH. These efforts assumed the same granularity for the data items on air channel and disk pages: if a data item is to be broadcast more frequently (replicated), the entire page has to be replicated. In addition, due to the plain structural nature of the page-based environment, the research looked at the pages as abstract entities and was not meant to consider the contents of the pages (data and its semantics) as a means to order the pages. In object-oriented systems, semantics among objects greatly influence the method in which objects are retrieved and, thus, have their direct impact on the ordering of these objects or pages. In addition, the replication should be performed at the data item granularity level.

An index is a mechanism that speeds up associative searching. An index can be formally defined as a function that takes a key value and provides an address referring to the location of the associated data. Its main advantage lies in the fact that it eliminates the need for an exhaustive search through the pages of data on the storage medium. Similarly, within the scope of broadcasting, an index points to the location or possible availability of a data item on the broadcast, hence, allowing the mobile unit to predict the arrival time of the data item requested. The prediction of the arrival time enables the mobile unit to switch its operational mode into an energy-saving mode. As a result, an indexing mechanism facilitates data retrieval from the air channel(s), minimizing response time while reducing power consumption. Table 3 summarizes the advantages and disadvantages of indexing schemes.

The literature has addressed several indexing techniques for a single broadcast channel as well as parallel broadcast channels with special attention to signature-based indexing and tree-based indexing (Boonsiriwattanakul, Hurson, Vijaykrishnan, & Chehadeh, 1999; Chehadeh, Hurson, & Miller, 2000; Chehadeh, Hurson, &

Table 3. Advantages and disadvantages of indexing schemes

Advantages	Disadvantages
Provides auxiliary information that allows mobile users to predict arrival time of objects	Longer broadcast
Enables utilization of different operational modes (active, nap, doze, etc.)	Longer response time
Reduces power consumption (less tune-in time)	Computational overhead due to complexity in retrieval, allocation, and maintenance of the indexes

Tavangarian, 2001; Hu & Lee, 2000, 2001; Imielinski et al., 1997; Juran, Hurson, & Vijaykrishnan, 2004; Lee, 1996).

Signature-Based Indexing

A signature is an abstraction of the information stored in a record or a file. The basic idea behind signatures on a broadcast channel is to add a control part to the contents of an information frame (Hu & Lee, 2000, 2001; Lee, 1996). This is done by applying a hash function to the contents of the information frame, generating a bit vector, and then superimposing it on the data frame. As a result, a signature partially reflects the data content of a frame. Different allocations of signatures on a broadcast channel have been studied; among them, three policies, namely, *single signature*, *integrated signature*, and *multilevel signature*, are studied in Hu and Lee (2000) and Lee (1996).

During the retrieval, a query is resolved by generating a signature based on the user's request. The query signature is then compared against the signatures of the data frames in the broadcast. A successful match indicates a possible hit. Consequently, the content of the corresponding information frame is checked against the query to verify that it corresponds to the user's demands. If the data of the frame corresponds to the user's request, the data is recovered; otherwise, the corresponding information frame is ignored. In general, this scheme reduces the access time and the tune-in time when pulling information from the air channel.

Tree-Based Indexing

Two kinds of frames are broadcast on the air channel: data frames and index frames. The index frame contains auxiliary information representing one or several data attributes pointing to the location of data collection (i.e., information frames) sharing the same common attribute value(s). This information is usually organized as a tree in which the lowest level of the tree points to the location of the information frames on the broadcast channel.

A broadcast channel is a sequential medium and, hence, to reduce the mobile unit's active and tune-in time, and consequently to reduce the power consumption, the index frames are usually replicated and interleaved with the data frames. Two index replication schemes (namely, *distributed indexing* and *(1, m) indexing*) have been studied in Imielinski et al. (1997). In distributed indexing, the index is partitioned and interleaved in the broadcast cycle (Hu & Lee, 2000, 2001; Lee, 1996). Each part of the index in the broadcast is followed by its corresponding data frame(s). In (1, *m*) indexing, the entire index is interleaved *m* times during the broadcast cycle (Imielinski et al., 1997; Lee 1996) — the whole index is broadcast before every 1/*m* fraction of the cycle.

Previous work has shown that the tree-based indexing schemes are more suitable for applications where information is accessed from the broadcast channel randomly, and the signature-based indexing schemes are more suitable in retrieving sequentially structured data elements (Hu & Lee, 2000, 2001). In addition, tree-based indexing schemes have shown superiority over the signature-based indexing schemes when the user request is directed towards interrelated objects clustered on the broadcast channel(s). Furthermore, tree-based indexing schemes relative to signature-based indexing schemes are more suitable in reducing the overall power consumption. This is due to the fact that a tree-based indexing provides global information regarding the physical location of the data frames on the broadcast channel. On the other hand, signature-based indexing schemes are more effective in retrieving data frames based on multiple attributes (Hu & Lee, 2000). Table 4 compares and contrasts the signature- and tree-based indexing.

Table 4. Signature-based versus tree-based indexing

Feature	Signature-Based Indexing	Tree-Based Indexing
Less power consumption		✓
Longer length of broadcast	✓	✓
Computational overhead	✓	✓
Longer response time	✓	✓
Shorter tune-in time		✓
Random data access		✓
Sequentially structured data	✓	
Clustered data retrieval		✓
Multi-attribute retrieval	✓	

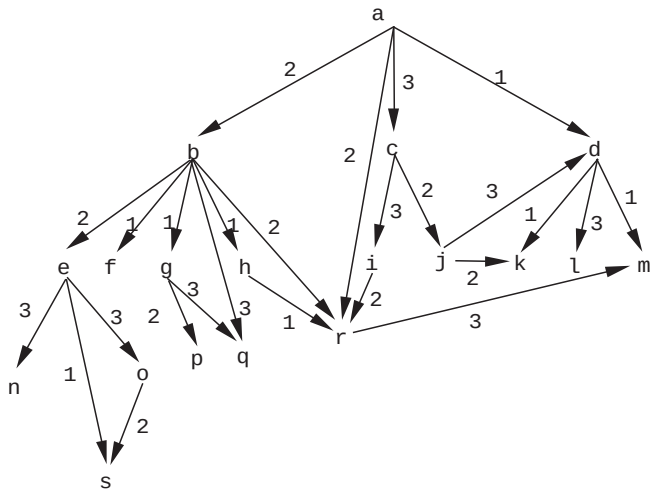
Data Organization on a Single Channel

An appropriate data placement algorithm should attempt to detect data locality and cluster related data close to one another. An object-clustering algorithm takes advantage of semantic links among objects and attempts to map a complex object into a linear sequence of objects along these semantic links. It has been shown that such clustering can improve the response time by an order of magnitude (Banerjee, Kim, Kim, & Garza, 1988; Chang & Katz, 1989; Chehadeh, Hurson, Miller, Pakzad, & Jamoussi, 1993; Cheng & Hurson, 1991a). In the conventional computing environment, where data items are stored on disk(s), the clustering algorithms are intended to place semantically connected objects physically along the sectors of the disk(s) close to one another (Cheng & Hurson, 1991a). The employment of broadcasting in the mobile computing environment motivates the need to study the proper data organization along the sequential air channel. Figure 2 depicts a weighted directed acyclic graph (DAG) and the resulting clustering sequences achieved when different clustering techniques are applied.

In order to reduce the response time, the organization of data items on an air channel has to meet the following three criteria.

- **Linear ordering.** The one-dimensional sequential access structure of the air channel requires that the object ordering be linear. In a DAG representation of a complex object, an edge between two nodes could signify an access pattern among the two nodes. The *linearity* property is defined as follows: If an edge exists between two objects, o_1 and o_2, and in the direction $o_1 \rightarrow o_2$, then o_1 should be placed prior to o_2.
- **Minimum linear distance between related objects.** In a query, multiple objects might be retrieved following their connection patterns. Intuitively, reducing the distance among these objects along the broadcast reduces the response time and power consumption.

Figure 2. Graph and various clustering methods

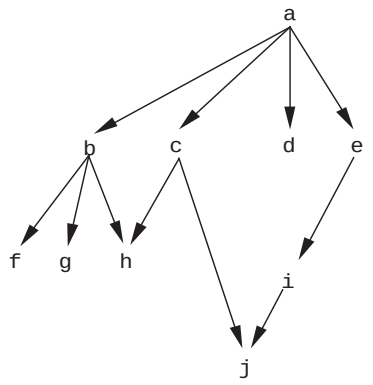


Clustering Method	Resulting Sequence
Depth First	abensofgpqhrmcijdkl
Breadth First	abrcdefgqhijklmnsop
Children-Depth First	abrcdefgqhmnsopijkl
Level Clustering	acibgqprmenosjdlkfh

- **More availability for popular objects.** In a database, not all objects are accessed with the same frequency. Generally, requests for data follow the 20/80 rule — a popular, small set of the data (20%) is accessed the majority of the time (80%). Considering the sequential access pattern of the broadcast channel, providing more availability for popular objects can be achieved by simply replicating such objects.

Figure 3 depicts a directed graph and multiple linear sequences that satisfy the linear ordering property. The middle columns represent the cost of delays between every two objects connected via an edge. For the sake of simplicity and without loss of generality, a data unit is used as a unit of measurement. Furthermore, it is assumed that all data items are of equal size. The cost associated with an edge between a pair of data items is calculated by counting the number of data items that separate these two in the linear sequence. For example, in the abfgchdeij sequence, data items a and d are separated by the sequence bfgch and thus have a cost of 6. The rightmost column represents the total cost associated with each individual linear sequence. An optimal sequence is the linear sequence with the minimum total sum. In a query where multiple related objects are retrieved, a reduced average linear distance

Figure 3. Graph, linear sequences, and costs



	Linear Sequence	Individual Costs											Total Cost
		ab	ac	ad	ae	bf	bg	bh	ch	cj	ei	ij	
1	abfgchdeij	1	4	6	7	1	2	4	1	5	1	1	33
2	abfgcheijd	1	4	9	6	1	2	4	1	4	1	1	34
3	abcdefghij	1	2	3	4	4	5	6	5	7	4	1	42
4	abgfeichjd	1	6	9	4	2	1	6	1	2	1	3	36
5	acdeijbhgf	6	1	2	3	3	2	1	6	4	1	1	30
6	adeicjbhgf	6	4	1	2	3	2	1	3	1	1	2	26
7	adecbihgfj	4	3	1	2	4	3	2	3	6	3	4	35
8	adecbhgfij	4	3	1	2	3	2	1	2	6	6	1	31
9	adecijbhgf	6	3	1	2	3	2	1	4	2	2	1	27
10	adbfgcheij	2	5	1	7	1	2	4	1	4	1	1	29
11	adceijbhgf	6	2	1	3	3	2	1	5	3	1	1	28
12	aeidcjbhgf	6	4	3	1	3	2	1	3	1	1	3	28
13	aedcbihgfj	4	3	2	1	4	3	2	3	6	4	4	36
14	aedcijbhgf	6	3	2	1	3	2	1	4	2	3	1	28

translates into smaller average response time. In this example, the best linear sequence achieves a total sum of 26.

Data Organization on Parallel Channels

The broadcast length is a factor that affects the average response time in retrieving data items from the air channel — reducing the broadcast length could also reduce the response time. The broadcast length can be reduced if data items are broadcast along parallel air channels.

Formally, we attempt to assign the objects from a weighted DAG onto multiple channels, while (a) preserving dependency implied by the edges, (b) minimizing the overall broadcast time (load balancing), and (c) clustering related objects close

to one another (improving the response time). As one could conclude, there are trade-offs between the second and third requirements: Achieving load balancing does not necessarily reduce the response time in accessing a series of data items.

Assuming that all channels have the same data rate, one can draw many analogies between this problem and static task scheduling in a homogeneous multiprocessor environment — tasks are represented as a directed graph $D \equiv (N, A)$, with nodes (N) and directed edges (A) representing processes and dependence among the processes, respectively. Compared to our environment, channels can be perceived as processors (PEs), objects as tasks, and the size of a data item as the processing cost of a task. There is, however, a major distinction between the two environments. In the multiprocessor environment, information is normally communicated among the PEs, while in the multichannel environment there is no data communication among channels.

The minimum makespan problem, in static scheduling within a multiprocessor environment, attempts to find the minimum time in which n dependent tasks can be completed on m PEs. An optimal solution to such a problem is proven to be NP hard. Techniques such as graph reduction, max-flow min-cut, domain decomposition, and priority list scheduling have been used in search of suboptimal solutions. Similar techniques can be developed to assign interrelated objects closely over parallel channels.

Distribution of data items over the broadcast parallel air channels brings the issue of access conflicts between requested data items that are distributed among different channels. The access conflict is due to two factors:

- the receiver at the mobile host can only tune into one channel at any given time, and
- the time delay to switch from one channel to another.

Access conflicts require the receiver to wait until the next broadcast cycle(s) to retrieve the requested information. Naturally, multiple passes over the broadcast channels will have a significant adverse impact on the response time and power consumption.

Conflicts in Parallel Air Channels

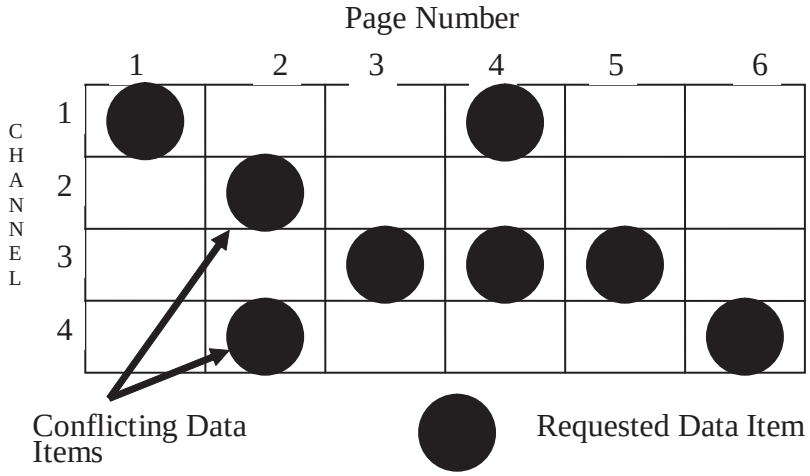
Definition 1. A K-data item request is an application request intended to retrieve K data items from a broadcast.

It is assumed that each channel has the same number of pages (frames) of equal length and, without loss of generality, each data item is residing on only a single page. A single broadcast can be modeled as an N x M grid, where N is the number of pages per broadcast and M is the number of channels. In this grid, K data items ($0 \leq K \leq MN$) are randomly distributed throughout the MN positions of the grid. Based on the common page size and the network speed, the time required to switch from one channel to another is equivalent to the time it takes for one page to pass in the broadcast. Thus, it is impossible for the mobile unit to retrieve both the ith page on Channel A and $(i + 1)$th page on Channel B (where $A \neq B$). Figure 4 is a grid model that illustrates this issue.

Definition 2. Two data items are defined to be in conflict if it is impossible to retrieve both on the same broadcast.

In response to a user request, the access latency is then directly dependent on the number of passes over the broadcast channels. One method of calculating the number of required passes over the broadcast channels is to analyze the conflicts between data items. For any particular data item, all data items in the same or succeeding page (column) and on a different row (channel) will be in conflict. Thus, for any specific page (data object) in the grid, there are $(2M - 2)$ conflicting pages (data items) in the broadcast (The last column has only $M - 1$ conflict positions, but it is assumed that N is sufficiently large to make this

Figure 4. Sample broadcast with M = 4, N = 6, and K = 8



difference insignificant.) These $(2M - 2)$ positions are known as the conflict region.

For any particular data item, it is possible to determine the probability of exactly i conflicts occurring, or $P(i)$. Because the number of conflicts for any particular data item is bounded by $(M - 1)$, the weighted average of these probabilities can be determined by summing a finite series. This weighted average is the number of broadcasts (passes) required to retrieve all K data items if all conflicts between data items are independent.

$$B = \sum_{i=0}^{M-1} (i + 1) * P(i) \qquad (1)$$

Access Patterns

In order to reduce the impact of conflicts on the access time and power consumption, retrieval procedures should be enhanced by a scheduling protocol that determines data retrieval sequence during each broadcast cycle. The scheduling protocol we proposed is based on the following three prioritized heuristics:

1) Eliminate the number of conflicts
2) Retrieve the maximum number of data items
3) Minimize the number of channel switches

The scheme determines the order of retrieval utilizing a forest - an *access forest.* An access forest is a collection of trees (*access trees*), where each access tree represents a collection of access patterns during a broadcast cycle. Naturally, the structure of the access forest, that is, the number of trees and the number of children that any parent can have, is a function of the number of broadcast channels.

Definition 3. An access tree is composed of two elements: nodes and arcs.

- **Node.** A node represents a requested data item. The nodes are labeled to indicate its

conflict status: mnemonically, C_1 represents when the data item is in conflict with another data item(s) in the broadcast and C_0 indicates the lack of conflict.

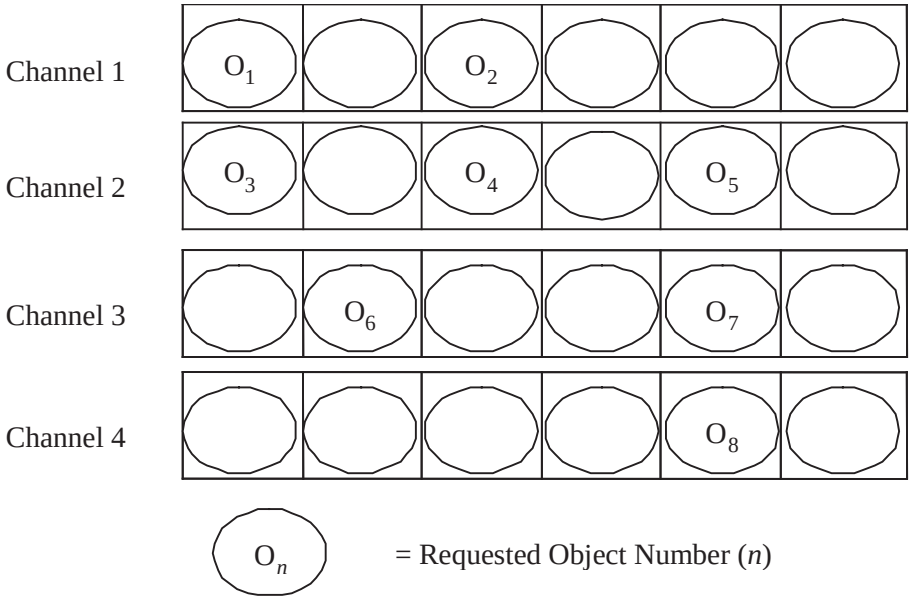
Each access tree in the access forest has a different node as a root-the root is the first accessible requested data item on a broadcast cycle. This simply implies that an access forest can have at most *n* trees where *n* is the number of broadcast channels.

- **Arcs.** The arcs of the trees are weighted arcs. A weight denotes whether or not channel switching is required in order to retrieve the next scheduled data item in the access pattern. A branch in a tree represents a possible access pattern of data items during a broadcast cycle with no conflicts. Starting from the root, the total number of branches in the tree represents all possible access patterns during a broadcast cycle.

This scheme allows one to generate all possible nonconflicting, weighted access patterns from all channels. The generated access patterns are ranked based on their weights-a weight is set based on the number of channel switches-and then the one(s) that allows the maximum number of data retrievals with minimum number of channel switches is selected. It should be noted that the time needed to build and traverse the access forest is a critical factor that must be taken into account to justify the validity of this approach. The following working example provides a detailed guide to illustrate the generation of the access patterns for each broadcast cycle.

1) **Search.** Based on the user's query, this step determines the offset and the chan-

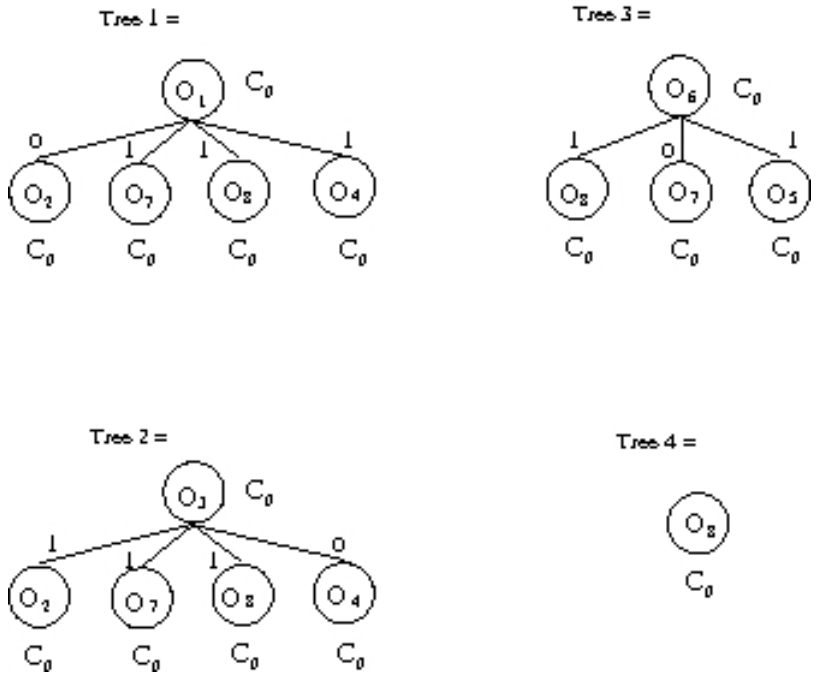
Figure 5. A parallel broadcast of four channels with eight requested data items



nel number(s) of the requested objects on the broadcast channels. Figure 5 depicts a request for eight data items from a parallel broadcast channel of four channels.

2) **Generation of the access forest.** For each broadcast channel, search for the requested data item with the smallest offset (these objects represent the roots of an access tree). For the example, the data items with the smallest offsets are O_1, O_3, O_6 and O_8. Note that the number of access trees is upper bounded by the number of broadcast channels.
3) **Root assignment.** For each channel with at least one data item requested, generate a tree with root node as determined in Step 2. The roots are temporarily tagged as C_0.
4) **Child assignment.** Once the roots are determined, it is necessary to select the child or children of each rooted access tree: For each root, and relative to its position on the air channel, the algorithm determines the closest nonconflicting data items on each channel. With respect to a data item $O_{i,x}$ at location X on air channel i ($1 \le i \le n$), the closest nonconflicting data item is either the data item $O_{i,x+1}$ or the data item $O_{j,x+2}$, $j \ne i$. If the child is in the same broadcast channel as the root, the arc is weighted as 0; otherwise it is weighted as 1. Each added node is temporarily tagged as C_0. Figure 6 shows a snapshot of the example after this step.
5) **Root label update.** Once the whole set of requested data items is analyzed and the access forest is generated, the conflict labels of the nodes of each tree are updated. This

Figure 6. Children of each root



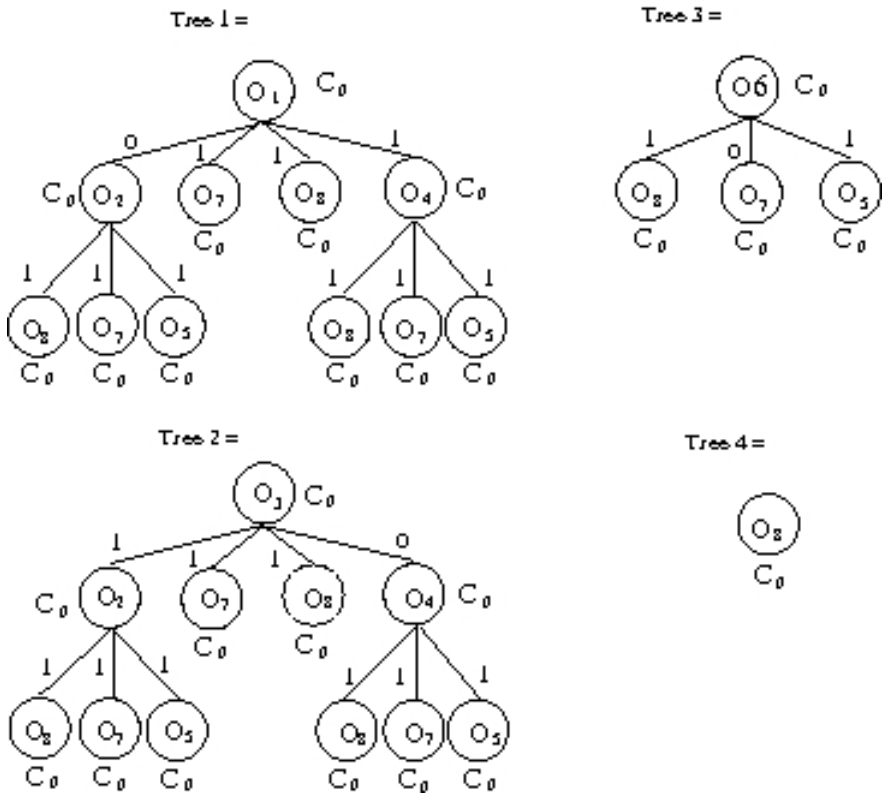
process starts with the root of each tree. If a root is in conflict with any other root(s), a label of C_1 is assigned to all the roots involved in the conflict, otherwise the preset value of C_0 is maintained.

6) **Node label update.** Step 5 will be applied to the nodes in the same level of each access tree in the access forest. As in Step 5, a value of C_1 is assigned to the nodes in conflict. Figure 7 shows the example with the updated labels.

7) **Sequence selection.** The generation of the access forest then allows the selection of the suitable access patterns in an attempt to reduce the network latency and power consumption. A suitable access pattern is equivalent to the selection of a tree branch that:
   - has the most conflicts with other branches,
   - allows more data items to be pulled off the air channels, and
   - requires the least number of channel switches.

The O_3, O_4, and O_5 sequence represents a suitable access pattern for our running example during the first broadcast cycle. Step 7 will be repeated to generate access patterns for different broadcast cycles. The algorithm terminates when all the requested data items are covered in different access patterns. The data item sequence

Figure 7. Final state of the access forest



O_1, O_2, and O_7 and data item sequence O_6 and O_8 represent the last two patterns for retrieving all of the data items requested in the example.

DATA ORGANIZATION ON A SINGLE CHANNEL

As noted in the literature, the object-oriented paradigm is a suitable methodology for modeling public data that are by their very nature in multimedia format (Atkinson, Bancilhon, DeWitt, Dittrich, Maier, & Zdonik, 1989; Fong, Kent, Moore, & Thompson, 1991; Hurson, Pakzad, & Cheng, 1993; Kim, 1990). In addition, object-oriented methodology provides a systematic mechanism to model a complex object in terms of its simpler components.

In this section, without loss of generality, we model information units as objects. Object clustering has proven to be an effective means of data allocation that can reduce response times (Banerjee et al., 1988; Chang & Katz 1989; Chehadeh et al., 1993; Cheng & Hurson, 1991b; Lim, Hurson, Miller, & Chehadeh, 1997). In our research, we investigated two heuristic allocation strategies. The first strategy assumes a strict linearity requirement and deals with nonweighted DAGs. The second approach relaxes such restriction in favor of clustering strongly related objects closer to one another and consequently deals with weighted DAGs.

Strict Linearity: ApproximateLinearOrder Algorithm

Definition 4. An independent node is a node that has either one or no parent. A graph containing only independent nodes makes up a forest.

Heuristic Rules

1) Order the children of a node based on their number of descendants in ascending order.
2) Once a node is selected, all of its descendants should be visited and placed on the sequence in a depth-first manner, without any interruptions from breadth siblings.
3) If a node has a nonindependent child, with all of its parents already visited, the nonindependent child should be inserted in the linear sequence before any independent child.

The ApproximateLinearOrder algorithm implements these heuristics and summarizes the sequence of operations required to obtain a linear sequence. The algorithm assumes a greedy strategy and starts by selecting a node with an in-degree of zero and out-degree of at least one.

ApproximateLinearOrder Algorithm

1) traverse DAG using DFS traversal and as each node is traversed
2) append the traversed node *N* to the sequence
3) remove *N* from {nodes to be traversed}
4) **if** {nonindependent children of *N* having all their parents in the sequence} $\neq \varnothing$
5) *Set* $\leftarrow$ {nonindependent children of *N* having all their parents in the sequence}
6) **else**
7) **if** {independent children of *N*} $\neq \varnothing$
8) *Set* $\leftarrow$ {independent children of *N*}
9) *NextNode* $\leftarrow$ node $\in$ *Set* | node has least # of descendants among the nodes in *Set*

Applying this algorithm to the graph of Figure 3 generates either the 5th or 11th sequence — dependent on whether c or d was chosen first as the child with the least number of independent children. As one can observe, neither of these sequences is the optimal sequence. However, they are reasonably better than other sequences and can practically be obtained in polynomial time. It should be noted that nodes not connected to any other nodes — nodes with in-degree and out-degree of zero — are considered harmful and thus

are not handled by the algorithm. Having them in the middle of the sequence introduces delays between objects along the sequence. Therefore, we exclude them from the set of nodes to be traversed and handle them by appending them to the end of the sequence. In addition, when multiple DAGs are to be mapped along the air channel, the mapping should be done with no interleaving between the nodes of the DAGs.

Varying Levels of Connectivity: PartiallyLinearOrder Algorithm

In a complex object, objects are connected through semantic links with different degrees of connectivity. The different access frequency of objects in an object-oriented database reveals that some patterns are more frequently traversed than others (Fong et al., 1991). This observation resulted in the so-called PartiallyLinearOrder algorithm that assumes a weighted DAG as its input and produces a linear sequence. It combines the nodes (single_node) of the graph into multi_nodes in descending order of their connectivity (semantic links). The insertion of single_nodes within a multi_node respects the linear order at the granularity level of the single_nodes. The multi_nodes are merged (with multi_nodes or single_nodes) at the multi_node granularity, without interfering with internal ordering sequences of a multi_node. Figure 8 shows the application of the PartiallyLinearOrder algorithm.

PartiallyLinearOrder Algorithm

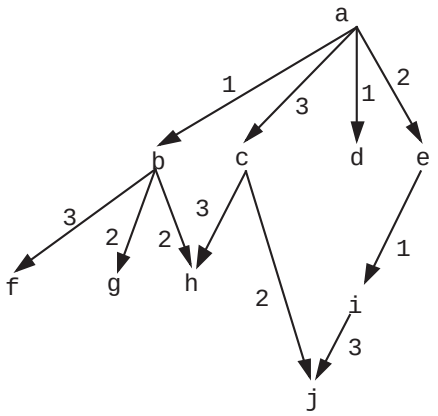
1) **for** every weight w_s in descending order
2) **for** every two nodes N_1 & N_2 connected by w_s
3) merge N_i & N_j into one multi_node
4) **for** every multi_node *MN*
5) $w_m = w_s - 1$
6) **for** every weight w_m in descending order
7) **while** ∃ adjacent_node *AN* connected to *MN*
8) **if** ∃ an edge in both directions between *MN* & *AN*
9) compute $WeightedLinearDistance_{MN\_AN}$ & $WeightedLinearDistance_{AN\_MN}$
10) merge *MN* & *AN* into one multi_node, based on the appropriate direction

Performance Evaluation

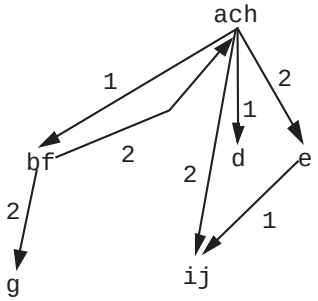
Parameters

A simulator was developed to study the behavior of the proposed mapping algorithms based on a

Figure 8. Process of PartiallyLinearOrder



(a) Original Graph



(b) First and Second Iterations

bfgacheijd

(c) Third Iteration

Table 5. Description of parameters

Parameter	Description
Input Parameters	
Number of Nodes	Number of objects within the graph (excluding replication)
Object Size	Sizes of objects (small/medium/large)
Object-Size Distribution	Distribution of the sizes of objects within the database
Next-Node Ratio	Connectivity to next node (random or connection)
Out-Degree Distribution	Distribution of the type of nodes based on their out-degrees
Level Distribution	Semantic connectivity of two objects (weak/normal/strong)
Percentage of Popular Objects	Percentage of objects requested more often than others
Replication Frequency	The number of times a popular object is to be replicated
Output Parameter	
Average Access Delay	In a single query, the average delay between accessing two objects

set of rich statistical parameters. Our test bed was an object-oriented financial database. The OO7 benchmark was chosen to generate the access pattern graphs. We used the NASDAQ exchange (NASDAQ, 2002) as our base model, where data is in both textual and multimedia (graphics — i.e., graphs and tables) formats. Table 5 shows a brief description of the input and output parameters. The simulator is designed to measure the average access delay for the various input parameters. Table 6 provides a listing of the input parameters along with their default values and possible ranges.

Table 6. Input parameter values

Parameter	Default Value	Ranges
Number of Nodes	5,000	400-8,000
Object Size (in Bytes) • Small • Medium • Large	 $2 \le o < 20$ $20 \le o < 7K$ $7K \le o < 50K$	 2-20 20-7K 7K-50K
Object-Size Distribution [S:M:L]	1:1:1	0-6:0-6:0-6
Next-Node Ratio [C:R]	8:2	0-10:10-0
Out-Degree Distribution [0:1:2:3]	3:3:2:1	1-6:1-6:1-6:1-6
Level Distribution [W:N:S]	1:1:1	1-4:1-4:1-4
Percentage of Popular Objects	20%	10-50%
Replication Frequency	2	1-10

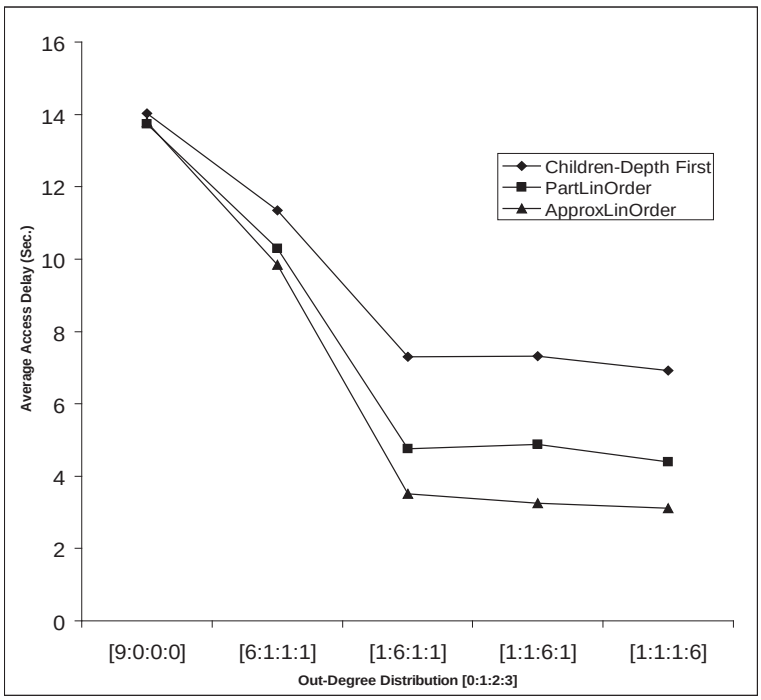
The default values are set as the value of the parameter when other parameters are varied during the course of the simulation. The ranges are used when the parameter itself is varied.

Results

The simulator operates in two stages.

- Structuring the access-pattern object graph, based on certain statistical parameters, and mapping it along the air channel using various mapping algorithms. To get a wide spectrum of possible graphs, parameters such as (a) the percentage of nonfree nodes, (b) the depths of the trees within the graph, and (c) the amount of sharing that exists between trees through nonfree nodes that were varied. Varying these statistical parameters, we generated 500 access graphs that were used as part of our test bed. In addition, we simulated three mapping algorithms: a nonlinear, children-depth-first clustering algorithm (Banerjee et al., 1988), and the PartiallyLinearOrder and ApproximateLinearOrder algorithms.
- Generating queries and accessing the requested objects from the air channel. During each run, each query on average accesses 20 objects either through their semantic links or randomly (following the [C:R] value of the next-node ratio). The simulator measured the average access delay. Each point in the curves (Figure 9) is the average result of running the simulator 100,000 times. Finally, we assumed a broadcast data rate of 1Mbit/sec and showed the results in terms of seconds.

Figure 9. Average access delay versus connectivity



Impact of Number of Objects. ApproximateLinearOrder and PartiallyLinearOrder schemes performed better than the conventional children-depth first by taking the linearity issue into consideration. As expected in all three cases, the average access delay increased as the total number of objects increased. The mapping of additional nodes on the broadcast introduced extra delays between the retrievals of two consecutive objects. Taking a closer look at this effect, we observed that this extra delay is mainly due to an increase in the distance for objects that are retrieved randomly (not based on their semantic links) since the goal of both algorithms is to cluster semantically related objects close to one another. The ApproximateLinearOrder algorithm outperformed the PartiallyLinearOrder algorithm since the latter attempts to cluster strongly connected objects closer to one another than loosely connected ones and, hence, compromises the linearity property for the loosely connected objects. This compromise overshadows the benefit and is amplified as the number of objects increased. To get a better insight on how our proposed schemes compare with the optimal case, two graphs with 10 nodes were constructed and the optimal sequences exhaustively generated. Using the same set of input values, the average access delay for both proposed schemes were simulated and compared against the average access delay for the optimal sequence. The results of ApproximateLinearOrder and PartiallyLinearOrder were 79% and 76%, respectively, of the access delay of the optimal case.

Size Distribution. In this experiment, we observed that the smallest average access delay took place when the air channel contained smaller data items. However, as the population of data items shifted toward the larger ones, the average access delay increased.

Next-Node Ratio. During the course of a query, objects are either accessed along the semantic links or in a random fashion. At one extreme, when all objects were accessed along the semantic links, the average access delay was minimal. The delay, however, increased for randomly accessed objects. Finally, where all the accesses are on a random basis, clustering (and linearity) does not improve the performance, and all mapping algorithms perform equally.

Out-Degree Distribution. This parameter indicates the number of children of a node within the graph — an out-degree of 0 indicates a sink node. Figure 9 shows the effect of varying the out-degree distribution within the graph structure. The point [9:0:0:0] indicates that all the nodes within the graph have an out-degree of 0, with no semantic link among the objects. This is similar to stating that any access to any object within the graph is done on a random basis. In general, the average access delay is reduced as more connectivity is injected in the access graph. It is interesting to note that it would be more desirable to deal with more, but simpler, objects than with few complex objects on the air channel.

In separate simulation runs, the simulator was also used to measure the effect of varying the percentage of popular objects and the replication frequency. These two parameters have the same effect on the total number of objects on the air channel, however, from the access pattern perspective, the semantic of the accesses are different. In both cases, the average access delay increased as either parameter increased. We also observed and measured the average access delay for different degrees of connectivity among objects. The average access delay for objects connected through strong connections is about 4.3 seconds, whereas it is 7.3 and 7.6 seconds for normally and weakly connected objects, respectively. As would be expected, these results show that the improvement is considerable for the objects connected by a strong connection, but for a normal connection, the performance was close to that of the weak-connection case since the algorithm performs its best optimization for strongly connected objects.

Section Conclusion

In this section, two heuristically based mapping algorithms were discussed, simulated, and analyzed. Performing the mapping in polynomial time was one of the major issues of concern while satisfying linearity, locality, and replication of popular objects. The ApproximateLinearOrder algorithm is a greedy-based approximation algorithm that guarantees the linearity property and provides a solution in polynomial time. The PartiallyLinearOrder algorithm guarantees the linearity property for the strongest related objects and relaxes the linearity requirement for objects connected through looser links. Finally, it was shown that the proposed algorithms offer higher performance than the traditional children-depth-first algorithm.

DATA ORGANIZATION ON PARALLEL CHANNELS

Reducing the broadcast length is one way to satisfy timely access to the information. This could be achieved by broadcasting data items along parallel air channels. This problem can be stated formally as follows: Assign the data items from a weighted DAG onto multiple channels while (a) preserving dependency implied by the edges, (b) minimizing the overall broadcast time (load balancing), and (c) clustering related data items close to one another (improving the response time). Realizing the similarities between these objectives and the task-scheduling problem in a multiprocessor environment, we proposed two heuristic–based, static scheduling algorithms, namely the largest object first (LOF) algorithm and the clustering critical-path (CCP) algorithm.

The Largest Object First Algorithm

This algorithm relies on a simple and localized heuristic by giving priority to larger data items. The algorithm follows the following procedure: For each collection of data items, recursively, a "proper" node with in-degree of 0 is chosen and assigned to a "proper" channel; a "proper" channel is the one with the smallest overall size and a "proper" node is the largest node with in-degree of 0. The assigned node along with all of its out-edges are eliminated from the object DAG. This results in a set of nodes with in-degree of 0. These nodes are added to the list of free nodes and then are selected based on their sizes. This process is repeated until all the nodes of the DAG are assigned.

Definition 5. A free node is a node that either has an in-degree of 0 (no parent) or has all of its parents allocated on a channel. A free node is a candidate node available for allocation.

Assuming that there are n nodes in the graph, the algorithm requires the traversal of all the nodes and thus requires n steps. At each step, the algorithm searches for the largest available node whose parents have been fully allocated. This would require at most $O(n^2)$. Therefore, the overall running time of the algorithm is $O(n^3)$. The LOF algorithm respects the dependency among the nodes, if any, and achieves a better load balancing by choosing the largest object first. This algorithm, however, does not allocate objects based on the degree of connectivity and/or the total size of the descendent objects that could play a significant role in balancing the loads on the channels. In addition, this algorithm does not necessarily cluster related object on the parallel air channels.

LOF Algorithm

1) **repeat (2-4)** until all nodes are assigned
2) assign a free node with the largest weight whose parents are fully allocated to the least-loaded channel
3) remove all out-edges of the assigned node from the DAG

4) insert resulting free nodes into the list of free nodes

The Clustering Critical-Path Algorithm

A critical path is defined as the longest sequence of dependent objects that are accessed serially. A critical path is determined based on the weights assigned to each node. A weight is defined based on several parameters such as the size of the data item, the maximum weight of the descendents, the total weight, and the number of descendents.

Definition 6. A critical node is a node that has a child with an in-degree greater than 1.

Load Balancing

Critical Node effect. Allocate a critical node with the highest number of children with in-degrees greater than 1 first.

Number of children with in-degrees of 1. Allocate nodes with the highest number of children with in-degrees of 1 first. This could free up more nodes to be allocated in parallel channels.

Clustering Related Objects

The weight of a node should be made a function of the weights of the incoming and outgoing edges. The weight of each node is calculated based on Equation 2. It should be noted that:

- There is a trade-off between load balancing and clustering related objects: The allocation strategy for the purpose of load balancing could upset the clustering of related objects and vice versa. Therefore, we propose a factor to balance the two requirements. This factor takes a constant value $\in [0,1]$ and can be assigned to favor either requirement over the other.
- The size of a data item is a multiple of a constant value.
- The weight of an edge is a multiple of a constant value.

$$W = MWC + F\left[S + NCID1 + \sum_{i=1}^{NCIDM} SPC_i - NCIDM(S)\right] + (1-F)\left[(NMIW)MIW + \frac{1}{(NMOW)MOW}\right] \quad (2)$$

where

W weight of a node
MWC maximum weight among the node's children
F factor of optimizing for load balancing versus clustering related objects
S size of a node (object)
NCID1 number of children with in-degrees of 1
NCIDM number of children with in-degrees greater than 1
SPC size of all parent objects
MIW maximum weight of incoming edges
NMIW number of maximum-weighted incoming edges
MOW maximum weight of outgoing edges
NMOW number of maximum-weighted outgoing edges

The algorithm required to assign the weight of every node in the graph with time complexity of $O(n^2)$ (n is the number of nodes in the DAG) is as follows.

AssignWeights(DAG) Algorithm

1) for every node *i* (Starting at the leaf nodes and traversing the DAG in a breadth-first manner)
2) Calculate SPC_i
3) Calculate W_i

The CCP algorithm takes a DAG as its input and calls the AssignWeights Algorithm. The running time of the CCP algorithm is equal to the running time of AssignWeights plus the running time of the *repeat* loop. The loop has to be repeated n times and Line 4 can be done in $O(n)$.

Therefore, the overall running time of the CCP algorithm is $O(n^2)$.

CCP(DAG) Algorithm

1) AssignWeights(DAG)
2) **repeat** until all the nodes have been processed
3) Select the free node N with the largest weight
4) **if** all parents of N are fully allocated on the channels
5) place it on the currently least-loaded channel
6) **else**
7) Fill up the least-loaded channel(s) with nulls up to the end of the last allocated parent of *N* then place *N* on it.

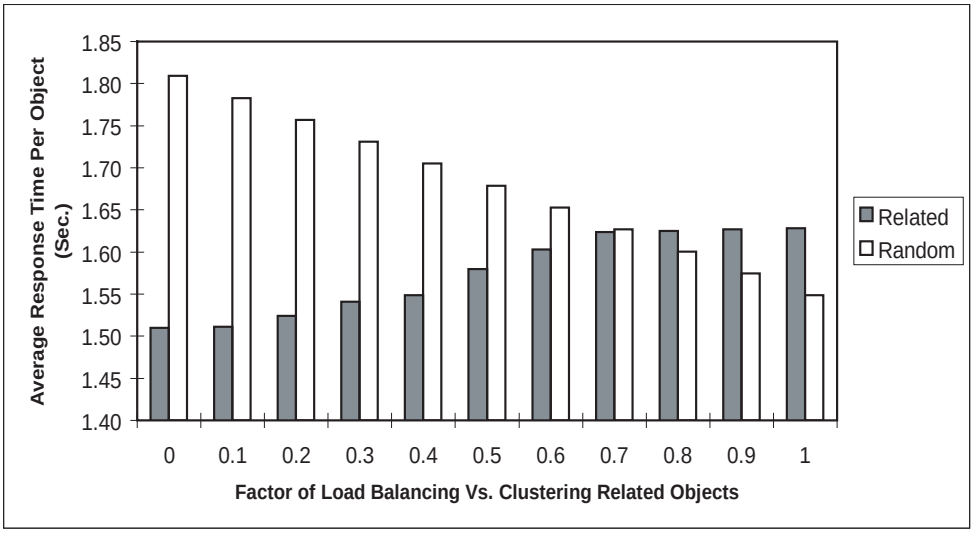
Performance Evaluation

To evaluate the performance of the proposed algorithms, our simulator was extended to measure the average response time per data item retrieval. To measure the effectiveness of the algorithms across a more unbiased test bed, the degree of connectivity among the data items in the DAG was randomly varied, and 100 different DAGs were generated. In every DAG, the out-degrees of the nodes were determined within the range of 0 and 3. To limit the experimentation running time, a decision was made to limit the number of nodes of each DAG to 60. The weights connecting the nodes, similar to the experiment reported in previously, were categorized as strong, normal, and weak, and were uniformly distributed along the edges of a DAG.

The simulation is accomplished in two steps: In the first step, every DAG is mapped onto the air channels using the LOF and CCP algorithms. In the second step, the simulator simulates the process of accessing the air channels in order to retrieve the data items requested in a query. Among the requested data items, 80% were selected based on their semantic relationship within the DAG and 20% were selected randomly. Finally, the average response time was calculated for 100,000 runs.

1) **Number of Air Channels.** As anticipated, increasing the number of channels resulted in a better response time for both the LOF and CCP. However, this improvement tapered off as the number of channels increased above a certain threshold value, since, additional parallelism provided by the number of channels did not match the number of free nodes available to be allocated, simultaneously. In addition, as expected, the CCP method outperformed the LOF method—the CCP heuristics attempt to smooth the distribution of the objects among the air channels while clustering the related objects.
2) **Out-Degree Distribution.** In general, the CCP method outperformed the LOF method. When the out-degree distribution is biased to include nodes with larger out-degrees (i.e., making the DAG denser), the LOF performance degrades at a much faster rate than the CCP method. This is due to the fact that such bias introduces more critical nodes and a larger number of children per node. The CCP method is implicitly capable of handling such cases.
3) **Factor of Load Balancing versus Clustering Related Objects.** To get a better insight on the operations of the CCP method, we analyzed its behavior by varying the load balancing and degree of clustering (F; Equation 2). In this experiment, 80% of the data items requested by each query were related through certain semantic links and the rest were selected randomly. As can be seen (Figure 10), biasing in favor of clustering degrades the average response time for randomly selected data items. Optimization based on clustering increases the overall length of the broadcast, thus, contributing to larger response time for randomly accessed objects. For semantically related data

Figure 10. Load balancing versus clustering



items, however, decreasing *F* influenced the broadcast to favor the allocation of related data items closer to one another, thus improving the average response time. Such rate of improvement, however, declined as *F* reached a certain threshold value (0.2 in this case). At this point the behavior of the system reaches a steady state (the objects cannot be brought closer to one another). In different simulation runs, the ratio of randomly selected and semantically related data items varied in the ranges between 30/70% and 70/30% and the same behavior was observed. This figure can be productive in tuning the performance of the CCP method. Assuming a feedback channel is to be used to collect the statistics of the users' access pattern, *F* can be adjusted adaptively to match the access pattern. As an example, if the frequency of accessing data items based on their connection is equivalent to that of accessing data items randomly, then a factor value of 0.7 would generate the best overall response time.

Section Conclusion

This section concentrated on the proper mapping of data items on multiple parallel air channels. The goal was to find the most appropriate allocation scheme that would (a) preserve the connectivity among the data items, (b) provide the minimum overall broadcast time (load balancing), and (c) cluster related data items close to one another (improving the response time). Applying the LOF heuristic showed an improvement in load balancing. However, it proved short in solving the third aforementioned requirement. The CCP algorithm was presented to compensate this shortcoming. Relying on the critical path paradigm, the algorithm assumed several heuristics and showed better performance.

ENERGY-EFFICIENT INDEXING

In this section, we investigate and analyze the usage of indexing and indexed-based retrieval techniques for data items along the single and parallel broadcast channel(s) from an energy-efficient point of view. In general, index-based channel access protocols involve the following steps.

1) **Initial probe.** The client tunes into the broadcast channel to determine when the next index is broadcast.
2) **Search.** The client accesses the index and determines the offset for the requested data items.
3) **Retrieve.** The client tunes into the channel and pulls all the required data items.

In the initial probe, the mobile unit must be in active, operational mode. As soon as the mobile unit retrieves the offset of the next index, its operational mode could change to doze mode. To perform the *Search* step, the mobile unit must be in active mode, and when the unit gets the offset of the required data items, it could switch to doze mode. Finally, when the requested data items are being broadcast (*Retrieve* step), the mobile unit changes its operational mode to active mode and tunes into the channel to download the requested data. When the data is retrieved, the unit changes to doze mode again.

Object-Oriented Indexing

Object-oriented indexing is normally implemented as a multilevel tree. We can classify the possible implementation techniques into two general schemes: single-class indexing and hierarchical indexing. In the single-class scheme, multiple multilevel trees are constructed, each representing one class. In this case, the leaf nodes of each tree point to data items belonging only to the class indexed by that tree. A query requesting all objects with a certain ID has to navigate all these trees. On the other hand, the hierarchical scheme constructs one multilevel tree representing an index for all classes. The same query has to only navigate the common tree.

Data Indexing on a Single Air Channel

We assume an *air-channel page* as the storage granule on the air channel. Due to the sequential nature of the air channel, the allocation of the nodes of a multilevel tree has to follow the navigational path used to traverse the tree, starting at the root. Therefore, an ordering scheme is used to sequentially map the nodes on the air channel. Similarly, data items are allocated onto air channel pages following their index.

Storage Requirement

The overall storage requirement is the sum of the storage required by the inner and leaf nodes. For both schemes, the structure of the inner node is the same (Figure 11). An inner node is a collection of records, where each record is composed of a [*Key*, *pointer*] pair. Assume the order of the tree is o and the fan-out of every node is f ($o \leq f \leq 2o$, except for the root where $2 \leq f \leq 2o$). The leaf node structures of both schemes are shown in Figure 12. As can be seen, the main difference between the two schemes is that the hierarchical scheme requires a list of classes that have data items indexed by the index.

For the sake of simplicity, and without loss of generality, we assume that there are no overflow pages, furthermore assuming the following notations.

- P size of air-channel page
- K average number of distinct keys for an attribute
- S average size of a leaf-node index record in a single-class index

Figure 11. Inner-node structure of single-class and hierarchical schemes

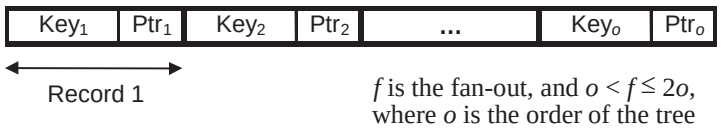
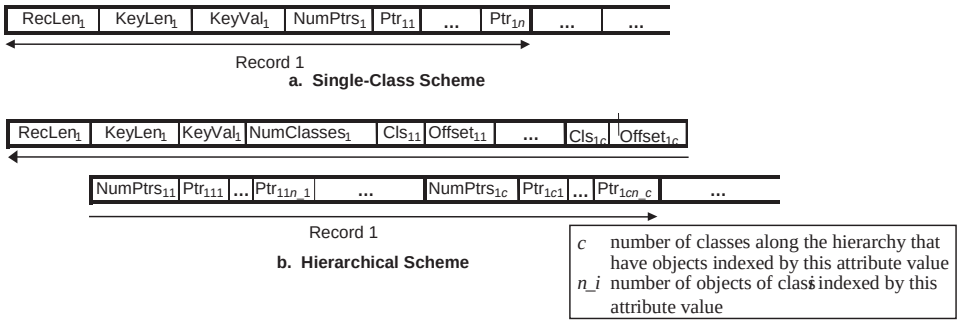


Figure 12. Leaf-node structure of single-class and hierarchical schemes



H average size of a leaf-node index record in a hierarchical index

L number of leaf-node pages

IN number of inner-node pages for either scheme

$$L_{Single-Class} = \lceil K / \lfloor P / S \rfloor \rceil \quad (3)$$

$$L_{Hierarchical} = \lceil K / \lfloor P / H \rfloor \rceil \quad (4)$$

$$IN = 1 + \lfloor L / f \rfloor + \lfloor \lfloor L / f \rfloor / f \rfloor + \ldots \quad (5)$$

It should be noted that in the case of a single-class scheme, Equations 3 and 5 should be calculated for all the classes.

Timing Analysis

To perform the timing analysis, one has to consider the domain of a query. The cardinality of the domain of a query is the number of classes to be accessed by the query along the hierarchy. Our timing analysis evaluates the *response* and *active* time as the performance metrics. The response time is defined as the time elapsed between the first user access to the air channel and when the required information is retrieved. The active time is defined as the time during which the mobile unit has to be active accessing the channel. In the timing analysis, we use the number of pages as our unit of measurement. Finally, to support our protocols, we assume that every air-channel

page contains control information indicating the location of the first page of the next index. This can simply be implemented as an offset (2 or 4 bytes).

a) **Hierarchical Method.** In this scheme, whether the domain of the query covers one class or all classes along the hierarchy, the same index structure has to be traversed. The protocol is shown below.

Hierarchical Protocol

1) Probe onto channel and get offset to the next index *active*
2) Reach the index *doze*
3) Retrieve the required index pages *active*
4) Reach the required data pages *doze*
5) Retrieve required data pages *active*

- **Response Time.** Assume I_H and D denote the size of the index and data, respectively. On average, it takes half the broadcast (the size of the broadcast is I + D) to locate the index from the initial probe. Once the index is reached, it has to be completely traversed before data pages appear on the broadcast. On average, it takes half the size of the data to locate and retrieve the required data items. Thus, the response time is proportional to:

$$\frac{I_H + D}{2} + I_H + \frac{D}{2} = \frac{3I_H}{2} + D = Broadcast + \frac{I_H}{2} \quad (6)$$

- **Active Time.** The mobile unit's modules have to be active to retrieve a page. Once the index is reached, a number of inner-node pages have to be accessed in order to get and retrieve a leaf-node page. The number of pages to be retrieved at the index is equal to the height of the index tree ($\log_f(D)$). Finally, the amount of the data pages to be read is equal to the number of data items to be retrieved that reside on distinct pages (NODP). Therefore, the active time is:

$$1 + \log_f(D) + NODP \quad (7)$$

b) **Single-Class Method.** In this scheme, we assume that the first page of every index contains information indicating the location of each index class. This structure can be implemented by including a vector of pairs [class_id, offset]. Assuming that the size of the offset and the class_id is 4 bytes each, the size of this structure would be 8c, where c is the number of class indexes on the broadcast.

Single-Class Protocol

1) Probe onto channel and get offset to the next index *active*
2) Reach the index *doze*
3) Retrieve offsets to the indexes of required classes *active*
4) for every required class
5) Reach the index *doze*
6) Retrieve the required index pages *active*
7) Reach the required data *doze*
8) Retrieve required data pages *active*

- **Response Time.** The size of a single index and its associated data are labeled as I_i and D_i, respectively. Since the total number of objects to be indexed is the same in single-class and hierarchical indexes, the sum of all D_i for all classes is equal to D. Assume a query references a set of classes where x and y stand for the first and last classes to be accessed. The average distance to be covered to get to x is half the distance covering the indexes and data between the beginning of y and the beginning of x. Once the index x is located, then all the indexes and data of all the classes between x and y (including

those of x) have to be traversed. Once y is reached, its index and half of its data (on average) have to be traversed. Thus, the response time is proportional to:

$$\frac{\sum_{i=y}^{x-1}(I_i + D_i)}{2} + \sum_{i=x}^{y}(I_i + D_i) - \frac{D_y}{2} \qquad (8)$$

Equation 8 provides a general means for calculating the average response time. However, the results are dependent on the location of the probe and the distance between x and y. It has been shown that the response time is lower bounded by half the size of the broadcast and upper bounded by slightly above the size of the broadcast. Further discussion on this issue is beyond the scope of this chapter and the interested reader is referred to Chehadeh, Hurson, and Kavehrad (1999).

- **Active Time.** Similar to the hierarchical case, the active time is dependent on the number of index pages and data pages to be retrieved. Therefore, the active time is the sum of the height of the trees for all the indexes of classes to be retrieved plus the number of the corresponding data pages. This is shown in the Equation 9. The 2 in the front accounts for the initial probe plus the additional page containing the index of classes (Line 3 in the protocol).

$$2 + \sum_{i=x}^{y}\left[\log_f(D_i) + NODP_i\right] \qquad (9)$$

Performance Evaluation

Our simulator was extended to study both the response time and energy consumption with respect to the two allocation schemes. The overall structure of the schema graph determines the navigational paths among the classes within the graph. The relationships of the navigational paths within the graph influence the number and structure of indexes to be used.

- **Inheritance Relationship.** Within an inheritance hierarchy, classes at the lower level of the hierarchy inherit attributes of the classes at the upper level. Therefore, data items belonging to the lower-level classes tend to be larger than those within the upper levels. The distribution of the number of data items is application dependent. In our analysis, and without loss of generality, we assumed the data items to be equally distributed among the classes of the hierarchy.
- **Aggregation Relationship.** In an aggregation hierarchy, data items belonging to lower classes are considered "part of" data items and those at the higher ends are the "collection" of such parts. Therefore, data items belonging to higher classes are generally larger than those belonging to the lower ones. In addition, the cardinality of a class at the upper end is smaller than a class at the lower end.

As a result, the organization of classes within the schema graph has its influence on the distribution of both the number and size of data items among the classes of the database. We assumed an average of eight classes for each hierarchy and categorize the sizes of data items as small, medium, large, and very large. Furthermore, 60% of the data items have distinct keys and the value of any attribute is uniformly distributed among the data items containing such attribute. Table 7 shows a list of all the input parameters assumed for this case.

The information along the broadcast channel is organized in four different fashions: the hierarchical and single-class methods for the inheritance and aggregation relationships. Table 8 shows the data and index page sizes for these organizations. Note that it is the number of data items (not data

Table 7. Input parameters

Parameter	Value (Default/Range)
Number of Data Items on Broadcast	5,120
Average Number of Classes Along Hierarchy	8
Percentage Distribution of Number of Data Items in Inheritance Hierarchy	25,25,25,25%
Percentage Distribution of Number of Data Items in Aggregation Hierarchy	40,30,20,10%
Distribution of Data Size [S,M,L,VL]	16,512,3K,6K bytes
Distribution of the Data Sizes in Inheritance Hierarchy	VL,L,M,S
Distribution of the Data Sizes in Aggregation Hierarchy	S,M,L,VL
Percentage of Classes to be Retrieved (Default/Range)	70% / [10-100%]
Average Number of Data Items to Retrieve per Class	2
Fan-Out in Index Tree	5
Average Number of Data Items with Distinct Key Attribute per Class	60% of data items per class
Size of Air-Channel Page	512 bytes
Broadcast Data Rate	1 M bits/sec
Power Consumption Active Mode	130 mW
Power Consumption Doze Mode	6.6 mW

Table 8. Number of index and data pages

	Aggregation/ Hierarchical	Aggregation/ Single [Eight Classes]	Inheritance/ Hierarchical	Inheritance/ Single [Eight Classes]
Index Pages	2,343	67,63,49,39,36,32,18,16	2,343	40,40,40,40,40,40,40,40
Data Pages	13,562	73,75,652,769,2504,28140,3206,3502	34,015	12517,11520,5253,4252,637,440,25,20

Table 9. Response time degradation factor relative to the no-index scheme

Aggregation/ Hierarchical	Aggregation/ Single	Inheritance/ Hierarchical	Inheritance/ Single
1.17	1.05	1.1	1.02

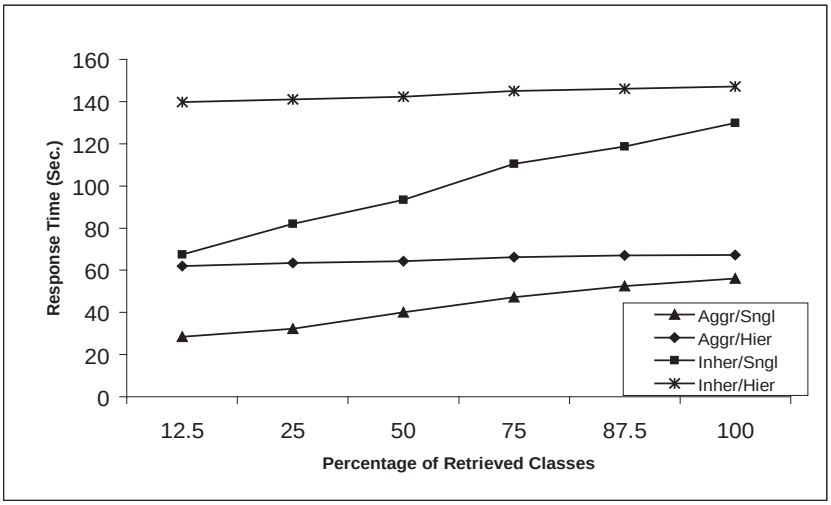
pages) that controls the number of index blocks. Within each indexing scheme, for each query, the simulator simulates the process of probing the air channel, getting the required index pages, and retrieving the required data pages. In each query, on average, two data items from each class are retrieved. The simulation measures the response time and amount of energy consumed.

a) **Response Time.** Placing an index along the air channel contributes to extra storage overhead and thus longer response time. Hence, the best response time is achieved when no index is placed, and the entire broadcast is searched. Table 9 shows the degradation factor in the average response time due to the inclusion of an index in the broadcast. The factor is proportional to the ratio of the size of the index blocks to that of the entire broadcast.

Figure 13 shows the response time for all four different broadcast organizations. From the figure, one could conclude that for both the inheritance and aggregation cases, the response time of the hierarchical organization remained almost constant (with a slight increase, as the number of classes to retrieve increases). This is due to the fact that regardless of the number of classes and the location of the initial probe, all accesses have to be directed to the beginning of the index (at the beginning of the broadcast). The slight increase is attributed to the increase in the total number of objects to be retrieved — assuming that the objects to be retrieved are distributed uniformly along the broadcast. It should be noted, however, that such an increase is only minor since the response time is mainly influenced by the initial procedure.

Two observations can be made: (a) the single-class method offers a better response time than the hierarchical case and (b) the response time for the single-class method increases as the number of retrieved classes increase. The first observation is due to the fact that in the single-class method, accesses do not have to be directed to the beginning

Figure 13. Response time versus number of retrieved classes



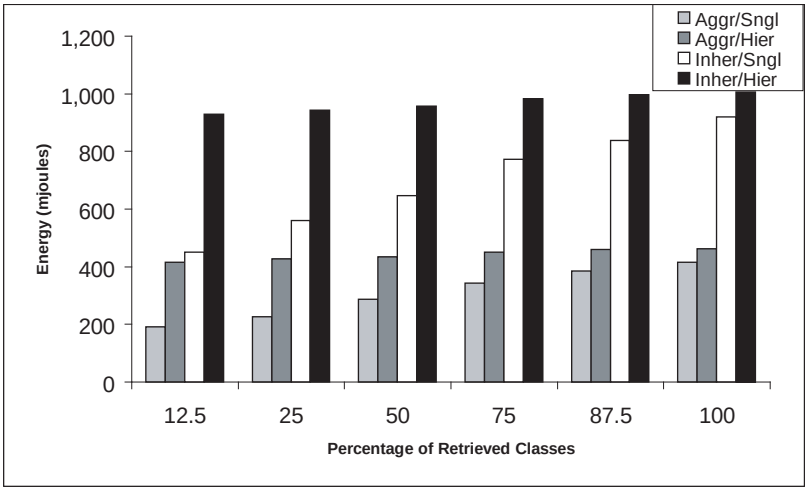
of the broadcast. The second observation is due to the fact that an increase in the number of classes to be retrieved directly increases the number of index and data pages to be accessed.

Indexing based on the aggregation relationship offers lower response time than indexing based on the inheritance relationship since the distribution of the number of objects in the inheritance relationship is more concentrated on the larger objects. Having larger objects results in a longer broadcast, and, hence, it takes longer to retrieve the objects.

b) **Energy.** For each query, the amount of energy consumed is the sum of the energy consumed while the unit is in both active and doze modes. In the case where no index is provided, the mobile unit is in active mode during the entire probe. However, in the case where an index is provided, the active time is proportional to the number of index and data pages to be retrieved. As expected, the active time increases as the number of retrieved classes increases. The hierarchical method searches only one large index tree, whereas the single-class method searches through multiple smaller index trees. The number of pages to retrieve per index tree is proportional to the height of the tree. For a query spanning a single class, the single-class method produces a better active time than the hierarchical method. As the number of classes to be retrieved increases, the hierarchical tree is still traversed only once. However, more single-class trees have to be traversed, and, hence, results in an increase in the active time.

In both the single-class and the hierarchical methods, the aggregation case requires lower active time than the inheritance case since the inheritance case has larger objects, thus requiring the retrieval of more pages. For the sake of practicality, we utilized the power consumption data of the Hitachi SH7032 processor: 130 mW when active and 6.6 mW when in doze. Since power is the amount of energy consumed per unit

Figure 14. Detailed energy consumption



of time, the total energy can be calculated directly using Equation 10:

$$Energy = (ResponseTime - ActiveTime)DozeModePower + (ActiveTime)ActiveModePower \quad (10)$$

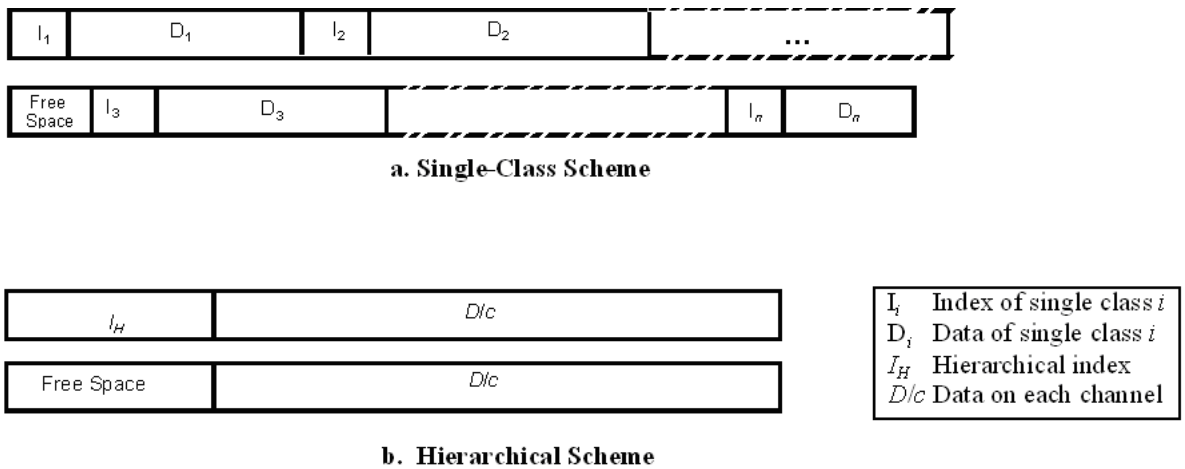
Figure 14 details the energy consumed during the entire query operation in mjoules. The power consumption of the mobile unit is much higher while the unit is in active mode, with a ratio of 19.7. However, our experiments showed that, in general, the duration of the active-mode operations was much smaller than doze-mode operations. As a result, the energy consumed during doze time was the dominating factor. As can be seen from Figure 14, the single-class method is superior to the hierarchical method. This is very similar to the results obtained for the response time, and similarly, the power consumption of the single-class method is lower than that of the hierarchical method.

Data Indexing on Parallel Air Channels

Allocation of Object-Oriented Indexing on Parallel Air Channels

Figure 15 shows the allocation of the single-class and hierarchical-based schemes on the two parallel air channels. For the single-class indexing scheme (Figure 15a), the index and data of each class are distributed and placed along the channels. The hierarchical indexing scheme (Figure 15b) places the index on one channel, and divides and distributes the data among the channels. The most popular data items can be put in the free space. Note that in both cases, similar to the single air channel, it is possible to interleave and distribute the index pages and associated data pages using a variety of methods.

Figure 15. Allocation of single-class and hierarchical indexes on two parallel air channels



Storage Requirement

In the case of broadcasting along parallel air channels, the storage requirement is the same as that for a single air channel.

Timing Analysis

In the case of parallel air channels, one has to account for switching between channels when analyzing access time and power consumption. During the switching time, the pages that are being broadcast on different channels cannot be accessed by the mobile unit. In addition, the mobile unit at each moment of time can tune into one channel—*overlapped page range*. By considering the average page size (512 bytes), communication bandwidth (1Mbit/sec), and switching time (the range of microseconds), we assume that the overlapped page range equals two pages. Finally, we assumed that the power consumption for switching between two channels is 10% of the power consumed in active mode. Equation 11 calculates the power consumption.

a) **Hierarchical Method.** The following protocol shows the sequence of operations.

Hierarchical Protocol

1) Probe onto channel and retrieve offset to the next index *active*
2) Do {Reach the next index *doze*
3) Retrieve the required index pages *active*
4) Do {Reach the next possible required data page *doze*
5) Retrieve the next possible required data page *active*
6) }while every possible required data page is retrieved from the current broadcast
7) } while there are unaccessed data items becuse of overlapped page range

$$EnergyConsumption = (ResponseTime - ActiveTime)* DozeModePower + (ActiveTime)* ActiveModePower + TheNumberOfSwitching *10% * ActiveModePower \quad (11)$$

- **Response Time.** I_H and D are used to denote the size of the hierarchy index and data, respectively. For the c-channel environment, the average size of data on each channel is D/c. To locate the index from the initial probe, it takes half the broadcast of one channel (the size of the broadcast is $I_H + D/c$). Once the index is reached, it has to be completely traversed before data pages appear on the broadcast and, on average, it takes half the size of the data to locate and retrieve the required data items. Thus, the response time from the initial probe to the first complete broadcast is proportional to

$$(I_H + D/c)/2 + I_H + (D/c)/2 = 3\,I_H/2 + D/c \quad (12)$$

Because of the overlapped page range, the mobile units may not be able to get all of the required data during one complete broadcast (e.g., because of conflicts). Therefore, it has to scan the next broadcast. Let P be the probability of the data that are in the same overlapped page range. The distance from the last location to the next index is also half the size of the data of one channel. Once the index is reached, the same process will occur. Thus, the response time from the last location of the previous broadcast until the mobile unit can acquire all of the required data is proportional to

$$P * (D/2c + I_H + D/2c) = P * (I_H + D/c) \quad (13)$$

As a result, on average, the response time is proportional to

$$(1.5 + P)\, I_H + (1 + P)\, D/c \qquad (14)$$

- **Active Time.** The mobile unit has to be active during the first probe (to retrieve a page). Once the index is reached, a number of nonleaf node pages have to be accessed in order to get and retrieve a leaf-node page. The number of pages to be retrieved at the index is equal to the height of the index tree ($\log_f(D)$). The amount of the data pages to be read is equal to the *NODP*. Again, because of the overlapped page range among parallel air channels, the probability of accessing the index of the next broadcast has to be included. Therefore, the active time is proportional to

$$1 + \log_f(D) + NODP + P * \log_f(D) \qquad (15)$$

b) **Single-Class Indexing Scheme**

Single-Class Protocol

1) Probe onto channel and retrieve offset to the next index *active*
2) Do {Reach the next index *doze*
3) Retrieve offsets to the indexes of required classes *active*
4) Reach the next possible index *doze*
5) Retrieve the next possible required index page *active*
6) Do {Reach the next possible index or data page *doze*
7) Retrieve the next possible index or data page *active*
8) } while not (all indexes and data of required classes are scanned)
9) } while there are some data pages which are not retrieved because of overlapped page range

- **Response Time.** As before, the size of a single index and its associated data are labeled as I_i and D_i, respectively. The response time is simply driven by dividing Equation 8 by the number of the air channels (Equation 16):

$$\frac{\sum_{i=y}^{x-1}(I_i + D_i)}{2c} + \frac{\sum_{i=x}^{y}(I_i + D_i)}{c} - \frac{Dy}{2c} \qquad (16)$$

Let P be the probability of the data that are in the same overlapped page range. Thus, the response time for getting the remaining required data items on the second broadast probe is proportional to

$$P*\left(\frac{\sum_{i=y}^{x-1}(I_i + D_i)}{2c} + \frac{\sum_{i=x}^{y}(I_i + D_i)}{c} - \frac{Dy}{2c}\right) \qquad (17)$$

As a result, the response time is proportional to

$$(1+P)*\left(\frac{\sum_{i=y}^{x-1}(I_i + D_i)}{2c} + \frac{\sum_{i=x}^{y}(I_i + D_i)}{c} - \frac{D_y}{2c}\right) \qquad (18)$$

- **Active Time.** Similar to the hierarchical case, the active time is the sum of the height of the trees for all the indexes of the classes to be retrieved plus the number of the corresponding data pages. This is shown in Equation 19. The *2* at the beginning of the equation accounts for the initial probe plus the additional page containing the index of classes. Because of the overlapped page range among parallel air channels, the probability of accessing the index of the next broadcast has to be included. Therefore, the active time is proportional to

$$2 + \sum_{i=x}^{y}\left[\log_f(D_i) + NODP_i\right] + \frac{P}{2}\sum_{i=x}^{y}\log_f(D_i) \qquad (19)$$

Performance Evaluation

Once again, our simulator was extended to study the response time and energy consumption of the single-class and hierarchical indexing schemes in parallel air channels based on the input parameters presented in Table 7.

a) **Response Time.** In the case of no indexing, the response time was constant and independent of the number of channels. This is due to the fact that without any indexing mechanism in place, the mobile unit has to scan every data page in sequence until all required data pages are acquired. Moreover, when indexing schemes are in force, the response time lessens as the number of channels increases.

For the inheritance and aggregation cases, the response time decreases as the number of channels increases. This is due to the fact that, as the number of channels increases, the length of the broadcast becomes shorter. However, the higher the number of channels, the higher the probability of conflicts in accessing data residing on different channels in the overlapped page range. As a result, doubling the number of channels will not decrease the response time by half.

For both the inheritance and aggregation indexing schemes, the single-class method offers a better response time than the hierarchical method. The single-class method accesses do not have to be started at the beginning of the broadcast. For the hierarchical method, on the other hand, any access has to be started from the beginning of the broadcast, which makes the response time of the hierarchical method longer. Indexing based on the aggregation relationship offers a lower response time than that of the inheritance relationship because the distribution of data items in the inheritance relationship is more concentrated on the larger data items.

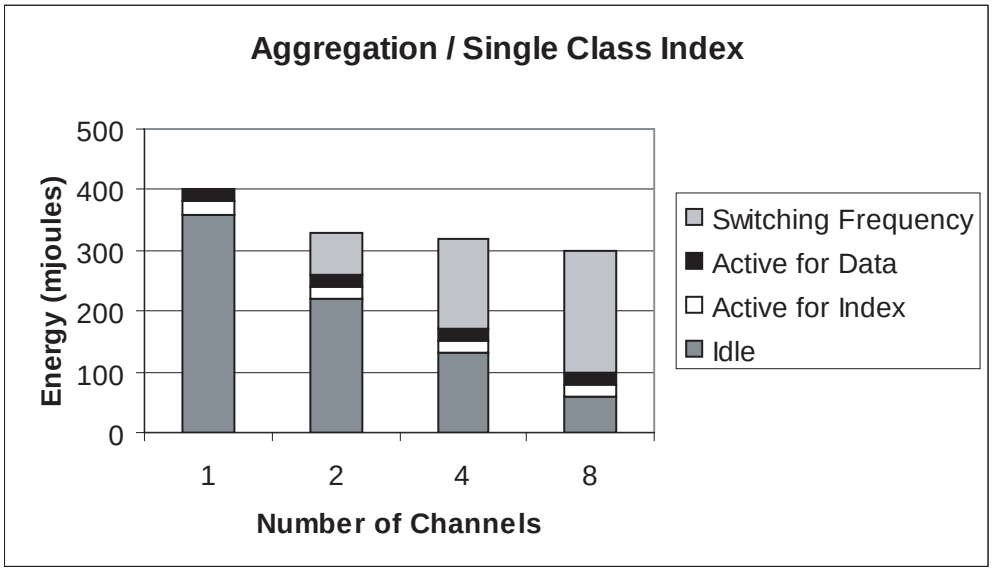
b) **Energy.** The active time is proportional to the number of index and data pages to be retrieved. For broadcast data without an index, the active time is the same as the response time. In addition, for all four indexing schemes, the active time remains almost constant and independent of the number of air channels. This is because the active time is proportional to the number of index and data pages to be retrieved.

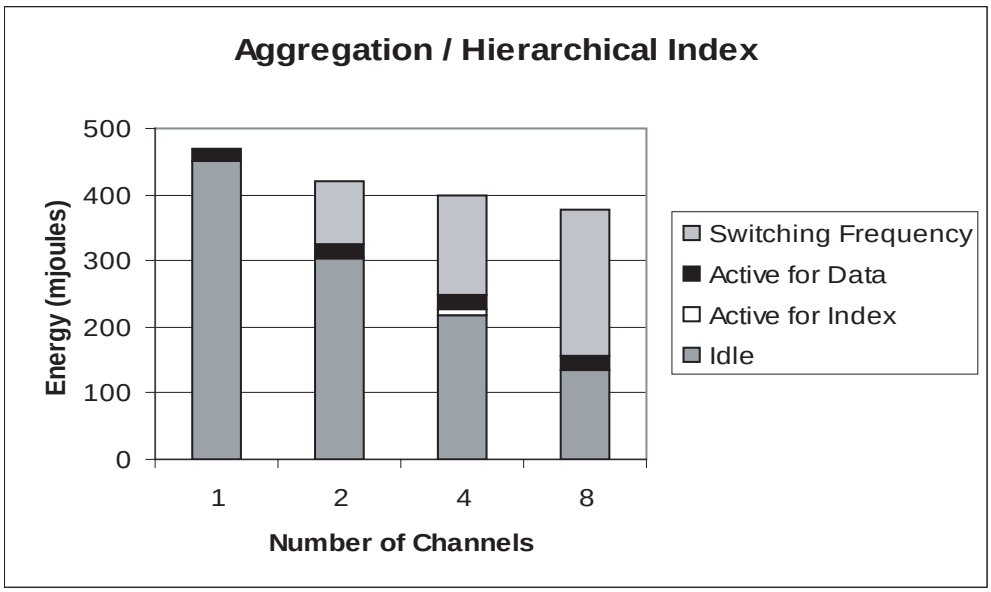
In general, the hierarchical method requires less active time than the single-class method. The hierarchical method searches only one large index tree, whereas the single-class method searches through multiple smaller index trees, and the number of pages to be retrieved per index tree is proportional to the height of the tree. In both the single-class and the hierarchical methods, the indexing based on an aggregation relationship requires lower active time than the inheritance method. This is simply due to the fact that the inheritance relationship resulted in larger data items, thus requiring the retrieval of more pages.

In a separate simulation run we observed the total energy consumption. It was concluded that the total energy consumption of broadcasting without any indexing schemes is much higher than that of broadcasting supported by indexing, and the energy consumption of the single-class method is lower than that of the hierarchical method. This is very similar to the results obtained for the response time. When indexing was supported, energy consumption, on average, decreased about 15 to 17 times in the case of the aggregation relationship and the inheritance relationship, respectively.

Figure 16 shows the detail of energy consumption for the aggregation relationship. As the number of channels increases, the energy

Figure 16. Detailed energy consumption





consumption during idle time decreases. The energy consumption for retrieving indexes increases because the probability of the data being in the same overlapped page range increases. The higher this probability, the more the mobile unit has to get the index from the next broadcast. Finally, the energy consumption for switching between two

different channels increases because the required data are distributed among the channels. The larger the number of channels, the more distributed is the data among the channels, and, consequently, the more frequent switching between channels.

Section Conclusion

This section investigated an energy-efficient solution by the means of applying indexing schemes to object-oriented data broadcast over single and parallel air channels. Two methods, namely, the hierarchical and single-class methods, were explored. Timing analysis and simulation were conducted to compare and contrast the performance of different indexing schemes against each other. It was shown that including an index degrades the response time moderately, however, such degradation is greatly offset by the improvement in energy consumption. For a single air channel, broadcasting with supported indexing schemes increased the response time when compared with broadcasting without indexing support. However, the response time is reduced by broadcasting data with an index along parallel air channels. Moreover, the response time decreased as the number of air channels is increased. Relative to nonindexed broadcasting, the mobile unit's energy consumption decreased rather sharply when indexing is supported. For a set of queries retrieving data items along the air channel(s), the single-class indexing method resulted in a faster response time and lower energy consumption than the hierarchical method.

CONFLICTS AND GENERATION OF ACCESS PATTERNS

One of the problems associated with broadcasting information on parallel air channels is the possibility for conflicts between accessing data items on different channels. Because the mobile unit can tune into only one channel at a time, some data items may have to be retrieved on subsequent broadcasts. In addition, during the channel switch time, the mobile unit is unable to retrieve any data from the broadcast. Conflicts will directly influence the access latency and, hence, the overall execution time. This section is intended to provide a mathematical foundation to calculate the expected number of passes required to retrieve a set of data items requested by an application from parallel air channels by formulating this problem as an *asymmetric traveling salesman problem* (TSP). In addition, in an attempt to reduce the access time and power consumption, we propose heuristic policies that can reduce the number of passes over parallel air channels. Analysis of the effectiveness of such policies is also the subject of this section.

Enumerating Conflicts

Equation 1 showed the number of broadcasts (passes) required to retrieve K data items from M parallel channels if conflicts between data items are independent. To calculate $P(i)$, it is necessary to count the number of ways the data items can be distributed while having exactly i conflicts, then divide it by the total number of ways the K data items can be distributed over the parallel channels. In order to enumerate possible conflicting cases, we classify the conflicts as single or double conflicts as defined below.

Definition 7. A single conflict is defined as a data item in the conflict region that does not have another data item in the conflict region in the same row. A double conflict is a data item that is in the conflict region and does have another data item in the conflict region in the same row.

The number of data items that cause a double conflict, d, can range from 0 (all single conflicts) up to the number of conflicts, i, or the number of remaining data items, $(K - i - 1)$. When counting combinations, each possible value of d must be considered separately. The number of possible

combinations for each value of d is summed to determine the total number of combinations for the specified value of i. When counting the number of ways to have i conflicts and d double conflicts, four factors must be considered.

- Whether each of the $(i - d)$ data items representing a single conflict is in the left or right column in the conflict region. Because each data item has two possible positions, the number of variations due to this factor is $2^{(i-d)}$.
- Which of the $(M - 1)$ rows in the conflict region are occupied by the $(i - d)$ single conflicts. The number of variations due to this factor is $\binom{M-1}{i-d}$.
- Which of the $(M-1)-(i-d)$ remaining rows in the conflict region are occupied by the d double conflicts; $(i-d)$ is subtracted because a double conflict cannot occupy the same row as a single conflict. The number of variations due to this factor is $\binom{(M-1)-(i-d)}{d}$
- Which of the $(MN - 2M + 1)$ positions not in the conflict region are occupied by the $(K - i - d - 1)$ remaining data items. The number of variations due to this factor is $\binom{MN-2M+1}{K-i-d-1}$.

Note that these sources of variation are independen from each other and, hence:

$$P(i) = \frac{\sum_{d=0}^{d \le MIN(i,K-i-1)} 2^{(i-d)} \binom{M-1}{i-d} \binom{(M-1)-(i-d)}{d} \binom{MN-2M+1}{K-i-d-1}}{\binom{MN-1}{K-1}} \quad (20)$$

If the conflicts produced by one data item are independent from the conflicts produced by all other data items, then Equation 20 will give the number of passes required to retrieve all K requested data items. However, if the conflicts produced by one data item are not independent of the conflicts produced by other data items, additional conflicts will occur which are not accounted for in our analysis. Equation 20 will thus underestimate the number of broadcasts required to retrieve all K data items.

Retrieving Data from Parallel Broadcast Air Channels in the Presence of Conflicts

The problem of determining the proper order to retrieve the requested data items from the parallel channels can be modeled as a TSP. Making the transformation from a broadcast to the TSP requires the creation of a complete directed graph G with K nodes, where each node represents a requested object. The weight w of each edge (i, j) indicates the number of broadcasts that must pass in order to retrieve data item j immediately after retrieving data item i. Since any particular data item can be retrieved in either the current broadcast or the next broadcast, the weight of each edge will be either 0 or 1. A weight of 0 indicates that the data item j is after data item i in the broadcast with no conflict. A weight of 1 indicates that data item j is either before or in conflict with data item i.

Simulation Model

The simulation models a mobile unit retrieving data items from a broadcast. A broadcast is represented as an N x M two-dimensional array, where N represents the number of data items in each channel of a broadcast and M represents the number of parallel channels. For each value of K, where K represents the number of requested data items $(1 \le K \le MN)$, the simulation randomly generates 1,000 patterns representing the uniform distribution of K data items among the broadcast channels. The K data items from each randomly generated pattern are retrieved using various retrieval algorithms. The number of passes is recorded and compared. To prevent the randomness of the broadcasts from affecting the

comparison of the algorithms, the same broadcast is used for each algorithm in a particular trial and the mean value is reported for each value of *K*. Finally, several algorithms for ordering the retrieval from the broadcast, both TSP related and non-TSP related, were analyzed.

Data Retrieval Algorithms

Both exact and approximate TSP solution finders and two heuristic based methods were used to retrieve the data items from the broadcast.

a) **TSP Methods.** An exact TSP solution algorithm was used to provide a basis for comparison with the other algorithms. These algorithms are simply too slow and too resource intensive. While a better implementation of the algorithm may somewhat reduce the cost, it cannot change the fact that finding the exact solution will require exponential time for some inputs. Knowing the exact solution to a given TSP does, however, allow us to evaluate the quality of a heuristic approach. A TSP heuristic based on the assignment problem relaxation requires far less CPU time and memory than the optimal tour finders, so it is suitable for use on a mobile unit. A publicly available TSP solving package named TspSolve (Hurwitz& Craig, 1996) was used for all TSP algorithm implementations.

b) **Next Data Item Access.** The strategy used by this heuristic is simply to always retrieve the next available data item in a broadcast. This can be considered as a greedy approach. It is also similar to the nearest neighbor approach to solving TSP problems.

c) **Row Scan.** A simple *row scan* heuristic was also used. This algorithm simply reads all the data items from one channel in each pass. If a channel does not have any requested data in it, it is skipped. This algorithm will always require as many passes as there are channels with requested data items in them. The benefit of this algorithm is that it does not require any time to decide on an ordering. It can thus begin retrieving data items from a broadcast immediately. This is especially important when a large percentage of the data items in a broadcast are requested.

Results

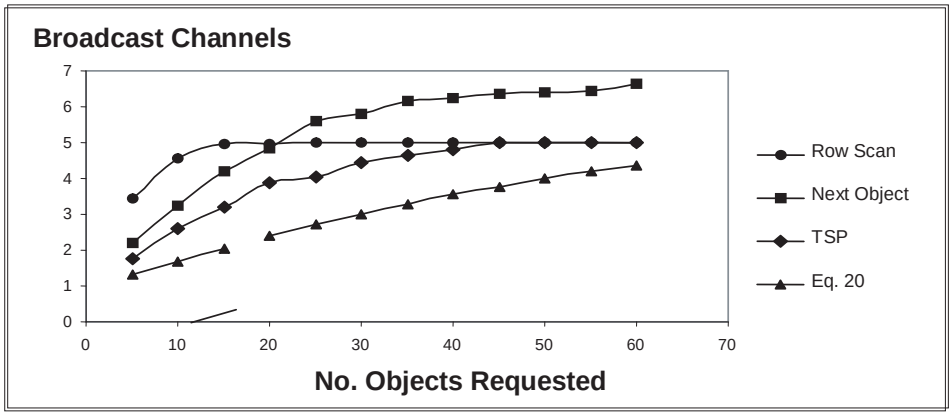
As expected, the TSP methods provide much better results than both the two heuristic-based algorithms. Our simulations showed that the TSP heuristic performed almost exactly as well as the optimal TSP algorithm. This is a very interesting observation because it means that one can use a fast heuristic to schedule retrievals of data items from the broadcast without any performance degradation.

In Figure 17, the TSP methods show that the number of broadcasts required to retrieve all *K* requested data items from a broadcast is much greater than the number of broadcasts predicted by Equation 20 — Equation 20 was based on the assumption that the conflicts among the requested data items are independent. Figure 17 used five parallel channels and 20 pages per channel. It is also interesting to note that the straightforward row scan nearly matches the performance of the TSP-based algorithms when more than about 45% of the total number of data items is requested. In this case, there are so many conflicts that it is virtually impossible to avoid having to make as many passes as there are parallel channels. When this occurs, it is better to do the straightforward row scan than to spend time and resources running a TSP heuristic.

Optimal Number of Broadcast Channels

More channels mean that a given amount of information can be made available in a shorter period of time at the expense of more conflicts.

Figure 17. Comparison of several algorithms for retrieving objects from parallel channels



The simulation results showed that it is always advantageous to use more broadcast channels. While there will be more conflicts between data items, this does not quite counteract the shorter broadcast length of the many-channel broadcasts. This was especially evident when only a few data items in a broadcast were being accessed.

Ordered Access List

The scope of the general access protocol for indexed parallel-channel configuration in the presence of conflicts was extended in order to use heuristics that can generate the ordered access list of requested data items that reduces

- the number of passes over the air channels and
- the number of channel switches.

During the *Search* step, the index is accessed to determine the offset and the channel of the requested data items. Then, a sequence of access patterns is generated. Finally, the *Retrieval* step is performed following the generated access patterns.

Extended Retrieval Protocol

1) Probe the channel and retrieve the offset to the next index
2) Access the next index
3) Do {Search the index for the requested object
4) Calculate the offset of the object
5) Get the channel on which the object will be broadcast
6) } while there is an unprocessed requested object
7) Generate access patterns for the requested objects (using retrieval scheme)
8) Do {Wait for the next broadcast cycle
9) Do {Reach the first object as indicated by the access pattern
10) Retrieve the object
11) } while there is an unretrieved object in the access pattern

12) } while there is an unprocessed access pattern

Performance Evaluation

We extended the simulator to emulate the process of accessing data from a hierarchical indexing scheme in parallel air channels. Moreover, the simulator also analyzes the effect of conflicts on the average access time and power consumption.

Our retrieval scheme, based on the user request, generates a retrieval forest representing all possible retrieval sequences. However, as expected, the generated retrieval forest grows exponentially with the number of requested data items. The key observation needed to reduce the size of the tree is to recognize that each requested data item has a unique list of children, and the number of children for a particular data item is limited to the number of channels. The simulator takes advantage of these observations to reduce the size of the retrieval tree and the calculation time without sacrificing accuracy.

The generation of the user requests was performed randomly, representing a distribution of K data items in the broadcast. In various simulations runs, the value of K was varied from one to $N \times M$—in a typical user query of public data, K is much less than $N \times M$. Finally, to take into account future technological advances, parameters such as transmission rate and power consumption in different modes of operation were fed to the simulator as variable entities.

The simulator calculates the average active time, the average idle time, the average query response time, the average number of broadcast passes, the number of channel switches, and the energy consumption of the retrieval process. As a final note, the size of the index was 13.52% of the size of the broadcast (not including the index) and the number of channels varied from 1 to 16 (2 to 17 when an independent channel was used for transmitting the index).

Simulation Model

For each simulation run, a set of input parameters, including the number of parallel air channels, the broadcast transmission rate, and the power consumption in different operational modes, was passed to the simulator. The simulator was run 1,000 times and the average of the designated performance metrics was calculated. The results of the simulations where an indexing scheme was employed were compared against a broadcast without any indexing mechanism. Two indexing scenarios were simulated.

- **Case 1.** The index was transmitted with the data in the first channel (index with data broadcast).
- **Case 2.** The index was transmitted over a dedicated channel in a cyclic manner.

Results

A comparison between the extended retrieval protocol against the row scan algorithm was performed. The index transmission was performed in a cyclic manner on an independent channel, and the number of requested data items was varied between 5 and 50 out of 5,464 securities within the NASDAQ exchange database. The simulation results showed that, regardless of the number of parallel air channels, the proposed algorithm reduces both the number of passes and the response time compared to the row scan algorithm. Moreover, the energy consumption was also reduced, but only when the number of data items retrieved was approximately 15 or less (Table 10).

Relative to the row scan algorithm, one should also consider the expected overhead of the proposed algorithm. The simulation results showed that in the worst case, the overhead of the proposed algorithm was slightly less than the time required to transmit one data page.

Table 10. Improvement of proposed algorithm versus row scan (10 data items requested)

# of Channels	# of Passes	Response Time	Energy
2	48.0%	28.0%	2.7%
4	68.0%	43.6%	3.1%
8	72.3%	46.5%	3.3%
16	71.8%	40.8%	3.4%

a) **Response Time.** Figures 18 to 20 show the response times in terms of the number of data items requested and the number of broadcast channels. Three cases were examined.

- **Case 1. Data and index are intermixed on broadcast channel(s).** Figure 18 shows the response times for different numbers of broadcast channels when retrieving the full range of existing data items from the broadcast. It can be concluded that when a few data items are requested, the response time decreases as the number of channels increases. After a certain threshold point, the response time increases as the number of channels increases. This is due to an increase in the number of conflicts and hence an increase in the number of passes over the broadcast channels to retrieve the requested data.
- **Case 2. Index is broadcast over a dedicated channel in cyclic fashion.** Similar to Case 1, Figure 19 depicts the simulation results when retrieving the full range of existing data from the broadcast. Again, as expected, the

Figure 18. Response time (Case 1)

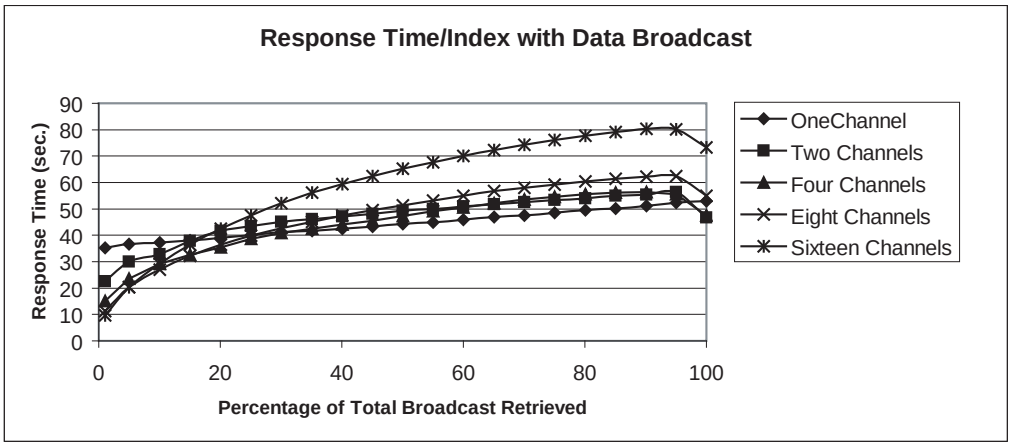


Figure 19. Response time (Case 2)

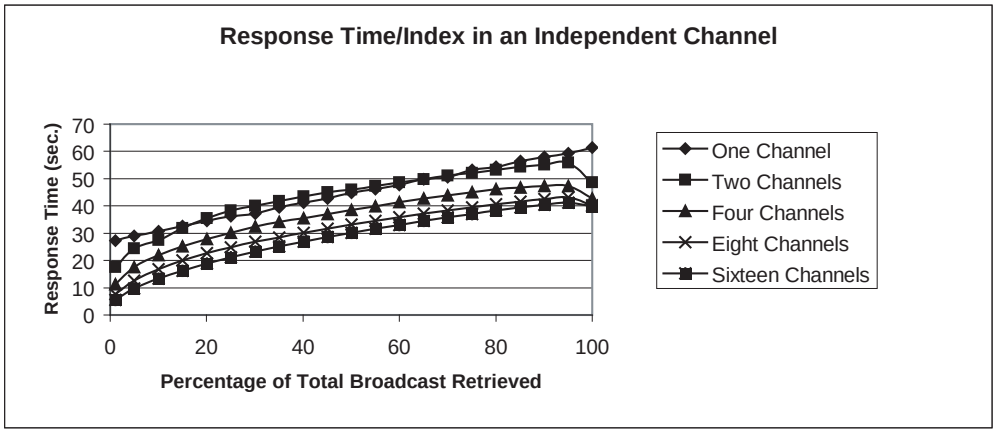
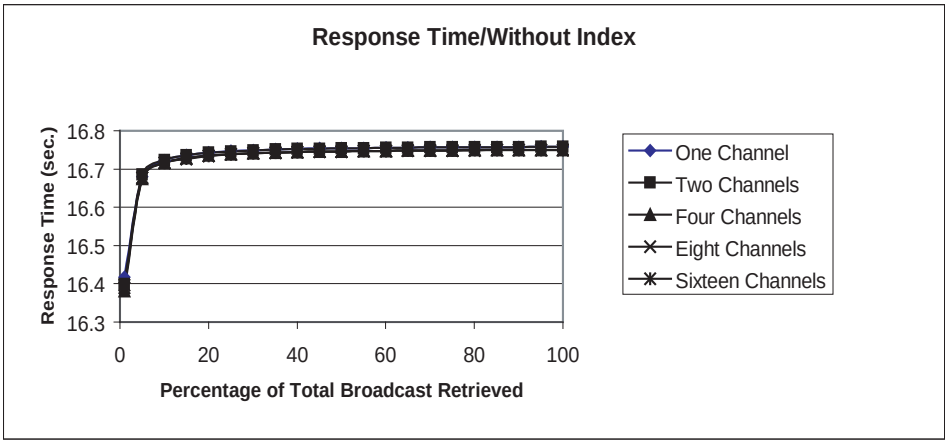


Figure 20. Response time (Case 3)



response time decreases as the number of channels increases. Comparing the results for one-channel and two-channel configuration, we can conclude that in some instances the two-channel configuration is not as effective — there is the possibility of conflicts, many of which unavoidably cause an increase in the number of passes and hence longer response time.

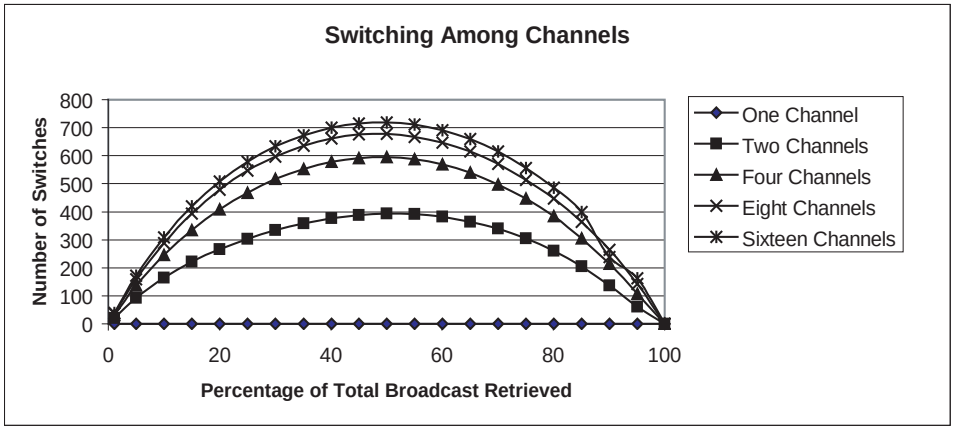
- **Case 3. No indexing is employed.** From Figure 20 one can conclude that the response time remains relatively constant regardless of the number of channels used. In this organization, the user must scan the same amount of data regardless of the user query and number of parallel channels.

In general, employment of an indexing scheme reduces the response time when retrieving a relatively small number of data items. As the percentage of data items requested increases, the number of conflicts increases as well. The proposed retrieval protocol tries primarily to reduce the conflicts in each pass of the broadcast; however, when the number of potential conflicts increases considerably, some conflicts become unavoidable, causing an increase in the number of passes and hence an increase in the response time. When the percentage of requested data approaches 100%, the response time reduces. This proves the validity of the proposed scheduling algorithm since it generates the same retrieval sequence as the row scan method.

b) **Switching Frequency.** Again, three cases were examined.

- **Case 1 & Case 2. Employment of indexing schemes.** Figure 21 shows the switching frequency for Case 1 and Case 2—the switching pattern is not affected by the indexing policy employed. From this figure one can conclude that the switching frequency increases as the number of channels and number of data items retrieved increase. This can be explained by an increase in the number of conflicts; as the proposed method tries to reduce the number of conflicts, the switching frequency will increase. Also, as stated previously, an increase in the number of channels increases the number of conflicts as well. One can notice that when the percentage of data items requested exceeds 50%, the switching frequency begins to decrease. This is due to the fact that the proposed method does not attempt to switch channels as often to avoid the conflicts as the number of conflicts increases substantially.

Figure 21. Switching frequency (Case 1 and Case 2)



- **Case 3. No indexing is employed.** When no indexing technique is utilized, the row scan method is employed, producing a constant switching frequency independent of the number of data items requested. The switching frequency is, at the most, equal to the number of total channels employed in the simulation.

c) **Energy Consumption.** Figures 22 and 23 depict detailed energy consumption when 1% of data items on the broadcast are requested. It can be observed that the energy consumption is almost the same; however, Case 1 consumes more energy than Case 2 in doze mode.

Figure 22. Energy consumption (Case 1)

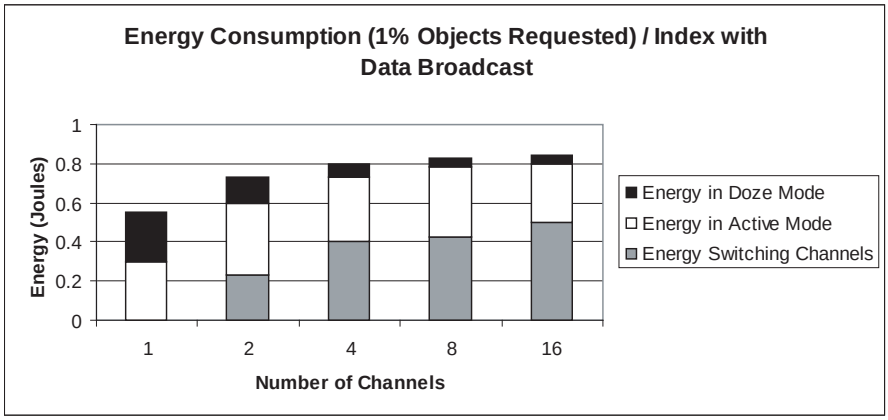
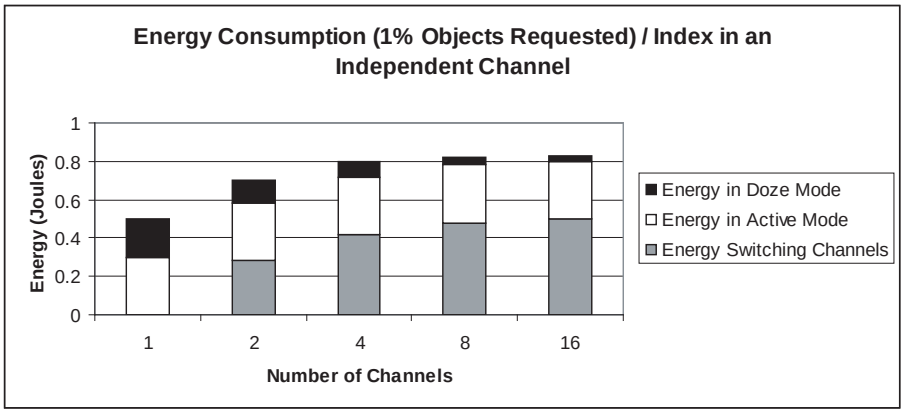


Figure 23. Energy consumption (Case 2)



- **Case 1 & Case 2. Employment of indexing schemes.** In general, due to the increase in the number of channel-switching frequency, the energy consumption increases as the number of channels increases. In addition, we noted that the energy consumption increases when up to 50% of the broadcast data items are requested, then it decreases as the number of requested data items increases. This is directly related to the channel-switching frequency. In Case 1, in many instances, the mobile unit must wait in doze mode while the index is retransmitted.
- **Case 3. No indexing is employed.** When no indexing technique is used (Figure 24), the energy consumption varies only minimally due to the nature of the row scan algorithm employed.

From these figures we can observe that both Case 1 and Case 2 consume less power than Case 3 when a small percentage of data items is retrieved (around 1%). When the percentage of data items requested increases, the number of conflicts, the switching frequency, and, consequently, the energy consumption increase.

d) **Number of Passes.** As a note, the number of passes is independent of the index allocation scheme. Therefore, the number of passes for Cases 1 and 2 is the same.

- **Case 1 & Case 2. Employment of indexing schemes.** The increase in the number of passes is directly related to the increase in the number of channels and increase in the number of data items requested (Figure 25). An increase in the number of channels implies an increase in the number of conflicts, and, hence, the higher possibility of unavoidable conflicts, resulting in an increase in the number of passes. It can be noticed that when the number of data items requested is large, the number of passes exceeds the number of channels available. This is due to the priority order of the heuristics used in the proposed retrieval algorithm. In general, it is improbable that a query for public data requests a lot of data

Figure 24. Energy consumption (Case 3)

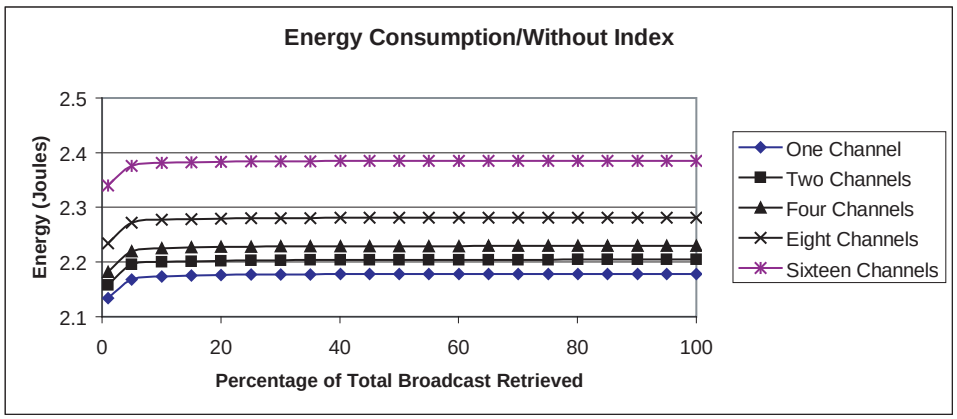
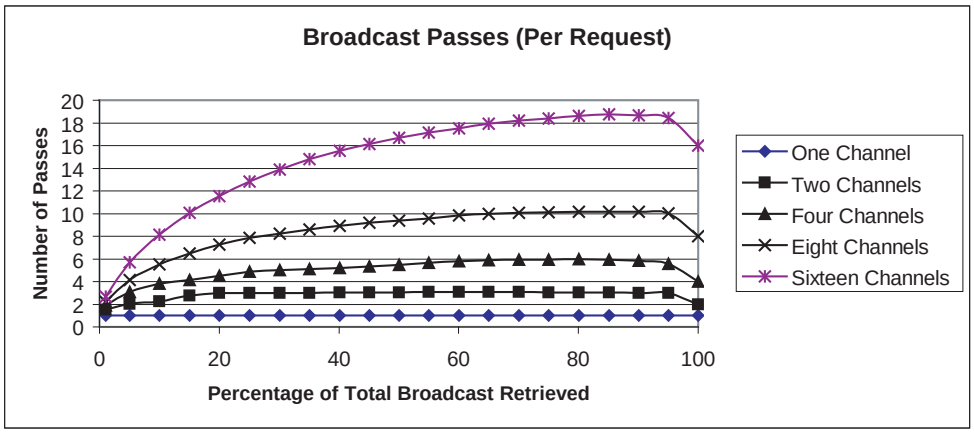


Figure 25. Number of passes over the parallel channels



items from the broadcast channels. Our experience showed that for a query requesting up to 50 data items, the proposed method reduces the number of passes compared to Case 3.

- **Case 3. No indexing is employed.** In contrast, when no indexing technique is employed, the number of passes required is a function of the number of air channels. In the worst case we need *N* passes where *N* is the number of broadcast channels.

Section Conclusion

Conflicts directly influence the access latency and, hence, the overall execution time. This section provided a mathematical foundation to calculate the expected number of passes required to retrieve a set of data items requested by an application from parallel air channels. In addition, in an attempt to reduce the access time and power consumption, heuristics were used to develop access policies that reduce the number of passes over the parallel air channels. Analysis of the effectiveness of such policies was also the subject of this section.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This chapter aims to address the applicability and effectiveness of data broadcasting from two viewpoints: energy and response time. Within the scope of data broadcasting, we discussed different data allocation schemes, indexing approaches, and data retrieval methods for both single and parallel air channels. Comparisons of different algorithms were demonstrated through simulation results.

The scope of this research can be extended in many directions. For instance, we assumed that the resolution of queries happens on an individual basis at the mobile unit. It may be possible to reduce computation by utilizing a buffer and bundling several queries together, processing them as a whole. Our proposed scheduling scheme

was based on three prioritized heuristics. It is interesting to investigate a new set of heuristics that can reduce the switching frequency while retrieving a large percentage of data items from the broadcast.

ACKNOWLEDGMENT

This work would have not been possible without the sincere effort of many students who participated in the development of conceptual issues as well as simulation results. We would like to thank them. In addition, this work in part has been supported by the Office of Naval Research and the National Science Foundation under the contracts N00014-02-1-0282 and IIS-0324835, respectively.

REFERENCES

Acharya, S., Alonso, R., Franklin, M., & Zdonik, S. (1995). Broadcast disks: Data management for asymmetric communication environments. *Proceedings of ACM SIGMOD International Conference on the Management of Data*, (pp. 199-210).

Alonso, R., & Ganguly, S. (1992). Energy efficient query optimization. *Technical Report MITL-TR-33-92*, Princeton, NJ: Matsushita Information Technology Laboratory.

Alonso, R., & Korth, H. F. (1993). Database system issues in nomadic computing. *Proceedings of ACM SIGMOD Conference on Management of Data*, (pp. 388-392).

Atkinson, M., Bancilhon, F., DeWitt, D., Dittrich, K., Maier, D., & Zdonik, S. (1989). The object-oriented database system manifesto. *Proceedings of Conference on Deductive and Object-Oriented Databases*, (pp. 40-57).

Badrinath, B. R. (1996). Designing distributed algorithms for mobile computing networks. *Computer Communications, 19*(4), 309-320.

Banerjee, J., Kim, W., Kim, S.-J., & Garza, J. F. (1988). Clustering a DAG for CAD databases. *IEEE Transactions on Software Engineering, 14*(11), 1684-1699.

Boonsiriwattanakul, S., Hurson, A. R., Vijaykrishnan, N., & Chehadeh, C. (1999). Energy-efficient indexing on parallel air channels in a mobile database access system. *Proceedings of the Third World Multiconference on Systemics, Cybernetics, and Informatics, and Fifth International Conference on Information Systems Analysis and Synthesis, IV,* (pp. 30-38).

Bowen, T. F. (1992). The DATACYCLE architecture. *Communication of ACM, 35*(12), 71-81.

Bright, M. W., Hurson, A. R., & Pakzad, S. (1992). A taxonomy and current issues in multidatabase systems. *IEEE Computer, 25*(3), 50-60.

Bright, M. W., Hurson, A. R., & Pakzad, S. (1994). Automated resolution of semantic heterogeneity in multidatabases. *ACM Transactions on Database Systems, 19*(2), 212-253.

Chang, E. E., & Katz, R. H. (1989). Exploiting inheritance and structure semantics for effective clustering and buffering in an object-oriented DBMS. *Proceedings of ACM SIGMOD Conference on Management of Data*, (pp. 348-357).

Chehadeh, Y. C., Hurson, A. R., & Tavangarian, D. (2001). Object organization on single and parallel broadcast channel. *Proceedings of High Performance Computing*, (pp. 163-169).

Chehadeh, Y. C., Hurson, A. R., & Kavehrad, M. (1999). Object organization on a single broadcast channel in the mobile computing environment [Special issue]. *Multimedia Tools and Applications Journal, 9*, 69-94.

Chehadeh, Y. C., Hurson, A. R., & Miller L. L. (2000). Energy-efficient indexing on a broadcast channel in a mobile database access system. *Proceedings of IEEE Conference on Information Technology*, (pp. 368-374).

Chehadeh, Y. C., Hurson, A. R., Miller, L. L., Pakzad, S., & Jamoussi, B. N. (1993). Application of parallel disks for efficient handling of object-oriented databases. *Proceedings of the Fifth IEEE Symposium on Parallel and Distributed Processing*, (pp. 184-191).

Cheng, J.-B. R., & Hurson, A. R. (1991a). Effective clustering of complex objects in object-oriented databases. *Proceedings of ACM SIGMOD Conference on Management of Data*, (pp. 22-27).

Cheng, J.-B. R., & Hurson, A. R. (1991b). On the Performance issues of object-based buffering. *Proceedings of International Conference on Parallel and Distributed Information Systems*, (pp. 30-37).

Chlamtac, I., & Lin, Y.-B. (1997). Mobile computing: When mobility meets computation. *IEEE Transactions on Computers*, *46*(3), 257-259.

Comer, D. C. (1991). *Internetworking with TCP/IP Volume I: Principles, Protocols, and Architecture* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Demers, A., Pertersen, K., Spreitzer, M., Terry, D., Theier, M., & Welch, B. (1994). The bayou architecture: Support for data sharing among mobile users. *Proceedings of IEEE Workshop on Mobile Computing Systems and Applications*, (pp. 2-7).

Fong, E., Kent, W., Moore, K., & Thompson, C. (1991). *X3/SPARC/DBSSG/OODBTG Final Report*. Available from NIST.

Fox, A., Gribble, S. D., Brewer, E. A., & Amir, E. (1996). Adapting to network and client variability via on-demand dynamic distillation. *Proceedings of ASPLOS-VII*, Boston, Massachusetts, (pp. 160-170).

Honeyman, P., Huston, L., Rees, J., & Bachmann, D. (1992). The LITTLE WORK project. *Proceedings of the Third IEEE Workshop on Workstation Operating Systems*, (pp. 11-14).

Hu, Q.L., & Lee, D. L. (2000). Power conservative multi-attribute queries on data broadcast. *Proceedings of IEEE International Conference on Data Engineering (ICDE 2000)*, (pp. 157-166).

Hu, Q. L., & Lee, D. L. (2001). A hybrid index technique for power efficient data broadcast. *Distributed and Parallel Databases Journal*, *9*(2), 151-177.

Hurson, A. R., Chehadeh, Y. C., & Hannan, J. (2000). Object organization on parallel broadcast channels in a global information sharing environment. *Proceedings of IEEE Conference on Performance, Computing, and Communications*, (pp. 347-353).

Hurson, A. R., Pakzad, S., & Cheng, J.-B. R. (1993). Object-oriented database management systems. *IEEE Computer, 26*(2), 48-60.

Hurwitz, C. & Craig, R. J. (1996). *Software Package Tsp_Solve 1.3.6*. Available from http://www.cs.sunysb.edu/~algorithm/implement/tsp/implement.shtml.

Imielinski, T., & Badrinath, B. R. (1994). Mobile wireless computing: Challenges in data management. *Communications of the ACM, 37*(10), 18-28.

Imielinski, T., & Korth, H. F. (1996). Introduction to mobile computing. In T. Imielinski and H. F. Korth (Eds.), *Mobile computing* (pp. 1-43). Boston: Kluwer Academic.

Imielinski, T., Viswanathan, S., & Badrinath, B. R. (1994). Energy efficient indexing on air. *Proceedings of ACM SIGMOD Conference on Management of Data*, (pp. 25-36).

Imielinski, T., Viswanathan, S., & Badrinath, B. R. (1997). Data on air: Organization and access.

IEEE Transactions on Computer, 9(3), 353-372.

Joseph, A. D., Tauber, J. A., & Kaashoek, M. F. (1997). Mobile computing with the rover toolkit [Special issue]. *IEEE Transactions on Computers, 46*(3), 337-352.

Juran, J., Hurson, A. R., & Vijaykrishnan, N. (2004). Data organization and retrieval on parallel air channels: Performance and energy issues. *ACM Journal of WINET, 10*(2), 183-195.

Kaashoek, M. F., Pinckney, T., & Tauber, J. A. (1994). Dynamic documents: Mobile wireless access to the WWW. *IEEE Workshop on Mobile Computing Systems and Applications,* 179-184.

Kim, W. (1990). *Introduction to object-oriented databases.* Cambridge, MA: MIT Press.

Lai, S. J., Zaslavsky, A. Z., Martin, G. P., & Yeo, L. H. (1995). Cost efficient adaptive protocol with buffering for advanced mobile database applications. *Proceedings of the Fourth International Conference on Database Systems for Advanced Applications.*

Lee, D. L. (1996). Using signatures techniques for information filtering in wireless and mobile environments [Special issue]. *Distributed and Parallel Databases, 4*(3), 205-227.

Lee, M. T., Burghardt, F., Seshan, S., & Rabaey, J. (1995). InfoNet: The networking infrastructure of InfoPad. *Proceedings of Compcon*, (pp. 779-784).

Lim, J.B., & Hurson, A. R. (2002). Transaction processing in mobile, heterogeneous database systems. *IEEE Transactions on Knowledge and Data Engineering, 14*(6), 1330-1346.

Lim, J. B., Hurson, A. R., Miller, L. L., & Chehadeh, Y. C. (1997). A dynamic clustering scheme for distributed object-oriented databases. *Mathematical Modeling and Scientific Computing,* 8, 126-135.

Munoz-Avila, A., & Hurson, A. R. (2003a). Energy-aware retrieval from indexed broadcast parallel channels. *Proceedings of Advanced Simulation Technology Conference (High Performance Computing),* (pp. 3-8).

Munoz-Avila, A., & Hurson, A. R. (2003b). Energy-efficient objects retrieval on indexed broadcast parallel channels. *Proceedings of International Conference on Information Resource Management,* (pp. 190-194).

NASDAQ World Wide Web Home Page. (2002). Retrieved May 11, 2004, from http://www.nasdaq.com

Satyanarayanan, M. (1996). Fundamental challenges in mobile computing. *Proceedings* of *15th ACM Symposium on Principles of Distributed Computing,* (pp. 1-7).

Satyanarayanan, M., Noble, B., Kumar, P., & Price, M. (1994). Application-aware adaptation for mobile computing. *Proceedings of the Sixth ACM SIGOPS European Workshop,* (pp. 1-4).

Weiser, M. (1993). Some computer science issues in ubiquitous computing. *Communications of the ACM, 36*(7), 75-84.

Zdonik, S., Alonso, R., Franklin, M., & Acharya, S. (1994). Are disks in the air just pie in the sky? *Proceedings of Workshop on Mobile Computing Systems and Applications, (pp. 1-8).*

This work was previously published in Wireless Information Highways, edited by D. Katsaros, A. Nanopoulos, and Y. Manalopoulos, pp. 96-154, copyright 2005 by IRM Press (an imprint of IGI Global).

Chapter 7.37
Multimedia over Wireless Mobile Data Networks

Surendra Kumar Sivagurunathan
University of Oklahoma, USA

Mohammed Atiquzzaman
University of Oklahoma, USA

ABSTRACT

With the proliferation of wireless data networks, there is an increasing interest in carrying multimedia over wireless networks using portable devices such as laptops and personal digital assistants. Mobility gives rise to the need for handoff schemes between wireless access points. In this chapter, we demonstrate the effectiveness of transport layer handoff schemes for multimedia transmission, and compare with Mobile IP, the network layer-based industry standard handoff scheme.

I. INTRODUCTION

Mobile computers such as personal digital assistants (PDA) and laptop computers with multiple network interfaces are becoming very common. Many of the applications that run on a mobile computer involve multimedia, such as video conferencing, audio conferencing, watching live movies, sports, and so forth. This chapter deals with multimedia communication in mobile wireless devices, and, in particular, concentrates on the effect of mobility on streaming multimedia in wireless networks.

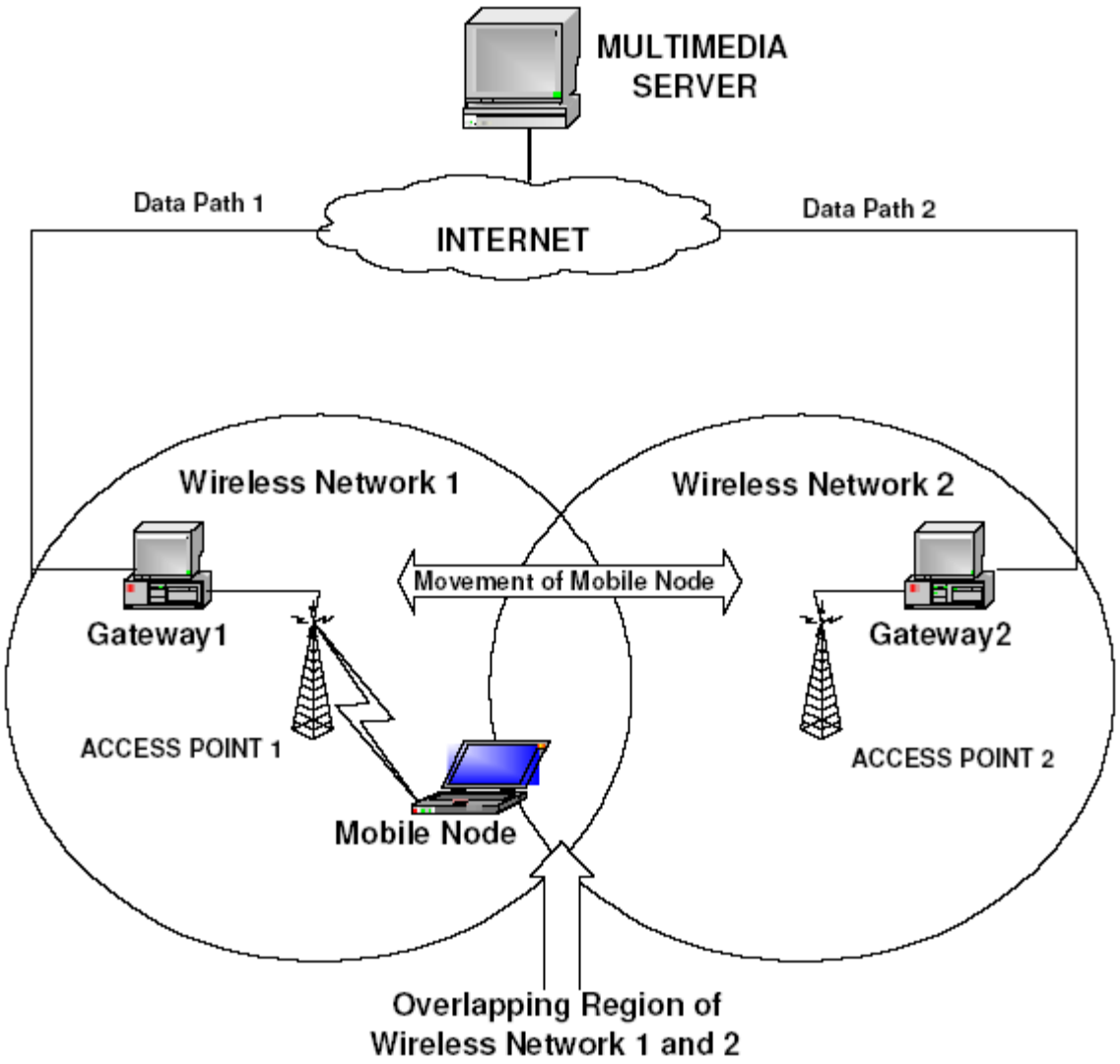
Streaming multimedia over wireless networks is a challenging task. Extensive research has been carried out to ensure a smooth and uninterrupted multimedia transmission to a mobile host (MH) over wireless media. The current research thrust is to ensure an uninterrupted multimedia transmission when the MH moves between networks or subnets. Ensuring uninterrupted multimedia transmission during handoff is challenging because the MH is already receiving multimedia from the network to which it is connected; when it moves into another network, it needs to break the connection with the old network and establish a connection with the new network. Figure 1 shows an MH connected to Wireless Network

1; when it moves, it has to make a connection with the new network, say Wireless Network 2. The re-establishment of a new connection takes a considerable amount of time, resulting in the possibility of interruption and resulting loss of multimedia.

The current TCP/IP network infrastructure was not designed for mobility. It does not support handoff between IP networks. For example, a device running a real-time application, such as video conference, cannot play smoothly when the user hands off from one wireless IP network to another, resulting in unsatisfactory performance to the user.

Mobile IP (MIP) (Perkins, 1996), from the Internet Engineering Task Force (IETF), addresses

Figure 1. Illustration of handoff with mobile node connected to Wireless Network 1



the mobility problem. MIP extends the existing IP protocol to support host mobility, including handoff, by introducing two network entities: home agent (HA) and foreign agent (FA). The HA and FA work together to achieve host mobility. The correspondent node (CN) always communicates with the mobile node (MN) via its home network address, even though MH may not dwell in the home network. For CN to have seamless access to MN, the MH has to be able to handoff in a timely manner between networks.

Handoff latency is one of the most important indicators of handoff performance. Large handoff latency degrades performance of real-time applications. For example, large handoff latency will introduce interruption in a video conference due to breaks in both audio and video data transmission. In addition to high handoff latency, MIP suffers from a number of other problems including triangle routing, high signaling traffic with the HA, and so forth. A number of approaches to reduce the MIP handoff latency are given next.

Mobile IP uses only one IP; a certain amount of latency in data transmission appears to be unavoidable when the MH performs a handoff. This is because of MN's inability to communicate with the CN through either the old path (because it has changed its wireless link to a new wireless network) or the new path (because HA has not yet granted its registration request). Thus, MH cannot send or receive data to or from the CN while the MH is performing registration, resulting in interruption of data communication during this time interval. This interruption is unacceptable in a real-world scenario, and may hinder the widespread deployment of real-time multimedia applications on wireless mobile networks. Seamless IP-diversity based generalized mobility architecture (SIGMA) overcomes the issue of discontinuity by exploiting multi-homing (Stewart, 2005) to keep the old data path alive until the new data path is ready to take over the data transfer, thus achieving lower latency and lower loss during handoff between adjacent subnets than Mobile IP.

The *objective* of this chapter is to demonstrate the effectiveness of SIGMA in reducing handoff latency, packet loss, and so forth, for multimedia transmission, and compare with that achieved by Mobile IP. The *contribution* of this chapter is to describe the implementation of a real-time streaming server and client in SIGMA to achieve seamless multimedia streaming during handoff. SIGMA *differs* from previous work in the sense that all previous attempts modified the hardware, infrastructure of the network, server, or client to achieve seamless multimedia transmission during handoff.

The rest of this chapter is organized as follows. Previous work on multimedia over wireless networks is described in the next section. The architecture of SIGMA is described in the third section, followed by the testbed on which video transmission has been tested for both MIP and SIGMA in the fourth section. Results of video over MIP and SIGMA and presented and compared in the fifth section, followed by conclusions in the last section.

BACKGROUND

A large amount of work has been carried out to improve the quality of multimedia over wireless networks. They can be categorized into two types:

- Studies related to improving multimedia (e.g., video or audio) over wireless networks. They do not consider the mobility of the MN, but attempt to provide a high quality multimedia transmission within the same wireless network for stationary servers and clients.
- Studies related to achieving seamless multimedia transmission during handoffs. They

consider mobility of the MH and try to provide a seamless and high quality multimedia transmission when the MH (client) moves from one network to another.

Although our interest in this chapter is seamless multimedia transmission during handoffs, we describe previous work on both categories in the following sections.

Multimedia over Wireless Networks

Ahmed, Mehaoua, and Buridant (2001) worked on improving the quality of MPEG-4 transmission on wireless using differentiated services (Diffserv). They investigated QoS provisioning between MPEG-4 video application and Diffserv networks. To achieve the best possible QoS, all the components involved in the transmission process must collaborate. For example, the server must use stream properties to describe the QoS requirement for each stream to the network. They propose a solution by distinguishing the video data into important video data and less important video data (such as complementary raw data). Packets which are marked as less important are dropped in the first case if there is any congestion, so that the receiver can regenerate the video with the received important information.

Budagavi and Gibson (2001) improved the performance of video over wireless channels by multiframe video coding. The multiframe coder uses the redundancy that exists across multiple frames in a typical video conferencing sequence so that additional compression can be achieved using their multiframe-block motion compensation (MF-BMC) approach. They modeled the error propagation using the Markov chain, and concluded that use of multiple frames in motion increases the robustness. Their proposed MF-BMC scheme has been shown to be more robust on wireless networks when compared to the base-level H.263 codec which uses single frame-block motion compensation (SF-BMC).

There are a number of studies, such as Stedman, Gharavi, Hanzo, and Steele (1993), Illgner and Lappe (1995), Khansari, Jalai, Dubois, and Mermelstein (1996), and Hanzo and Streit (1995), which concentrate on improving quality of multimedia over wireless networks. Since we are only interested in studies that focus on achieving seamless multimedia transmission during handoff, we do not go into details of studies related to multimedia over wireless networks. Interested readers can use the references given earlier in this paragraph.

Seamless Multimedia over Mobile Networks

Lee, Lee, and Kim (2004) achieved seamless MPEG-4 streaming over a wireless LAN using Mobile IP. They achieved this by implementing packet forwarding with buffering mechanisms in the foreign agent (FA) and performed pre-buffering adjustment in a streaming client. Insufficient pre-buffered data, which is not enough to overcome the discontinuity of data transmission during the handoff period, will result in disruption in playback. Moreover, too much of pre-buffered data wastes memory and delays the starting time of playback. Find the optimal pre-buffering time is, therefore, an important issue in this approach.

Patanapongpibul and Mapp (2003) enable the MH to select the best point of attachment by having all the reachable router advertisements (RA) in a RA cache. RA cache will have the entire router's link whose advertisements are heard by the mobile node. These RAs are arranged in the cache according to a certain priority. The priority is based on two criteria: (1) the link signal strength, that is, signal quality and SNR level, and (2) the

time since the RA entry was last updated. So the RAs with highest router priority are forwarded to the IP packet handler for processing. The disadvantage of this method includes extra memory for the RA cache.

Pan, Lee, Kim, and Suda (2004) insert four components in the transport layer of the video server and the client. These four components are: (1) a path management module, (2) a multipath distributor module at the sender, (3) a pair of rate control modules, and (4) a multipath collector module at the receiver. They achieve a seamless video by transferring the video over multiple paths to the destination during handoffs. The overhead of the proposed scheme is two-fold: reduction in transmission efficiency due to transmission of duplicated video packets and transmission of control packets associated with the proposed scheme, and processing of the proposed scheme at the sender and receiver.

Boukerche, Hong, and Jacob (2003) propose a two-phase handoff scheme to support synchronization of multimedia units (MMU) for wireless clients and distributed multimedia systems. This scheme is proposed for managing MMUs to deliver them to mobile hosts on time. The two-phase scheme consists of: setup handoff and end handoff. In the first phase, setup handoff procedure has two major tasks: updating new arrival BSs and maintaining the synchronization for newly arrived mobile hosts (MHs). If an MH can reach another BS, then MH reports "new BS arrived" to its primary BS. End handoff procedure deals with the ordering of MMUs and with the flow of MMUs for a new MH. Any base station can be a new primary base station. The algorithm notifies MHs, BSs, and servers, and then chooses the closest common node from the current primary base station and new base stations. This method suffers from the disadvantage of additional overhead of updating the base station (BS) with newly arrived BSs and ordering of MMUs.

SIGMA FOR SEAMLESS MULTIMEDIA IN MOBILE NETWORKS

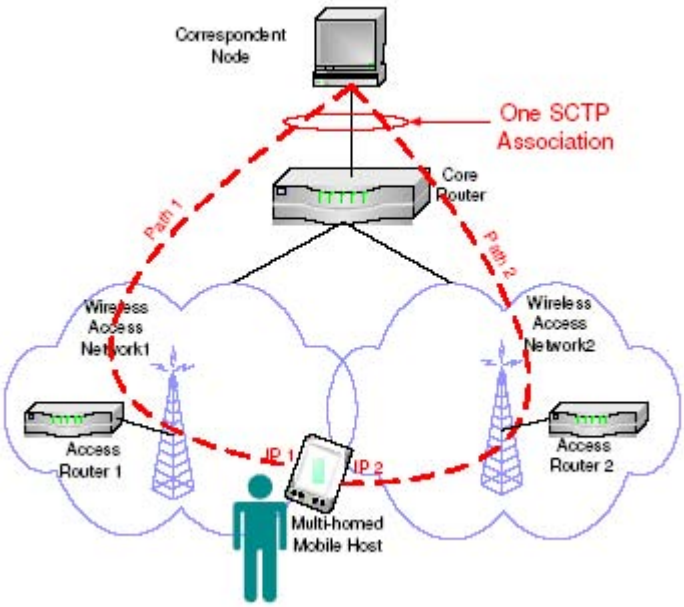
Limitations of previously proposed schemes in achieving seamless multimedia transmission during handoff in a wireless environment have been discussed in the previous section. In this section, we will discuss our proposed handoff scheme, called SIGMA, which has been designed for seamless multimedia transmission during handoffs, followed by its advantages over previous schemes.

Introduction to SIGMA

To aid the reader in getting a better understanding of SIGMA, in this section, we describe the various steps involved in a SIGMA handoff. A detailed description of SIGMA can be found in Fu, Ma, Atiquzzaman, and Lee (2005). We will use the stream control transmission protocol (Stewart, 2005), a new emerging transport layer protocol from IETF, to illustrate SIGMA.

Stream control transmission protocol's (SCTP) multi-homing (see Figure 2) allows an association between two endpoints to span across multiple IP addresses or network interface cards. One of the addresses is designated as the primary while the other can be used as a backup, in the case of failure of the primary address, or when the upper layer application explicitly requests the use of the backup. Retransmission of lost packets can also be done over the secondary address. The built-in support for multi-homed endpoints by SCTP is especially useful in environments that require high-availability of the applications, such as Signaling System 7 (SS7) transport. A multi-homed SCTP association can speedup recovery from link failure situations without interrupting any ongoing data transfer. Figure 2 presents an example of SCTP multi-homing where two nodes,

Figure 2. An SCTP association featuring multi-homing



CN and MH, are connected through two wireless networks, with MH being multi-homed. One of MN's IP addresses is assigned as the primary address for use by CN for transmitting data packets; the other IP address can be used as a backup in case of primary address failure.

STEP 1: Obtain New IP Address

Referring to Figure 2, the handoff preparation procedure begins when the MH moves into the overlapping radio coverage area of two adjacent subnets. Once the MH receives the router advertisement from the new access router (AR2), it should initiate the procedure of obtaining a new IP address (IP2 in Figure 2). This can be accomplished through several methods: DHCP, DHCPv6, or IPv6 Stateless Address Autoconfiguration (SAA) (Thomson & Narten, 1998). The main difference between these methods lies in whether the IP address is generated by a server (DHCP/DHCPv6) or by the MH itself (IPv6 SAA). For cases where the MH is not concerned about its IP address but only requires the address to be unique and routable, IPv6 SAA is a preferred method for SIGMA to obtain a new address since it significantly reduces the required signaling time.

STEP 2: Add IP Addresses to Association

When the SCTP association is initially setup, only the CN's IP address and the MH's first IP address (IP1) are exchanged between CN and

MH. After the MH obtains another IP address (IP2 in STEP 1), MH should bind IP2 into the association (in addition to IP1) and notify CN about the availability of the new IP address (Fu, Ma, Atiquzzaman, & Lee, 2005).

SCTP provides a graceful method to modify an existing association when the MH wishes to notify the CN that a new IP address will be added to the association and the old IP addresses will probably be taken out of the association. The IETF Transport Area Working Group (TSVWG) is working on the "SCTP Address Dynamic Reconfiguration" Internet draft (Stewart, 2005), which defines two new chunk types (ASCONF and ASCONF-ACK) and several parameter types (Add IP Address, Delete IP address, Set Primary Address, etc.). This option will be very useful in mobile environments for supporting service reconfiguration without interrupting on-going data transfers.

In SIGMA, MH notifies CN that IP2 is available for data transmission by sending an ASCONF chunk to CN. On receipt of this chunk, CN will add IP2 to its local control block for the association and reply to MH with an ASCONF-ACK chunk indicating the success of the IP addition. At this time, IP1 and IP2 are both ready for receiving data transmitted from CN to MH.

STEP 3: Redirect Data Packets to New IP Address

When MH moves further into the coverage area of wireless access network2, data path2 becomes increasingly more reliable than data path1. CN can then redirect data traffic to the new IP address (IP2) to increase the possibility of data being delivered successfully to the MH. This task can be accomplished by the MH sending an ASCONF chunk with the Set-Primary-Address parameter, which results in CN setting its primary destination address to MH as IP2.

STEP 4: Updating the Location Manager

SIGMA supports location management by employing a location manager that maintains a database which records the correspondence between MH's identity and current primary IP address (Reaz, Atiquzzaman, & Fu, 2005). MH can use any unique information as its identity, such as the home address (as in MIP), domain name, or a public key defined in the public key infrastructure (PKI).

Following our example, once the Set-Primary-Address action is completed successfully, MH should update the location manager's relevant entry with the new IP address (IP2). The purpose of this procedure is to ensure that after MH moves from the wireless access network1 into network2, further association setup requests can be routed to MH's new IP address IP2. This update has no impact on existing active associations.

We can observe an important difference between SIGMA and MIP: the location management and data traffic forwarding functions are coupled together in MIP, whereas they are *decoupled in SIGMA to speedup handoff and make the deployment more flexible*.

STEP 5: Delete or Deactivate Obsolete IP Address

When MH moves out of the coverage of wireless access network1, no *new* or *retransmitted* data packets should be directed to address IP1. In SIGMA, MH can notify CN that IP1 is out of service for data transmission by sending an ASCONF chunk to CN (Delete IP Address). Once received, CN will delete IP1 from its local association control block and reply to MH with an ASCONF-ACK chunk indicating the success of the IP deletion.

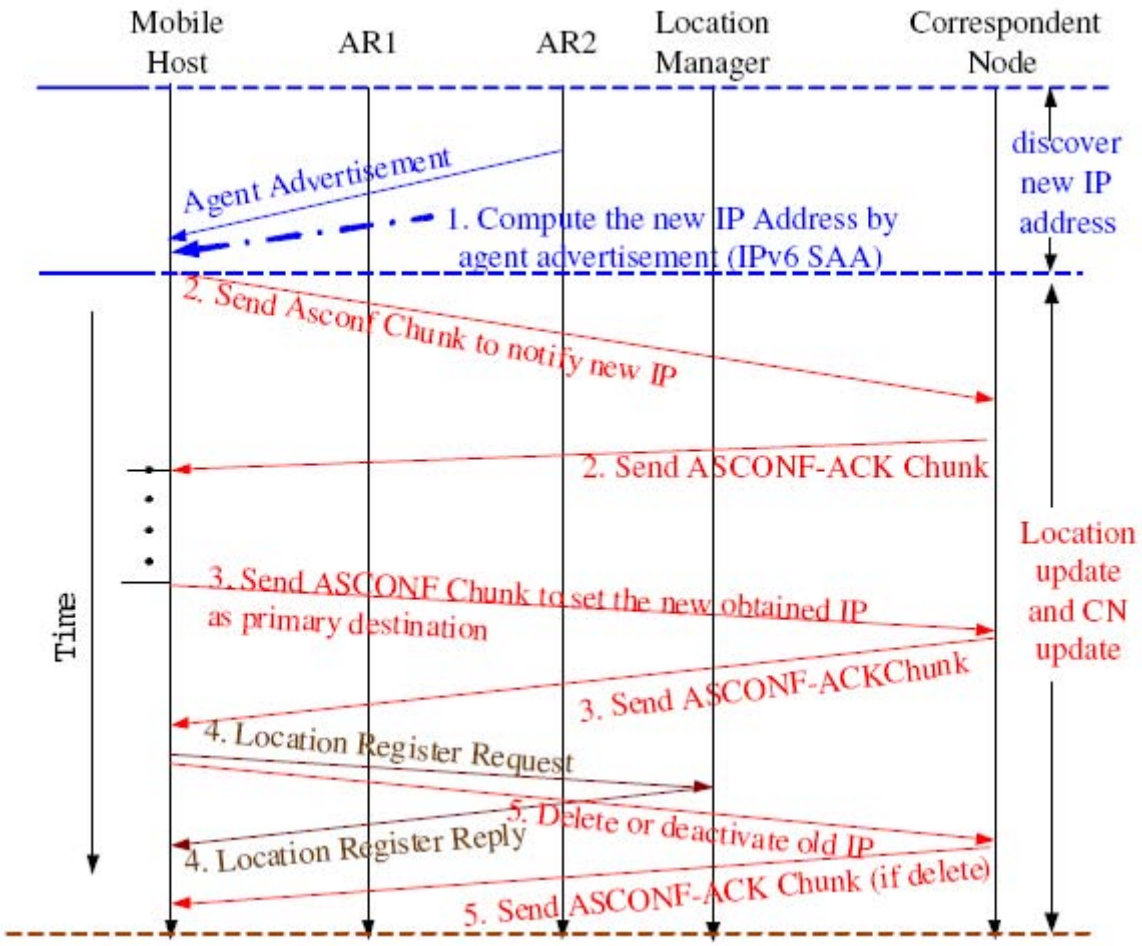
A less aggressive way to prevent CN from sending data to IP1 is for the MH to advertise a

zero receiver window (corresponding to IP1) to CN (Goff, Moronski, Phatak, & Gupta, 2000). This will give CN an impression that the interface (on which IP1 is bound) buffer is full and cannot receive any more data. By deactivating instead of deleting the IP address, SIGMA can adapt more gracefully to MH's zigzag (often referred to as ping pong) movement patterns and reuse the previously obtained IP address (IP1), as long as the lifetime of IP1 has not expired. This will reduce the latency and signaling traffic that would have otherwise been caused by obtaining a new IP address.

Timing Diagram of SIGMA

Figure 3 summarizes the signaling sequences involved in SIGMA. Here we assume IPv6 SAA and MH initiated Set-Primary-Address. Timing diagrams for other scenarios can be drawn similarly, but are not shown here because of space limitations. In this figure, the numbers before

Figure 3. Timeline of signaling in SIGMA



the events correspond to the step numbers in the previous sub-sections, respectively.

Advantages of SIGMA over the Previous Works

A number of previous work have considered seamless multimedia transmission during handoff, as mentioned in the second section, which have their own disadvantages. Here, we discuss the advantages of SIGMA over previous work. Lee et al. (2004) performed pre-buffering adjustment in client. Playback disruption may occur if the pre-buffered data is not enough to overcome the discontinuity of data transmission that occurs during handoff. Moreover, excessive pre-buffered data wastes memory usage and delays the starting time of playback. Find the optimal pre-buffering time is an important issue in this approach. Since SIGMA does not pre-buffer any data in the client, such optimization issues are not present in SIGMA.

Patanapongpibul et al. (2003) use the router advertisement (RA) cache. The disadvantage of this method is that it needs extra memory for RA cache; SIGMA does not involve any caching and hence does not suffer from such memory problems. Pan et al. (2004) use multipath (as discussed earlier), which suffers from (1) reduction in bandwidth efficiency due to transmission of duplicated video packets and transmission of control packets associated with the proposed scheme, and (2) processing overhead at the sender and receiver. Absence of multipaths or duplicate video packets in SIGMA results in higher link bandwidth efficiency.

Boukerche et al. (2003) proposed a two-phase handoff scheme which has additional overhead of updating the base station (BS) with newly arrived BSs, and also ordering of multimedia units (MMUs). In SIGMA, there is no feedback from MH to any of the base stations, and hence does not require ordering of multimedia units or packets.

EXPERIMENTAL TESTBED

Having reviewed the advantages of SIGMA over other schemes for multimedia transmission in the previous section, in this section, we present experimental results for SIGMA as obtained from an experimental setup we have developed at the University of Oklahoma. We compare the results of handoff performance during multimedia transmission over both SIGMA and Mobile IP. To make a fair comparison, we have used the same test bed for both MIP and SIGMA. Figure 4 (to be described later) shows the topology of our test bed, which has been used by a number of researchers—Seol, Kim, Yu, and Lee (2002), Wu, Banerjee, Basu, and Das (2003), Onoe, Atsumi, Sato, and Mizuno (2001)—for measurement of handoff performance. The difference in data communication between the CN and the MH for MIP and SIGMA lies in the lower layer sockets: the file sender for MIP is based on the regular TCP socket, while that for SIGMA is based on SCTP socket. We did not use the traditional *ftp* program for file transfer because it was not available for the SCTP protocol. To obtain access to the SCTP socket, we used Linux 2.6.2 kernel with Linux Kernel SCTP (LKSCTP) version 2.6.2-0.9.0 on both CN and MN. A number of MIP implementations, such as HUT Dynamics (HUT), Stanford Mosquito (MNET), and NUS Mobile IP (MIP), are publicly available. We chose HUT Dynamics for testing MIP in our test bed due to the following reasons: (1) Unlike Stanford Mosquito, which integrates the FA and MN, HUT Dynamics implements HA, FA, and MH daemons separately. This architecture is similar to SIGMA where the two access points and MH are separate entities. (2) HUT Dynamics implements hierarchical FAs, which will allow future comparison between SIGMA and hierarchical Mobile IP. Our MIP testbed consists four nodes: correspondent node (CN), foreign agent (FA), home agent (HA), and mobile node (MN). All the nodes run corresponding agents developed by HUT Dynamics.

The hardware and software configuration of the nodes are given in Table 1.

The CN and the machines running the HA and FA are connected to the Computer Science (CS) network of the University of Oklahoma, while the MH and access points are connected to two separate private networks. The various IP addresses are shown in Table 2. IEEE 802.11b is used to connect the MH to the access points.

The network topology of SIGMA is similar to the one of Mobile IP except that there is no HA or FA in SIGMA. As shown in Figure 4, the machines which run the HA and FA in the case of MIP act as gateways in the case of SIGMA. Table 1 shows the hardware and software configuration for the SIGMA experiment. The various IP addresses are shown in Table 2. The experimental procedure of Mobile IP and SIGMA is given next:

1. Start with the MH in Domain 1.
2. **For Mobile IP:** Run HUT Dynamics daemons for HA, FA, and MN. **For SIGMA:** Run the SIGMA handoff program, which has two functions: (1) monitoring the link

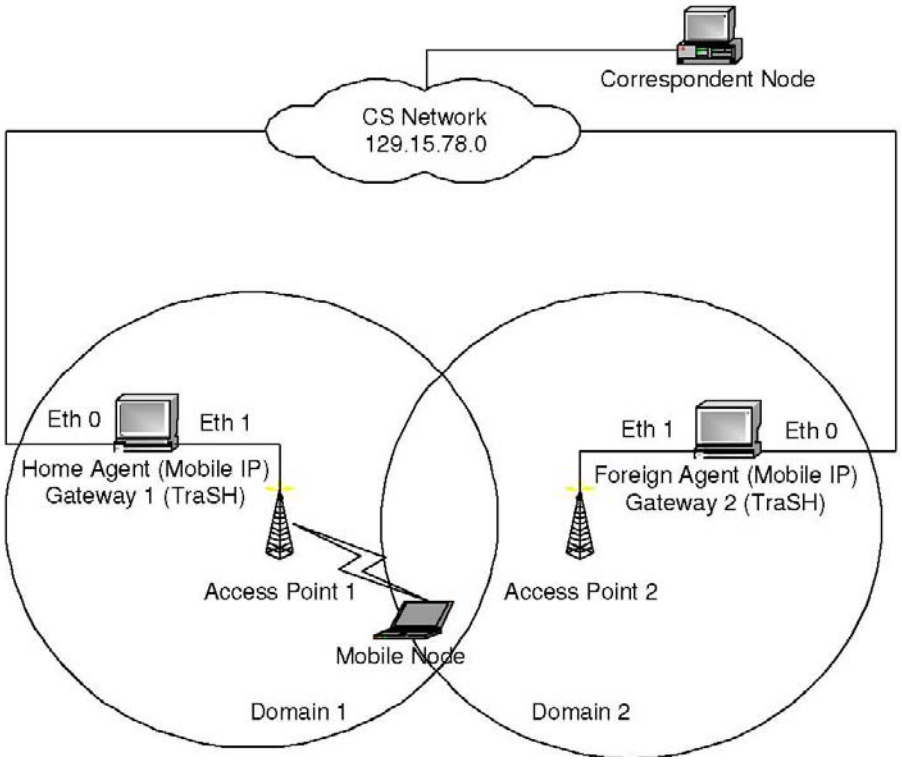
Table I. Mobile IP and SIGMA testbed configurations

Node	**Hardware**	**Software**	**Operating System**
Home Agent(MIP) Gateway1 (SIGMA)	Desktop, two NICs	HUT Dynamics 0.8.1 Home Agent Daemon (MIP)	Redhat Linux 9 kernel 2.4.20
Foreign Agent (MIP) Gateway2 (SIGMA)	Desktop, two NICs	HUT Dynamics 0.8.1 Foreign Agent Daemon (MIP)	Redhat Linux 9 kernel 2.4.20
Mobile Node	Dell Inspiron- 1100 Laptop, one Avaya 802.11b wireless card	HUT Dynamics 0.8.1 Mobile Node Daemon (MIP), File receiver	Redhat Linux 9 kernel 2.4.20
Correspondent Node	Desktop, one NIC	File sender	Redhat Linux 9 2.6.20

Table 2. Mobile IP and SIGMA network configurations

Node	**Network Configuration**
Home Agent (MIP) Gateway1 (SIGMA)	eth0: 129.15.78.171, gateway 129.15.78.172; eth1:10.1.8.1
Foreign Agent (MIP) Gateway2 (SIGMA)	eth0: 129.15.78.172 gateway 129.15.78.171; eth1: 10.1.6.1
Mobile Node	Mobile IP's Home Address: 10.1.8.5 SIGMA's IP1: 10.1.8.100 SIGMA's IP2 : 10.1.6.100
Correspondent Node	129.15.78.150

Figure 4. SIGMA and Mobile IP testbed



layer signal strength to determine the time to handoff, and (2) carrying out the signaling shown in Figure 4.

3. Run file sender/video server and file receiver/video client (using TCP sockets for Mobile IP, using SCTP sockets for SIGMA) on CN and MN, respectively.
4. Run Ethereal (ETHEREAL) on the CN and MH to capture packets.
5. Move MH from Domain 1 to Domain 2 to perform handoff by Mobile IP and SIGMA. Capture all packets sent from CN and received at MN.

RESULTS

Various results were collected on the experimental setup and procedure described earlier. In this section, we present two kinds of results: file transfer and multimedia transmission. The reason for showing the results of file transfer is to prove that SIGMA achieves seamless handoff not only for multimedia but also for file transfers.

Results for File Transfer

In this section, we present and compare the results of handoffs using MIP and SIGMA for file

transfer. For comparison, we use throughput, RTT, and handoff latency as the performance measures. *Throughput* is measured by the rate at which packets are received at the MN. *RTT* is the time required for a data packet to travel from the source to the destination and back. We define *handoff latency* as the time interval between the MH receiving the last packet from Domain 1 (previous network) and the first packet from Domain 2 (the new network). The experimental results are described next.

Results from Mobile IP Handoff

Figure 5 shows the throughput during Mobile IP handoff between Domain 1 and Domain 2. The variations in throughput within HA (from 20 second to 30 second) and within FA (from 37 second to 60 second) are due to network congestion arising from cross traffic in the production CS network.

The average throughput before, during and after handoff are 2.436 Mbps, 0 Mbps and 2.390 Mbps, respectively. Figure 6 shows the packet trace during MIP handoff. The actual handoff latency for MIP can be clearly calculated by having a zoomed-in view of the packet trace graph. Figure 7 shows a zoomed-in view of the packet trace, where the calculated handoff latency is eight seconds for Mobile IP. Figure 8 shows the RTT for the MIP handoff. As we can see, the RTT is high for eight seconds (the handoff latency time), during the handoff.

The registration time (or registration latency) is also a part of the handoff latency. Registration latency, the time taken by the MH to register with the agent (HA or FA), is calculated as follows. Ethereal capture showed that the MH sent a

Figure 5. Throughput during MIP handoff

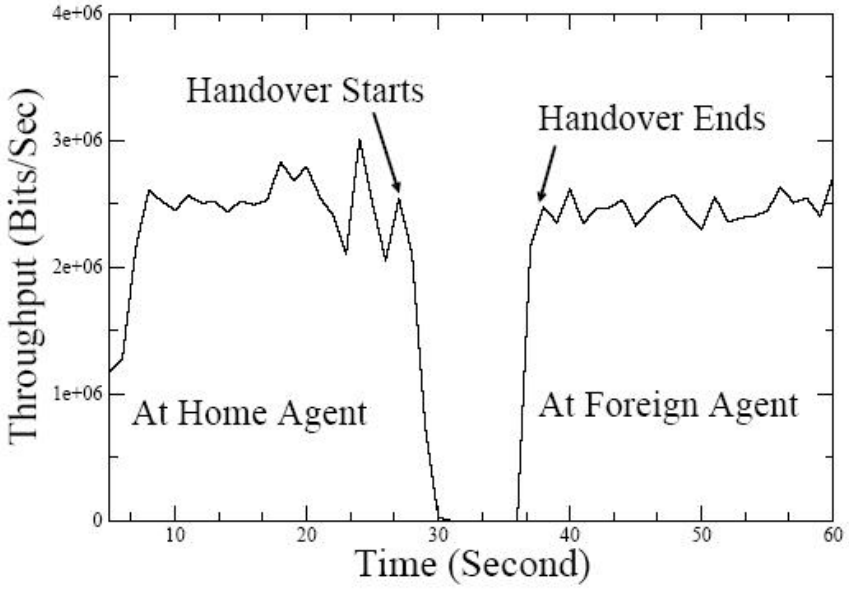


Figure 6. Packet trace during MIP handoff

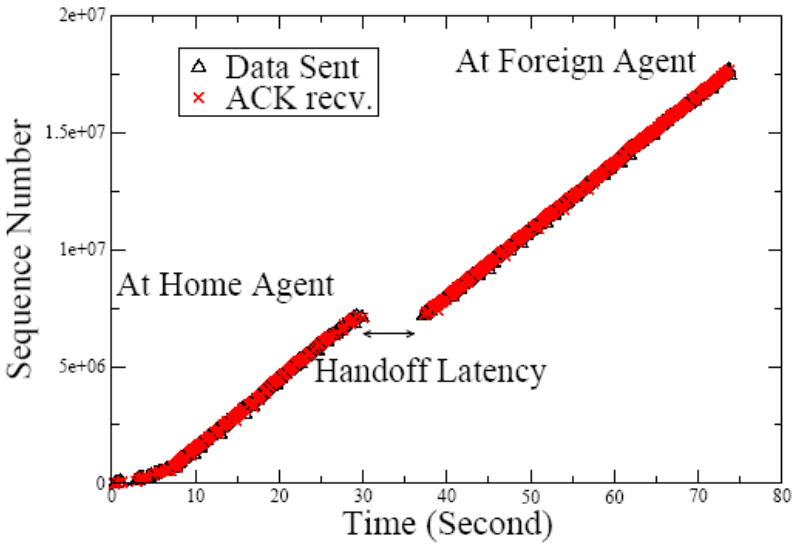


Figure 7. Zoomed in view during MIP handoff instant

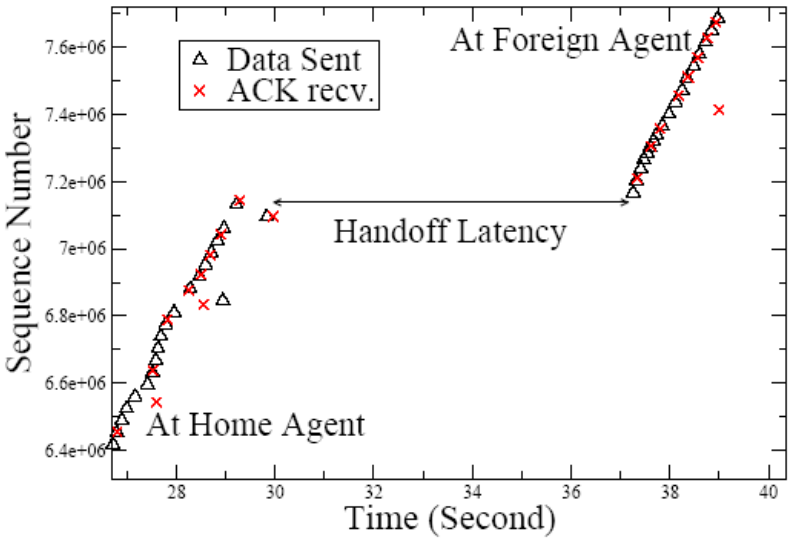
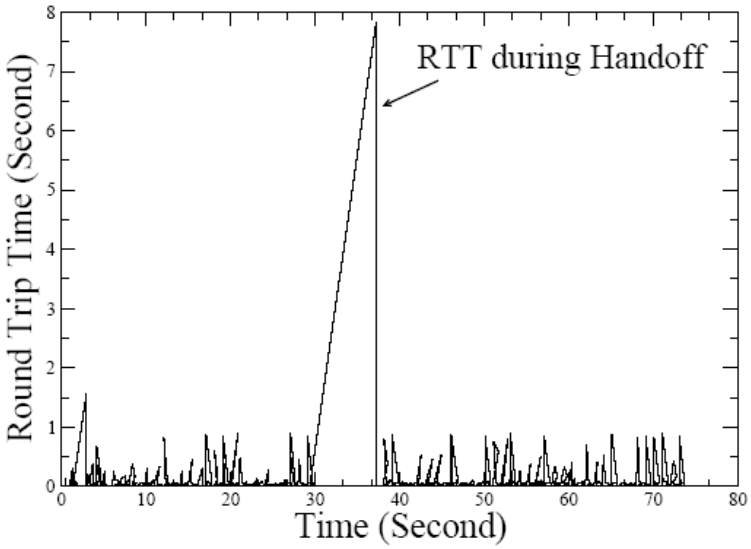


Figure 8. RTT during MIP handoff



registration request to the HA at time t = 14.5123 second and received a reply from the HA at t = 14.5180 second. Hence, the calculated registration time for registering with HA is 5.7 milliseconds. Similarly, during MIP handoff, Ethereal capture showed that the MH sent a registration request to FA at time t =7.1190 second and received a reply from the FA at t =7.2374, resulting in a registration time of 38.3 milliseconds. This is due to the fact that after the MH registers with the HA, it can directly register with the HA. On the other hand, if it registers with the FA, the MH registers each new care-of-address with its HA possibly through FA. The registration latency is, therefore, higher when the MH is in the FA.

Results from SIGMA Handoff

Figure 9 shows the throughput during SIGMA handoff where it can observed that the throughput does not go to zero. The variation in throughput is due to network congestion arising from cross traffic in the production CS network. Although we cannot see the handoff due to it being very small, it should be emphasized that the ethereal capture showed the handoff starting and ending at t = 60.755 and t = 60.761 seconds, respectively, that is, a handoff latency of six milliseconds.

Figure 10 shows the packet trace during SIGMA handoff. It can be seen that packets arrive at the MH without any gap or disruption; this is also a powerful proof of SIGMA's smoother handoff as compared to handoff in Mobile IP. This experimentally demonstrates that *a seamless handoff can be realized with SIGMA.* Figure 11 shows a zoomed-in view of the packet trace during the SIGMA handoff period; a handoff latency of six milliseconds can be seen between the packets arriving at the old and new paths.

Figure 9. Throughput during SIGMA handoff

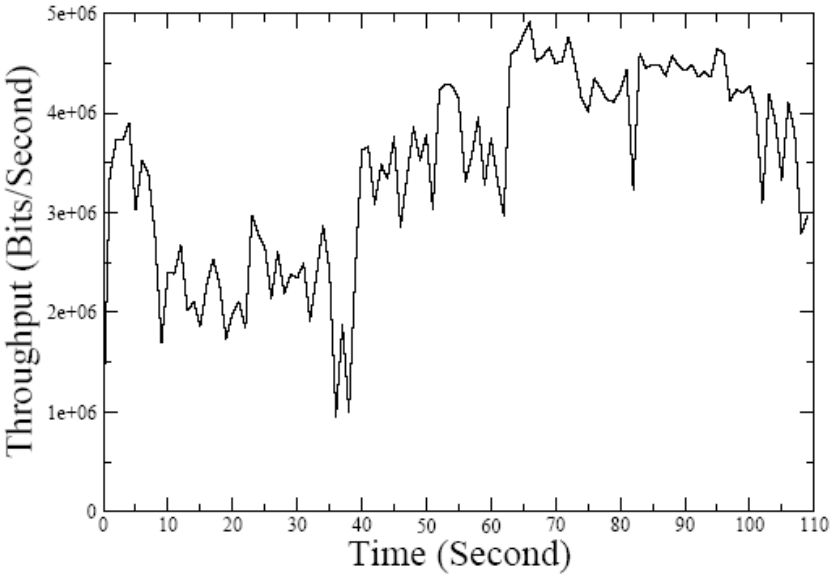


Figure 10. Packet trace during SIGMA handoff

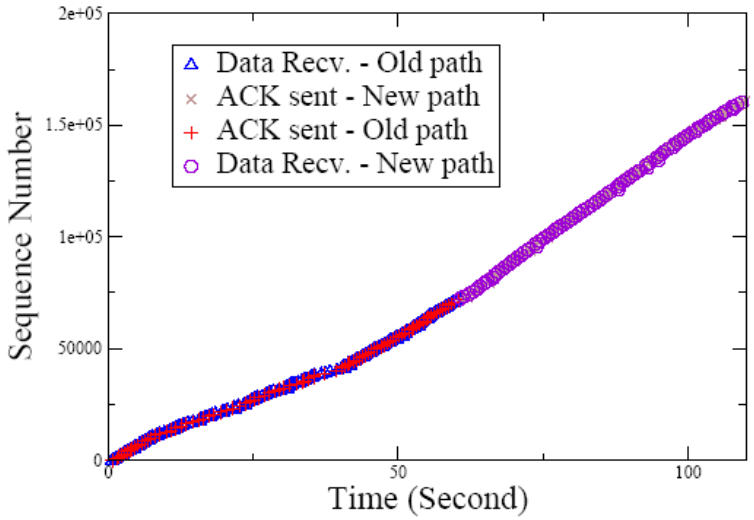


Figure 11. Zoomed in view during SIGMA handoff

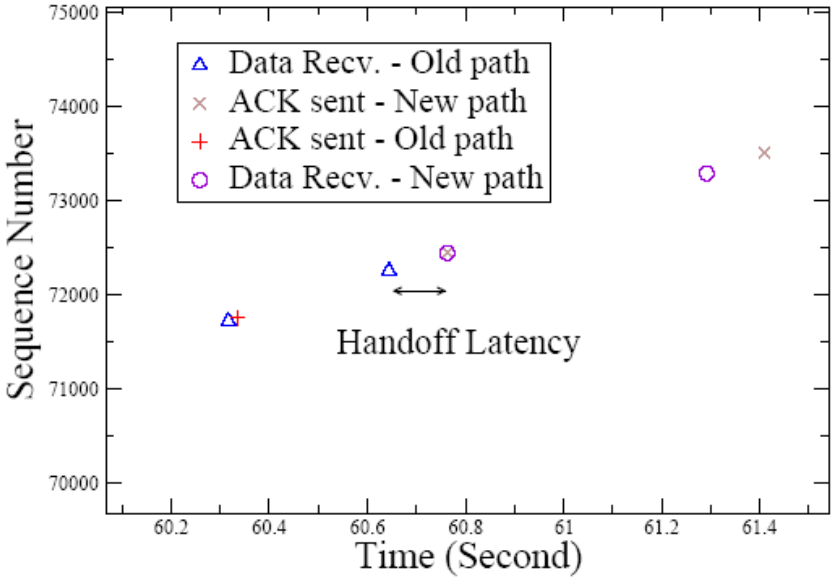


Figure 12. RTT during SIGMA handoff

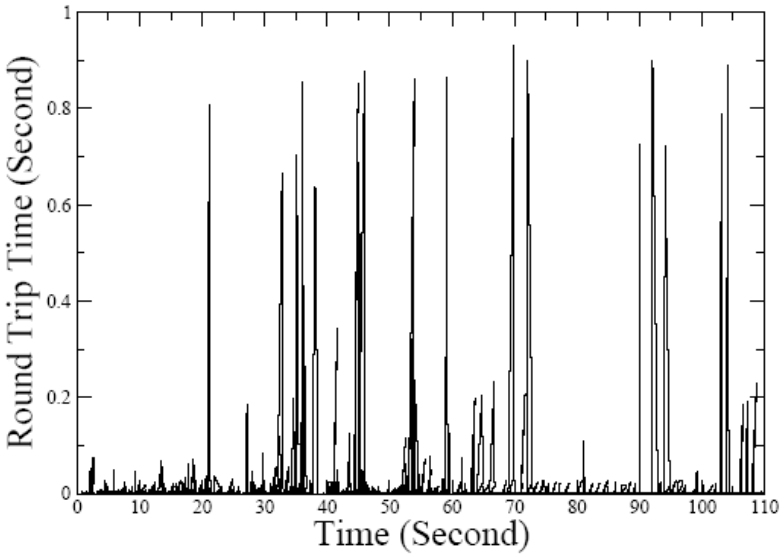


Figure 12 shows the RTT during SIGMA handoff. A seamless handoff is evident from the absence of any sudden RTT increase during handoff.

Result of Multimedia Data Transfer

To test the handoff performance for multimedia over SIGMA, we used a streaming video

Figure 13. Throughput of video during SIGMA handoff

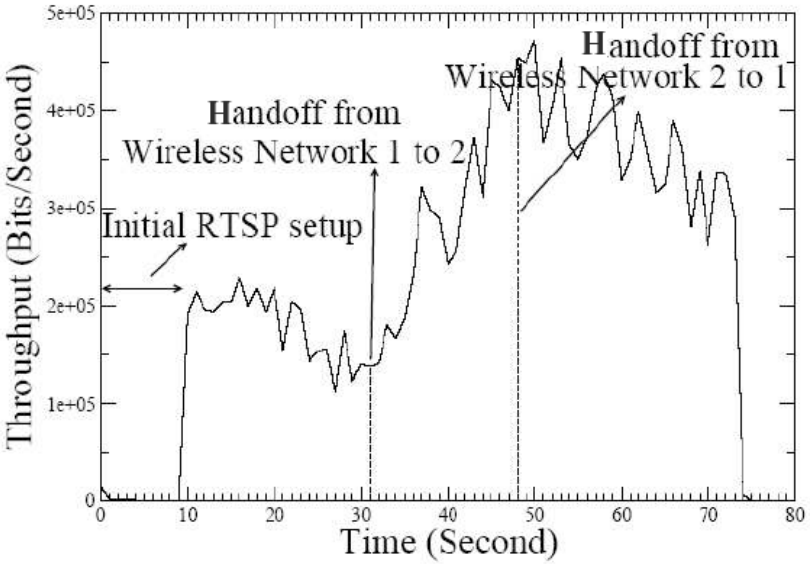
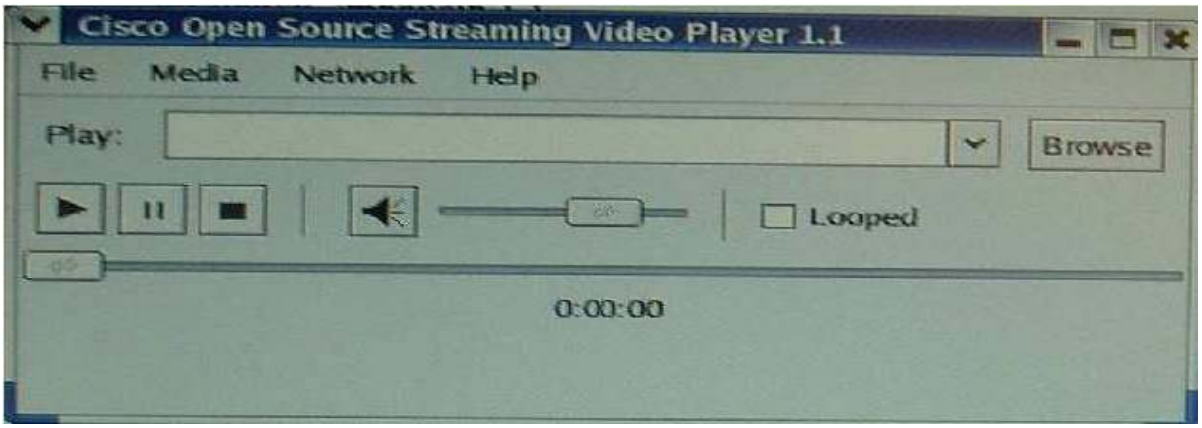


Figure 14. Screen shot of MPEG4-IP player



client and a streaming server at the MH and CN, respectively (details in the fourth section). Apple's Darwin Streaming Server (DARWIN) and CISCO's MPEG4IP player (MPEG) were modified to stream data over SCTP. A seamless handoff, with no interruption in the video stream, was achieved with SIGMA.

Figure 13 shows the throughput of multimedia (video) data, when the MH moves between subnets. The connection request and setup between the client and server is carried out during the first 10 seconds. It can be seen that the throughput does not drop during handoff at time = 31 second when MH moves from wireless network 1 to 2. A second handoff takes place when the MH moves from network 2 to network 1 at time = 48. It is seen that seamless handoff is achieved by SIGMA for both the handoffs.

Figure 14 shows a screen capture of the MPEG4IP player used in our experiment. Figure 15 shows the video playing in the player during handoff, where "rtsp://129.15.78.139/fta.sdp" represents the server's IP address and the streaming format (SDP).

Comparison of SIGMA and MIP Handoffs

We observed previously that the registration time of MIP was only 0.1 second, and the handoff latencies of MIP and SIGMA were eight seconds and six milliseconds, respectively. We describe the reasons for the MIP handoff latency being much longer than its registration time in the following:

Figure 15. Screen-shot of MPEG4-IP player playing streaming video

1. In HUT Dynamics, the MIP implementation used in this study, the MH obtains a registration lifetime after every successful registration. It originates another registration on expiry of this lifetime. So it is possible for the MH to postpone registration even after it has completed a link layer handoff and received FA advertisements. This may introduce some delay which can be up to the duration of a life time.
2. As mentioned in the previous section, the registration of MH also costs some time, measured as 38.3 milliseconds in our testbed.

The handoff latency in MIP comes from three factors: (1) remaining home registration lifetime after link layer handoff which can be from zero to a lifetime, (2) FA advertisement interval plus the time span of last time advertisement which is not listened by MN, and (3) registration latency. During these three times, the CN cannot communicate through either the previous path because it has completed link layer handoff, or the new path because MH has not yet completed the registration. As a result, the throughput was zero during this time. Obviously, such shortcoming has been eliminated in SIGMA through multi-homing and decoupling of registration and data transfer. Consequently, data continue to flow between the CN and MH during the handoff process.

CONCLUSION AND FUTURE TRENDS

We have shown that SIGMA achieves seamless multimedia transmission during handoff between wireless networks. As future work, video streaming can be tested over SIGMA during vertical handoffs, that is, between wireless LANs, cellular, and satellite networks.

ACKNOWLEDGMENT

The work reported in this chapter was funded by National Aeronautics and Space Administration (NASA) grant no. NAG3-2922.

REFERENCES

Ahmed, T., Mehaoua, A., & Buridant, G. (2001). Implementing MPEG-4 video on demand over IP differentiated services. *Global Telecommunications Conference, GLOBECOM*, San Antonio, TX, November 25-29 (pp. 2489-2493). Piscataway, NJ: IEEE.

Boukerche, A., Hong, S., & Jacob, T., (2003). A two-phase handoff management scheme for synchronizing multimedia units over wireless networks. *Proc. Eighth IEEE International Symposium on Computers and Communication*, Antalya, Turkey, June-July (pp. 1078-1084). Los Alamitos, CA: IEEE Computer Society.

Budagavi, M., & Gibson, J. D. (2001, February). Multiframe video coding for improved performance over wireless channels. *IEEE Transactions on Image Processing, 10*(2), 252-265.

DARWIN. Retrieved June 23, 2005, from http://developer.apple.com/darwin/projects/streaming/

ETHEREAL. Retrieved June 30, 2005, from www.ethereal.com

Fu, S., Atiquzzaman, M., Ma, L., & Lee, Y. (2005, November). Signaling cost and performance of SIGMA: A seamless handover scheme for data networks. *Journal of Wireless Communications and Mobile Computing, 5*(7), 825-845.

Fu, S., Ma, L., Atiquzzaman, M., & Lee, Y. (2005). Architecture and performance of SIGMA: A seamless mobility architecture for data networks. *40th IEEE International Conference on Com-*

munications (ICC), Seoul, Korea, May 16-20 (pp. 3249-3253). Institute of Electrical and Electronics Engineers Inc.

Goff, T., Moronski, J., Phatak, D. S., & Gupta, V. (2000). Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments. *IEEE INFOCOM*, Tel Aviv, Israel, March 26-30 (pp. 1537-1545). NY: IEEE.

Hanzo, L., & Streit, J. (1995, August). Adaptive low-rate wireless videophone schemes. *IEEE Trans. Circuits Syst. Video Technol., 5*(4), 305-318.

HUT. Retrieved June 1, 2005, from http://www.cs.hut.fi/research/dynamics/

Illgner, R., & Lappe, D. (1995). Mobile multimedia communications in a universal telecommunications network. *Proc. SPIE Conf. Visual Communication Image Processing*, Taipei, Taiwan, May 23-26 (pp. 1034-1043). USA: SPIE.

Khansari, M., Jalai, A., Dubois, E., & Mermelstein, P. (1996, February). Low bit-rate video transmission over fading channels for wireless microcellular system. *IEEE Trans. Circuits Syst. Video Technol., 6*(1), 1-11.

Lee, C. H., Lee, D., & Kim, J. W. (2004). Seamless MPEG-4 video streaming over Mobile-IP enabled wireless LAN. *Proceedings of SPIE, Multimedia Systems and Applications*, Philadelphia, Pennsylvania, October (pp. 111-119). USA: SPIE.

LKSCTP. Retrieved June 1, 2005, from http://lksctp.sourceforge.net

MIP. Retrieved June 1, 2005, from opensource.nus.edu.sg/projects/mobileip/mip.html

MNET. Retrieved June 1, 2005, from http://mosquitonet.stanford.edu/

MPEG. Retrieved June 1, 2005, from http://mpeg4ip.sourceforge.net/faq/index.php

Onoe, Y., Atsumi, Y., Sato, F., & Mizuno, T. (2001). A dynamic delayed ack control scheme on Mobile IP networks. *International Conference on Computer Networks and Mobile Computing*, Los Alamitos, CA, October 16-19 (pp. 35-40). Los Alamitos, CA: IEEE Computer Society.

Pan, Y., Lee, M., Kim, J. B., & Suda, T. (2004, May). An end-to-end multipath smooth handoff scheme for streaming media. *IEEE Journal on Selected Areas in Communications, 22*(4), 653-663.

Patanapongpibul, L., & Mapp, G. (2003). A client-based handoff mechanism for Mobile IPv6 wireless networks. *Proc. Eighth IEEE International Symposium on Computers and Communications*, Antalya, Turkey, June-July (pp. 563-568). Los Alamitos, CA: IEEE Computer Society.

Perkins, C. (1996). IP mobility support. *IETF RFC 2002*, October.

Reaz, A. S., Atiquzzaman, M., & Fu, S. (2005). Performance of DNS as location manager. *IEEE Globecom*, St. Louis, MO, November 28-December 2 (pp. 359-363). USA: IEEE Computer Society.

Seol, S., Kim, M., Yu, C., & Lee., J. H. (2002). Experiments and analysis of voice over MobileIP. *13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Lisboa, Portugal, September 15-18 (pp. 977-981). Piscataway, NJ: IEEE.

Stedman, R., Gharavi, H., Hanzo, L., & Steele, R. (1993, February). Transmission of subband-coded images via mobile channels. *IEEE Trans. Circuit Syst. Video Technol., 3*, 15-27.

Stewart, R. (2005, June). *Stream control transmission protocol (SCTP) dynamic address configuration.* IETF DRAFT, draft-ietf-tsvwgaddip-sctp-12.txt.

Thomson, S., & Narten, T. (1998, December). *IPv6 stateless address autoconfiguration.* IETF RFC 2462.

Wu, W., Banerjee, N., Basu, K., & Das, S. K. (2003). Network assisted IP mobility support in wireless LANs. *Second IEEE International Symposium on Network Computing and Applications, NCA'03*, Cambridge, MA, April 16-18 (pp. 257-264). Los Alamitos, CA: IEEE Computer Society.

This work was previously published in Mobile Multimedia Communications: Concepts, Applications, and Challenges, edited by G. Karmakar and L. Dooley, pp. 24-44, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.38
High Performance Scheduling Mechanism for Mobile Computing Based on Self-Ranking Algorithm

Hesham A. Ali
Mansoura University, Egypt

Tamer Ahmed Farrag
Mansoura University, Egypt

ABSTRACT

Due to the rapidly increasing number of mobile devices connected to the Internet, a lot of research is being conducted to maximize the benefit of such integration. The main objective of this article is to enhance the performance of the scheduling mechanism of the mobile computing environment by distributing some of the responsibilities of the access point among the available attached mobile devices. To this aim, we investigate a scheduling mechanism framework that comprises an algorithm that provides the mobile device with the authority to evaluate itself as a resource. The proposed mechanism is based on the "self ranking algorithm" (SRA), which provides a lifetime opportunity to reach a proper solution. This mechanism depends on an event-based programming approach to start its execution in a pervasive computing environment. Using such a mechanism will simplify the scheduling process by grouping mobile devices according to their self-ranking value and assigning tasks to these groups. Moreover, it will maximize the benefit of the mobile devices incorporated with the already existing Grid systems by using their computational power as a subordinate value to the overall power of the system. Furthermore, we evaluate the performance of the investigated algorithm extensively, to show how it overcomes the connection stability problem of the mobile devices. Experimental results emphasized that

the proposed SRA has a great impact in reducing the total error and link utilization compared with the traditional mechanism.

INTRODUCTION

Mobile computing and commerce are spreading rapidly, replacing or supplementing wired computing. Moreover, the wireless infrastructure upon which mobile computing is built may reshape the entire information technology (IT) field. Therefore, it is fair to say that nowadays, mobile devices have a remarkable high profile in the most common communication devices. Individuals and organizations around the world are deeply interested in using wireless communication, because of its flexibility and its unexpected and fast development. The first solution to the need for mobile computing was to make computers small enough so they could be easily carried. First, the laptop computer was invented; later, smaller and smaller computers, such as 3G, personal digital assistants (PDAs) and other handhelds, appeared. Portable computers, from laptops to PDAs and others, are called mobile devices. In recent years, a great development took place on the Internet and with mobile technologies. Consequently, the next step will be merging these two technologies, leading to the Wireless Internet. The Wireless Internet will be much more than just Internet access from mobile devices; the Wireless Internet will be almost invisible, as people will use mobile services and applications directly. On the other hand, these services and applications will be acting as our agents, conducting searches and communicating with other services and applications to satisfy our needs. Not only will the integration of mobile technology and the Internet paradigm reinforce the development of the new context-aware applications, but it also will sustain traditional features, such as user preferences, device characteristics, properties of connectivity and the state of service and usage history. Furthermore, the context includes features strictly related to user mobility, such as a user's current geospatial location (time and/or space). As direct use of existing Internet applications in a mobile environment is usually unsatisfactory, services and applications need to take into account the specific characteristics of mobile environments. The next section will provide an overview of mobile devices as well as the present relation model between mobile devices and the Grid.

Mobile Devices' Development

The number of individuals and organizations relying on wireless devices is continually increasing. Table 1 represents a statistical study of current and future increase in the sales of wireless equipment and the considerable growth in the sales of mobile phones.

Table 1 shows the rapid growth in sales rates of wireless equipment, and they serve the purpose of being a good metric of the flourishing future of mobile computing. From 2001 to 2005, investments on mobile devices are expected to increase by 41% and reach $31 billion. In 2004, the laptops on the market reached 39.7 million. On the other hand, not only did the number of mobile devices and wireless equipment increase, but also the computational power and memory storage. As a result, mobile computing and wireless Internet became a very important research area. This article will approach it from the computational Grid viewpoint.

Mobile Devices and the Computational Grid

The interaction between mobile devices and the computational Grid, such as depicted in Figure 1, can be classified into two models:

1. **Mobile as a user of Grid resources:** The development in the computational power of mobile devices, such as smart phones,

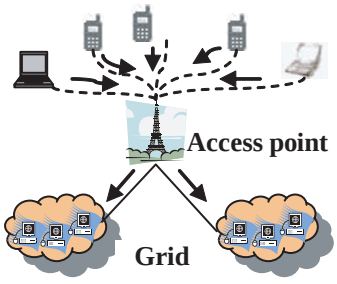
Table 1. Worldwide wireless LAN equipment shipments (1000s of units) (Navrati Saxena, 2005)

Product Segment	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
Adapters	6890	12599	21333	30764	41417	50415
Access Points	1437	1965	3157	3919	4851	5837
Broadband Gateways	552	850	1906	3365	5550	7941
Other WLAN Equipment	47	59	82	105	132	158
Total	8926	15473	26478	38153	51950	64351

PDAs, and so forth, will be limited due to its size, battery life, bandwidth and storage of data. However, when this integration occurs, all of the huge computational power and stored data of the Grid will be available to the mobile client. The mobile clients send their requests to the access point (AP), which can be considered as the Grid gateway; the scheduler is responsible for finding a suitable resource to perform the incoming request (Sanver, 2004).

2. **Mobile as a Grid resource:** When one mobile device is considered a resource, it will be a very inferior and low-ranking resource compared with a personal computer (PC). Meanwhile, because of the large number of mobile devices that can be used, it can be a worthwhile computational power. Also, because of its large geographical distribution, it can be considered a very excellent data collector, which can be used in many applications, such as geographical information systems, weather news, and so forth. Relatively, there are two approaches to integrate the mobile device into the existing Grid; *the first* is that all the information of every mobile device is recorded in the scheduler, so every device is considered to be one Grid resource. *The second approach* is one in which the information of the mobile devices is hidden from the scheduler; it considers all the devices connected to an access point as one Grid resource, and the access point responsible for scheduling tasks on the mobile devices is also connected to it.

Figure 1. An overview of integration of mobile devices with computational grid



This article introduces SRA, which will be used to build a mobile computing scheduling mechanism. Before introducing the proposed algorithm, an overview about related work in the scheduling mechanism in the Grid is given,

followed by a detailed description of the targeted problem at hand and a proposed framework. Moreover, the proposed SRA will be introduced, along with the simulation used to state the proposed algorithm and, finally, the results of that simulation will be analyzed.

RELATED WORK

Before elaborating on the problem, five of the most recent systems, especially on scheduling algorithms, are studied (He, 2003; Buyya, 2003; Somasundara, 2004; Berman & Casanova, 2005). Although the researchers have very different parameters and concepts, all of them have two main objectives. The first is to increase the utilization of the system, while the second is to find a suitable resource (as the economic cost, quality of services [QoS], deadline, etc.).

Table 2 shows a comparison between the most recent systems. Undoubtedly, one of the common problems that face any system when dealing with a large number of resources is "Load Balancing." Due to the fact that the ranking value of the resources is different, each of these systems

Table 2. Comparison between referenced systems

Project Features	Condor (Arun A. Somasundara, 2004)	Sphinx (Jang-uk In , 2004)	DBC (Rajkumar Buyya, 2003)	Disconnected operation service (Sang-Min Park, 2003)	QoS Guided Scheduling (Xiaoshan He ,2003)
Mobility	----	----	----	Job Proxy	----
Load Balancing	Backfilling	Resource usage Accounts and Users Quotas	Improved by considering Time deadline addition to Cost	----	QoS Guided improve but not direct solution
Long Beginning Time	Backfilling	Resource usage Accounts (Quotas)	Improved by considering Time deadline and Cost	----	----
Resource Ranking Parameters	By The User	Percent of resource usage account used	Cost	disconnection rate and the reconnection rate	Availability of required QoS
Multi Scheduler	supported	----	----	supported	----
Resource Reservation	Future work	supported	Future work	----	----
QoS support	----	supported	Future work	----	supported
Scheduling Constrains	FIFO, user priorities	user priorities	Budget, deadline	----	QoS (one dimensional)

endeavors to solve the problem, as illustrated in Table 2. Another problem is how the system will deal with the mobility of clients and resources. Noticeable is the limitation of research that takes into account the mobility of the resources (Nurmi, 2004; Park, 2003). The study of these five systems shows that they are based on different parameters to rank the resource, but the most popular are QoS and the economic cost (He, 2003; Buyya, 2003). The expressions used in Table 2 are explained here:

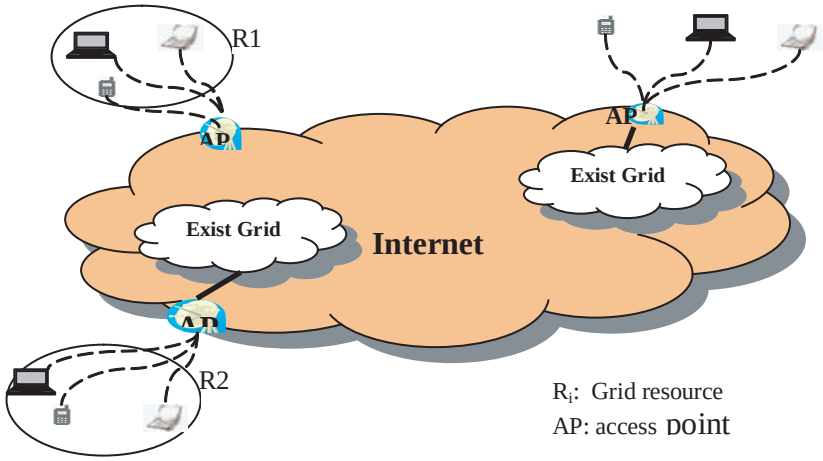
- **Backfilling:** A technique that tries to fill the gaps in the scheduling operation by executing the low-priority functions in the low-ranking resources that have not been used for a long time. This increases the system's overall utilization and makes a kind of load balancing between the resources (Somasundara, 2004).
- **Resource usage accounts (quotas):** Each resource must be assigned to certain functions according to its usage account. Preventing the resource from not being used can be caused by the presence of high QoS resources. This approach gives the scheduler force more functions to be assigned to a certain resource by maximizing its quota (In, 2004).
- **Job proxy:** Created when the mobile user submits a job, it is responsible for the interaction between the mobile device and the system. It can also simulate mobile action in case of mobile disconnecting. It does this until the mobile is connected again. If the mission is accomplished and the mobile is still disconnected, it stores the result for a certain time-out duration (Park, 2003).
- **QoS guided:** The QoS Guided scheduler has a kind of intelligence as not to consume the high QoS resource in performing the jobs that need low QoS. It does this to save its power to the other tasks that need this high QoS (He, 2003).

SCHEDULING AND THE CONNECTION STABILITY PROBLEM

The new approach in the computing area is *Internet computing.* It uses the already existing infrastructure of the Internet and builds its own Grid using devices interconnected to the Internet (Frontier, 2004). This is a very economical approach, because there is no need to build a special infrastructure. On the other hand, a lot of questions and issues raise, such as: "Do we need to build a new infrastructure of a grid to integrate the mobile devices as a grid user or as a grid resource?" and "What about the already existing grid projects?" (Gradwell, 2003; Dail, 2002; Frey, 2001; Berman, 2005). Figure 2 shows how the already existing infrastructure can be ordered and organized to create an infrastructure that helps to integrate the mobile devices with existing Grid systems like Condor, GriPhyN and Grid2003. This infrastructure aims at using huge computational power due to the large number of Internet users. It also aims at using the different services and resources available in the already existing Grid projects. Above all, the main objective is to use the Internet network to connect the mobile devices to the other parts of this infrastructure and to put all these services and computational power available to the mobile device. Finally, it aims at increasing the computational power and number of services of the system by integrating the large number of mobile devices distributed around the world (Saxena, 2005).

The most important problem that can face any Grid system is to develop a scheduling mechanism to manage such integration. The previous scheduling mechanisms depended on QoS (He, 2003), cost (Buyya, 2002; Barmouta, 2003) or a hybrid between other parameters (In, 2004; Takefusa, 2001) to select the best scheduling decision. Due to this integration and the mobility of the device, a new parameter appeared. This parameter represents the stability of the connection established between the devices and the access

Figure 2. The system infrastructure organization



point; in other words, the rate of disconnecting and the rate of reconnecting. All of the already existing systems make the scheduler monitor and evaluate the performance and availability of its attached resources. This was acceptable with PCs, but because of the huge number of mobile devices expected to attach to the scheduler, a very high overload on the scheduler can happen. So, the scheduler slows down more and more as the number of the attached resources increases.

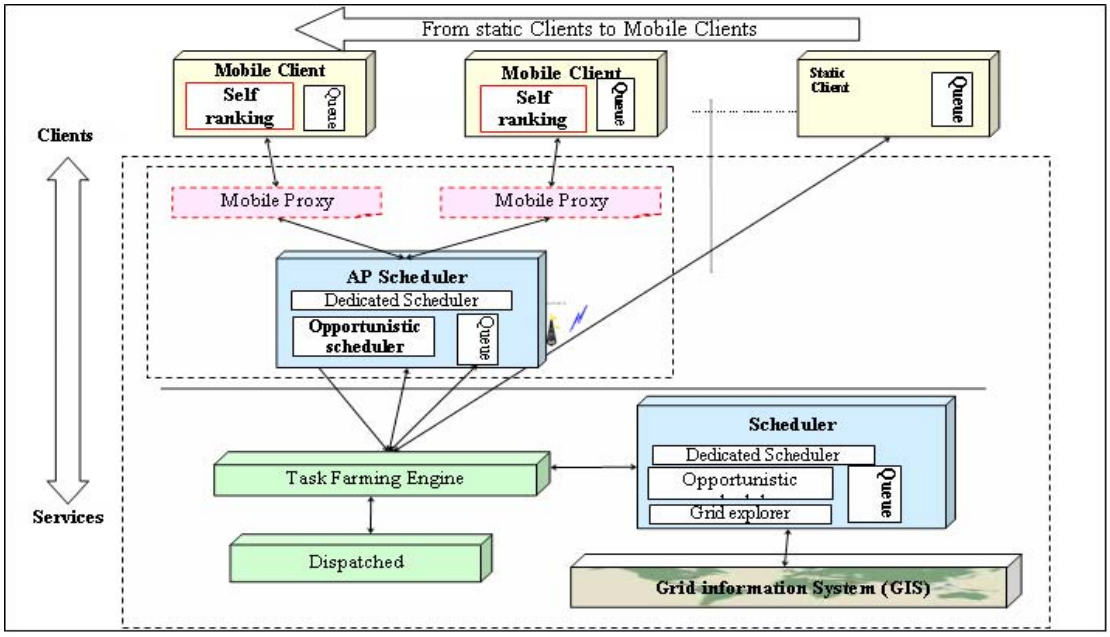
Plan of Solution

To overcome the overhead resulted from collecting the data at the access point scheduler and storing the historical ones of the mobile device performance, an SRA will be investigated. This algorithm has two key points: The *first* is to provide the mobile device with the authority to evaluate and rank itself and remove this task from the central point (scheduler). *Second* is considering the mobility of the resource as important metrics in such an environment. Therefore, the main aim of this algorithm is to calculate a ranking value for each attached mobile device that may be considered as a metric of the mobile performance. Moreover, it will be used to classify the mobile devices into groups to make the process of scheduling simpler and faster.

PROPOSED FRAMEWORK

Figure 3 depicts the framework and system components relationship for the given organization in Figure 2. The following design guidelines must be adhered to: (1) Use opportunistic schedulers introduced in the Condor (Somasundara, 2004), because it is an excellent idea to make a good load balance between high-ranking resources and low-ranking ones (e.g., mobile devices); (2) Use the mobile proxy introduced in Park (2003), but we changed its name from job proxy to our proposed name "mobile proxy" which will be the interface between the mobile client and the other components of the system; and (3) Use

Figure 3. Mobile device scheduling framework and components relationship



multi-schedulers because of the distribution of the considered infrastructure.

Proposed Framework Entities

In the following, the entities participating in the given framework are defined and their functions explained, as well as how they interact with each other.

- **The Task Farming Engine (TFE):** Responsible for partitioning the requested job into small tasks that will be assigned to resources to perform them using the scheduler and dispatcher.
- **The Scheduler**: Responsible for resource discovery, resource trading, resource selection and tasks assignment.
- **The Dispatcher:** Responsible for the actual assigning of tasks to the resources decided by the scheduler, monitoring execution of the tasks and controlling the process of collecting the different partitions of the job. Finally, it sends the overall result to the job requester.
- **Grid Information System (GIS):** Can be considered as the resources characteristics database used by the scheduler to find a suitable resource to perform the requested tasks using the resource QoS, cost, rank.
- **Dedicated Scheduler:** Each resource is assigned to one dedicated scheduler who has all rights to use the resource at any time except if the resource owner needs his resource. This monopoly may lead to non-functioning of some resources because

they are in the resources' list of certain Dedicated Schedulers, besides other high-ranked resources. So, these high-ranked resources will be preferred to the scheduler. This problem may be resolved by the temporary claiming of the resource to another type of scheduler named "opportunistic scheduler." This problem causes holes in the scheduling operation.

- **Opportunistic scheduler:** When the dedicated scheduler claims some of its resources because they were idle for a long time or they had a low-ranking value, which made them useless for a long time. The opportunistic scheduler tries to use this resource to execute some small tasks that may end before the dedicated scheduler needs the resource again. This operation is named "Backfilling." Note that this method will maximize utilization of the overall system.

If a mobile client is connected to an access point, the first step is to create a mobile proxy object, which will be considered as a simulation of the mobile device. So, it may store the hardware specification of the mobile and its current location, and it may also monitor the movement of the mobile from one access point to another. This mobile proxy information will be the base knowledge on which the scheduler builds its work.

Figure 4. Request processing flow

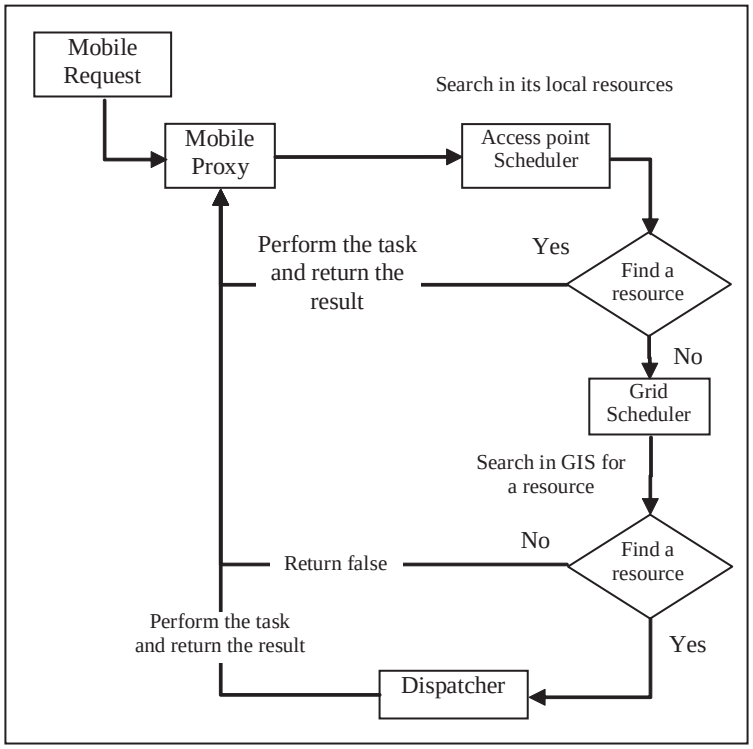


Figure 4 depicts a request-processing scenario. If the mobile client makes a request, this request will be stored in the mobile proxy. Then, it goes to a scheduler using the scheduling mechanism, trying to find a suitable resource to perform this request from its local connected resources. If the access point scheduler does not find a suitable resource, it forwards this request to a higher-level scheduler, which usually has static PCs with more computational power. This scheduler uses the GIS to find a suitable resource. When the resource is located, the dispatch assigns the requested task to this resource. When the task is performed, the outcome returns to the mobile proxy, which is responsible for sending the result to the mobile client in its current location.

PROPOSED SRA

The idea of the SRA is to reduce the dependability on the access point scheduler and distribute this overhead among the attached mobile devices. This can be done by making every mobile able to evaluate itself. Then, the access point can use this ranking value in the process of scheduling.

The trigger to start this algorithm execution depends on the event-based programming approach. The events that were taken into account are: (1) the event of disconnecting the mobile device and its scheduler, because this event means the end of the last connected period; (2) the event of reconnecting the mobile device to its scheduler, because this event means the end of the last disconnected period; and (3) the event of finishing a task, because this event changes the value of the mobile utilization. The self ranking value (R) has two parts: First is the Connectivity metric (M_{CD}), which can be considered as a metric of performance and connectivity of the mobile device, as well. The second part is the utilization metric (U), which can be considered as a metric of the success of the mobile device in performing the assigned task. When the mobile client has a new ranking value, this value must be sent to the mobile proxy to be entered as a parameter in the scheduling process.

Figure 5. Rank metric map

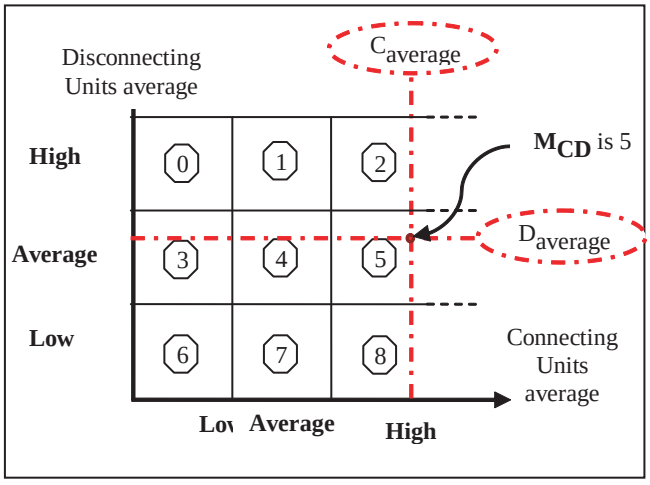


Figure 6. SRA for determining mobile device ranking value

0 Imports the prestored values :
t_s : the start of the last reconnection event. t_e: the end of the connection period .
T_u :Time unit defined by the scheduler N_c : number of reconnecting event occurring.
N_D: number of disconnecting event occurring.
$C_{average}$: The average number of time units to being connecting continually
$D_{average}$: the average number of time units being disconnecting continually
N_s : the number of successfully performed tasks N : total number of tasks assigned to the device.

1 Wait for incoming event and check it.

2 *if the event is Disconnecting Event at time (t) then :*

2.1 Replace the prestored value t_e with the new value t : $t_e = t$

2.2 Calculate the Connection Period P_c by using the stored value of t_s and t_e: $P_c = t_e - t_s$

2.3 Calculate number of time units X_c of the connection period P_c by using T_u provided by the Scheduler :

2.4 Calculate the new value of $C_{average}$ by using the prestored value of $C_{average}$ and the prestored N_c:

$$C_{average(new)} = C_{average\ (old)} * N_c + X_c / N_c + 1$$

2.5 $N_c = N_c + 1$.

2.6 Calculate the connectivity metric M_{CD} by using the new calculated $C_{average}$ and the prestored $D_{average}$ as coordinates of a point in the "rank metric map".

3 *if the event is reconnecting Event at time (t) then :*

3.1 Replace the prestored value t_s with the new value t : $t_s = t$

3.2 Calculate the disconnection Period P_D by using the stored value of t_s and t_e: $P_D = t_s - t_e$

3.3 Calculate number of time units X_D of the disconnection period P_D by using T_u provided by the Scheduler :

$$X_D = P_D / T_u$$

3.4 Calculate the new value of $D_{average}$ by using the prestored value of $D_{average}$ and the prestored N_D:

$$D_{average(new)} = D_{average\ (old)} * N_D + X_D / N_D + 1$$

3.5 $N_D = N_D + 1$.

3.6 Calculate the connectivity metric M_{CD} by using the new calculated $D_{average}$ and the prestored $C_{average}$ as coordinates of a point in the "rank metric map".

4 *if the event is Task finish notification Event at time (t) then :*

4.1 if this is a success notification then : $N_s = N_s + 1$.

4.2 N=N+1.

4.3 Calculate the Utilization metric U using the stored values of N_s and N : $U = (N_s / N) * 2$

5 Finally Using the M_{CD} and U values to calculate the Rank Value R as the following:

$$R = M_{CD} + U$$

The considered parameters to be used in the SRA are: the average number of time units being connected continually ($C_{average}$), the average number of time units being disconnected continually ($D_{average}$) and the previous utilization history metric (U). The calculated values of $C_{average}$ and $D_{average}$ will be used as a key to the proposed ranking map, which is used to calculate *the first* part of the rank value that measures the mobile performance and connectivity. The overall ranking value is assumed to be between 1 and 10. This part represents 80% of this value; this percentage can be changed according to the schedulers' administrators. Figure 5 shows the rank metric map, which is based on two roles, first as the $C_{average}$ value increases, the rank must increase also. Second, as $D_{average}$ value increases, the rank must decrease. The $C_{average}$ and $D_{average}$ is used to calculate values. It works as a coordinator of the connectivity metric (MCD) on the rank metric map. The second part of the ranking value is the metric of the utilization of the mobile devices. So, it is calculated by the ratio between the number of the successful tasks and the number of all tasks. Summation of the two parts will generate the overall ranking value of the mobile device. Figure 6 shows the proposed algorithm.

SIMULATION MODEL

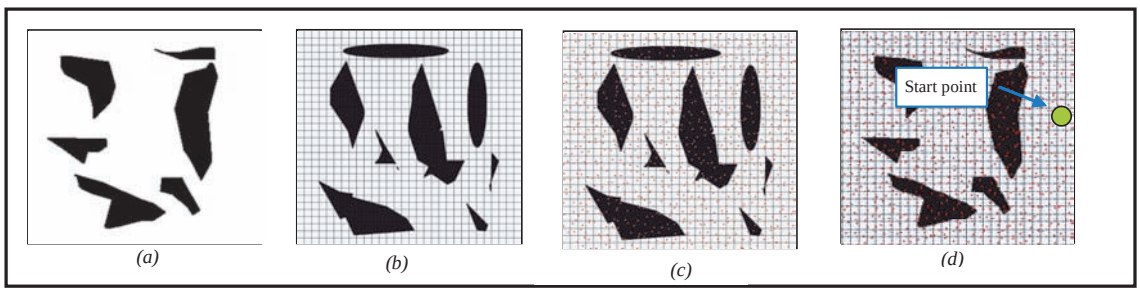
Validation of the proposed algorithm is done via simulation. The investigated simulation program is composed of three modules. The first is responsible for generating a random movement path for the mobile devices, while the second is responsible for tracking the generated path, and this will be done through the access point. Finally, the third is responsible for tuning critical parameters values and collecting outputs parameters, which are required to calculate $C_{average}$ and $D_{average}$.

Mobile Device Movement Mechanism

The mobile device movement path that will be generated is based on a mechanism that guarantees a random path as follows:

1. Generate random black-and-white areas as shown in Figure 7a. White areas imply that there is an available connection between the mobile device and the scheduler access point; black areas depict disconnection.
2. Divide the whole area into small rectangular areas, as shown in Figure 7b.
3. Generate a point within each rectangle at a random position, as shown in Figure 7c.

Figure 7. Steps of random movement path generation



4. Save the position of the generated points in an array.
5. Select one point from the previous array in random fashion to be the starting point of the movement path, as shown in Figure 7d.
6. Select one of the possible eight directions shown in Figure 8 for the next hop.
7. Continue the movement towards the previous selected direction for a random number of hops.
8. Repeat steps 6 and 7 until the required length of movement path is acquired.
9. Store all the selected points in steps 6, 7 and 8 to represent a path for mobile device movement.
10. Repeat steps 2-9 to generate another mobile movement path.

Figure 9 illustrates some examples of the generated random mobile movement paths based on the previous mechanism.

AP and Monitoring the Mobile Device

This module simulates the AP monitoring of the mobile device movement process. In such a process, the AP sends an "Are you alive?" message. If there is an available connection, the mobile device responses with an "I'm alive" message. The time between sending and receiving is called the response time T_r; this time can

Figure 8. Choose a random direction from eight possible

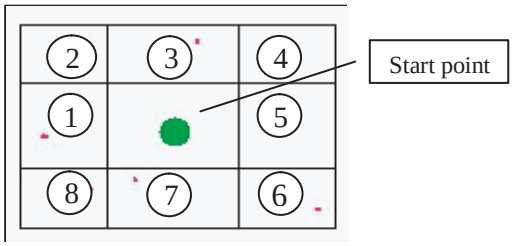


Figure 9. Examples of the random mobile movement paths

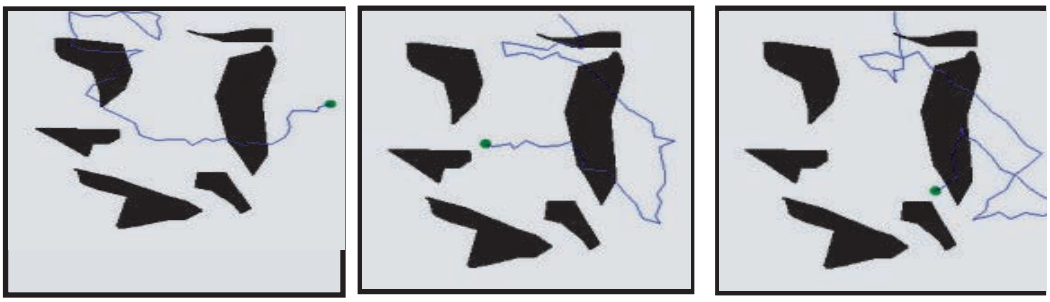
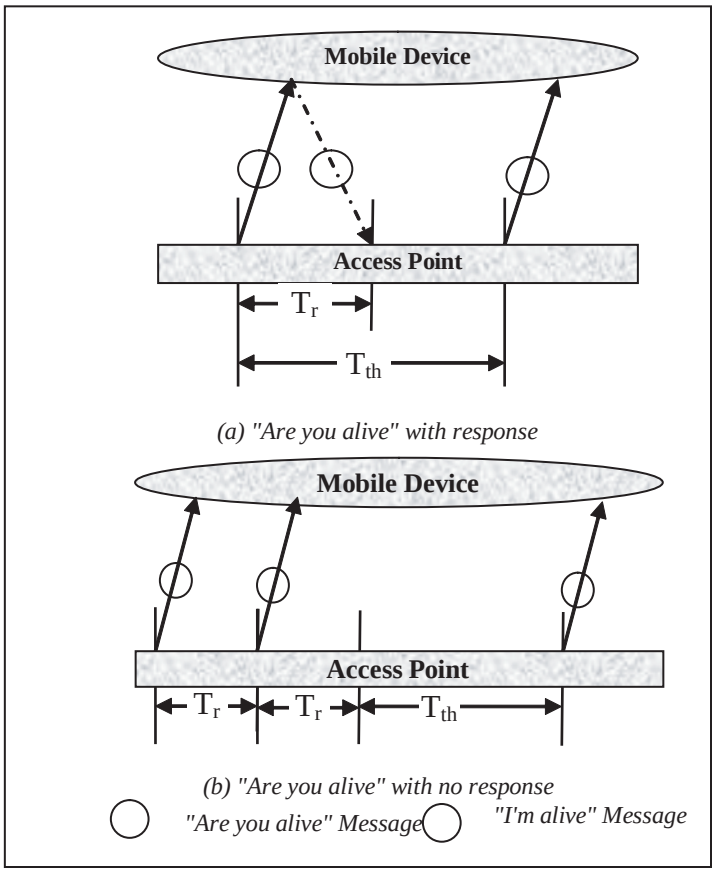


Figure 10. AP monitoring of the mobile device movement process



be determined experimentally. The AP waits for another threshold time T_{th} before sending the next monitoring message. On the other hand, if there is no response for T_r, the access point will send a message again. According to the response of the previous simulation, the AP reports the mobile device status. Figure 10 shows this process.

At this point, we have to notice that reducing T_{th} will lead to more accurate results, but on the other hand, the number of messages will increase. This means high-link usage, which is considered from the application point of view to be a bad usage.

Parameters Setting, Collecting, and Calculating

The different parameters, which are required for comparing the self-ranking against the traditional AP ranking from a network utilization and accuracy point of view, are calculated in this module.

First, the speed of the mobile device movement and T_{th} and T_r is tuned. Some parameters from the first and second modules are collected and stored, including: the length of the generated path, the number of connections and disconnections during the movement on the path, the total number of "Are you alive?" messages, and the number of messages with and without response. So, $C_{average}$ and $D_{average}$ can be calculated.

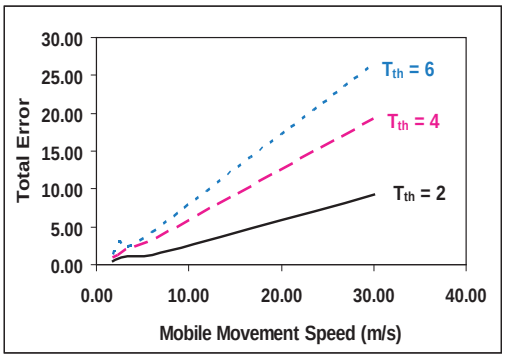
PERFORMANCE ANALYSIS AND DISCUSSION

Based on the previous discussion, on the change of the number of mobile devices used during the experiment (50, 75 and 100 mobiles) or on the change of the value of T_{th} (2, 4 and 6 seconds), various experiments are performed. Two factors were constant: the length of the movement path, which was selected to be relatively long (10000 hop); and T_r, which was selected to be relatively small (0.5 second). Each of these experiments will be repeated for different movement speeds, from low mobility (with average movement speed of 2 m/s) to high mobility (with average movement speed of 30 m/s).

The average error in calculating the $C_{average}$ and $D_{average}$ has been calculated for each experiment at each used speed, and their summation represents the total error in the experiment. Also, the number of network messages exchanged between AP and the mobile device, in both the AP monitoring and self monitoring, has been counted.

Figures 11, 12, and 13 show that the percentage of the total error increases rapidly as the movement speed of mobile device increases. This result is expected, because as movement speed increases, the ability of AP to sense the change in the mobile connectivity will be more and more limited. Also, the figures show that when the value of T_{th} increases, the percentage of the total error increases also, while the number of exchanged network messages decreases. This result is expected, because T_{th} represents the time between two monitoring messages; as this time increases, that means reduction in the ability of AP to sense the change in mobile connectivity. The figures show the comparison between the number of exchanged messages between AP and mobile devices in the case of self monitoring and case of AP monitoring. Note that, in the case of AP monitoring, the number approximately

Figure 11. Total error and link utilization at number of mobiles = 50



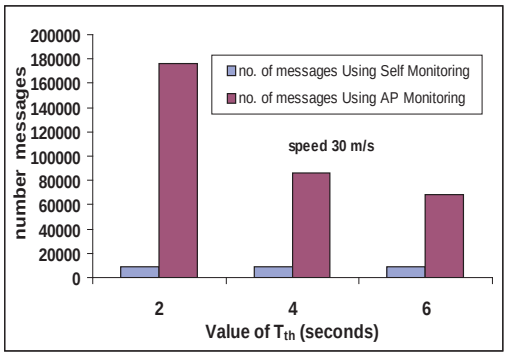
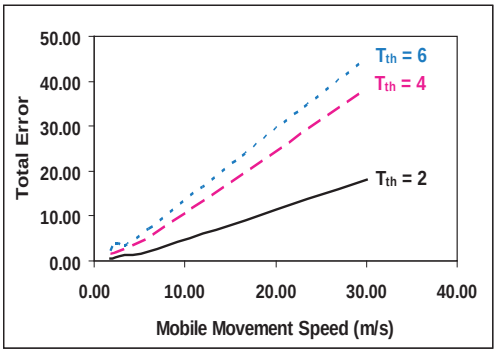


Figure 12. Total error and link utilization at number of mobiles = 75



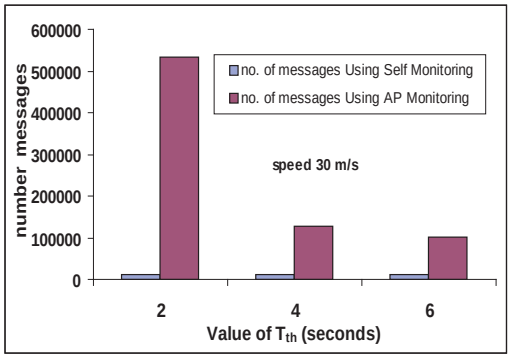
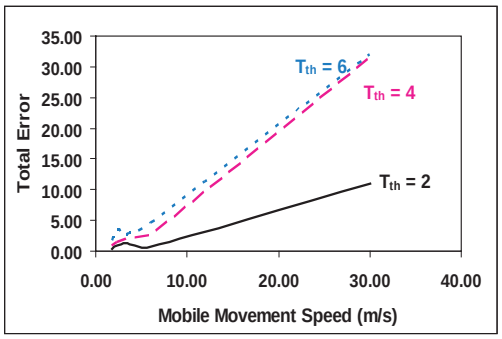
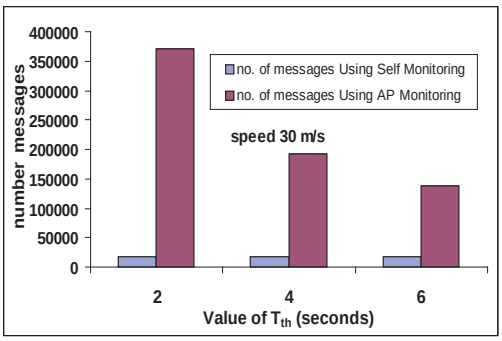


Figure 13. Total error and link utilization at no of mobiles = 100





doubled more than 70 times compared to the case of self-monitoring.

CONCLUSION

This article points out an overview of the issues of mobile devices integration with the existing Grid. It shows that when some authorization is impeded within the mobile client, every mobile can evaluate its own performance. The traditional method makes an overhead on the scheduler to perform a historical evaluation to the mobile performance, which makes it busy in a secondary task and leaves its main task of scheduling. So, the SRA will be the base of a scheduling mechanism that will schedule the tasks on the mobile devices. The originality of the proposed mechanism concentrates on mobile cooperating with services at the AP. Using such a mechanism

will lead to *minimizing* the calculation time consumed in mobile ranking and evaluating before starting the scheduling process. Moreover, it will lead also to *minimizing* the amount of stored data at the scheduler and *simplifying* the scheduling process by grouping the mobile devices according to their self-ranking value and assigning tasks to these groups. Finally, it will result in *maximizing* the profit of the mobile devices integrated with the already existing Grid systems by using their computational power as an addition to the system's overall power. In brief, the outcome will be *maximizing* system utilization and making the system more flexible to integrate any new devices without any need to increase the system complexity.

In this article, we present the newly emerging technical issues for realizing this mobile Grid system, and particularly focus on the job scheduling algorithm to achieve more reliable performance. However, there are still challenging problems, such as limited energy, device heterogeneity, security and so on. We will tackle these issues in future works and develop a prototype of a mobile Grid system.

REFERENCES

Barmouta, A., & Buyya, R. (2003, April 22-26). GridBank: A Grid Accounting Services Architecture (GASA) for distributed systems sharing and integration. In *Proceedings of the Parallel and Distributed Processing Symposium (IPDPS 2003)*, Nice, France (p. 245). IEEE Computer Society.

Berman, F., & Casanova, H. (2005). New Grid scheduling and rescheduling methods in the GrADS project. *International Journal of Parallel Programming, 32*(2-3), 209-229.

Buyya, R., Abramson, D., & Giddy, J. (2002). Economic models for resource management and scheduling in Grid computing. *Concurrency and Computation: Practice and Experience (CCPE) Journal, 14*(13-15), 1507-1542.

Buyya, R., & Murshed, M. (2002). GridSim: A toolkit for the modeling and simulation of distributed resource management and scheduling for Grid computing. *Concurrency and Computation: Practice and Experience (CCPE) Journal, 14*(13-15), 1175-1220.

Buyya, R., Murshed, M., & Abramson, D. (2003, June 24-27). A deadline and budget constrained cost-time optimization algorithm for scheduling task farming applications on global Grids. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'02)*, Las Vegas, NV.

Dail, H., Casanova, H., & Berman, F. (2002). A decoupled scheduling approach for the GrADS program development environment. In *Proceedings of the 2002 ACM/IEEE Conference on Supercomputing. Conference on High Performance Networking and Computing*, Baltimore, MD (pp. 1-14).

Frey, J. (2001, August 7-9). Condor-G: A computation management agent for multi-institutional Grids. In *Proceedings of the Tenth IEEE Symposium on High Performance Distributed Computing (HPDC10)*, San Francisco, CA..

Frontier. (2004). *The premier Internet computing platform* (White Paper). Retrieved from http://www.parabon.com/clients/clientWhitePapers.jsp

Gradwell, P. (2003). *Overview of Grid scheduling systems*. Retrieved from http://www.peter.me.uk/phd/writings/computing-economy-review.pdf

He, X. (2003). A QoS guided scheduling algorithm for Grid computing [Spcial issue Grid computing]. *Journal of Computer Science and Technology (JCS&T), 18*(4).

In, J.-u., & Avery, P. (2004, April). Policy based scheduling for simple quality of service in Grid

computing. In *Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS 2004),* Santa Fe, NM.

Nurmi, D., Wolski, R., & Brevik, J. (2004). *Model based checkpoint scheduling for volatile resource environments* (Technical Report). Santa Barbara: University of California Santa Barbara, Department of Computer Science.

Park, S.-M., Ko, Y.-B., & Kim, J.-H. (2003, December 15-18). Disconnected operation service in mobile Grid computing. In *Proceedings of Service-Oriented Computing — ICSOC 2003: First International Conference,* Trento, Italy.

Sanver, M., Durairaju, S.P., & Gupta, A. (2004). Should one incorporate mobile-ware in parallel and distributed computation? In *Proceedings of the 10th International Conference on High Performance Computing (HiPC 2003),* Hyderabad, India.

Saxena, N. (2005, April 3-7). New hybrid scheduling framework for asymmetric wireless rnvironments with request repetition. In *Proceedings of the Third International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt'05),* Riva del Garda, Trentino, Italy (pp. 368-376).

Saxena, N., Basu, K., Das, S.K., & Pinotti, C.M. (2005, April). A dynamic hybrid scheduling algorithm with clients' departure for impatient clients in heterogeneous environments. In *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05),* Rhodes Island, Greece.

Somasundara, A.A., Ramamoorthy, A., & Srivastava, M.B. (2004, December 5-8). Mobile element scheduling for efficient data collection in wireless sensor networks with dynamic deadlines. In *Proceedings of the 25th IEEE International Real-Time Systems Symposium (RTSS'04),* Lisbon, Portugal (pp. 296-305).

Sulistio, A., Yeo, C.S., & Buyya, R. (2003, June). Visual modeler for Grid Modeling and Simulation (GridSim) toolkit. In *Proceedings of the International Conference on Computational Science (ICCS 2003), Part III,* Melbourne, Australia (pp. 1123-1132).

Takefusa, A. (2001, August). A study of deadline scheduling for client-server systems on the computational Grid. In *Proceedings of the 10th IEEE Symposium on High Performance and Distributed Computing (HPDC'01),* San Francisco, CA (p. 406).

Yu, D. (2003, November 13-15). Divisible load scheduling for Grid computing. In *Proceedings of the 16th International Conference on Parallel and Distributed Computing Systems (PDCS 2003),* Marina del Rey, CA.

This work was previously published in the International Journal of Information Technology and Web Engineering, edited by D. Rine and G. Alkhatib, Volume 1, Issue 2, pp. 43-59, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 7.39
Multilayered Approach to Evaluate Mobile User Interfaces

Maria de Fátima Queiroz Vieira Turnell
Universidade Federal de Campina Grande (UFCG), Brazil

José Eustáquio Rangel de Queiroz
Universidade Federal de Campina Grande (UFCG), Brazil

Danilo de Sousa Ferreira
Universidade Federal de Campina Grande (UFCG), Brazil

ABSTRACT

This chapter presents a method for the evaluation of user interfaces for mobile applications. The method is based upon an approach that combines user opinion, standard conformity assessment, and user performance measurement. It focuses on the evaluation settings and techniques employed in the evaluation process, while offering a comparison between the laboratory evaluation and field evaluation approaches. The method's presentation and the evaluation comparison will be supported by a discussion of the results obtained from the method's application to a case study involving a Personal Digital Assistant (PDA). This chapter argues that the experience gained from evaluating conventional user interfaces can be applied to the world of mobile technology.

INTRODUCTION

As proposals for new techniques and methods emerge for the evaluation of mobile device usability, it becomes more difficult for practitioners to choose among them. To be able to evaluate the efficacy of these techniques and methods, as well as to reproduce their steps, they have to be described in a level of detail not often found in the literature. Claims are often made without solid statistical results and are usually based on superficial descriptions. This makes it difficult, if not impossible, to compare alternative choices. Given the features of these new devices (such as mobility, restrictive resources for information input and output, and dynamic contexts of use), HCI specialists may question the efficacy of the methods, techniques, and settings already known

to them from previous experiences. Thus, the major question that is addressed is whether it is possible to adapt the methods, techniques, and settings from previous evaluation experiences to this new class of devices, given their distinctive features.

The most frequent question raised in the vast majority of studies presented in the literature is whether to adopt a field approach or a laboratory approach. However, little is discussed in terms of which techniques are best suited for the specific evaluation target and its context of use. While this polemic subject may represent to the HCI specialist an import concern, it is equally important to consider the efficacy of the method, which accompanies this choice of approach (efficacy meaning the quality of the answers to the questions formulated as the basis of the evaluation). This is because the efforts employed in the evaluation may not pay off if a method is not well chosen or well employed.

This chapter presents a method for evaluating mobile devices based upon a set of techniques already known to the HCI specialist community. Each technique evaluates the problem from different perspectives: the user perspective (expressed as views on the product obtained through a questionnaire), the specialist's perspective (expressed when analyzing the user performance during the usability evaluation), and the usability community perspective (expressed in the form of standards conformity assessment). Each of these perspectives identifies evaluation problems and, when overlaid, they lead to a more reliable and complete product appraisal.

The remainder of this chapter is structured as follows. The second section gives a brief overview of the evaluation approaches currently in use for mobile devices, according to the literature review. The third section outlines the multi-layered approach. The fourth section illustrates the application of the multi-layered approach by means of a case study involving a Personal Digital Assistant (PDA). The fifth section discusses the results of the case study and their implications for the questions posed in this chapter. Finally, the sixth section concludes with the discussion of future trends in evaluation methods and how to apply the existing experience to the evaluation of this new class of products.

USER INTERFACE EVALUATION FOR MOBILE DEVICES

In the context of user-centered design processes, a significant portion of usability work involves the coordinated acquisition of valid and reliable data by a team of professionals. These specialists have varied backgrounds and skills and employ a number of evaluation methods. The expected result is an improved system design. This is achieved by the successful identification of a system's usability problems that might impact the interaction quality for a range of users.

Usability data consists of any information that can be used to measure or identify factors affecting the usability of a system being evaluated (Hilbert & Redmiles, 2000). These data are crucial for designing successful systems intended for human use. Such data are gathered by usability evaluation methods and techniques that can assign values to usability dimensions (Rosson & Carroll, 2002) and/or indicate usability deficiencies in a system (Hartson, Andre, & Williges, 2003). According to the International Organization for Standardization (ISO, 1998), usability dimensions are commonly taken to include user efficiency, effectiveness, and subjective satisfaction with a system in performing a specified task in a specified context.

Usability data are gathered via either analytic or empirical methods (Nielsen, 1993; Mayhew, 1999; Rosson & Carroll, 2002). Analytic methods, in which a system is evaluated based on its interface design attributes, are usually conducted by HCI specialists and do not involve human participants performing tasks. This means that these

methods often rely on the specialists' judgment. Empirical methods, in which the system is evaluated based on observed performance in actual use, involve data collection of human usage.

Other classifications include direct methods (recording actual usage) and indirect methods (recording accounts of usage) (Holzinger, 2005). There are also formative and summative methods (Wixon & Wilson, 1997). The direct methods are used to generate new ideas and gather data during the development of a system in order to guide iterative design (Hix & Hartson, 1993). The indirect methods are used to evaluate existing systems and gather data to evaluate a completed system in use (Scriven, 1967). Discovery methods (also called qualitative methods) are used to discover how users work, behave, and think, and what problems they have. Decision methods (also called quantitative methods) are used in selecting a design among several alternatives or in picking elements of interface designs (Wixon & Wilson, 1997).

In essence, usability data have been classed in a number of other models and frameworks, often focusing on (1) the approach employed for gathering the data (including the resources expended and the degree of formality) (Danielson, 2006); (2) the context of use (including lighting, noise level, network connectivity, communication costs, communication bandwidth, and the social situation) (ISO, 1998; ISO, 1999; Jones & Marsden, 2006); (3) the nature and fidelity of the artifact being evaluated (EATMP, 2000); and (iv) the goal of the acquisition process (Kan, 2002).

It is a fact that usability evaluation for stationary computer systems has grown in the last two decades. In spite of debates still taking place within the HCI area, they are often based on a tacit understanding of basic concepts. One example of this understanding is in relation to the distinction between field and laboratory evaluation approaches and their importance to the area. Classical extensive guidelines were written that describe how usability evaluation in controlled environments should be conducted (e.g., Dumas & Reddish, 1999; Mayhew, 1999; Nielsen, 1993). Additionally, experimental evaluations of the relative strengths and weaknesses of different techniques are available that can be applied in a usability evaluation (e.g., Molich et al., 1998).

In the last decade, methodologies and approaches in HCI have been challenged by the increasing focus on systems for wearable, handheld, and mobile computing devices. One such move beyond office, home, and other stationary-use settings has pointed to the need for new approaches in designing and evaluating these systems (Kjeldskov, 2003). While the primarily task-centered evaluation approaches may be applicable to the desktop computing paradigm (often structured with relatively predictable tasks), they may not be directly applicable to the often-unpredictable continuous interaction possibilities and relatively unstable mobile settings. Additionally, it is not easy for evaluation methods to integrate completely or even adequately in real world or simulated settings contexts during the evaluation process. Authors argue that mobile computing demands not only real users but also a real or simulated context with device interaction tasks. It also demands real tasks or realistic task simulations.

There are a number of studies that discuss the question of whether the evaluation should be carried out in a laboratory or field context (e.g., Goodman et al., 2004; Kjeldskov & Stage, 2004; Kjeldskov et al., 2005; Po et al., 2004). All of these papers have a common theme, in that they apply a multi-method approach to performance measurement and discuss solutions for efficient data analysis. Nonetheless, it is important to note that the approach to usability evaluation depends on the relevance of the results presented as well as on the quality of the data analysis process. In general, the reports only present the results of the data analysis, omitting the details of the analysis process itself. While the data gathering method is critical for data quality, a more rigorous analysis

on user comments and problem reports could help specialists better assess their choices.

There is a lot of current human-computer interaction research on alternatives for data collection methods and techniques. However, adequate data analysis and validation are only presented in few cases (e.g., Nielsen, 1994; Dumas & Redish, 1999; Po et al., 2004). In general, this aspect of the HCI research is poorly described in the literature, there being only vague conclusions and little guidance for attempts at successfully replicating the findings in other evaluation contexts. Many methods and techniques have been employed in the analysis of empirical data gathered during usability evaluations. Examples are for field testing analysis, video data analysis (Sanderson & Fisher, 1994), expert analysis (Molich et al., 1998), and head-mounted video and cued recall (Omodei et al., 2002). Its time-consuming character and its poor applicability for industrial purposes can explain the absence of an in-depth usage data analysis when under resource constraints (Baillie & Schatz, 2005). Nonetheless, it is strongly recommended for research purposes as a means to support new findings. For the same reason, it is equally important to provide sufficient detail to allow for replication and a substantiated choice of methods with similar levels of description.

THE MULTILAYERED EVALUATION APPROACH

The method described here was originally proposed for evaluating desktop interfaces. It was then adapted to evaluate the usability of mobile devices. It is based upon a multi-layered approach that combines standard conformity assessment, user performance measurement, and user satisfaction measurement. Each one of these evaluation techniques detects problems from a specific point of view. The multilayered approach is based on the premise that the combination of techniques (triangulation) will produce complementary and more robust results.

Standard Conformity Assessment

According to the International Organization for Standardization (ISO), conformity assessment means checking whether products, services, materials, processes, systems, and personnel measure up to the requirements of standards (ISO, 2006).

In its original version, this evaluation method adopts the standard ISO 9241 (*Ergonomic Requirements for Office Work with Visual Display Terminals*).

In the PDA case study it was found that only some parts of this standard can be applied to this mobile device: Parts 11 (ISO 9241-11, 1998), 14 (ISO 9241-14, 1997), 16 (ISO 9241-16, 1999), and 17 (ISO 9241-17, 1998). There are also some other standards that apply to this kind of device such as the ISO/IEC 14754 (*Pen-based Interfaces—Common gestures for text editing with pen-based systems*) (ISO/IEC 14754, 1999) and others that, although applicable to mobile devices, do not apply in this specific case. Examples are the ISO/IEC 18021 (*User interfaces for mobile tools for management of database communications in a client-server model*), since it is for devices capable of performing data interchange with servers (ISO/IEC 18021, 2002); and ITU-T E.161 (*Arrangement of digits, letters, and symbols on telephones and other devices that can be used for gaining access to a telephone network*, also known as ANSI T1.703-1995/1999, and ISO/IEC 9995-8:1994) (ITU, 2001).

User Satisfaction Measurement

User satisfaction has received considerable attention from researchers since the 1980s as an important surrogate measure of information systems success (Aladwani & Palvia, 2002; Goodhue &

Thompson, 1995; Bailey & Pearson, 1983). While most user satisfaction measuring instruments were not Web-based at the time of development, others have been successfully validated in a Web-based environment (e.g., De Oliveira et al., 2005).

The user satisfaction diagnosis provides an insight into the level of user satisfaction with the product, highlighting the importance of the problems found and their impact on the product acceptance.

User Performance Measurement

The user performance measurement aims in general to provide data on the effectiveness and efficiency of a user's interaction with a product. It enables comparisons with similar products, or with previous versions of the same product along its development. Additionally, it can highlight areas where a product can be enhanced to improve usability. When used with the other methods, the evaluator can build a complete picture of the usability of a system.

The most significant user interface problems can be found by conducting experiments (usability tests) with representative users to observe how quickly, easily, and safely they can operate a product. The major change introduced in the original method concerns the introduction of field tests as a complement to the original laboratory tests.

The Experiment: Comparing Field and Laboratory Use of a PDA

The main objective of this study is to investigate the need for adapting the original evaluation method to the context of mobile devices, based on the analysis of the influence of the context (field versus laboratory and mobility versus stationary interaction) on the evaluation of mobile devices and applications.

The mobile device chosen as the target for this study was a PDA, the *Nokia 770 Internet Tablet* and some of its native applications. Tests were performed in a controlled environment (the usability laboratory) and also in the field. Twenty-four users took part in the experiment, divided into two groups of twelve.

Experiment Design

The study was designed to investigate the influence of the context (field and laboratory) and associated aspects such as mobility, settings, and so forth, and the user experience on the evaluation results. The independent variables are those that are not influenced by the context, by the test facilitator, or by external factors such as noise and lighting. An experiment plan was drawn from the study's objectives. The independent variables were chosen as follows:

- **Task context** comprises factors that may affect the users' behavior and their performance during the experiment (usability test). These factors may be internal or external to the user. The external factors originate in the field environment, examples being noise level and light intensity. The internal factors, on the other hand, are stress or other health conditions that may affect the user's mental and physical abilities.
- **User mobility** refers to the conditions under which the task is being performed. An example is if the user is required to work while being mobile, that is, moving between places or wandering while working.
- **User experience level** refers to the user's knowledge regarding mobile devices in particular and desktop computers systems in general.

The dependent variables are all dependant on the user's experience level:

- **Task time** represents the time taken by a device's user to perform a task.

- **Number of incorrect choices** measures how many times the user has made incorrect choices while selecting options in the interface through a menu dialogue.
- **Number of incorrect actions** measures how many times the same error (excluding the number of incorrect choices) was committed by the user while performing a task.
- **Number of accesses to the online help** and **number of accesses to the printed help** measure how many times the user accessed the online and printed help while performing a task.
- **Perceived usefulness** represents the user's opinion about the usefulness of the mobile application for the prescribed task.
- **Perceived ease of use** represents the user subjective satisfaction when using the mobile device.

Table 1 summarizes the experiment plan, which states the independent and dependent variables to be observed during the experiment and used as indicators to answer the research questions.

Test Environment

A software tool was used in the field environment to remotely capture the device's screen through a wireless connection to the lab. The user inputs (through keypad and stylus) were registered by a micro-camera coupled to the device and also

Table 1. Plan for the experiment with the device Nokia 770

EXPERIMENT PLAN	
Target-Problems	1. With the shape/dimensions of the product 2. With the mechanisms for information input/output 3. With the processing power 4. With the navigation between functions 5. With information legibility
Test Objec-tives	1. Investigating the target problems 2. Detecting other problems
Objective Indicators	1. Task execution time 2. Number of incorrect actions 3. Number of incorrect choices 4. Number of repeated errors 5. Number of accesses to the online help 6. Number of off-line help (printed manuals) accesses
Subjective Indicators	1. Product ease of use 2. Task completion easiness 3. Input mechanism ease of use 4. Text input modes ease of use 5. Understandability of terms and labels 6. Understandability of messages 7. Help mechanism efficiency

remotely connected to the laboratory through a wireless connection. The interaction was registered in the controlled environment using two video cameras installed in the laboratory. One was focused on the users´ facial expressions and the other registered the device screen. As in the field environment, software was used to remotely capture the device's screen. Since the field setting required a wireless network, the field experiment was performed in the area surrounding the university's computer department. In both cases, the test facilitator was a human interface specialist who remained within reach in case the user required any explanation on the test procedure.

Participants

Users participating in the PDA experiment were selected on the basis of having previous experience with mobile devices (such as mobile phones), computers, and the Internet. They were also required to have some familiarity with the English language, since this is the language adopted in the device's user interface and in its documentation. The user sample was then classed according to the users' experience level into the categories shown in Table 2.

The recruited users were divided into two groups of 12 to participate in the field and laboratory tests. Based on user experience level, both groups were then subdivided into three subgroups of four beginners, four intermediates and four experts.

Materials

Laboratory Test Materials

- **Hardware:** The Nokia 770 Internet Tablet; PC based Workstation (2); Video cameras (3); Microphones (2).
- **Software:** VNC (Virtual Network Computing) software to capture the screens during the interaction with the device; the WebQuest tool with the questionnaires pre-test (to gather the user profile) and post-test (to collect and process the user satisfaction level).
- **Miscellaneous:** The Nokia 770 Internet Tablet Manual; chronometer (1); CDs for video backup; participant registration form; test conditions acceptance forms on which the users declared their acceptance of the experiment conditions; task script that consists of a written task description to guide the user during the session (versions for the user and for the evaluator); Form for event log.

Table 2. User sample categorization

CATEGORY / CHARACTERISTIC	Beginner	Intermediate	Expert
Previous Computer Knowledge	Basic/ Intermediate	Intermediate/ Advanced	Intermediate/ Advanced
Previous Experience with *Nokia*	No	No	Yes

Field Test Materials

- **Hardware:** The Nokia 770 Internet Tablet; PC-based Portable (laptop) Workstation (1); wireless video micro-camera (1); apparatus to support the video micro-camera (1); television set (1); VCR equipment (1).
- **Software:** VNC (Virtual Network Computing) software to capture the screens during the interaction with the device; WebQuest tool with the questionnaires pre-test (to gather the user profile) and post-test (to collect and process the user satisfaction level).
- **Miscellaneous*:*** Chronometer (1); CDs and VHS tapes for video backup; participant registration form; test conditions acceptance forms on which the users declared to accept the experiment conditions; task script that consists of a written task description to guide the user during the session (versions for the user and for the evaluator); form for event log.

Camera Apparatus

The apparatus shown in Figure 1 was built to couple a video micro-camera to the mobile device. This allowed the recording of user interaction through a remote link with the laboratory computer.

The WebQuest Tool

A Web tool named *WebQuest* supports the method application (De Oliveira et al., 2005). This tool was developed to support the specialist during data collection, to provide automatic score computation, to perform statistical analysis, and to generate graphical results. *WebQuest* also enables the specialist to reach a more diverse and geographically widespread sample of users through the Internet. One of its features is a flexible questionnaire structure, which enables specific context adaptation and, by means of an estimation model, ensures a higher degree of confidence on the indicators of user satisfaction. Currently *WebQuest* supports two questionnaires: (i) a pre-test questionnaire, the *USer* (*User Sketcher*), conceived to raise the profile of the system users; and (ii) a post-test questionnaire, the *USE* (*User Satisfaction Enquirer*), conceived to raise the user degree of satisfaction with the system. The pre-test questionnaire incorporates a model to estimate the user's subjective satisfaction and can be answered directly on the Web. The questions are related to the users´ physical characteristics, knowledge, and skills. Both questions and answers are configurable.

As for the *USE* (*User Satisfaction Enquirer*), it allows gathering quantifiable variables on the user acceptance of the device. Three of its aspects are of special interest. Firstly, it incorporates a model to estimate user subjective satisfaction. Secondly, the questionnaires can be answered directly on the Web. Thirdly, the items are partially or totally configurable. The adoption of an estimation model by *USE* allowed us to establish

Figure 1. Apparatus to support video camera during experiment

a subjective satisfaction coefficient directly from the inspection of the respondents' samples. The *WebQuest* tool allows the specialist to easily edit the questionnaire's items. These items are organized into groups: (1) *fixed*, which are applicable to various evaluation contexts and thus are not allowed to be altered; (2) *semi configurable*, which allow for changes in the answer options; and (3) *configurable*, which can be fully configured (both the question and respective options of answers). *USE* supports the specialist from the data collection through to automatic score computation, performing statistical analysis, and generating graphics with the results.

Experiment Procedure

The techniques employed in the experiment procedure were the observation and subsequent video analysis for accumulating quantitative data (such as time spent and error rate). An automated video capturing tool recorded the interactions of the subjects during the field tests to ensure a non-intrusive observation method. During task execution, the users were asked for their consent before being filmed. The conditions of test-subject participation included a written commitment not to disclose any product information. The users were also asked to give consent so that their images and/or sound recordings made during the experiment could be used for research purposes or in a multimedia product evaluation report. On the other hand, the users were given assurances from the evaluation team that no personal information or individual performance results would be disclosed.

The first step in following the method consisted in defining the evaluation scope for the product as well as a scenario for the test. Table 3 illustrates the sequence of tasks performed during the experiment.

The decision was based on a heuristic evaluation performed by the evaluation team. This initial step also supports the definition of a general profile for the user sample and a classification into categories. Following, the method the evaluation objectives were defined. These became the basis for choosing the product evaluation scenario (product context of use and laboratory settings) and the corresponding tasks to be performed by the users during the experiment. Having planned the evaluation, a pilot test was conducted to verify the adequacy of the proposed experiment procedural, materials, and environment. Through this fine tuning procedure it was found, in the PDA case study, that the time to perform the tasks had been underestimated. This resulted in re-dimensioning the test scenario to six tasks, with a review of the tasks themselves to fit the established session time of sixty minutes to prevent user tiredness.

All subjects were submitted to the same procedure prescribed in the experiment protocol. The study was conducted first in a laboratory setting and then in the field environment. During the field tests the participants were taken outdoors, and the tasks were conducted in an environment that was as close to real-use conditions as possible.

The experiment conducted in the usability laboratory had the audio and video of each session recorded. In the field experiment, only the

Table 3. Test scenario and sequence of tasks to be performed during experiment

TASKS IN SCRIPT	
T01	**Initializing the device**
T02	**Searching for books in an online store**
T03	**Visualizing a PDF file**
T04	**Entering textual information**
T05	**Using the electronic mail**
T06	**Using the audio player**

video of the sessions was recorded, supplemented by comments written by the specialist. As described in the experiment protocol, each session consisted of the following steps: (1) introducing the user to the test environment by explaining the test purpose, the procedure to be followed and the ethics involved in terms of the conditions of participation; (2) applying the pre-test questionnaire; (3) performing the task script; (4) applying the post-test questionnaire; and (5) performing a non-structured interview.

At the time of the experiment, the *Nokia 770 Internet Tablet* device was not yet widely known in the Brazilian market. The users who claimed to have had no previous contact with it were given a quick introduction. This introduction consisted of an instructional material given to the recruited users and also a quick explanation about the device's input and output modes and its main resources.

Results

The results obtained from the experiment in which the multi-layered method was applied support the original assumption that, in spite of the distinctive features of this class of devices, it is possible to adapt from the evaluation experience with conventional devices. This conclusion is supported by the evidence that the evaluation context did not significantly influence the user performance or the opinion about the device's usability, given through the analysis of the objective and subjective indicators associated with the experiment.

Standard Conformity Assessment Results

The results of the conformity assessment to the standards ISO 9241 Parts 14 and 16 and ISO 14754 are illustrated in Table 4. According to ISO, conformity assessment results can be summarized by computing an *adherence rate* (AR). This is the percentage of the applicable recommendations (Ar) that were successfully adhered to (Sar).

In spite of the device's characteristics that limit the number of applicable recommendations, these results corroborate the idea that the standards inspection is still applicable in the evaluation process. The efficacy of this technique can be considerably improved if it is based upon standards conceived specifically for mobile devices, which could evidence more usability problems.

Table 4. Nokia 770 conformity assessment with standards

Standard	#Sar	#Ar	AR (%)
ISO 9241 Part 14	**45,0**	**53,0**	**84,9**
ISO 9241 Part 16	**26,0**	**33,0**	**78,8**
ISO 14754	**4,0**	**11,0**	**36,4**

Sar—Successfully adhered recommendations
Ar—Applicable recommendations
AR—Adherence Rate

$$AR = \frac{Sar}{Ar} x100\%$$

User Satisfaction Measurement Results

For the PDA case study context, both questions and answers of the *USE* questionnaire were configured. The questionnaire was applied soon after the usability test and answered using the mobile device itself. As mentioned before, its purpose was to collect information on the user's degree of satisfaction with the device and on aspects such as interface navigation, documentation, and overall impressions.

The *USE* was composed of three sections. The first section is relative to "the product Use and Navigation." It is composed of 17 items and focuses on aspects such as menu items, navigation between functions, understandability of the messages, ease of use of the basic functionalities, and of the device's input and output mechanisms. The second section consists of six questions related to the online and off-line (printed manuals) documentation. The last section ("You and the product") consists of 15 items and aims to get the user's impressions and product acceptance level. The first 23 items use a 5-point semantic scale (1: *very easy*; 2: *easy*; 3: *not easy nor difficult*; 4: *difficult*; and 5: *very difficult*). The last 15 items use another 5-point semantic scale (1: completely agree; 2: agree; 3: do not agree nor disagree; 4: disagree; and 5: completely disagree). The users were asked to answer the questions and to assign an importance level to each one of them, on a scale from 0 to 10.

For the post-test questionnaire, *USE* adopts the model proposed by Bailey and Pearson (Bailey & Pearson, 1983) for measuring the overall user's sense of satisfaction. The following adaptations to the dimensions were considered: (1) the association of only one (1) semantic differential scale to the items, instead of the four (4) semantic differential scales, as proposed in the original model; (2) the adoption of a 5-point Likert scale, delimited by the ends -2 and 2 (instead of the 7-point scales delimited by the ends -3 and 3 as originally proposed); and (3) the incorporation of a 11-point importance scale (0 corresponding to *non applicable*), varying from 0.0 to 1.0 in intervals of 0.1 (instead of the original 7-point scales, which varied from 0.1 to 1.0 in intervals of 0.15).

The user's subjective satisfaction indicators for the PDA case study were 0.330 for the laboratory experiment and 0.237 for the field experiment. The normalized value ranges of the user satisfaction concerning a product are 0.67 to 1.00 (*Extremely Satisfied*), 0.33 to 0.66 (*Very satisfied*), 0.01 to 0.32 (*Fairly satisfied*), 0.00 (*Neither satisfied nor unsatisfied*), 0.01 to 0.32 (*Fairly dissatisfied*), 0.33 to 0.66 (*Very dissatisfied*), and 0.67 to 1.00 (*Extremely dissatisfied*). This is in accordance with the Bailey and Pearson model (Bailey & Pearson, 1983). The results obtained correspond respectively to *Very satisfied* and *Fairly satisfied*.

Performance Measurement Results

The User Sample Profile

The user sample profile was drawn with the support of the questionnaire *USer*. It was composed of 13 male and 11 female users, of which eight were *undergraduate students*, 12 *post-graduate students*, two *graduate level*, and two *post-graduate level*. The ages varied between *18* and *29* years. They were mainly *right handed* and mostly used some sort of reading aid (either *glasses* or *contact lenses).* All of them had at least one year of *previous experience of computer systems* and were currently using computers on a *daily* basis.

User Performance Data Analysis

After having analyzed the data gathered during the experiment on the user performance and having analyzed the list of problems found with this technique, it was possible to evaluate their impact and class them as: minor (50%), medium (50%), major (0%), consistency (35.7%), recurrent (64.3%), and general (0%).

The data analysis consisted of a statistical processing and had two main purposes: (1) to investigate the influence of the context on the results of the evaluation method (through the comparison of the results obtained from both environments); and (2) to investigate the influence of the user experience with the mobile device on the test results within each context. For the latter purpose, the three categories illustrated in Table 2 were used.

The statistic analysis performed consisted of: (1) building a report with univariance statistics; (2) generating the covariance matrices for the objective and subjective indicators that were previously defined; (3) applying the one-way F ANOVA test (Tabachnick & Fidell, 2006) to the data obtained from the previous step in order to investigate possible differences; and (4) applying the Tukey-Kramer process (Tabachnick & Fidell, 2006) to the one-way F ANOVA results aiming to investigate if the found differences were statistically significant to support inferences from the selected sample. The result of this technique was the identification of 13 problems, of which 92.3% were found in the laboratory and 61.5% in the field as: Laboratory (38.5%); Field (7.7%); and Laboratory & Field (53.8%).

Overlaying Results

Since the multi-layered evaluation is based upon a triangulation of results, Table 5 summarizes the usability problem categories identified by the three techniques.

The numbers correspond to the identification of each problem from a list of problems found through each technique. As can be seen from Table 5, some of the usability problem categories were more related to the performance measurement (e.g., hardware aspects, help mechanisms, processing capacity) whereas others (e.g., menu navigation, presentation of menu options) were identified by the conformity assessment. It was possible to identify 66.7% of the problems found by other methods when combining the results from the post-test questionnaire with the user comments made during the experiment and the informal interview at the end of the experiment. This confirms the importance of combining techniques to obtain a more complete result when performing usability evaluation. It must be pointed out that 29.62% of the problems based on the user opinion about the product were in disagreement with the results of the other two evaluation dimensions (specialist and the community points of view). This discrepancy can originate from the users' perception of product quality and the perception of their own skills to perform the task, accepting full responsibility over the difficulties that might arise during the interaction. When overlaying the problems in Table 5, in the category *Menu navigation*, the same problem was found by the techniques Standards Inspection and Performance Measurement.

DISCUSSION

From this study's data analysis it became evident that certain problem categories are better found by specific techniques, as shown in Table 5. For instance, problems associated to the device's physical characteristics are better found by means of conformity assessment, whereas the user performance located problems associated to the device's applications.

The analysis of the pre-test and post-test questionnaires and the informal interviews showed that domain knowledge and computer literacy have significant influence on user performance with mobile devices. This was true both under laboratory conditions and in the field, in relation to the incidence of errors. The univariate analyses of variance of the performance variables: *Time*, *Errors*, and *Accesses to help, are* presented in Table 6.

From this table, it can be seen that the user experience level had a more significant effect on

Table 5. Overlay of results obtained with the three evaluation techniques

PROBLEM CATEGORY	SI	PM	SM
Location and sequence of menu options	✓ **(05)**		× **(05)**
Menu navigation	✓ **(02)**	✓ **(01)**	
Presentation of menu options	✓ **(02)**		
Information feedback	✓ **(01)**		
Object manipulation	✓ **(05)**		
Symbols and icons	✓ **(02)**		× **(02)**
Text entry via stylus (Writing recognition)	✓**(07)**	✓ **(01)**	✓ **(08)**
Text entry via virtual keyboard		✓ **(01)**	✓ **(01)**
Processing power		✓ **(02)**	✓ **(02)**
Hardware issues		✓ **(03)**	✓ **(03)**
Fluent tasks execution		✓ **(05)**	✓ **(05)**
Online and offline help		✓**(01)**	✓ **(01)**

Legend:

SI—Standards Inspection
PM—Performance Measurement
× - Contradictory findings
✓ - Consistent findings

Table 6. Influence of the user experience on the performance indicators: Time, Number of errors, and accesses to help

Independent variable	Dependent variable	p-Value (Lab)	p-Value (Field)	Significance (α=0.05)
Experience	Task Time	0.081	0.081	Not significant
Experience	Errors	0.011	0.002	Significant
Experience	Help Accesses	0.427	-	Not significant

the number of errors in the field experiment than in laboratory experiment.

The studies in the literature fit basically into two categories: (1) user mobility, which means moving while using the device (inside of a laboratory or outdoors) and (2) user attention division. However, this study considers both aspects as part of the task context. In this experiment, the field test subjects were free to choose between moving or remaining still as they performed the task with the mobile device. During the informal interview the users stated that in a real context they would not perform the experiment tasks on the move**,** since they demanded too much attention. The specialist encouraged users to wander around the environment, although they could choose to enter a room in the building, sit down, or even lay the device on a table (which they did in most cases, under the argument that this setting was more comfortable). The movement registered was limited to situations in which the user waited for some device processing. (e.g., Web page downloads). There was a clear interference of the environment on the user attention during the field tests while moving.

The device's physical characteristics affected the user performance and the data gathering during the experiment**.** Outdoors, in ambient light, the device's legibility was reduced and aggravated by the reflections on the screen. According to the user's opinion stated during the informal interview, the camera apparatus did not interfere with the task execution, but the majority decided to lay the device down during task execution.

As for the entry of text information, the users showed a preference for the virtual keyboard instead of hand written character recognition. Based on their comments, as well as on the informal interview, it was concluded that writing long messages is very cumbersome both using the virtual keyboard and using the handwriting recognition application. Confirming previous findings, the experiment demonstrated that applications that require a lot of interaction and user attention are inappropriate for performing while walking due to attention division. This conclusion reinforces that, for the device targeted in this study, in spite of its mobility, the evaluation settings did not need to differ substantially from the one employed in the evaluation of stationary devices since the users tend not to wander while performing tasks that demand their attention or consisted of text input.

Until recently, studies have been published which deal with new paradigms and evaluation techniques for mobile devices. Few of the proposed new techniques are really innovative if compared to the ones traditionally employed. On the other hand, the main argument for proposing new techniques concerns the user and device mobility and the influence of this mobility on user performance. In contrast, this study evaluated the effect of mobility not only from the user performance perspective but also from user opinion point of view and the user level of satisfaction. From the application of the multi-layered approach, the data gathered and analyzed support the initial assumption that minor adaptations in the traditional evaluation techniques and respective settings are adequate to accommodate the evaluation of the category of mobile devices targeted by this study.

The conclusions corroborate with the views of the authors and that of Po (Po, 2003) that the laboratory and field evaluations do not diverge but are complimentary. As shown in this study, they both add to the evaluation process, producing data that is significant to the process and reinforcing the relevance of a multi-layered approach for the usability evaluation of mobile devices.

FUTURE TRENDS

Mobile devices impose challenges to the usability evaluation that are unique in respect to the observation strategies and the conception of test scenarios. With the continuous technological advances, a wider variety of new devices is being

released into the market, challenging users with the complexity of the interaction. In this scenario, the importance of the product usability is undisputable as is also the correct choice of evaluation methods, techniques, and tools.

One emerging trend in the mobile devices evaluation field is the possibility of gathering data in an unobtrusive way, using tools for remote, and automatic data capture that are transparent to the user. Developing those tools is a challenging activity given the inherent restrictions presented by the mobile devices (such as their limited processing power and limited storage capacity). But, in spite of the current limitations, it was shown in this study that the tools are becoming available to provide a great contribution to the evaluation setup and that these tools would benefit from further development.

REFERENCES

Aladwani, A., & Palvia, P. (2002). Developing and validating an instrument for measuring user-perceived Web quality. *Information & Management, 39,* 467-476.

Bailey, James E., & Pearson, S. W. (1983). Development of a tool for measuring and analyzing computer user satisfaction. *Management Science, 29*(5), 530-545.

Baillie, L., & Schatz, R. (2005). Exploring multimodality in the laboratory and the field. In *Proceedings of the 7th International Conference on Multimodal Interfaces* (pp. 100–107).

Danielson, D. R. (2006). Usability data quality. In C. Ghaoui (Ed.), *Encyclopedia of human-computer interaction* (pp. 661-667). Hershey, PA: Idea Group Reference.

De Oliveira, R. C. L., De Queiroz, J. E. R., Vieira Turnell, M. F. Q. (2005). WebQuest: A configurable Web tool to prospect the user profile and user subjective satisfaction. In G. Salvendy (Ed.), *Proceedings of the 2005 Human-Computer Interaction Conference. (The management of information: E-business, the Web, and mobile computing)* (Vol. 2) Nevada: Lawrence Erlbaum Associates (U.S. CD-ROM Multi Platform).

Dumas, J. S., & Redish, J. C. (1999). A practical guide to usability testing (revised ed.). Exeter, UK: Intellect.

EATMP. (2000). *Human factors integration in future ATM systems—Methods and tools* (Tech. Rep. HRS/HSP-003-REP-03). European Organization for the Safety of Air Navigation—European Air Traffic Management Programme. Retrieved August 13, 2006, from http://www.eurocontrol.int/humanfactors/gallery/content/public/docs.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213-236.

Goodman, J., Brewster, S., & Gray, P. (2004). Using field experiments to evaluate mobile guides. In *Proceedings of 3rd Annual Workshop on HCI in Mobile Guides* (pp. 1533-1536).

Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction, 15*(1), 145-181.

Hilbert, D. M., & Redmiles, D. F. (2000). Extracting usability information from user interface events. *ACM Computing Surveys, 32*(4), 384-421.

Hix, D., & Hartson, H. R. (1993). *Developing user interfaces: Ensuring usability through product & process.* New York: John Wiley and Sons, Inc.

Holzinger, A. (2005). Usability engineering methods (UEMs) for software developers. *Communications of the ACM, 48*(1), 71-74.

ISO 9241-11. (1998). *Ergonomic requirements for office work with visual display terminals*

(VDTs)—Part 11: Guidance on usability. International Organization for Standardization, Geneva, Switzerland.

ISO 9241-14. (1997). *Ergonomic requirements for office work with visual display terminals (VDTs)—Part 14: Menu dialogues.* International Organization for Standardization, Geneva, Switzerland.

IS09241-16. (1999). *Ergonomic requirements for office work with visual display terminals (VDTs)—Part 16: Direct manipulation dialogues.* International Organization for Standardization, Geneva, Switzerland.

IS09241-17. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs)—Part 17: Form filling dialogues.* International Organization for Standardization, Geneva, Switzerland.

ISO 13407. (1999). *Human-centered design processes for interactive systems.* International Organization for Standardization, Geneva, Switzerland.

ISO. (2006). *ISO and conformity assessment.* Retrieved September 23, 2006, from http://www.iso.org/iso/en/prods-services/otherpubs/pdf/casco_2005-en.pdf.

ISO/IEC 14754. (1999). Information technology —pen-based interfaces—common gestures for text editing with pen-based systems. International Organization for Standardization, Geneva, Switzerland.

ISO/IEC 18021. (2002). *Information technology—user interfaces for mobile tools for management of database communications in a client-server model.* International Organization for Standardization, Geneva, Switzerland.

ITU-T E.161. (2001). *Arrangement of digits, letters and symbols on telephones and other devices that can be used for gaining access to a telephone network.* International Telecommunications Union-telecommunications, Geneva, Switzerland.

Jones, M., Marsden, G. (2006). *Mobile interaction design.* Chichester, West Sussex: John Wiley and Sons, Inc.

Kan, S. H. (2002). *Metrics and models in software quality engineerin*g (2nd ed.). Reading, MA: Addison-Wesley Professional.

Kjeldskov, J., Graham, C., Pedell, S., Vetere, F., Howard, S., Balbo, S., & Davies, J. (2005). Evaluating the usability of a mobile guide: The influence of location, participants and resources. *Behavior and Information Technology, 24*(1), 51–65.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal on Human and Computer Studies, 60*(5-6), 599–620.

Mayhew, D. J. (1999). *The usability engineering lifecycle.* San Francisco: Morgan Kaufmann Publishers Inc.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., et al. (1998). Comparative evaluation of usability tests. In *Proceedings of the Usability Professionals Association Conference* (pp. 189-200).

Nielsen, J. (1993). *Usability engineering.* Boston: Academic Press.

Omodei, M. A., Wearing, J., & McLennan, J. P. (2002). Head-mounted video and cued recall: A minimally reactive methodology for understanding, detecting and preventing error in the control of complex systems. In *Proceedings of the 21st European Annual Conference of Human Decision Making and Control.*

Po, S., Howard, S., Vetere, F., & Skov, M. B. (2004). Heuristic evaluation and mobile usability: Bridging the realism gap. In *Proceedings of the Mobile HCI* (pp. 49–60).

Rosson, M. B., & Carroll, J. M. (2002). *Usability engineering: Scenario-based development of human-computer interaction*. San Diego, CA: Academic Press.

Sanderson, P., & Fisher, C. (1994). Usability testing of mobile applications: A comparison between laboratory and field testing. *Human-Computer Interaction, 9,* 251–317.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives in curriculum evaluation* (pp. 39- 83). Skokie, IL: Rand McNally.

Tabachnick, B. G., & Fidell, L. S. (2006). *Experimental designs using ANOVA* (1st ed.). Duxbury Applied Series. Pacific Grove, CA: Duxbury Press.

Wixon, D., & Wilson, C. (1997). The usability engineering framework for product design and evaluation. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (2nd ed.) (pp. 653-688). New York: John Wiley and Sons, Inc.

KEY TERMS

Conformity Assessment: A collective term used for a number of techniques used to determine if a product, system, or process (including design) meets a defined specification.

Device Mobility during a Usability Evaluation: The ability to interact with the user and continue to perform its functions while being transported.

Efficacy of an Evaluation Method or Technique: Translated into the number of problems found, gravity of those problems versus the time, and cost of performing the experiments.

Likert Scale: An attitude scale in which respondents indicate their degree of agreement/disagreement with a given proposition concerning some object, aspect, person, or situation.

Multi-Layered Evaluation Approach: A product or prototype usability evaluation method that combines techniques for data gathering and analysis based on multiple perspectives (the user's, the specialist's, and the usability community). The results are overlaid in order to find discrepancies and offer more robust results.

User Mobility during the Usability Evaluation: The ability to move while performing a task with a product.

User Performance Measurement: The process of gathering actual data from users as they work with a system and its documentation. Usually, the user is given a set of tasks to complete and the evaluator measures the relevant parameters such as the percentage of tasks or subtasks successfully completed, time required to perform each task or subtask, frequency and type of errors, duration of pauses, indications of user frustration, and the ways in which the user seeks assistance.

User Satisfaction Measurement: The process of obtaining qualitative and quantitative information which indicates the extent to which user expectations concerning some object, process, product, or situation are being met. Such information can be obtained in a variety of ways, both formally and informally.

Virtual Network Computing (*VNC*): A desktop sharing system that uses the RFB (Remote Frame Buffer) protocol to remotely control another computer. It transmits the keyboard presses and mouse clicks from one computer to another over a network, relaying the screen updates back in the other direction.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 847-862, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 7.40
Mobile Information Processing Involving Multiple Non-Collaborative Sources

Say Ying Lim
Monash University, Australia

David Taniar
Monash University, Australia

Bala Srinivasan
Monash University, Australia

ABSTRACT

As more and more servers appearing in the wireless environment provide accesses to mobile users, more and more demand and expectation is required by mobile users toward the available services. Mobile users are no longer satisfied with obtaining data only from one server, but require data from multiple servers either at the same or different locations. This eventually leads to the need for information gathering that spans across several non-collaborative servers. This article describes some of our researches in information gathering from multiple non-collaborative servers that may involve servers that not only accept direct queries from mobile users but also servers that broadcast data. We also look at how location dependent data plays an important role to mobile information gathering.

INTRODUCTION

The direction of the mobile technology industry is beginning to emerge and advance at a rapid pace as more mobile users have evolved (Myers & Beigl, 2003). Interests in mobile technology have grown exponentially over the last few years and are greatly influenced especially by the dramatic reduction in the cost of hardware and protocol standardization (Hurson & Jiao, 2005; Kapp, 2002). The increase in progression and advancement of

mobile technology has created a new paradigm of computing called mobile computing in which people are allowed to be connected wirelessly to access data anytime, anywhere without having to worry about the distance barrier (Lee, Zhu, & Hu, 2005; Lee et al., 2002; Madria, Bhargava, Pitoura, & Kumar, 2000). Users have also become more productive with the achievement of mobility since they are able to access a full range of resources regardless of where they are located and where they are able to get hold of real time information.

The emerging growth of the use of intelligent mobile devices (e.g., mobile phones and PDAs) opens up a whole new world of possibilities, which includes delivering information to mobile devices that are customized and tailored according to their current location (Gutting et al., 2000; Tsalgatidou, Veijalainen, Markkula, Katasonov, & Hadjiefthymiades, 2003; Xu et al., 2003). Mobile queries are requests for certain information that are initiated by mobile users to the appropriate servers from their mobile devices. Query processing in a mobile environment may involve join processing from either single or several different servers with the mobile devices (Liberatore, 2002; Lo, Mamoulis, Cheung, Ho, & Kalnis, 2003). In addition, mobile queries can be performed regardless of where the users are located and the results obtained are influenced by the location of the user. Data that are downloaded from different locations would be different and there is a need to bring together these data according to a user who may want to synchronize the data that are downloaded from different location to be consolidated into a single output. Thus, the intention is to take into account location dependent factors, which allow mobile users to query data without facing location problems (Song, Kang, & Park, 2005; Tse, Lam, Ng, & Chan, 2005; Xu, Tang, & Lee, 2003). This concept is associated with location dependent query.

One of the main objectives of this article is to demonstrate the importance of allowing mobile users who believe that obtaining data from a single server is not enough and may need further processing with data that are obtained from other servers. Furthermore, the user may get data from several servers that are from the same or different providers. In other words, there are times when the user has the desire to gather data from several non-collaborative servers into their mobile devices (Lo, et al, 2003; Malladi & Davis, 2002). Mobile devices have made it capable for mobile users to process and retrieve data from multiple remote databases by sending queries to the servers and then process the multiple data gathered from these sources locally on the mobile devices (Mamoulis, Kalnis, Bakiras, & Li, 2003; Ozakar, Morvan, & Hameurlain, 2005). By processing the data locally, mobile users would have more control over what they actually want as the final result of the query. They can therefore choose to query data from different servers and process them locally according to their requirements. Also, by being able to obtain specific data over several different sites, it would help bring optimum results to mobile user queries. Furthermore, by driving away the computation on the client device, the bandwidth computation may also be reduced.

Example 1: *A mobile user may want to know where the available vegetarian restaurants are in the city he or she is currently visiting. There are two major servers (e.g., tourist office and the vegetarian community) that may give information about the available vegetarian restaurants. First, using his or her wireless PDA, he or she would download information broadcast from the tourist office. Then, he or she would download the information provided by the vegetarian community. After obtaining the lists from the two information providers, he or she may perform an operation on his or her mobile device that joins the contents from the two relations obtained earlier from the two non-collaborative organizations. This illustrates the importance of assembling information obtained from multiple non-collaborative sources*

in a mobile device in order to obtain more comprehensive information.

This article investigates the need for information gathering spanning several non-collaborative servers that may bring to mobile users. Furthermore, due to the various nature of how a server may disseminate their data (e.g., through ad-hoc queries or data broadcasting), this article also evaluates the query processing methods involving the previously mentioned strategies.

In this article, we first present an insight of the background of mobile environment, non-collaborative servers, and prospective applications for information gathering from multiple sources. By formulating the taxonomy, it helps to give understanding of the possible database operations that can be performed on the mobile devices. We will also describe the process of information gathering that results from multiple servers that involves location dependent data, which are then followed by a system prototype. Finally, the last section concludes the article. Note that in this article we use the term mobile client, mobile user, and users interchangeably.

BACKGROUND AND PRELIMINARIES

Before discussing more details, the process of information gathering, and its rationale, this section would first introduce some background and preliminaries related to mobile query processing in a typical mobile environment, which involves multiple non-collaborative servers. Firstly, this section provides some introductory knowledge on the wireless environment covering what constitutes the architecture of a typical mobile computing environment, followed by the usefulness of obtaining information from multiple sources and lastly the prospective application of this study.

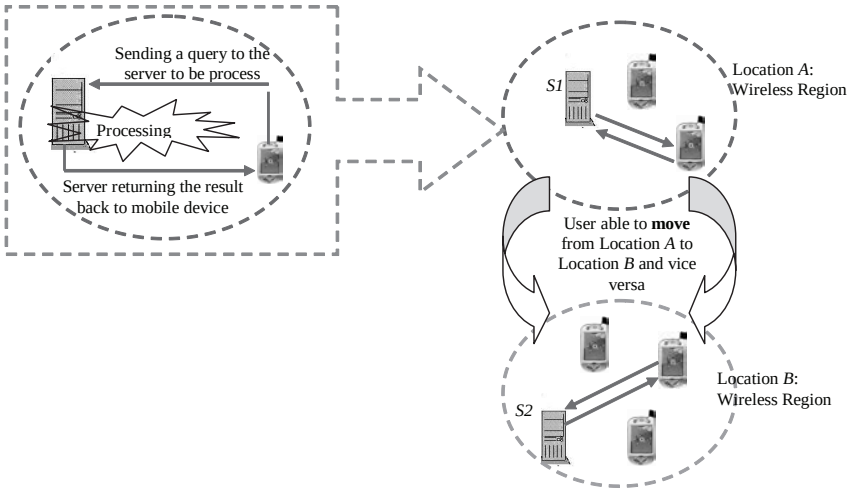
Mobile Computing Environment: A Background

Mobile computing has provided mobile users the ability to access information anytime, anywhere. It enables mobile users to query databases from their mobile devices over the wireless communication channels (Imielinski & Badrinath, 1994). In general, mobile users with their mobile devices and servers that store data are involved in a typical mobile environment (Lee et al., 2005; Madria et al., 2000; Wolfson, 2002). Each of these mobile users communicates with a single server or multiple servers that may or may not be collaborative with one another. This server is also known as mobile base station (MBS), which the mobile users communicate to in order to carry out any activities such as transaction and information retrieval. The servers supply its services to a wide range of users who are within the active region through a wireless interface. Thus, mobile users have to be within a specific region to be able to receive a signal in order to connect to the servers.

Figure 1 depicts a mobile environment architecture where two servers are involved, *S1* and *S2*, in location *A* and location *B* respectively. Mobile users move freely within the different region to obtain different data by accessing the different servers via sending a query and receiving the results back to the mobile device upon completion of processing.

Example 2: *A property investor, while driving his or her car, downloads a list of nearby apartments for sale from a real-estate agent. As he or she moves, he or she downloads the requested information again from the same real-estate agent. Because his or her position has changed since he or she first inquires, the two lists of apartments for sale would be different due to the relative location when this investor was inquiring the information. Based on these two lists, the investor would prob-*

Figure 1. Mobile environment architecture



ably like to perform an operation on his or her mobile device to show only those apartments that exist in the latest list, and not in the first list.

Hence, in a typical mobile environment, it is unacceptable to meet with situations where a mobile user is currently obtaining data from a current active region and is still not feeling satisfied with the result. This leads to the need of further processing with other data that can only be obtained from other servers that may or may not be collaborative to each other.

Overview of Non-Collaborative Servers

The term collaborative usually relates to the traditional distributed databases whereby the desire to integrate the data of a particular enterprise and to provide centralized and controlled access to that data (Bell & Grimson, 1992; Ceri & Pelagatti, 1984; Özsu & Valduriez, 1999). The technology of distributed database may not be appropriate for use in the mobile environment, which involves not only the nomadic clients, which move around, but also non-collaborative servers, which are basically servers that are maintained by different organizations (Lo et al., 2003).

Therefore, non-collaborative servers would refer to servers that do not know each other and do not have any relation between one another. There are basically just individual server providers, which disseminate data to the users and they do not communicate with one another. Since each server can just be an independent service provider, often these independent servers are specialized within the domain of the information they are providing.

An example of such a server that disseminates information on restaurants normally just focuses on the restaurants information and limited supporting information, which can sometimes be included (e.g., how to get there--the transportation is just supporting information since it does not exactly show the route on how to get there from a particular location the user is currently

at). Therefore, there is still a need to obtain full information from multiple servers, which in this case are the restaurant and transportation servers separately.

In addition, not all service providers are supported by the usage of a mediator (Lo et al., 2003). Therefore, information obtained from other independent non-related service providers needs to be processed individually. It is not a fair assumption that all service providers are linked through a mediator. Hence, in our research, we focus on independent service providers, which refers to non-collaborative servers. Thus, it is vital to consider gathering information from non-collaborative servers because it is often not enough to just get data from a single server.

Prospective Applications

Information gathering in the context of mobile environment is a source where collecting various data together regardless of whether it is related or not related as long as it is useful for the mobile users. There exists several significant influences of information gathering from multiple sources that lies on personal applications. Next, we show summarized lists of some promising applications that bring great impacts to mobile users.

- **Entertainment applications:** Shopping appears to be a popular trend and hobby. Often shoppers would prefer to go to shops that give the lowest price for the item they are interested in buying. Many shops in the same shopping complex may sell the same item, but they are different companies and they are not related with one another. Thus, with the ability of getting the information, especially the pricing for the desired item, separately from the various shops could aid users in deciding which shops to go to that offers a better price.

Example 3: *A mobile user who is currently in a shopping complex is interested in a buying a tennis racquet. There are two different sports shops in the complex, sports shop A and sports shop B, that sell the tennis racquet that the user wants to buy. So first, by sending a query to shop A, he or she obtains a list of the prices for the tennis racquet. Then he or she sends a query to shop B, which again he or she will get a list of prices for the tennis racquet. So with these two lists, the mobile user can do a local processing, which compares the matching racquets and displays the shop that gives a lower price for the respective racquets that are being matched.*

- **Tourism applications:** Tourism brings value added in terms of economical growth to not only the country, but also the physical relationship between the visitor and the producer of a good or service. Tourism is an important element to boost the country's reputation and economy. Thus, it is important to give both local and international travelers the best and most convenient. Giving the ability of information gathering from multiple sources tends to emerge as valuable services to the mobile travelers regardless of where their current geographical coordinates.

Example 4: *An international tourist, while traveling to a foreign country, does not know the where abouts of the tourist attraction spots. He or she looks for famous tourist spots recommended by both the transport office and tourism office. First, using his or her wireless PDA, he or she would download information broadcast from the tourism office to get a list of the famous tourist spots. Then, he or she would download the information provided by the transport office to get information on the available transportation. Once he or she obtains the lists from the two information providers, he or she may perform an operation on his*

or her mobile device that joins the contents to match the tourist spots together with transport information on how to get there.

- **Emergency responses applications:** There are times especially when someone on the highway is having trouble with their car and needs to find the nearest possible petrol station that offers car services as soon as possible. In this circumstance, the person on the highway can use his or her mobile device to make a query as he or she travels along the highway to look for a petrol station that offer car services. This comes into the category of emergency cases, as it is rare and is not needed all the time.

Example 5: *A traveler currently in location A wants to know where the nearest gas station is (petrol kiosk) and using the mobile device, they downloaded a list of available petrol kiosk nearby to his or her current surrounding location. As they travel further until they arrive in location B, he or she makes another query to get another list of petrol kiosk, but this time the list is somewhat different since he or she has been driving and the location has moved from A to B. Therefore, based on these two lists, the traveler wants to display only those petrol kiosks that provide car service regardless of whether it is in A or B.*

- **Double checking applications:** Data that are stored in the servers that are to be disseminate to the public can sometimes be outdated due to the company that manages the data being closed down or other undesired catastrophes. If the users are still able to query for the data, part of the data may not be accurate anymore since it has not been maintained and updated well. This will make the data worthless if the users download it. Therefore, a certain degree that allows the users to see the data that is downloaded are obtained from a reliable source or not may be useful if the ability of processing data that are obtained from one server together with another list of data obtained from another server as a double checking precaution.

Example 6: *This example requires an assumption of one property that can be handled by several estate agents. A user obtains a list of properties in the city that are ready for sale from real-estate agent A. Without knowing, real-estate agent A has just been declared bankrupt and the lists that are currently in the server have not been updated since. Thus, some of the properties that the user has downloaded have actually been sold. Without knowing all this, since the user is able to obtain another list of properties in the city from another agent, which is agent B, this list would have been able to be used as a reference list to the previous list that was obtained from agent A. Since one property can be handled by several agents, the properties for sale in the city between list A and list B should be the same. The only difference may be the price on whether the agent is selling it cheaper or more expensive than the other. Thus, by seeing the difference in the availability of the properties between the two lists, this information can appear to the mobile user that one or the other is not correct since we have to assume one property is to be handled by several agents.*

In summary, we can see from the previous sample application domain that obtaining from a single place is not sufficient enough to provide the desired results to the mobile users. The mobile users often require several data that are non-related with one another to be gathered and processed together so that a higher level and meaningful information can be obtained. By giving more flexibility to the mobile users to "mix-and-match" non-related data from several servers proves to return a more comprehensive result that is able to satisfy the needs of the users. Therefore, information gathering from multiple non-collaborative servers brings benefits and gives a good prospect

for users to achieve a higher quality and productive information.

MOBILE USER QUERIES

The context of mobile user queries in this article is that the mobile queries contain operations that are being carried out when multiple lists of data are obtained from multiple servers (Lim, Taniar, & Srinivasan, 2006). In this section, we will present a taxonomy of the mobile user queries in two elements namely (i) *non-location-based on-mobile queries* and (ii) *location-based on-mobile queries.*

- In non-location-based on-mobile queries, the need to obtain constructive information often requires mobile users to download lists from multiple sources to be integrated and processed together. In a mobile environment, joins are used to bring together information from two or more different information sources. It joins multiple data from different servers into a single output to be displayed on the mobile device. The idea of this is basically to ensure mobile users have the ability to reduce the query results with maximum return of satisfaction because with the additional post-processing, the output results can be greatly reduced based on the user's requirements and needs before the final display on the device.

Consider Example 1 presented earlier where it shows how a join operation is needed to be performed on a mobile device as the mobile user downloads information from two different sources, which are the tourist office and the vegetarian community. In this case, two pieces of information might be joined on the restaurant IDs from the two different lists. This therefore, illustrates a simple on-mobile join case, where it is basically a process of combining data from one relation to another.

- Location-based on-mobile queries have become a growing trend due to the constant behaviour of mobile users who move around. Location-dependent processing is of interest in a number of applications, especially those that involve geographical information systems (Cai & Hua, 2002; Cheverst, Davies, & Mitchell, 2000; Jung, You, Lee, & Kim, 2002; Tsalgatidou et al., 2003). An example query might be "to find the nearest petrol kiosk" or "find the three nearest vegetarian restaurants." As the mobile users move around, the query results may change and would therefore depend on the location of the issuer. This means that if a user sends a query and then changes his or her location, the answer of that query has to be based on the location of the user issuing the query (Seydim, Dunham, & Kumar, 2001; Waluyo, Srinivasan, & Taniar, 2005). Location dependent processing involves the circumstances when mobile users are in the situation where they download a list when in a certain location and then they move around and download another list in their new current location. Or another circumstance might be a mobile user might already have a list in his or mobile device but moves and needs to download the same list again but from a different location. In any case, there is a need to synchronize these lists that have been downloaded from a different location.

Consider Example 2 presented earlier. It shows an example of how location dependent queries processes are being carried out. With the two different lists on hand that the investor currently had based on the properties in the two different locations, the investor would probably like to perform some kind of database operation on his

or her mobile device. The difference in the list is due to the moving location from one point to another point by the investor.

MODELS OF INFORMATION PROCESSING

There are times when a user may need to query several non-collaborative servers in order to obtain a more comprehensive list of data. The user may need to perform some database operations locally on the mobile device based on the list of data that has been downloaded from the remote databases.

Figure 2 models the various strategies that the server can adopt. *Server strategy* involves mobile users sending queries to the server for processing (Seydim et al., 2001). It relates to processing to be taken by the server to process and return the results based on the mobile user queries. *On-air strategy* is similar to traditional broadcasting

Figure 2. Query processing strategies

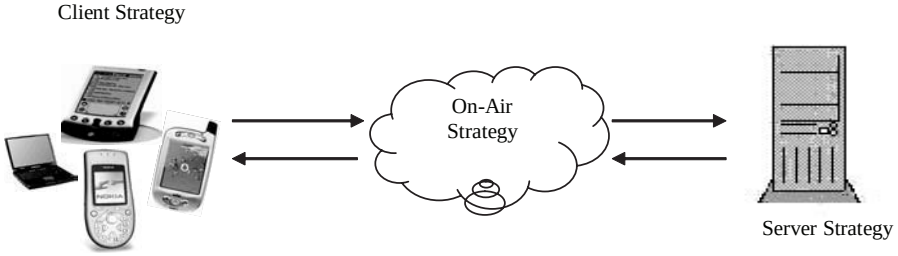
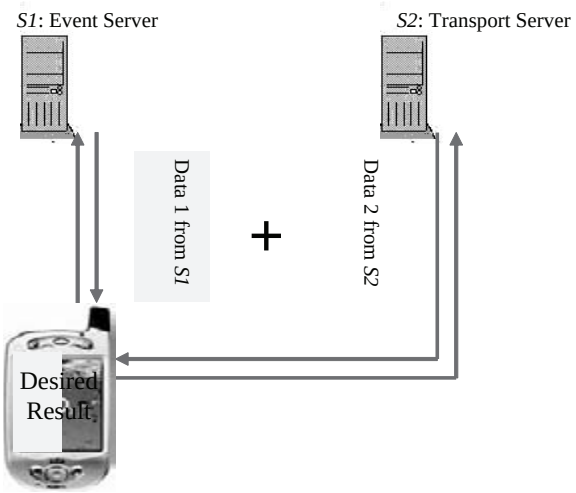


Figure 3. Example of assembling information from two different servers



techniques whereby the sets of database items are broadcasted through the air to a large number of mobile users over a single channel or multiple channel (Tran, Hua, & Jiang, 2001; Triantafillou, Harpantidou, & Paterakis, 2001). With the set of data on the air, mobile users can tune into one or more channel to get the data. *Client strategy* relates to maintaining cached data in the local storage and being able to have the ability to do local processing if queries results are being sent back to the mobile device and stored in the cache memory. Thus, efficient cache management is critical in mobile query processing (Cao, 2003; Elmargamid, Jing, Helal, & Lee, 2003; Xu, Hu, Lee, & Lee, 2004; Zheng, Xu, & Lee, 2002).

Lists of information can be obtained from servers that distribute their respective data using various strategies such as server strategy, on-air strategy, and client strategy. Each of the available servers has their associated query processing strategies and they can be processed together regardless of whether part of the servers use a different strategy.

Example 7: *Suppose a mobile user wants to know the timetable for the transportation services to a particular event. Each of the transportation timetables, as well as the event, is stored in different servers and maintained by two different organizations. Transport servers would deal with transportation data while an event server would deal with current events that are happening. Therefore, in order to know the transportation timetable for a particular event, the user has to gather data from the two different servers, which is first sending a query to obtain the event list into the mobile device, and then sending another query to the transport server to obtain the list of transportations. Now these two lists are in the mobile device and are ready to be processed locally to match the transportation timetable onto the respective events. This exemplifies the importance of assembling information from multiple servers into a single information, which is the desired result as the outcome on the PDA.*

Assuming that both transport and event servers are individual servers that accept direct query from the users, Figure 3 models an illustration of how two different lists are obtained from two different sources to be processed locally. This achieves the object of processing information obtained from multiple non-collaborative servers.

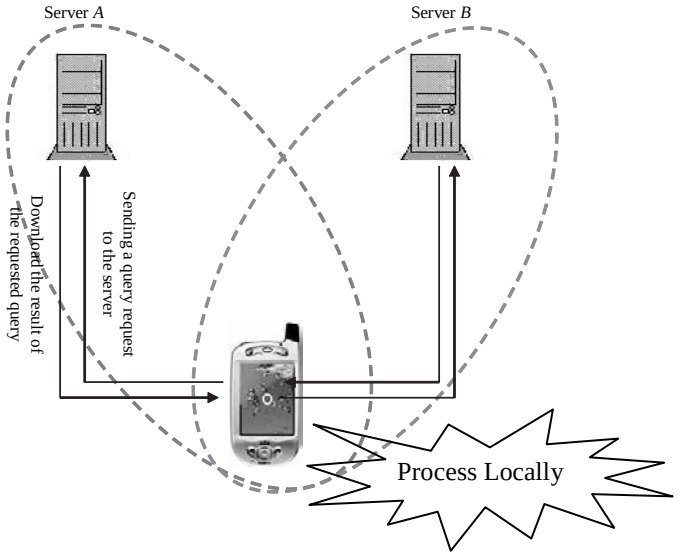
In the following sub section, we will uncover several case studies and explain how multiple non-collaborative servers that use the different strategies integrate its results into useful information for the mobile users. Just for illustration and simplicity purposes, we only illustrate situations where there are only two servers that are in use.

Case Study 1: Both Servers use Server Strategy

Without acknowledging the current standing location of the user, we would like to allow the user to be able to carry out simple database operations locally on the mobile device such as simple join between the different lists of data that are downloaded into the mobile device from the remote databases. We would first examine cases where the two different servers that the users need to obtain its information from to be integrated together are both using server strategies as being modeled in Figure 4.

Server strategy has limited functionalities since it provides dedicated point-to-point connections in accepting the mobile users request directly. This is due to the limited bandwidth that is available. Therefore, if suddenly there are many users wishing to send a request to the same server, the server may be congested. Thus, server strategy may cause an increase in exceeding usage of bandwidth especially when too many data requests are being sent out by the mobile users.

Figure 4. Example of on mobile query processing



The overwhelming mobile users requests may affect the query performance. This can easily cause a scalability bottleneck with a large mobile user population.

As far as the cost remains a major concern to a wide majority of mobile users, obtaining data via server strategy may be expensive or cost effective. This is because mobile users are establishing a direct communication to the server, which is how server strategy provides exclusive point-to-point communications between the user and server, which in this case the server processes the query that is being sent by the mobile users and returns the results back to the mobile users (Sun, Shi, & Shi, 2003).

In addition to the previous issues that users may face when obtaining data from servers that accept direct requests, there are several other additional complexities such as deciding which servers to download first in order to reduce memory consumptions and minimize transfer costs as well worthless or unnecessary data transfers. The techniques obtained from servers indicate there is a need to download in advance at least a list of data from one side of the server to the mobile device. Due the limitation of memory, it would be wise to use a technique that is able to utilize the minimum memory. Both response and access time are also a major concern because they may slow down the results from the query especially when the number of requesting queries is increasing.

Case Study 2: Server Strategy and On-Air Strategy

There are situations when certain data are broadcast on a public wireless channel, which requires the user to tune into the broadcast channel to filter out the relevant data. Users have no control in issuing queries directly to the servers. Therefore, we are concerned with how users can efficiently

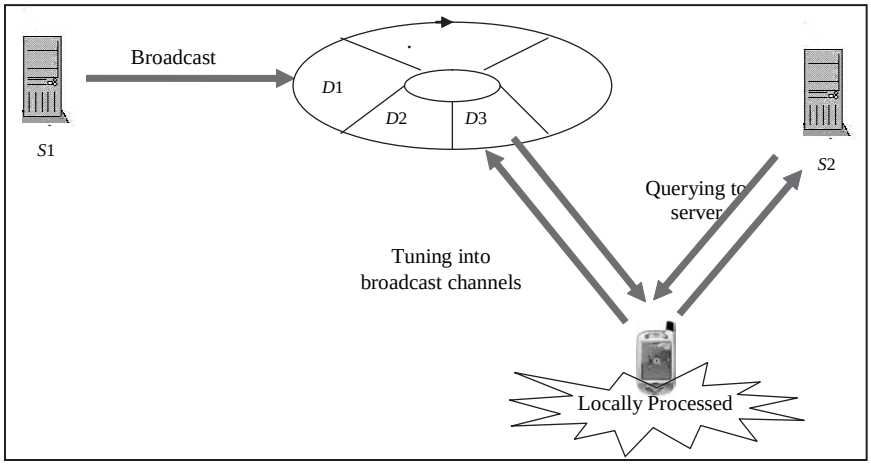
obtain their specific request without being able to send a query to the servers that use the broadcasting system. In this system, by broadcasting it actually lets an arbitrary number of users access the data simultaneously. Therefore, this may be acquainted with issued of over population in accessing a particular data that may slow down the access. In other words, when encountering a multiple non-collaborative servers setting, some servers may not able to accept queries that are directly issued by the mobile users but rather provide data broadcasting.

Figure 5 illustrates an example of a server that supports data broadcast in conjunction with another server that supports direct query. Example, in order to gather data from the two non-collaborative servers, whereby one server supports data broadcast and another server supports direct query, the mobile users would need to tune into the wireless channel to obtain the desired data from the server that supports data broadcast and issue a query to the other server, which accept direct query.

When determining the on-air strategy to be used over server strategy or client strategy, one important issue is to determine an optimal broadcast sequence for the data items that are to be distributed to the mobile users. This refers to data broadcast scheduling, which one must look at in order to have minimal access time and minimal tuning time in receiving the required data items. By prioritizing the data items and using a good selection mechanism can reduce the broadcast cycle length, which eventually is able to reduce the query response time (Chung et al., 2001; Lee, Lo, & Chen, 2002). The data items can be characterized as both "hot" and "cold" and this can be the determinant of which data item should be given a higher priority over the others (Zhang & Gruenwalk, 2002).

Another alternative to the selection mechanism, in taking into account of reducing response time, is to have more than one broadcast channel whereby the broadcast data can be distributed to more than one broadcast channel. In most cases, data items are broadcast over a single channel as it avoids additional issue of the organization of data and allocation while having more than one channel (Imielinski, Viswanathan, & Badrinath, 1997). Furthermore, the use of a single chan-

Figure 5. Example of on mobile querying broadcasted data



nel appears to be more problematic especially when there are a large number of data items to be broadcast, thus, with the adoption of multiple channels to broadcast data, the chance of reducing long delays before obtaining the desired data items can be achieved.

The next factor that could help reduce response time would be concerning the organization of the data items especially when retrieving multiple data items is required. An illustration of such a situation can be a mobile client wants to send a query to retrieve multiple stock prices concurrently. This is an example of multiple data items retrieval and in order to retrieve such query in a more efficient way, the need to consider the semantic relationship between the data items is required (Chung et al., 2001; Ren & Dunham, 2000). However, in order to predict which data item that the mobile client would be interested in next is difficult because there is not much knowledge of any future query that is available. Existing related work has investigated the use of access graphs to represent the dependency of the data items (Lee et al., 2002). Other existing algorithms that have been investigated to identify the most effective organization of the data items includes heuristics algorithm (Hurson et al., 2005) and randomized algorithm (Bar-Noy, Naor, & Schieber, 2000).

The last possible deciding factor for query response time can be determined by incorporating broadcast indexing scheme. Indexing scheme can reduce tuning time for the mobile client to access their required data item (Lee, Leong, & Si, 2002). By applying this scheme, mobile clients can conserve their battery life and thus, results in energy saving because the clients can switch to "doze" mode and back to "active" mode only when he or she knows the desired data item is about to arrive.

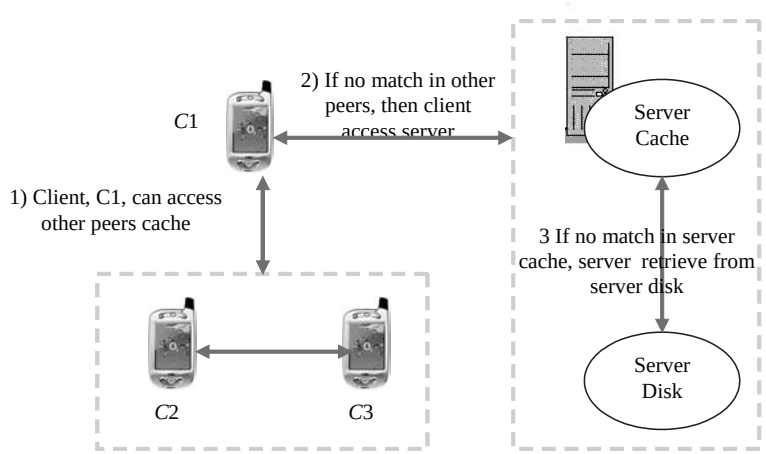
Although the on-air strategy appears to be more scalable in comparison to the server strategy, there are still limitations that it brings to the users because most users would find it easier to send a direct request to a specific server. In addition, since the data that are being broadcast are usually open to the public, there are privacy issues that may arise. Thus, if the mobile user wants to obtain private data, they would have to rely on the server strategy rather than the on-air strategy. Therefore, by being able to incorporate strategies that are able to accommodate the request of the users to bring it more flexibility like how server strategy does for the users maybe beneficial.

Case Study 3: Server Strategy and Client Strategy

Caching frequently accessed data in a client's local storage becomes prominent in improving the performance and data availability of data access queries (Chan, Si, & Leong, 1998). This is made available by caching the frequently accessed data items in the local mobile device storage as well as when frequent disconnection occurs, the query can still be partially processed from caches and at least some of the results from the previous queries can be returned to the users (Lee et al., 2002). This is because the mobile device is able to keep the existing data and if the user needs the same exact data, the downloading can be minimized if the mobile device recognizes that the data has been previously loaded into the device. Caching at the mobile client helps in relieving the low bandwidth constraints imposed in the mobile environment (Kara & Edwards, 2003). Issues that characterize the caching mechanism would include cache granularity, cache coherence, and cache replacement.

Figure 6 shows whenever a user issues a query, it first searches its cache and if a valid copy is found in the cache, it will return the results immediately. Otherwise it can also search the client's other local cache for the required results or it can be obtained either through the server or broadcast (on-air) strategy. Thus, it is often important to have cache management because often a user may download similar data repeatedly from the same source.

Figure 6. Example of on mobile querying with cache data



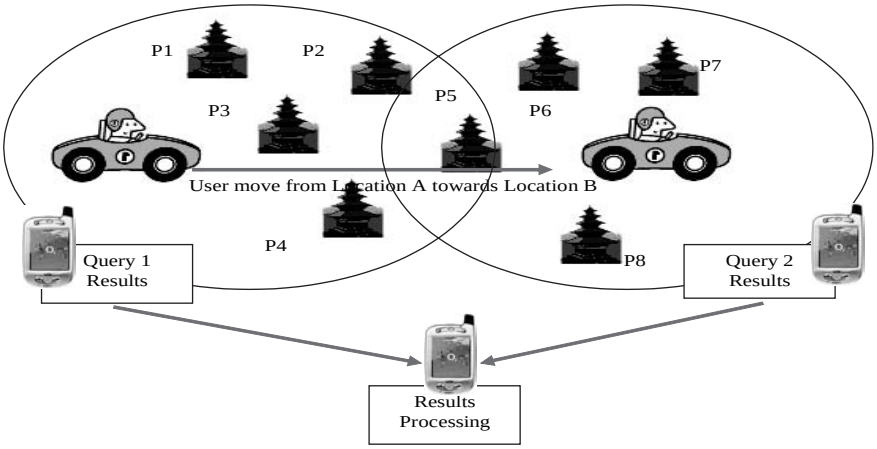
In general, the main limitation that limits the ability of having to cache everything on the mobile device mainly lies on the limited memory capacity. Therefore, one of the challenges that concerns the local cache memory is to exploit algorithms that can maximize the cache capacity to reduce the repeated transfer cost as much as possible and increase the respond time to the user's queries request. As we are concerned, the existing works on caching for mobile devices are still not sufficient for the new nomadic types of queries. A vast range of existing has been greatly being done on the issue of cache replacement and cache granularity and employing them into several possible cases in the real mobile environment situations. Index caching has been popular to save memory caching to improve query response time as well as managing space more efficiently, which is significant for location dependent queries. A few related works to this have also been done (Xu et al., 2003; Xu et al., 2004; Zheng et al., 2002). We also need to identify which mobile device has the request cache data that the other mobile users can access, as well as to make sure the cache data is still up to date before allowing the other interested users to access it (Elmargamid et al., 2003).

LOCATION DEPENDENT QUERY PROCESSING

Whenever a user moves from one location to another, the objects being queried can turn out to be different according to their geographical coordinates. Hence, location dependent plays an important role (Saltenis & Jensen, 2002; Sistla, Wolfson, Chamberlain, & Dao, 1997; Waluyo et al., 2005). It is important to show the mobile user, who is moving from one location to another location frequently, that the queries he or she sends, depends on the location that he or she is querying from.

Figure 7 shows that when a user is in location *A*, the query results are P1,P2, P3, P4, and P5, but as he or she moves toward location *B* and send the

Figure 7. Example of on mobile location dependent query processing



query again, the results now are P5, P6, P7, and P8. This shows that the change of user's query results are reflected by the change of the user's location. Basically with this, two query results that are obtained from Query 1 and Query 2 respectively, the user would like to process them together on the mobile device to obtain the just the desired data.

The main problem in this location dependent are the consistency problems that may arise, especially when the database is updated in the midst of processing a query. Furthermore, in order to ensure the shortest path, whereby we will first go for the server that stored the desired results, which can be obtained faster and easier or that can be obtained nearest to our current geographical coordinate compared to obtaining from another server. So, it is crucial to ensure the mobile users are able to obtain the shortest path to the desired destination. Some other issues would involve information processing when the results of the queries are obtained from different locations. We also need to learn how we are able to select the server to perform the query that is available from different locations so that we can answer the query more efficiently and correctly.

One important issue can be illustrated as follows. For instance, there are two locations, *A* and *B*. The user has a query in mind and based on the optimization in regards to the shorter path theory, the user will need to download information from location *B* and then send to location *A* for processing. But at the moment, the user is in location *A* and is going to location *B* soon. So, if following the optimization theory, this will require the user to first go to location *B* to download the desired data, and then move back to location *A* even though the current location of the user is now in *A*. In this method, there may arise a risk; maybe by the time the user goes to location *B*, the desired item is no longer available or there may have been network congestion and so on. If this occurs, the user would waste their time going to location *B* to get the desired data. Thus, we may need a different processing method where the user can start downloading from location *A* or just the partial key until the cache is full, and then when going to location *B*, send the request

to *B* to download the remaining items based on the key that is already in the cache.

For example, in Case Study 1, which involves querying to different servers that are utilizing the server strategy, it is obvious that in order to minimize the cost, we need to first download the list that has fewer records and then send it to the other server for matching. But if the location aspect is there, it might not be possible to choose which server to access first because it is dictated by the relative location. So this arises one important issue that needs to be looked at to what are the other aspect besides selecting servers that contain fewer records to be downloaded first.

Another example, based on Case Study 2, where we involve two servers in different locations, where the server in location *A* does not accept queries but the server in location *B* accepts queries. The issue may occur if the user is required to get the data from location *B* first, which accepts queries and then send to the server in location *A* that does not accept queries. This may create a problem because once we obtain the desired results from location *B* through sending a query to request the desired data, the results will then be sent to our mobile device. However, with this data, we are required to send to the server in location *A*, maybe to perform some comparison but that server does not accept queries, but only broadcasts data out to the mobile user. So our problem would be how can we then select another better technique that can reverse the situation so that we can get tune into the broadcast channel to get the data first and then only send the obtained data to the server to be process for the final results.

The last example based on Case Study 3 where caching is involved, a mobile user is currently in location *A* and wants to obtain data from location *A* before moving to location *B*. It might be a good idea while in location *A*, that the user requests the desired data from peer mobile users whether they have data from location *B* (maybe just partial), so that the user can "borrow" this data before having to personally go to location *B* to get those data. Or if they can only obtain location *B* data partially from the peer mobile user during their visit in location *A*, they can still benefit from less downloads transfer when they go to location *B* to obtain the desired data since they already have partial key, they can just request the exact information that is needed based on the partial key.

Existing related work has been done on computing the shortest path search by using compact exit hierarchy (CEH) and applying semantic location model (Lee et al., 2005), thus the issue is whether the method can be integrated into the situation when the need of combining the results with the other list of information are needed. Other previous work relates to investigating techniques in decreasing the access time and to also have the organizing time shorter especially when there are a number of updates at once and quick enough to respond to the users. Also being done in past research work about improving the consistency of data and response time for the mobile users in retrieving the required data. Since location dependent query deals with geographical coordinates of the mobile users, it is important to also look at minimizing movements of users. This refers to helping reduce the mobile user having to go back to the previous location to obtain additional data if he or she forgets when he or she was there before.

IMPLEMENTATION AND SYSTEM PROTOTYPE

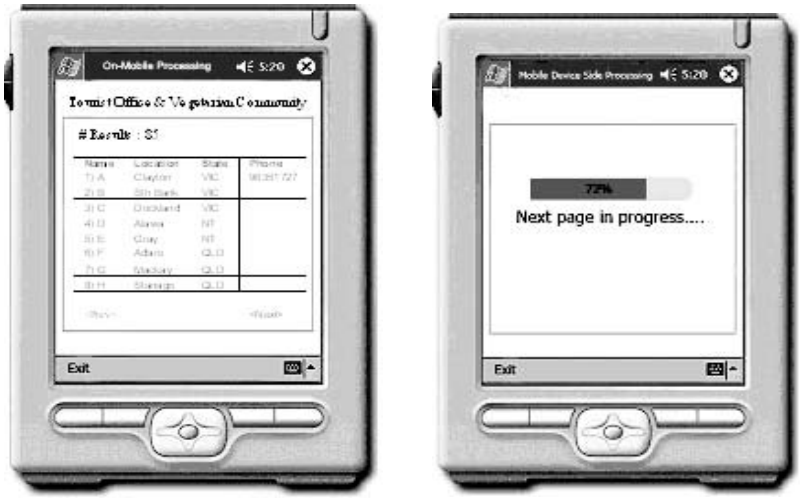
We have implemented our proposed methods on a mobile environment whereby the server and mobile device architecture consists of two desktop computers, which act as servers of two different locations, and a PDA as a mobile device, which wirelessly connects to both servers. Initially the PDA is connected to one server and requests the data from that server. When the PDA is moved, it requests data from the second server. In this environment, we have simulated location-based

query processing. The database in the servers uses MS SQL server, and both servers provides the mobile device with an access to the database. We use C# programming language within the .NET framework to program our proposed methods. Initially, the development is purely done at a desktop, and without involving the PDA. The testing was done using the Pocket PC emulator, which comes with the Visual Studio .Net. Once the testing is complete, the deployment is done by transferring down the executable from PC Emulator to the real PDA through the MS Active Sync, in which it copies not only the executable program, but also the visual .Net run time

In this section, we describe a query-processing prototype that involves multiple non-collaborative servers. This prototype is built on the idea from Example 1. Basically the prototype shows a sample interface design on how the final product would appear on the mobile user's PDA. Recalling Example 1, the mobile user is interested to know all vegetarian restaurants that are available in the country he or she is currently located. The mobile user is interested in obtaining recommendations from both servers, the tourist office and the vegetarian community servers. In this prototype design, we do not incorporate any location dependent processing and we assume that the current country (location) is Australia. Figure 8a models the final display on the PDA for the mobile user and Figure 8b models a sample screen shot of informing user the next page is being processing.

From this prototype, we can see that it only shows the basic features, which are the results in tabular format that is simple and straightforward. This can be a limitation especially when in today's world, multimedia has emerged as an important component in human interaction. A lot of people are demanding multimedia features now due to the vast benefits on what multimedia information can bring to the mobile users. Not only does the incorporation of multimedia elements enhance the appearance and make the information more

Figure 8. Final output of prototype design



a *b*

interesting, but it also gives a better interaction between the user and the device. For instance, a person who is blind may not read what is displayed on the PDA screen, but with the ability to incorporate multimedia feature, the results can be translated into "voice" talk instead of just displaying the results. This shows one importance on what multimedia information can bring around to the mobile users.

CONCLUSION AND FUTURE DIRECTION

In this article, we have presented possible applications that will be beneficial for information gathering from multiple non-collaborative sources. A brief taxonomy of the possible database operations involving multiple sources is also presented. We have also demonstrated not every server that is available in the wireless environment accepts direct queries from the users. There are some situations when the servers do not have the ability to accept direct queries and we need to process that data together with data that are obtained based on direct queries. As the wireless and mobile communication of mobile users has increased, location has become a very important constraint. A list of data obtained from different locations brings in different contents, and hence, there is a need to efficiently make these different lists of data into a single valuable piece of information for mobile users. All the issues and limitations have been outlined accordingly to where the lists of data are obtained via server, on-air, or client strategies. A sample prototype is being designed to demonstrate where the project may be applied in real life application.

Our future work is to further investigate the gathering processing techniques to further optimize the response and the data processing that is obtained from the non-collaborative servers. Since there are several issues that arise regardless of whether the lists are obtained from the server, on-air, or client strategy, they are difficult and very challenging to overcome. Thus, further investigation on choosing the right technique for each strategy according to situations should be done individually before processing several lists that are obtained from various strategies together. In addition, individually evaluating the best technique to obtain certain lists of data from each strategy should also be done to obtain the suitable technique. It is also beneficial to explore on the issues of scalability in terms of the servers that are needed to process together maintain the same efficiency or improve efficiency even though *n* servers are involved.

REFERENCES

Bell, D., & Grimson, J. (1992). *Distributed database systems.* Addison-Wesley.

Cai, Y., & Hua, K. A. (2002). An adaptive query management technique for real-time monitoring of spatial regions in mobile database systems. In *Proceedings of 21st IEEE International Conference on Performance, Computing, and Communications* (pp. 259-266).

Ceri, S., & Pelagatti, G. (1984). *Distributed databases: Principles and systems.* New York: McGraw-Hill.

Chan, B. Y., Si, A., & Leong, H. V. (1998). Cache management for mobile databases: Design and evaluation. In *Proceedings of the International Conference on Data Engineering (ICDE)* (pp. 54-53).

Cheverst, K., Davies, N., Mitchell, K., & A., F. (2000). Experiences of developing and deploying a context-aware tourist guide. In *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking* (pp. 20-31).

Elmargamid, A., Jing, J., Helal, A., & Lee, C. (2003). Scalable cache invalidation algorithms for

mobile data access. *IEEE Transactions on Knowledge and Data Engineering, 15*(6), 1498-1511.

Gutting, R. H., Bohlen, M. H., Erwig, M., Jensen, C. S., Lorentzos, N. A., Schneider, M., & Vazierginiannis, M. (2000). A foundation for representing and querying moving objects. *ACM Transactions on Database Systems Journal, 25*(1), 1-42.

Hurson, A. R., & Jiao, Y. (2005). Data broadcasting in mobile environment. In D. Katsaros, A. Nanopoulos, & Y. Manolopaulos (Eds.), *Wireless information highways.* London: IRM Press Publisher.

Jung, II, D., You, Y. H., Lee, J. J., & Kim, K. (2002). Broadcasting and caching policies for location-dependent queries in urban areas. In *Proceedings of the of the 2nd International Workshop on Mobile Commerce* (pp. 54-59).

Kapp, S. (2002). 802.11: Leaving the wire behind. *IEEE Internet Computing, 6.*

Lee, D. K., Xu, J., Zheng, B., & Lee, W. C. (2002). Data management in location-dependent information services. *IEEE Pervasive Computing, 2*(3), 65-72, July-Sept.

Lee, D. K., Zhu, M., & Hu, H. (2005). When location-based services meet databases. *Mobile Information Systems, 1*(2), 2005.

Lee, K. C. K., Leong, H. V., & Si, A. (2002). Semantic data access in an asymmetric mobile environment. In *Proceedings of the 3rd Mobile Data Management* (pp. 94-101).

Liberatore, V. (2002). Multicast scheduling for list requests". In *Proceedings of IEEE INFOCOM Conference* (pp. 1129-1137).

Lim, S. Y., Taniar, D., & Srinivasan, B. (2006). A taxonomy of database operations on mobile devices. *Handbook of Research on Mobile Multimedia*, accepted for publication, 2006.

Lo, E., Mamoulis, N., Cheung, D. W., Ho, W. S., & Kalnis, P. (2003). In *Processing ad-hoc joins on mobile devices.* Technical report, The University of Hong Kong (2003). Retrieved from http://www.csis.hku.hk/~dbgroup/techreport

Madria, S. K., Bhargava, B., Pitoura, E., & Kumar, V. (2000). Data organisation for location-dependent queries in mobile computing. In *Proceedings of ADBIS-DASFAA* (pp. 142-156).

Malladi, R., & Davis, K. C. (2002). Applying multiple query optimization in mobile databases. In *Proceedings of the 36th Hawaii International Conference on System Sciences* (pp. 294-303).

Mamoulis, N., Kalnis, P., Bakiras, S., & Li, X. (2003). Optimization of spatial joins on mobile devices. In *Proceedings of the SSTD.*

Myers, B. A., & Beigl M. (2003). Handheld computing. *IEEE Computer Magazine, 36*(9), 27-29.

Özsu, M. T., & Valduriez, P. (1999). Principles of distributed database systems (2nd ed.). Prentice Hall.

Ozakar, B., Morvan, F., & Hameurlain, A. (2005). Mobile join operators for restricted sources. *Mobile Information Systems, **1***(3).

Ren, Q., & Dunham, M. H. (2000). Using semantic caching to manage location-dependent data in mobile computing. In *Proceedings of the 6th International Conference on Mobile Computing and Networking* (pp. 210-221).

Seydim, A.Y., Dunham, M. H., & Kumar, V. (2001). Location-dependent query processing. In *Proceedings of the 2nd International Workshop on Data Engineering on Mobile and Wireless Access (MobiDE'01)* (pp. 47-53).

Si, A., & Leong, H. V. (1999). Query optimization for broadcast database. *Data and Knowledge Engineering, 29*(3), 351-380.

Sistla, A. P., Wolfson, O., Chamberlain, S., & Dao, S. (1997). Modeling and querying moving objects. In *Proceedings of the 13th International Conference on Data Engineering* (pp. 422-432).

Saltenis, S., & Jensen, C. S. (2002). Indexing of moving objects for location-based services. *Proceedings of ICDE* (pp. 463-472).

Song, M., Kang, S. W., & Park, K. (2005). On the design of energy-efficient location tracking mechanism in location-aware computing. *Mobile Information Systems, 1*(2), 109-127.

Tran, D. A., Hua, K. A., & Jiang, N. (2001). A generalized design for broadcasting on multiple physical-channel air-cache. In *Proceedings of the ACM SIGAPP Symposium on Applied Computing (SAC'01)* (pp. 387-392).

Triantafillou, P., Harpantidou, R., & Paterakis, M. (2001). High performance data broadcasting: A comprehensive systems "perspective." In *Proceedings of the 2nd International Conference on Mobile Data Management (MDM 2001)* (pp. 79-90).

Tsalgatidou, A., Veijalainen, J., Markkula, J., Katasonov, A., & Hadjiefthymiades, S. (2003). Mobile e-commerce and location-based services: Technology and requirements. In *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Services* (pp. 1-14).

Tse ,P. K. C., Lam, W. K., Ng, K. W., & Chan, C. (2005). An implementation of location-aware multimedia information download to mobile system. *Journal of Mobile Multimedia, 1*(1), 33-46.

Waluyo, A. B., Srinivasan, B., & Taniar, D. (2005). Research on location-dependent queries in mobile databases. *International Journal of Computer Systems Science & Engineering, 20*(3), 77-93, March.

Wolfson, O. (2002). Moving objects information management: The database challenge. In *Proceedings of the 5th Workshop on Next Generation Information Technology and Systems (NGITS)* (pp. 75-89).

Xu, J., Hu, Q., Lee, W. C., & Lee, D. L. (2004). Performance evaluation of an optimal cache replacement policy for wireless data dissemination. *IEEE Transaction on Knowledge and Data Engineering (TKDE), 16*(1), 125-139.

Xu, J., Tang, X., & Lee, D. L. (2003). Performance analysis of location-dependent cache invalidation schemes for mobile environments. *IEEE Transactions on Knowledge and Data Engineering (TKDE), 15*(2), 474-488.

Xu, J., Zheng, B., Lee, W. C., & Lee, D. L. (2003). Energy efficient index for querying location-dependent data in mobile broadcast environments. *Proceedings of the 19th IEEE International Conference on Data Engineering (ICDE '03)* (pp. 239-250).

Zheng, B., Xu, J., Lee, D. L. (2002). Cache invalidation and replacement strategies for location-dependent data in mobile environments. *IEEE Transactions on Computers, 51*(10), 1141-1153.

This work was previously published in the International Journal of Business Data Communications and Networking, edited by J. Gutierrez, Volume 3, Issue 2, pp. 72-93, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

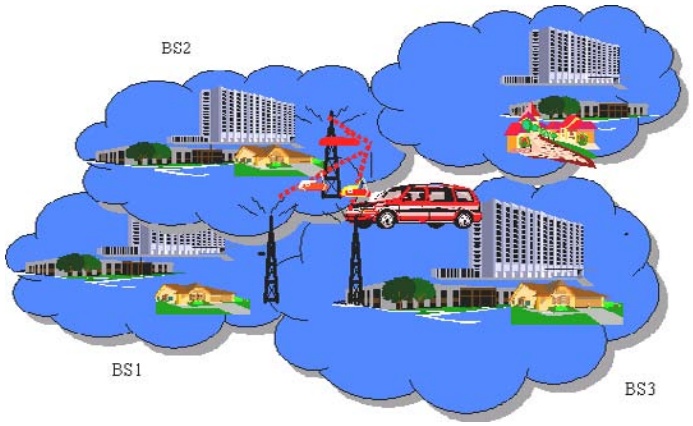
Chapter 7.41
A Bio-Inspired Approach for the Next Generation of Cellular Systems

Mostafa El-Said
Grand Valley State University, USA

INTRODUCTION

In the current 3G systems and the upcoming 4G wireless systems, *missing neighbor pilot* refers to the condition of receiving a high-level pilot signal from a Base Station (BS) that is not listed in the mobile receiver's neighbor list (LCC International, 2004; Agilent Technologies, 2005). This pilot signal interferes with the existing ongoing call, causing the call to be possibly dropped and

Figure 1. Missing pilot scenario



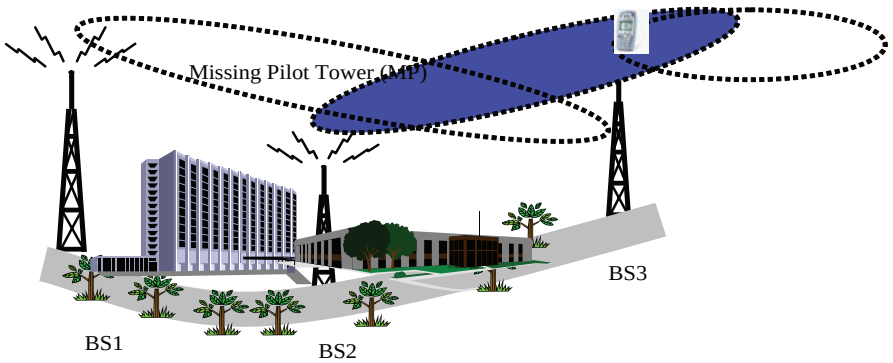
increasing the handoff call dropping probability. Figure 1 describes the missing pilot scenario where BS1 provides the highest pilot signal compared to BS1 and BS2's signals. Unfortunately, this pilot is not listed in the mobile user's active list.

The horizontal and vertical handoff algorithms are based on continuous measurements made by the user equipment (UE) on the Primary Scrambling Code of the Common Pilot Channel (CPICH). In *3G systems*, UE attempts to measure the quality of all received CPICH pilots using the Ec/Io and picks a dominant one from a cellular system (Chiung & Wu, 2001; El-Said, Kumar, & Elmaghraby, 2003). The UE interacts with any of the available radio access networks based on its memorization to the neighboring BSs. As the UE moves throughout the network, the serving BS must constantly update it with neighbor lists, which tell the UE which CPICH pilots it should be measuring for handoff purposes. In *4G systems*, CPICH pilots would be generated from any wireless system including the 3G systems (Bhashyam, Sayeed, & Aazhang, 2000). Due to the complex heterogeneity of the 4G radio access network environment, the UE is expected to suffer from various carrier interoperability problems. Among these problems, the missing neighbor pilot is considered to be the most dangerous one that faces the 4G industry.

The wireless industry responded to this problem by using an inefficient traditional solution relying on using antenna downtilt such as given in Figure 2. This solution requires shifting the antenna's radiation pattern using a mechanical adjustment, which is very expensive for the cellular carrier. In addition, this solution is permanent and is not adaptive to the cellular network status (Agilent Technologies, 2005; Metawave, 2005).

Therefore, a self-managing solution approach is necessary to solve this critical problem. Whisnant, Kalbarczyk, and Iyer (2003) introduced a system model for dynamically reconfiguring application software. Their model relies on considering the application's static structure and run-time behaviors to construct a workable version of reconfiguration software application. Self-managing applications are hard to test and validate because they increase systems complexity (Clancy, 2002). The ability to reconfigure a software application requires the ability to deploy

Figure 2. Missing pilot solution: Antenna downtilt



a dynamically hardware infrastructure in systems in general and in cellular systems in particular (Jann, Browning, & Burugula, 2003).

Konstantinou, Florissi, and Yemini (2002) presented an architecture called NESTOR to replace the current network management systems with another automated and software-controlled approach. The proposed system is inherently a rule-based management system that controls change propagation across model objects. Vincent and May (2005) presented a decentralized service discovery approach in mobile ad hoc networks. The proposed mechanism relies on distributing information about available services to the network neighborhood nodes using the analogy of an electrostatic field. Service requests are issued by any neighbor node and routed to the neighbor with the highest potential.

The autonomic computing system is a concept focused on adaptation to different situations caused by multiple systems or devices. The IBM Corporation recently initiated a public trail of its Autonomic Toolkit, which consists of multiple tools that can be used to create the framework of an autonomic management system. In this article, an autonomic engine system setting at the cellular base station nodes is developed to detect the missing neighbor (Ganek & Corbi, 2003; Haas, Droz, & Stiller, 2003; Melcher & Mitchell, 2004). The autonomic engine receives continuous feedback and performs adjustments to the cell system's neighboring set by requiring the UE to provide signal measurements to the serving BS tower (Long, 2001).

In this article, I decided to use this toolkit to build an autonomic rule-based solution to detect the existence of any missing pilot. The major advantage of using the IBM autonomic toolkit is providing a common system infrastructure for processing and classifying the RF data from multiple sources regardless of its original sources. This is a significant step towards creating a transparent autonomic high-speed physical layer in 4G systems.

PROPOSED SOLUTION

The proposed AMS relies on designing an autonomic high-speed physical layer in the smart UE and the BS node. *At the UE side,* continuous CPICH pilot measurements will be recorded and forwarded to the serving BS node via its radio interface. *At the BS node,* a scalable self-managing autonomic engine is developed using IBM's autonomic computing toolkit to facilitate the mobile handset's vertical/horizontal handover such as shown in Figure 3. The proposed engine is cable of interfacing the UE handset with different wireless technologies and detects the missing pilot if it is existed.

The autonomic engine relies on a generic log adapter (GLA), which is used to handle any raw measurements log file data and covert it into a standard format that can be understood by the autonomic manager. Without GLA, separate log adapters would have be coded for any system that the autonomic manager interfaced with. The BS node will then lump all of the raw data logs together and forward them to the Generic Log Adapter for data classification and restructuring to the common base event format. Once the GLA has parsed a record in real time to common base event format, the autonomic manager will see the record and process it and take any action necessary by notifying the BS node to make adjustments to avoid the missing pilot and enhance the UE devices' quality of service.

PERFORMANCE MEASUREMENTS AND KEY FINDINGS

To test the applicability of the proposed solution, we decided to use the system's response time, AS's service rate for callers experiencing missing pilot problem, and the performance gain as performance metrics. Also, we developed a Java class to simulate the output of a UE in a heterogeneous RF access network. Table 1 summarizes

Figure 3. Autonomic base station architecture

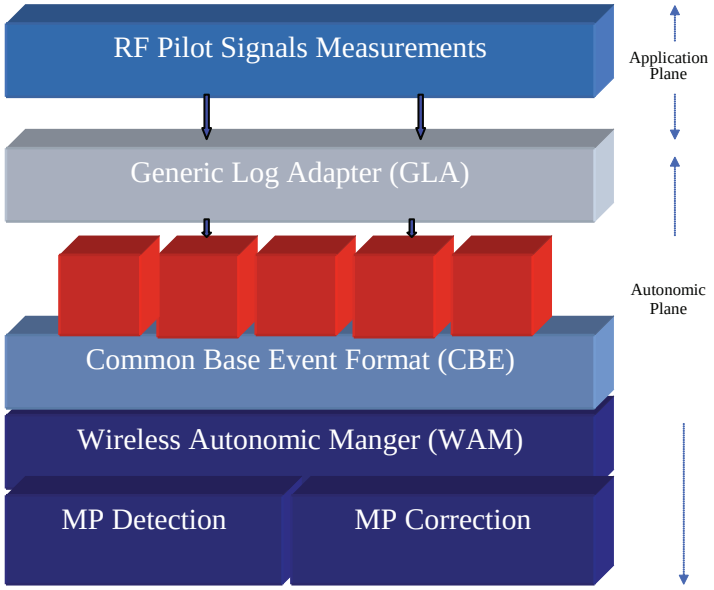


Table 1. Summary of the system performance analysis

	Log File Size in (# Records)	System Response Time in (Sec)	Processing Rate by the Base Station in (Records/Sec)
Trial Experiment 1	985	145	6.793103448
Trial Experiment 2	338	95	3.557894737
Trial Experiment 3	281	67	4.194029851
Trial Experiment 4	149	33	4.515151515
Average Processing Rate by the Base Station in (Records/Sec)	**4.765044888**		
Base Station Service Rate For callers experiencing missing pilot problem (Records/Sec)	**5.3**		
Performance Gain	**1.112266542**		

the simulation results for four simulation experiments with different log files size.

The results shown in Table 1 comply with the design requirements for the current 3G system. This is illustrated in the following simple example.

DESIGN REQUIREMENTS FOR 3G SYSTEMS

- The 3G cell tower's coverage area is divided into three sectors, with each sector having (8 traffic channel * 40 call/channel = 320 voice traffic per sector) and (2 control channels * 40 callers/channels = 80 control traffic per sector).
- The overlapped area between towers (hand-off zone) occupies 1/3 of the sector size and serves (1/3 of 320 = 106 callers (new callers and/or exciting ones)). If we consider having the UE report its status to the tower every 5 seconds, we could potentially generate 21.2 records in 1 second.
- It is practical to assume that 25% of the 21.2 reports/second accounts for those callers that may suffer from the missing pilot problem—that is, the tower's service rate for missing neighbor pilot callers is 21.2/4 = 5.3 records/second. This is the threshold level used by the tower to accommodate those callers suffering from the missing pilot problem.

ANALYSIS OF THE RESULTS

- Response time is the time taken by the BS to process, parse the incoming log file and detect the missing neighbor pilot. It is equal to (145, 95, 67, and 33) for the four experiment trials. All values are in seconds.
- Processing rate by the base station is defined as the total number of incoming records divided by the response time in (records/second). It is equal to (6.7, 3.5, 4.1, and 4.5) for the four experiment trials.
- The UE reports a missing pilot problem with an average rate of 4.7 records/second.
- The base station's service rate for callers experiencing the missing pilot problem = 5.3 records/second.
- The performance gain is defined as:

$$\frac{\text{Base Station Service Rate For callers experiencing missing pilot problem in (Records/Sec)}}{\text{Average Processing Rate by the Base Station in (Records/Sec)}}$$

$$= 5.3/4.7 = 1.1$$

- Here it is obvious that the service rate (5.3 records/second) is greater than the UE's reporting rate to the base station node (4.7 records/second). Therefore, the above results prove that the proposed solution does not overload the processing capabilities of the BS nodes and can be scaled up to handle a large volume of data.

FUTURE TRENDS

An effective solution for the interoperability issues in 4G wireless systems must rely on an adaptive and self-managing network infrastructure. Therefore, the proposed approach in this article can be scaled to maintain continuous user connectivity, better quality of service, improved robustness, and higher cost-effectiveness for network deployment.

CONCLUSION

In this article, we have developed an autonomic engine system setting at the cellular base station (BS) nodes to detect the missing neighbor. The autonomic engine receives continuous feedback and performs adjustments to the cell system's neighboring set by requiring the user equipment (UE) to provide signal measurements to the serving BS tower. The obtained results show that the proposed solution is able to detect the

missing pilot problem in any heterogeneous RF environment.

REFERENCES

Agilent Technologies. (2005). Retrieved October 2, 2005, from http://we.home.agilent.com

Bhashyam, S., Sayeed, A., & Aazhang, B. (2000). Time-selective signaling and reception for communication over multipath fading channels. *IEEE Transaction on Communications, 48*(1), 83-94.

Chiung, J., & Wu, S. (2001). Intelligent handoff for mobile wireless Internet. *Journal of Mobile Networks and Applications, 6*, 67-79.

Clancy, D. (2002). *NASA challenges in autonomic computing. Almaden Institute 2002, IBM Almaden Research Center,* San Jose, CA.

El-Said, M., Kumar, A., & Elmaghraby, A. (2003). Pilot pollution interference cancellation in CDMA systems. *Special Issue of Wiley Journal: Wireless Communication and Mobile Computing on Ultra Broadband Wireless Communications for the Future, 3*(6), 743-757.

Ganek, A., & Corbi, T. (2003). The dawning of the autonomic computing era. *IBM Systems Journal, l42*(1), 5-19.

Haas, R., Droz, P., & Stiller, B. (2003). Autonomic service deployment in networks. *IBM Systems Journal, 42*(1), 150-164.

Jann, L., Browning, A., & Burugula, R. (2003). Dynamic reconfiguration: Basic building blocks for autonomic computing on IBM pSeries servers. *IBM Systems Journal, 42*(1), 29-37.

Konstantinou, A., Florissi, D., & Yemini, Y. (2002). Towards self-configuring networks. *Proceedings of the DARPA Active Networks Conference and Exposition* (pp. 143-156).

LCC International. (2004). Retrieved December 10, 2004, from http://www.hitech-news.com/30112001-MoeLLC.htm

Lenders, V., May, M., & Plattner, B. (2005). Service discovery in mobile ad hoc networks: A field theoretic approach. *Special Issue of Pervasive and Mobile Computing, 1,* 343-370.

Long, C. (2001). *IP network design*. New York: McGraw-Hill Osborne Media.

Melcher, B., & Mitchell, B. (2004). Towards an autonomic framework: Self-configuring network services and developing autonomic applications. *Intel Technology Journal, 8*(4), 279-290.

Metawave. (2005). Retrieved November 10, 2005, from http://www.metawave.com

Whisnant, Z., Kalbarczyk, T., & Iyer, R. (2003). A system model for dynamically reconfigurable software. *IBM Systems Journal, 42*(1), 45-59.

KEY TERMS

Adaptive Algorithm: Can "learn" and change its behavior by comparing the results of its actions with the goals that it is designed to achieve.

Autonomic Computing: An approach to self-managed computing systems with a minimum of human interference. The term derives from the body's autonomic nervous system, which controls key functions without conscious awareness or involvement.

Candidate Set: Depicts those base stations that are in transition into or out of the active set, depending on their power level compared to the threshold level.

Missing Neighbor Pilot: The condition of receiving a high-level pilot signal from a base sta-

tion (BS) that is not listed in the mobile receiver's neighbor list.

Neighbor Set: Represents the nearby serving base stations to a mobile receiver. The mobile receiver downloads an updated neighbor list from the current serving base station. Each base station or base station sector has a unique neighbor list.

Policy-Based Management: A method of managing system behavior or resources by setting "policies" (often in the form of "if-then" rules) that the system interprets.

Virtual Active Set: Includes those base stations (BSs) that are engaged in a live communication link with the mobile user; they generally do not exceed three base stations at a time.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 63-67, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Section VIII
Emerging Trends

This section highlights research potential within the field of mobile computing while exploring uncharted areas of study for the advancement of the discipline. Chapters within this section highlight evolutions in mobile services, frameworks, and interfaces. These contributions, which conclude this exhaustive, multi-volume set, provide emerging trends and suggestions for future research within this rapidly expanding discipline.

Chapter 8.1
Bridging Together Mobile and Service-Oriented Computing

Loreno Oliveira
Federal University of Campina Grande, Brazil

Emerson Loureiro
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

The growing popularity of powerful *mobile devices*, such as modern cellular phones, smart phones, and PDAs, is enabling *pervasive computing* (Weiser, 1991) as the new paradigm for creating and interacting with computational systems. Pervasive computing is characterized by the interaction of mobile devices with embedded devices dispersed across *smart spaces*, and with other mobile devices on behalf of users. The interaction between user devices and smart spaces occurs primarily through services advertised on those environments. For instance, airports may offer a notification service, where the system registers the user flight at the check-in and keeps the user informed, for example, by means of messages, about flight schedule or any other relevant information.

In the context of smart spaces, *service-oriented computing* (Papazoglou & Georgakopoulos, 2003), in short SOC, stands out as the effective choice for advertising services to mobile devices (Zhu, Mutka, & Ni, 2005; Bellur & Narendra, 2005). SOC is a computing paradigm that has in services the essential elements for building applications. SOC is designed and deployed through *service-oriented architectures* (SOAs) and their applications. SOAs address the flexibility for dynamic binding of services, which applications

need to locate and execute a given operation in a pervasive computing environment. This feature is especially important due to the dynamics of smart spaces, where resources may exist anywhere and applications running on mobile clients must be able to find out and use them at runtime.

In this article, we discuss several issues on bridging mobile devices and service-oriented computing in the context of smart spaces. Since smart spaces make extensive use of services for interacting with personal mobile devices, they become the ideal scenario for discussing the issues for this integration. A brief introduction on SOC and SOA is also presented, as well as the main architectural approaches for creating SOC environments aimed at the use of resource-constrained mobile devices.

BACKGROUND

SOC is a distributed computing paradigm whose building blocks are distributed services. Services are self-contained software modules performing only pre-defined sets of tasks. SOC is implemented through the deployment of any software infrastructure that obeys its key features. Such features include loose coupling, implementation neutrality, and granularity, among others (Huhns & Singh, 2005). In this context, SOAs are software architectures complying with SOC features.

According to the basic model of SOA, service providers advertise service interfaces. Through such interfaces, providers hide from service clients the complexity behind using different and complex kinds of resources, such as databanks, specialized hardware (e.g., sensor networks), or even combinations of other services. Service providers announce their services in service registries. Clients can then query these registries about needed services. If the registry knows some provider of the required service, a reference for that provider is returned to the client, which uses this reference for contacting the service provider. Therefore, services must be described and published using some machine-understandable notation.

Different technologies may be used for conceiving SOAs such as grid services, Web services, and Jini, which follow the SOC concepts. Each SOA technology defines its own standard machineries for (1) service description, (2) message format, (3) message exchange protocol, and (4) service location.

In the context of pervasive computing, services are the essential elements of smart spaces. Services are used for interacting with mobile devices and therefore delivering personalized services for people. Owning to the great benefits that arise with the SOC paradigm, such as interoperability, dynamic service discovery, and reusability, there is a strong and increasing interest in making mobile devices capable of providing and consuming services over wireless networks (Chen, Zhang, & Zhou, 2005; Kalasapur, Kumar, & Shirazi, 2006; Kilanioti, Sotiropoulou, & Hadjiefthymiades, 2005). The dynamic discovery and invocation of services are essential to mobile applications, where the user context may change dynamically, making different kinds of services, or service implementations, adequate at different moments and places.

However, bridging mobile devices and SOAs requires analysis of some design issues, along with the fixing of diverse problems related to using resources and protocols primarily aimed at wired use, as discussed in the next sections.

INTEGRATING MOBILE DEVICES AND SOAS

Devices may assume three different roles in a SOA: service provider, service consumer, or service registry. In what follows, we examine the most representative high-level scenarios of how mobile devices work in each situation.

Consuming Services

The idea is to make available, in a wired infrastructure, a set of services that can be discovered and used by mobile devices. In this context, different designs can be adopted for bridging mobile devices and service providers. Two major architectural configurations can be derived and adapted to different contexts (Duda, Aleksy, & Butter, 2005): direct communication and proxy aided communication. In Figure 1 we illustrate the use of direct communication.

In this approach, applications running at the devices directly contact service registries and service providers. This approach assumes the usage of fat clients with considerable processing, storage, and networking capabilities. This is necessary because mobile clients need to run applications coupled with SOA-defined protocols, which may not be suited for usage by resource-constrained devices.

However, most portable devices are rather resource-constrained devices. Thus, considering running on mobile devices applications with significant requirements of processing and memory footprint reduces the range of possible client devices. This issue leads us to the next approach, proxy-aided communication, illustrated in Figure 2.

In this architectural variation, a proxy is introduced between the mobile device and the SOA infrastructure, playing the role of mobile device proxy in the wired network. This proxy interacts via SOA-defined protocols with registries and service providers, and may perform a series of content adaptations, returning to mobile devices results using lightweight protocols and data formats.

This approach has several advantages over the previous one. The proxy may act as a cache, storing data of previous service invocations as well as any client relevant information, such as bookmarks and profiles. Proxies may also help client devices by transforming complex data into lightweight formats that could be rapidly delivered through wireless channels and processed by resource-constrained devices.

Advertising Services

In a general way, mobile devices have two choices for advertising services (Loureiro et al., 2006): the push-based approach and the pull-based approach. In the first one, illustrated in Figure 3, service providers periodically send the descriptions of the services to be advertised directly to potential clients, even if they are not interested in such services (1). Clients update local registries with information about available services (2), and if some service is needed, clients query their own registries about available providers (3).

Figure 1. Direct communication between mobile client and SOA infrastructure

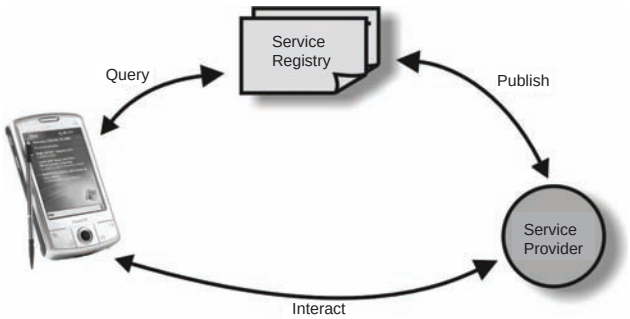


Figure 2. Proxy intermediating communication between mobile client and SOA

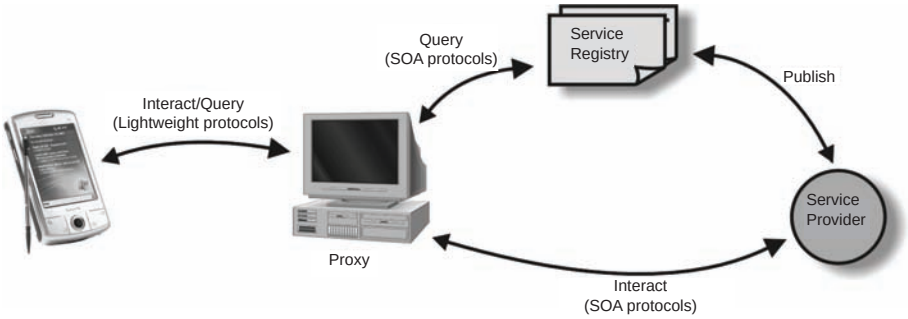


Figure 3. Push-based approach

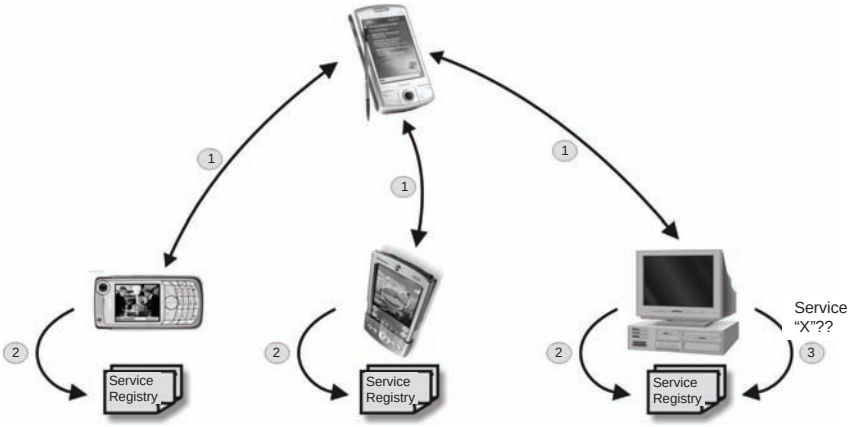


Figure 4. Pull-based approach with centralized registry

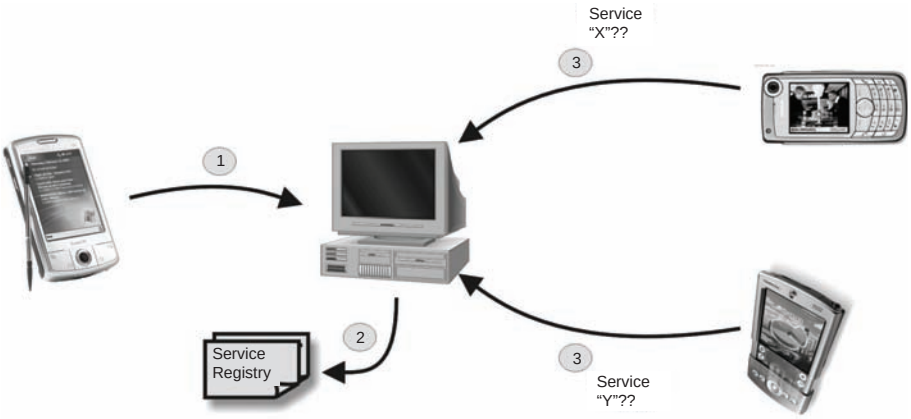
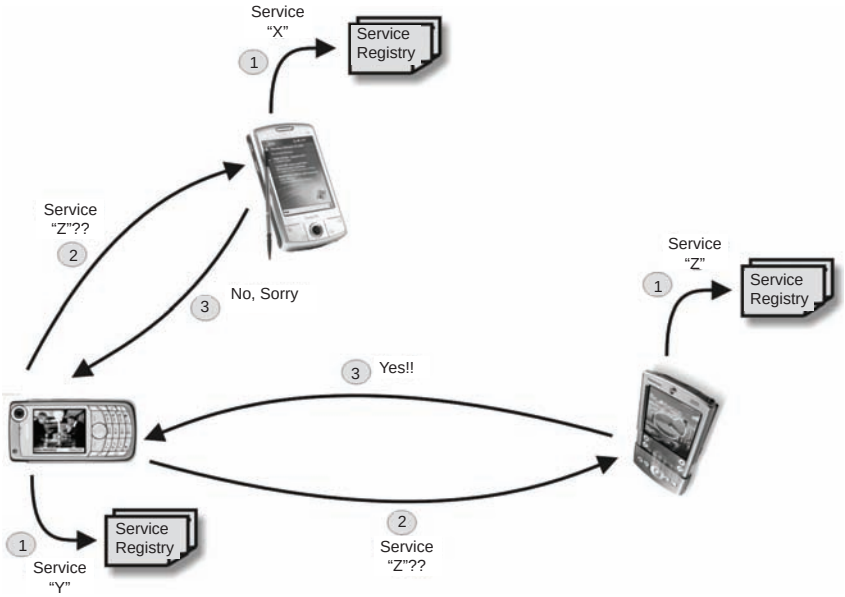


Figure 5. Pull-based approach with distributed registries



In the pull-based approach, clients only receive the description of services when they require a service discovery. This process can be performed in two ways, either through centralized or distributed registries. In the former, illustrated in Figure 4, service descriptions are published in central servers (1), which maintain entries about available services (2). Clients then query this centralized registry in order to discover the services they need (3).

In the distributed registry approach, illustrated in Figure 5, the advertisement is performed in a registry contained in each provider (1). Therefore, once a client needs to discover a service, it will have to query all the available hosts in the environment (2) until discovering some service provider for the needed service (3).

ISSUES ON INTEGRATING MOBILE DEVICES AND SOAs

Regardless of using mobile devices for either consuming or advertising services in SOAs, both *mobility* and the *limitations* of these devices are raised as the major issues for this integration. Designing and deploying effective services aimed at mobile devices requires careful analysis of diverse issues related to this kind of service provisioning.

Next, we depict several issues that arise when dealing with mobile devices in SOAs. This list is not exhaustive, but rather representative of the dimension of parameters that should be balanced when designing services for mobile use.

Suitability of Protocols and Data Formats

SOAs are primarily targeted at wired infrastructures. Conversely, small mobile devices are known by their well-documented limitations. Thus, protocols and formats used in conventional SOAs may be inadequate for use with resource-constrained wireless devices (Pilioura, Tsalgatidou, & Hadjiefthymiades, 2002; Kilanioti et al., 2005).

For instance, UDDI and SOAP are, respectively, standard protocols for service discovery and messaging in Web services-based SOAs. When using UDDI for service discovery, multiple costly network round trips are needed. In the same manner, SOAP messages are too large and require considerable memory footprint and CPU power for being parsed. Hence, these two protocols impact directly in the autonomy of battery-enabled devices.

Disconnected and Connected Services

In the scope of smart spaces, where disconnections are the norm rather than the exception, we can identify two kinds of services (Chen et at., 2005): disconnected and connected services. The first ones execute by caching the inputs of users in the local device. Once network connectivity is detected, the service performs some sort of synchronization. Services for messaging (e.g., e-mail and instant messages) and field research (e.g., gathering of data related to the selling of a specific product in different supermarkets) are some examples of services that can be implemented as disconnected ones.

Connected services, on the other hand, are those that can only execute when network connectivity is available in the device. Some examples of connected services include price checking, ordering, and shipment tracking. Note, however, that these services could certainly be implemented as disconnected services, although their users will generally need the information when demanded, neither before nor later. Therefore, there is no precise categorization of what kind of services would be connected or disconnected, as this decision is made by the system designer.

User Interface

User interfaces of small portable devices are rather limited in terms of screen size/resolution and input devices, normally touch screens or small built-in keyboards. This characteristic favors services that require low interaction to complete transactions (Pilioura et al., 2002). Services requiring many steps of data input, such as long forms, tend to: stress users, due to the use of non-comfortable input devices; reduce device autonomy, due to the extra time for typing data; and increase the cost of data transfer, due to larger amounts of data being transferred.

A possible alternative for reducing data typing by clients is the use of context-aware services (Patterson, Muntz, & Pancake, 2003). Context-aware services may reduce data input operations of mobile devices by inferring, or gathering through sensors, information about a user's current state and needs.

Frequent Temporary Disconnections

Temporary disconnections between mobile device and service provider are common due to user mobility. Thus, both client applications and service implementations must consider the design of mechanisms for dealing with frequent disconnections.

Different kinds of services require distinct solutions for dealing with disconnections. For instance, e-business applications need machineries for controlling state of transactions and data synchronization between mobile devices and service providers (Sairamesh, Goh, Stanoi, Padmanabhan, & Li, 2004). Conversely, streaming service requires seamless reestablishment and

transference of sessions between access points as the user moves (Cui, Nahrstedt, & Xu, 2004).

Security and Privacy

Normally, mobile devices are not shared among different users. Enterprises may benefit from this characteristic for authenticating employees, for instance. That is, the system knows the user and his/her access and execution rights based on profiles stored in his/her mobile devices. However, in commercial applications targeted at a large number of unknown users, this generates a need of anonymity and privacy of consumers. This authentication process could cause problems, for example in case of device thefts, because the device is authenticated and not the user (Tatli, Stegemann, & Lucks, 2005).

Security also has special relevance when coping with wireless networks (Grosche & Knospe, 2002). When using wireless interfaces for information exchange, mobile devices allow any device in range, equipped with the same wireless technology, to receive the transferred data. At application layer, service providers must protect themselves from opening the system to untrusted clients, while clients must protect themselves from exchanging personal information with service providers that can use user data for purposes different than the ones implicit in the service definition.

Device Heterogeneity and Content Adaptation

Modern mobile devices are quite different in terms of display sizes, resolutions, and color capabilities. This requires services to offer data suitable for the display of different sorts of devices. Mobile devices also differ in terms of processing capabilities and wireless technologies, which makes harder the task of releasing adequate data and helper applications to quite different devices.

Therefore, platform-neutral data formats stand out as the ideal choice for serving heterogeneous sets of client devices. Another possible approach consists of using on-demand data adaptation. Service providers may store only one kind of best-suited data format and transform the data, for example, using a computational grid (Hingne, Joshi, Finin, Kargupta, & Houstis, 2003), when necessary to transfer the data to client devices. Moreover, dynamic changes of conditions may also require dynamic content adaptation in order to maintain pre-defined QoS threshold values. For instance, users watching streamed video may prefer to dynamically reduce video quality due to temporary network congestion, therefore adapting video data, and to maintain a continuous playback instead of maintaining quality and experiencing constant playback freezing (Cui et al., 2004).

Consuming Services

As discussed before, system architects can choose between two major approaches for accessing services of SOAs from mobile devices: direct communication and proxy-aided communication. The two approaches have some features and limitations that should be addressed in order to deploy functional services. Direct communication suffers from the limitations of mobile devices and relates to other discussions presented in this article, such as adequacy of protocols and data formats for mobile devices and user interface.

If, on the one hand, proxy-aided communication seems to be the solution for problems of the previous approach, on the other hand it also brings its own issues. Probably the most noted is that proxies are single points of failures.

Furthermore, some challenges related to wired SOAs are also applicable to both approaches discussed. Service discovery and execution need to be automated to bring transparency and pervasiveness to the service usage. Moreover, especially in the context of smart spaces, services need to be personalized according to the current user

profile, needs, and context. Achieving this goal may require describing the semantics of services, as well as modeling and capturing the context of the user (Chen, Finin, & Joshi, 2003).

Advertising Services

A number of issues and technical challenges are associated with this scenario. The push-based approach tends to consume a lot of bandwidth of wireless links according to the number of devices in range, which implies a bigger burden over mobile devices.

Using centralized registries creates a single point of failure. If the registry becomes unreachable, it will not be possible to advertise and discover services. In the same manner, the discovery process is the main problem with the approach of distributed registries, as it needs to be well designed in order to allow clients to query all the hosts in the environment.

Regardless of using centralized or distributed registries, another issue rises with mobility of service providers. When service providers move between access points, a new address is obtained. This changing of address makes service providers inaccessible by clients that query the registry where it published its services. Mechanisms for updating the registry references must be provided in order for services to continue to be offered to their requestors.

FUTURE TRENDS

The broad list of issues presented in this article gives suggestions about future directions for integrating SOC and mobile devices. Each item depicted in the previous section is already an area of intensive research. Despite this, both SOC and mobile computing still lack really functional and mature solutions for the problems presented.

In particular, the fields of context-aware services and security stand out as present and future hot research fields. Besides, the evolution itself of mobile devices towards instruments with improved processing and networking power, as well as better user interfaces, will reduce the complexity of diverse challenges presented in this article.

CONCLUSION

In this article we have discussed several issues related to the integration of mobile devices and SOC. We have presented the most representative architectural designs for integrating mobile devices to SOAs, both as service providers and service consumers.

While providing means for effective integration of mobile devices and service providers, SOC has been leveraging fields such as mobile commerce and pervasive computing. Nonetheless, several issues remain open, requiring extra efforts for designing and deploying truly functional services.

REFERENCES

Bellur, U., & Narendra, N. C. (2005). Towards service orientation in pervasive computing systems. *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05)* (Vol. II, pp. 289-295).

Chen, H., Finin, T., & Joshi, A. (2003). An ontology for context-aware pervasive computing environments. *The Knowledge Engineering Review, 18*(3), 197-207.

Chen, M., Zhang, D., & Zhou, L. (2005). Providing Web services to mobile users: The architecture design of an m-service portal. *International Journal of Mobile Communications, 3*(1), 1-18.

Cui, Y., Nahrstedt, K., & Xu, D. (2004). Seamless user-level handoff in ubiquitous multimedia

service delivery. *Multimedia Tools Applications, 22*(2), 137-170.

Duda, I., Aleksy, M., & Butter, T. (2005). Architectures for mobile device integration into service-oriented architectures. *Proceedings of the 4th International Conference on Mobile Business (ICBM'05)* (pp. 193-198).

Grosche, S.S., & Knospe, H. (2002). Secure mobile commerce. *Electronics & Communication Engineering Journal, 14*(5), 228-238.

Hingne, V., Joshi, A., Finin, T., Kargupta, H., & Houstis, E. (2003). Towards a pervasive grid. *Proceedings of the 17th International Parallel and Distributed Processing Symposium (IPDPS'03)* (p. 207.2).

Huhns, M.N., & Singh, M.P. (2005). Service-oriented computing: Key concepts and principles. *IEEE Internet Computing, 9*(1), 75-81.

Kalasapur, S., Kumar, M., & Shirazi, B. (2006). Evaluating service oriented architectures (SOA) in pervasive computing. *Proceedings of the 4th IEEE International Conference on Pervasive Computing and Communications (PERCOMP'06)* (pp. 276-285).

Kilanioti, I., Sotiropoulou, G., & Hadjiefthymiades, S. (2005). A client/intercept based system for optimized wireless access to Web services. *Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05)* (pp. 101-105).

Loureiro, E., Bublitz, F., Oliveira, L., Barbosa, N., Perkusich, A., Almeida, H., & Ferreira, G. (2007). Service provision for pervasive computing environments. In D. Taniar (Ed.), *Encyclopedia of mobile computing and commerce*. Hershey, PA: Idea Group Reference.

Papazoglou, M. P., & Georgakopoulos, D. (2003). Service-oriented computing: Introduction. *Communications of the ACM, 46*(10), 24-28.

Patterson, C. A., Muntz, R. R., & Pancake, C. M. (2003). Challenges in location-aware computing. *IEEE Pervasive Computing, 2*(2), 80-89.

Pilioura, T., Tsalgatidou, A., & Hadjiefthymiades, S. (2002). Scenarios of using Web services in m-commerce. *ACM SIGecom Exchanges, 3*(4), 28-36.

Sairamesh, J., Goh, S., Stanoi, I., Padmanabhan, S., & Li, C. S. (2004). Disconnected processes, mechanisms and architecture for mobile e-business. *Mobile Networks and Applications, 9*(6), 651-662.

Tatli, E. I., Stegemann, D., & Lucks, S. (2005). Security challenges of location-aware mobile business. *Proceedings of the 2nd IEEE International Workshop on Mobile Commerce and Services (WMCS'05)* (pp. 84-95).

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 265*(3), 66-75.

Zhu, F., Mutka, M. W., & Ni, L. M. (2005). Service discovery in pervasive computing environments. *IEEE Pervasive Computing, 4*(4), 81-90.

KEY TERMS

Grid Service: A kind of Web service. Grid services extend the notion of Web services through the adding of concepts such as statefull services.

Jini: Java-based technology for implementing SOAs. Jini provides an infrastructure for delivering services in a network

Mobile Device: Any low-sized portable device used to interact with other mobile devices and resources from smart spaces. Examples of mobile devices are cellular phones, smart phones, PDAs, notebooks, and tablet PCs.

Proxy: A network entity that acts on behalf of another entity. A proxy's role varies since data

relays to the provision of value-added services, such as on-demand data adaptation.

Streaming Service: One of a number of services that transmit some sort of real-time data flow. Examples of streaming services include audio streaming or digital video broadcast (DVB).

Web Service: Popular technology for implementing SOAs built over Web technologies, such as XML, SOAP, and HTTP.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 71-77, copyright 2007 by Information Science Publishing (an imprint of IGI Global).

Chapter 8.2
Context-Awareness and Mobile Devices

Anind K. Dey
Carnegie Mellon University, USA

Jonna Häkkilä
Nokia Research Center, Finland

ABSTRACT

Context-awareness is a maturing area within the field of ubiquitous computing. It is particularly relevant to the growing sub-field of mobile computing as a user's context changes more rapidly when a user is mobile, and interacts with more devices and people in a greater number of locations. In this chapter, we present a definition of context and context-awareness and describe its importance to human-computer interaction and mobile computing. We describe some of the difficulties in building context-aware applications and the solutions that have arisen to address these. Despite these solutions, users have difficulties in using and adopting mobile context-aware applications. We discuss these difficulties and present a set of eight design guidelines that can aid application designers in producing more usable and useful mobile context-aware applications.

INTRODUCTION

Over the past decade, there has been a widespread adoption of mobile phones and personal digital assistants (PDAs) all over the world. Economies of scale both for the devices and the supporting infrastructure have enabled billions of mobile devices to become affordable and accessible to large groups of users. Mobile computing is a fully realized phenomenon of everyday life and is the first computing platform that is truly ubiquitous. Technical enhancements in mobile computing, such as component miniaturization, enhanced computing power, and improvements in supporting infrastructure have enabled the creation of more versatile, powerful, and sophisticated mobile devices. Both industrial organizations and academic researchers, recognizing the powerful combination of a vast user population and a sophisticated computing platform, have focused tremendous effort on improving and enhancing the experience of using a mobile device.

Since its introducion in the mid-1980s, the sophistication of mobile devices in terms of the numbers and types of services they can provide has increased many times over. However, at the same time, the support for accepting input from users and presenting output to users has remained relatively impoverished. This has resulted in slow interaction, with elongated navigation paths and key press sequences to input information. The use of predictive typing allowed for more fluid interaction, but mobile devices were still limited to using information provided by the user and the device's service provider. Over the past few years, improvements to mobile devices and back-end infrastructure has allowed for additional information to be used as input to mobile devices and services. In particular, context, or information about the user, the user's environment and the device's context of use, can be leveraged to expand the level of input to mobile devices and support more efficient interaction with a mobile device. More and more, researchers are looking to make devices and services *context-aware*, or adaptable in response to a user's changing context.

In this chapter, we will define context-awareness and describe its importance to human-computer interaction and mobile devices. We will describe some of the difficulties that researchers have had in building context-aware applications and solutions that have arisen to address these. We will also discuss some of the difficulties users have in using context-aware applications and will present a set of design guidelines that indicate how mobile context-aware applications can be designed to address or avoid these difficulties.

What is Context-Awareness

The concept of context-aware computing was introduced in Mark Weiser's seminal paper 'The Computer for the 21st Century' (Weiser, 1991). He describes ubiquitous computing as a phenomenon '*that takes into account the natural human environment and allows the computers themselves to vanish into the background*.' He also shapes the fundamental concepts of context-aware computing, with computers that are able to capture and retrieve context-based information and offer seamless interaction to support the user's current tasks, and with each computer being able to *'adapt its behavior in significant ways'* to the captured context.

Schilit and Theimer (1994a) first introduce the term *context-aware computing* in 1994 and define it as software that "adapts according to its location of use, the collection of nearby people and objects, as well as changes to those objects over time." We prefer a more general definition of context and context-awareness:

Context is any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves, and by extension, the environment the user and applications are embedded in. A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task. (Dey, 2001)

Context-aware features include using context to:

- Present information and services to a user
- Automatically execute a service for a user and
- Tag information to support later retrieval

In supporting these features, context-aware applications can utilize numerous different kinds of information sources. Often, this information comes from sensors, whether they are software sensors detecting information about the networked, or virtual, world, or hardware sensors detecting information about the physical world. Sensor data can be used to recognize the usage situation for instance from illumination, temperature, noise level, and device movements (Gellersen, Schmidt & Beigl, 2002; Mäntyjärvi &

Seppänen, 2002). Typically, sensors are attached to a device and an application on the device locally performs the data analysis, context-recognition, and context-aware service.

Location is the most commonly used piece of context information, and several different location detection techniques have been utilized in context-awareness research. Global positioning system (GPS) is a commonly used technology when outdoors, utilized, for example, in car navigation systems. Network cellular ID can be used to determine location with mobile phones. Measuring the relative signal strengths of Bluetooth and WLAN hotspots and using the hotspots as beacons are frequently used techniques for outdoors and indoors positioning (Aalto, Göthlin,Korhonen et al., 2004; Burrell & Gay, 2002; Persson et al., 2003). Other methods used indoors include ultrasonic or infrared-based location detection (Abowd et al., 1997; Borriello et al., 2005).

Other commonly used forms of context are time of day, day of week, identity of the user, proximity to other devices and people, and actions of the user (Dey, Salber & Abowd, 2001; Osbakk & Rydgren, 2005). Context-aware device behavior may not rely purely on the physical environment. While sensors have been used to directly provide this physical context information, sensor data often needs to be interpreted to aid in the understanding of the user's goals. Information about a user's goals, preferences, and social context can be used for determining context-aware device behavior as well. Knowledge about a user's goals helps prioritize the device actions and select the most relevant information sources. A user's personal preferences can offer useful information for profiling or personalizing services or refining information retrieval. The user may also have preferences about quality of service issues such as cost-efficiency, data connection speed, and reliability, which relate closely to mobile connectivity issues dealing with handovers and alternative data transfer mediums. Finally, social context forms an important type of context as mobile devices are commonly used to support communication between two people and used in the presence of other people.

Relevance to HCI

When people speak and interact with each other, they naturally leverage their knowledge about the context around them to improve and streamline the interaction. But, when people interact with computers, the computing devices are usually quite ignorant of the user's context of use. As the use of context essentially expands the conversational bandwidth between the user and her application, context is extremely relevant to human-computer interaction (HCI). Context is useful for making interaction more efficient by not forcing users to explicitly enter information about their context. It is useful for improving interactions as context-aware applications and devices can offer more customized and more appropriate services than those that do not use context. While there have been no studies of context-aware applications to validate that they have this ability, anecdotally, it is clear that having more information about users, their environments, what they have done and what they want to do, is valuable to applications. This is true in network file systems that cache most recently used files to speed up later retrieval of those files, as well as in tour guides that provide additional information about a place of interest the user is next to.

Relevance to Mobile HCI

Context is particularly relevant in mobile computing. When users are mobile, their context of use changes much more rapidly than when they are stationary and tied to a desktop computing platform. For example, as people move, their location changes, the devices and people they interact with changes more frequently, and their goals and needs change. Mobility provides additional opportunities for leveraging context but also requires additional context to try and understand how the user's goals are changing. This places extra burden on the mobile computing platform, as it needs to sense potentially rapidly changing context, synthesize it and act upon it. In the next section, we will discuss the difficulties that application builders have had with building

context-aware applications and solutions that have arisen to address these difficulties.

BUILDING MOBILE CONTEXT-AWARE APPLICATIONS

The first context-aware applications were centered on mobility. The Active Badge location system used infrared-based badges and sensors to determine the location of workers in an indoor location (Want et al., 1992). A receptionist could use this information to route a phone call to the location of the person being called, rather than forwarding the phone call to an empty office. Similarly, individuals could locate others to arrange impromptu meetings. Schilit, Adams and Want,(1994b) also use an infrared-based cellular network to location people and devices, the PARCTAB, and describe 4 different types of applications built with it (Schilit et al., 1994b). This includes:

- **Proximate selection:** Nearby objects like printers are emphasized to be easier to select than other similar objects that are further away from the user;
- **Contextual information and commands:** Information presented to a user or commands parameterized and executed for a user depend on the user's context;
- **Automatic contextual reconfiguration:** Software is automatically reconfigured to support a user's context; and
- **Context-triggered actions:** If-then rules are used to specify what actions to take based on a user's context.

Since these initial context-aware applications, a number of common mobile context-aware applications have been built: tour guides (Abowd et al., 1997; Cheverst et al. 2000; Cheverst, Mitchell & Davies, 2001), reminder systems (Dey & Abowd, 2000; Lamming & Flynn, 1994) and environmental controllers (Elrod et al., 1993; Mozer et al., 1995). Despite the number of people building (and re-building) these applications, the design and implementation of a new context-aware application required significant effort, as there was no reusable support for building context-aware applications. In particular, the problems that developers faced are:

- Context often comes from non-traditional devices that developers have little experience with, unlike the mouse and keyboard.
- Raw sensor data is often not directly useful to an application, so the data must be abstracted to turn it into useful context.
- Context comes from multiple distributed and heterogeneous sources, and this context often needs to be combined (or fused) to be useful. This process often results in uncertainty that needs to be handled by the application.
- Context is, by its very nature, dynamic, and changes to it must be detected in real time and applications must adjust to these constant changes in order to provide a positive user experience to users.

These problems resulted in developers building every new application from scratch, with little reuse of code or design ideas between applications.

Over the past five years or so, there has been a large number of research projects aimed at addressing these issues, most often trying to produce a reusable toolkit or infrastructure that makes the design of context-aware applications easier and more efficient. Our work, the Context Toolkit, used a number of abstractions to ease the building of applications. One abstraction, the context widget is similar to a graphical user interface widget in that it abstracts the source of an input and only deals with the information the source produces. For example, a location widget could receive input from someone manually entering information, a GPS device, or an infrared positioning system, but an application using a location widget does not have to deal with the details of the underlying sensing technology, only with the information the sensor produces: identity of the object being located, its location and the time when the object was located. Context interpreters support the

interpretation, inference and fusion of context. Context aggregators collect all context-related to a specific location, object or person for easy access. With these three abstractions, along with a discovery system to locate and use the abstractions, an application developer no longer needs to deal with common difficulties in acquiring context and making it useful for an application, and instead can focus on how the particular application she is building can leverage the available context. Other similar architectures include JCAF (Bardram, 2005), SOCAM (Gu, Pung & Zhang, 2004), and CoBRA (Chen et al., 2004).

While these architectures make mobile context-aware applications easier to build, they do not address all problems. Outstanding problems needing support in generalized toolkits include representing and querying context using a common ontology, algorithms for fusing heterogeneous context together, dealing with uncertainty, and inference techniques for deriving higher level forms of context such as human intent. Despite these issues, these toolkits have supported and continue to support the development of a great number of context-aware applications. So, now that we can more easily build context-aware applications, we still need to address how to design and build *usable* mobile context-aware applications. We discuss this issue in the following section.

USABILITY OF MOBILE CONTEXT-AWARE APPLICATIONS

With context information being provided as implicit input to applications and with those applications using this context to infer human intent, there are greater usability concerns than with standard applications that are not context-aware. Bellotti and Edwards discuss the need for context-aware applications to be *intelligible*, where the inferences made and actions being taken are made available to end-users (Bellotti & Edwards, 2001). Without this intelligibility, users of context-aware applications would not be able to decide what actions or responses to take themselves (Dourish, 1997).

To ground our understanding of these abstract concerns, we studied the usability and usefulness of a variety of context-aware applications (Barkhuus & Dey, 2003a; 2003b). We described a number of real and hypothetical context-aware applications and asked subjects to provide daily reports on how they would have used each application each day, whether they thought the applications would be useful, and what reservations they had about using each application. All users were given the same set of applications, but users were split into three groups with each group being given applications with a different level of proactivity. One group was given applications that they would personalize to determine what the application should do for them. Another group was provided with information about how their context was changing, and the users themselves decided how to change the application behavior. The final group was evaluating applications that autonomously changed their behavior based on changing context. Additional information was also gathered from exit interviews conducted with subjects.

Users indicated that they would use and prefer applications that had higher degrees of proactivity. However, as the level of proactivity increased, users had increasing feelings that they were losing control. While these findings might seem contradictory, it should be considered that owning a mobile phone constitutes some lack of control as the user can be contacted anywhere and at anytime; the user may have less control but is willing to bear this cost in exchange for a more interactive and smoother everyday experience. Beyond this issue of control, users had other concerns with regards to the usability of context-aware applications. They were concerned by the lack of feedback, or intelligibility, that the applications provided. Particularly for the more proactive versions of applications, users were unclear how they would know that the application was performing some action for them, what action was being performed, and why this action was being performed. A third concern was privacy. Users were quite concerned that the context data that was being used on mobile platforms could

be used by service providers and other entities to track their location and behaviors. A final concern that users had was related to them evaluating multiple context-aware applications. With potentially multiple applications vying for a user's attention, users had concerns about information overload. Particularly when mobile and focusing on some other task, it could be quite annoying to have multiple applications on the mobile device interrupting and requesting the user's attention simultaneously or even serially.

In the remainder of this chapter, we will discuss issues for designing context-aware applications that address usability concerns such as these.

Support for Interaction Design

Despite all of the active research in the field of context-aware computing, much work needs to be done to make context-awareness applications an integral part of everyday life. As context-awareness is still a very young field, it does not have established design practices that take into account its special characteristics. The development of applications has so far been done primarily in research groups that focus more on proof-of-concept and short-term use rather than deployable, long-term systems. For most of these applications, the interaction design has rarely been refined to a level that is required for usable and deployable applications. Particularly for applications aimed at consumers and the marketplace, robustness, reliability and usability must be treated more critically than they are currently, as these factors will have a significant impact on their success.

Currently, the lack of existing high-quality, commercial, and publicly available applications limits our ability to assess and refine the best practices in interaction design of context-aware mobile applications. As there is very little experience with real-life use of these applications, the ability of developers to compare and iterate on different design solutions is very restricted. As user groups for a particular application mostly do not exist yet, much of the current research is based on hypothesized or simulated systems rather that actualized use situations. Knowledge of what device features people fancy and which they just tolerate, and when application features become insignificant or annoying, are issues that are hard to anticipate without studies of long-term real-life usage.

As with any other novel technology, bringing it to the marketplace will bring new challenges. Bringing context-awareness to mobile devices as an additional feature may lead to situations where the interaction design is performed by people with little experience in context-aware computing. Using well-established commercial platforms such as mobile phones or PDAs often means that user interface designers only have experience with conventional mobile user interfaces. On the other hand, the technical specifications of an application are often provided by people who have no expertise in human-computer interaction issues. When entering a field that involves interdisciplinary elements, such as mobile context-awareness, providing tools and appropriate background information for designers helps them to recognize the risks and special requirements of the technology.

Hence, there are several factors which make examining context-awareness from the usability and interaction design perspective relevant. Failures in these may lead not only to unprofitable products, but may result in an overall negative effect—they may slow down or prevent the underlying technology from penetrating into mass markets.

Usability Risks for Mobile Context-Aware Applications

A system and its functionality are often described with mental models that people form from using the system. According to Norman (1990), one can distinguish between the designer's mental model and the user's mental model. The designer's model represents the designer's understanding and idea of the artefact being constructed, whereas the user's model is the user's conceptual model of the same artefact, its features and functionality, which has developed through her interaction with the system. In order to respond to the user's

needs, efficiently fulfil the user's goals and satisfy the user's expectations, the designer's and user's understanding of the device or application should be consistent with each other, in other words, the user's model and designer's model should be the same (Norman, 1990).

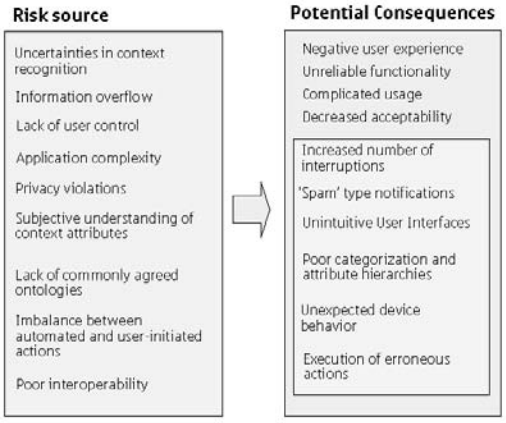
To ensure the best possible result, the mental models of different stakeholders in application development and use have to meet each other. First, the mental models of the application's technical designer and user interface designer should be consistent. This means that the user interface designer should have a basic understanding of the special characteristics of context-aware technology. Second, the designer's and user's mental models of the application should be the same. People's perception of context may differ significantly from each other, and both attributes and the measures used to describe context may vary greatly (Hiltunen, Häkkilä & Tuomela, 2005; Mäntyjärvi et al., 2003). The relationship between the designer's and the user's mental models should be checked with user tests several times during the design process. Without this careful design, there are two significant usability risks that may result: users will be unable to explain the behavior of the context-aware application, nor predict how the system will respond given some user action. While this is true of all interactive systems, it is especially important to consider for context-aware systems as the input to such systems is often implicit.

Context-awareness has several characteristics that can be problematic in interaction design. Figure 1 summarizes potential usability risks with context-aware applications.

A fundamental cause of potential usability risks is *uncertainty in context recognition*, which can be due to different reasons, such as detection accuracy, information fusion, or inferring logic. This is a key issue for designing the user interface for a mobile context-aware application, as it affects the selected features, their functionality and accuracy. In practice, features such as the proactivity level may be designed differently if the confidence level in context recognition can be estimated correctly. Uncertainty is a part of the nature of context-aware applications. Thus, it is important that the application and UI designers share a common understanding of the matter and take it into account when designing both the application and its user interface.

Application complexity has a tendency to grow when functions are added and it forms a potential risk for context-aware applications, as they use a greater number of information sources than

Figure 1. Sources of usability risks and their potential consequences related to context-aware mobile applications. Consequences that are unique to context-aware mobile applications are in the smaller rectangle on the right.



traditional mobile applications. Hiding the complex nature of the technology while maintaining a sufficient level of feedback and transparency so that the user can still make sense of the actions the device is performing (i.e., intelligibility) is a challenging issue. Here, the involvement of user-centric design principles is emphasized. Usability testing and user studies performed in an authentic environment combined with iterative design are key elements to producing well-performing user interface solutions.

Poor interoperability of services and applications relates to the absence of standardization in this maturing field and it limits the application design, available services, and seamless interaction desired across a wide selection of devices and users. Interoperability issues have gained much attention with the current trend of mobile convergence, where different mobile devices resemble each other more and more, yet providing services for them must be performed on a case-by-case basis.

Subjective understanding of context attributes creates a problem for user interface design, as the measures, such as the light intensity or noise level in everyday life are not commonly understood by end-users in terms of luxes or decibels but in relative terms such as 'dark,' 'bright,' 'silent' or 'loud.' This issue is connected to the *lack of commonly agreed ontologies*, which would guide the development of context-aware applications. The difficulties in categorizing context attributes and modeling context is evident from the literature (Hiltunen, Häkkilä & Tuomela, 2005; Mäntyjärvi et al., 2003).

As indicated earlier, *privacy violations* are possible with mobile context-aware systems collecting, sharing and using a tremendous amount of personal information about a user. When such information is shared with a number of different services, each of which will be contacting the user, *information overflow* often results. One can imagine a potential flow of incoming advertisements when entering a busy shopping street, if every shop within a radius of one hundred meters was to send an advertisement to the device. Information overflow is particularly a problem for the small screens that are typical with handheld devices.

As our earlier studies illustrated, the *lack of user control* can easily occur with mobile device automation, when context-triggered actions are executed proactively. However, the promise of context-awareness is that it provides "ease of use" by taking over actions that the user does not want to do or did not think to do for themselves. Any solution for correcting the *imbalance between the set of automated actions and user-initiated actions*, must take user control into account.

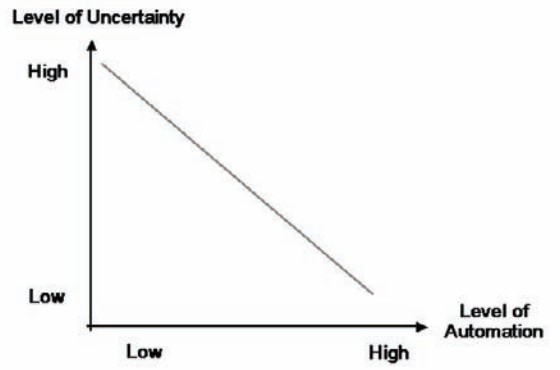
The consequences resulting from these usability risks are numerous. The general outcome can be a negative user experience. This may result from an increased number of interruptions, spam, and the execution of erroneous or otherwise unintuitive device behavior. Unreliable device functionality, and unintelligible user interfaces can lead to reduced acceptability of context-aware applications in the marketplace.

Design Guidelines for Mobile Context-Aware Applications

Context-awareness typically contains more risks than conventional, non-context-aware technology. At the same time, context-awareness can offer much added value to the user. In order to provide this value to end-users and avoid these negative design consequences and minimize usability risks, we have sought to provide a set of design guidelines that can offer practical help for designers who are involved in developing context-aware mobile applications (Häkkilä & Mäntyjärvi, 2006). These general guidelines have been validated in a series of user studies (Häkkilä & Mäntyjärvi, 2006) and should be taken into account when selecting the features of the application and during the overall design process.

GL1. Select appropriate level of automation. A fundamental factor with context-awareness is that it incorporates uncertainty. Uncertainty in context-recognition is caused by several different sources, such as detection accuracy, information fusion, or inferring logic. This is a key issue in designing user interfaces, as it affects the selected

Figure 2. How uncertainty in context-recognition should affect the selected level of automation/proactivity



features, their functionality and accuracy. In practice, features such as the automation level or level of proactivity may be designed differently if the confidence level of context recognition can be estimated correctly. The relationship between uncertainty and selected application automation level is illustrated in Figure 2. As shown in Figure 1, *uncertainties in context recognition* create significant usability risks, however, by selecting an appropriate level of automation, an application designer can acknowledge this fact and address it appropriately. The greater the uncertainty is in the context-recognition, the more important it is not to automate actions. The automation level has also a direct relationship with user control, and its selection has a large impact on the number of expected interruptions the system creates for the user. The level of automation must be considered in relation to the overall application design, as it affects numerous issues in the user interface design.

GL 2. Ensure user control. The user has to maintain the feeling that he is in the control over the device. The user, who normally has full control over his mobile device, has voluntarily given some of it back to the device in order to increase the ease of use of the device. To address this *lack of user control*, an important usability risk, the user must be able to take control of the device and context-aware application at any time. The desire to take control can happen in two basic circumstances—either the device is performing erroneous actions and the user wants to take a correcting action, or the user just wishes to feel in control (a feeling that users often have). The user has to have enough knowledge of the context-aware application and the device functionality in order to recognize malfunctioning behavior, at least in the case where context-recognition errors lead to critical and potentially unexpected actions. The perception of user control is diminished if the device behaves in unexpected manner or if the user has a feeling that the device is performing actions without him knowing it. User control can be implemented, for example, with confirmation dialogues however, this must be balanced with the need to minimize unnecessary interruptions, our next guideline.

GL3. Avoid unnecessary interruptions. Every time the user is interrupted, she is distracted from the currently active task, impacting her performance and satisfaction with the system. In most cases, the interruption leads to negative consequences, however if the system thinks that the interruption will provide high value or benefit to the user, allowing the interruption is often seen as positive. Examples of this are reminders and alarm clocks. The user's interruptibility depends on her context and the user's threshold for putting up with intrusion varies with each individual and her situation. Some context-aware functionality is so important that the user may want the application

to override all other ongoing tasks. This leads to a tension between avoiding unnecessary interruptions and supporting user control (GL2).

GL4. Avoid information overflow. The throughput of the information channel to each user is limited, and users can fully focus only on a small number of tasks at one time. In order to address the usability risk of *information overflow* where several different tasks or events compete for this channel, a priority ordering needs to be defined. Also, the threshold for determining the incoming event's relevancy in the context must be considered in order to avoid unnecessary interruptions (GL3). Systems should not present too much information at once, and should implement filtering techniques for to avoid messages that may appear to be spam to users. Also, information should be arranged in a meaningful manner to maintain and maximize the understandability of the system.

GL 5. Appropriate visibility level of system status. The visibility level of what the system is doing has to be sufficient for the user to be aware of the application's actions. While this guideline has been co-opted from Nielsen and Molich's user interface heuristics (1990), it has special meaning in context-aware computing. The implicit nature of context-awareness and natural *complexity of these types of applications* means that users may not be aware of changes in context, system reasoning or system action. When uncertainty in context-awareness is involved, there must be greater visibility of system state in order to allow the user to recognize the risk level and possible malfunctions. Important actions or changes in context should also be made visible and easily understandable for the user, despite the fact that users may have *subjective understandings of context attributes* and that there may be *no established ontology.* System status need not be overwhelming and interrupting to the user but can be provided in an ambient or peripheral fashion, where information is dynamically made more visible as the importance value grows, and may eventually lead to an interruption event to the user if its value is high enough.

GL 6. Personalization for individual needs. Context-awareness should allow a device or application to respond better to the individual user's personal needs. For instance, an application can implement filtering of interruptions according to the user's personal preferences. Personalization may also be used to improve the subjective understanding of context attributes. Allowing the user to name or change context attributes, such as location names or temperature limits, may contribute to better user satisfaction and ease of use. User preferences may change over time, and their representation in the application can be adjusted, for example implicitly with learning techniques or explicitly with user input settings.

GL 7. Secure user's privacy. *Privacy* is a central theme with personal devices, especially with devices focused on supporting personal communication, and impacts, for example trust, frequency of use, and application acceptability. Special care should be taken with applications that employ context sharing. Privacy requirements often vary between who is requesting the information, the perceived value of the information being requested and what information is being requested, so different levels of privacy should be supported. If necessary, users should have the ability to easily specify that they wish to remain anonymous with no context shared with other entities.

GL 8. Take into account the impact of social context. The social impact of a context-aware application taking an action must be part of the consideration in deciding whether to take the action or not. The application and its behavior reflects on users themselves. In some social contexts, certain device or user behavior may be considered awkward or even unacceptable. In such situations, there must be an appropriate *balance of user-initiated and system-initiated actions.* Social context has also has an effect on interruptibility. For example, an audible alert may be considered as inappropriate device behavior in some social contexts.

Once an application has been designed with these guidelines, the application must still be

evaluated to ensure that the usability risks that have been identified for mobile context-aware systems have been addressed. This evaluation can take place in the lab, but is much more useful when conducted under real, *in situ*, conditions.

SUMMARY

Context-aware mobile applications, applications that can detect their users' situations and adapt their behavior in appropriate ways, are an important new form of mobile computing. Context-awareness has been used to overcome the deficit of the traditional problems of small screen sizes and limited input functionalities of mobile devices, to offer shortcuts to situationally-relevant device functions, and to provide location sensitive device actions and personalized mobile services.

Context-awareness as a research field has grown rapidly during recent years, concentrating on topics such as context-recognition, location-awareness, and novel application concepts. Several toolkits for enabling building context-aware research systems have been introduced. Despite their existence, there exist very few commercial or publicly available applications utilizing context-awareness. However, the multitude of research activities in mobile context-awareness allow us to make reasonable assumptions about tomorrow's potential applications. For example, navigation aids, tour guides, location-sensitive and context-sensitive notifications and reminders, automated annotation and sharing of photographs, use of metadata for file annotation, sharing or search are topics which frequently appear in the research literature and will likely be relevant in the future. In addition, using context-awareness to address the needs of special user groups, for example in the area of healthcare also appears to be a rich area to explore.

Despite the active research in context-awareness, there is much that remains to be addressed in interaction design and usability issues for context-aware mobile applications. Due the novelty of the field and lack of existing commercial applications, design practices for producing usable and useful user interfaces have not yet evolved, and end-users' experiences with the technology are not always positive. We have presented a set of 8 design guidelines which have been validated and evaluated in a series of user studies, which point to areas where user interface designers must focus efforts in order to address the usability issues that are commonly found with mobile context-aware applications.

While context-aware applications certainly have more usability risks than traditional mobile applications, the potential benefits they offer to end-users are great. It is important that application designers and user interface designers understand each other's perspectives and the unique opportunities and pitfalls that context-aware systems have to offer. With context-aware applications, careful application and interface design must be emphasized. The consequences resulting from usability risks include an overall negative user experience. Unsuccessful application design may result in diminished user control, increased number of interruptions, spam, and the execution of erroneous device actions or otherwise unintuitive behaviour. Unreliable device functionality and an unintuitive user interface can lead to decreased acceptability of the context-aware features in the marketplace.

In this chapter we have discussed the notion of context-awareness and its relevance to both mobile computing and interaction design in mobile computing. We have described technical issues involved in building context-aware applications and the toolkits that have been built to address these issues. Despite the existence of these toolkits in making context-aware applications easier to build, there are several additional issues that must be addressed in order to make mobile context-aware applications usable and acceptable to end-users. We have presented a number of design guidelines that can aid the designers of mobile context-aware applications in producing applications with both novel and useful functionality for these end-users.

REFERENCES

Aalto, L, Göthlin, N., Korhonen, J., & Ojala T. (2004). Bluetooth and WAP Push based location-aware mobile advertising system. In *Proceedings of the 2nd International Conference on Mobile Systems, Applications and Services* (pp. 49-58).

Abowd, G. D., Atkeson, C. G., Hong, J., Long, S., Kooper, R.. & Pinkerton, M. (1997). Cyberguide: a mobile context-aware tour guide. *ACM Wireless Networks, 3,* 421-433.

Bardram, J. (2005). The java context awareness framework (JCAF)—A service infrastructure and programming framework for context-aware applications. In *Proceedings of Pervasive 2005* (pp. 98-115).

Barkhuus, L., & Dey, A.K. (2003a). Is context-aware computing taking control away from the user? Three levels of interactivity examined. In *Proceedings of UBICOMP 2003* (pp. 149-156).

Barkhuus, L., & Dey, A.K. (2003b). Location-based services for mobile telephony: A study of users' privacy concerns. In *Proceedings of INTERACT 2003* (pp. 709-712).

Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: Human considerations in context-aware systems. *HCI Journal, 16,* 193-212.

Borriello, G., Liu, A., Offer, T., Palistrant, C., & Sharp, R. (2005). WALRUS: Wireless, Acoustic, Location with Room-Level Resolution using Ultrasound. In *Proceedings of the 3rd International Conference on Mobile systems, application and services (MobiSys'05),* (pp. 191-203).

Burrell, J., & Gay, G. K. (2002). E-graffiti: Evaluating real-world use of a context-aware system. *Interacting with Computers, 14,* 301-312.

Chen, H., Finin, T. and Joshi, A. Chen, H., Finin, T., & Joshi, A. (2004). Semantic Web in the Context Broker Architecture. (2004). In *Proceedings of the Second IEEE international Conference on Pervasive Computing and Communications (Percom'04),* (pp. 277-286).

Cheverst, K., Davies, N., Mitchell, K., & Friday, A. (2000). Experiences of developing and deploying a context-aware tourist guide: The GUIDE project. In *Proceedings of the 6th annual international conference on Mobile computing and networking (MobiCom),* (pp. 20-31).

Cheverst, K., Mitchell, K., & Davies, N. (2001). Investigating Context-Aware Information Push vs. Information Pull to Tourists. In *Proceedings of MobileHCI'01,*

Dey, A.K., & Abowd, G.D. (2000). CybreMinder: A context-aware system for supporting reminders. In *Proceedings of the International Symposium on Handheld and Ubiquitous Computing* (pp. 172-186).

Dey, A.K., Salber, D., & Abowd, G.D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction Journal 16*(2-4), (pp. 97-166).

Dourish, P. (1997). Accounting for system behaviour: Representation, reflection and resourceful action. In Kyng and Mathiassen (Eds.), *Computers and design in context* (pp. 145-170). Cambridge, MA: MIT Press.

Elrod, S., Hall, G., Costanza, R., Dixon, M., & des Rivieres, J. Responsive office environments. *Communications of the ACM* 36(7), 84-85.

Gellersen, H.W., Schmidt, A., & Beigl, M. (2002). Multi-sensor sontext-awareness in mobile devices and smart artefacts. *Mobile Networks and Applications, 7,* 341-351.

Gu, T., Pung, H.K., & Zhang, D.Q. (2004). A middleware for building context-aware mobile services. In *Proceedings of IEEE Vehicular Technology Conference* (pp. 2656-2660).

Häkkilä, J., & Mäntyjärvi, J. (2006). Developing design guidelines for context-aware mobile applications. In *Proceedings of the IEE International*

Conference on Mobile Technology, Applications and Systems.

Hiltunen, K.-M., Häkkilä, J., & Tuomela, U. (2005). Subjective understanding of context attributes – a case study. In *Proceedings of Australasian Conference of Computer Human Interaction (OZCHI) 2005*, (pp. 1-4).

Lamming, M., & Flynn. M. (1994). Forget-me-note: Intimate computing in support of human memory. In *Proceedings of Friend21: International Symposium on Next Generation Human Interface* (pp. 125-128).

Mäntyjärvi, J., & Seppänen, T. (2002). Adapting applications in mobile terminals using fuzzy context information. In *Proceedings of Mobile HCI 2002* (pp. 95-107).

Mäntyjärvi, J., Tuomela, U., Känsälä, I., & Häkkilä, J. (2003). Context Studio—Tool for Personalizing Context-Aware Application in Mobile Terminals. In *Proceedings of Australasian Conference of Computer Human Interaction (OZCHI) 2003* (pp. 64-73).

Mozer, M.C., Dodier, R.H., Anderson, M., Vidmar, L., Cruickshank III, R.F., & Miller, D. The Neural Network House: An Overview. In L. Niklasson & M. Boden (Eds.), *Current trends in connectionism*, (pp. 371-380). Hillsdale, NJ: Erlbaum.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of CHI 1990* (pp. 249-256).

Norman, D. A. (1990). *The design of everyday things.* New York, NY: Doubleday.

Osbakk, P., & Rydgren, E. (2005). Ubiquitous computing for the public. In *Proceedings of Pervasive 2005 Workshop on Pervasive Mobile Interaction Devices (PERMID 2005)*, (pp. 56-59).

Persson, P., Espinoza, F., Fagerberg, P., Sandin, A., & Cöster, R. (2003). GeoNotes: A Location-based information System for Public Spaces. In K. Hook, D. Benyon & A. Munro (Eds.), *Readings in Social Navigation of Information Space* (pp. 151-173). London, UK: Springer-Verlag.

Schilit, B., & Theimer, M. (1994a). Disseminating active map information to mobile hosts. *IEEE Computer 8*(5), 22-32.

Schilit, B., Adams, N., & Want, R. (1994b). Context-aware computing applications. In *Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications* (pp. 85-90).

Want, R., Hopper, A., Falcao, V., & Gibbons, J. (1992). The Active Badge Location System. *ACM Transactions on Information Systems, 10*(1), 91-102.

Weiser, M. (1991). The computer for 21st century. *Scientific American,* 265(3), 94-104.

KEY TERMS

Context: Any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves, and by extension, the environment the user and applications are embedded in.

Context-Awareness: A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task.

Design Guidelines: Guidelines or principles that, when followed, can improve the design and usability of a system.

Interaction Design: The design of the user interface and other mechanism that support the user's interaction with a system, including providing input and receiving output.

Mobile Context-Awareness: Context-awareness for systems or situations where the user and her devices are mobile. Mobility is particularly relevant for context-awareness as the user's context changes more rapidly when mobile.

Usability Risks: Risks that result from the use of a particular technology (in this case, context-awareness) that impact the usability of a system.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 205-217, copyright 2008 by Information Science Publishing (an imprint of IGI Global).

Chapter 8.3
Policy-Based Mobile Computing

S. Rajeev
PSG College of Technology, India

S. N. Sivanadam
PSG College of Technology, India

K. V. Sreenaath
Arizona State University, USA

ABSTRACT

Mobile computing is associated with mobility of hardware, data and software in computer applications. With growing mobile users, dynamicity in catering of mobile services becomes and important issue. Polices define the overall behavior of the system. Policy based approaches are very dynamic in nature because the events are triggered dynamically through policies, thereby suiting mobile applications. Much of the existing architectures fail to address important issues such as dynamicity in providing service, Service Level provisioning, policy based QoS and security aspects in mobile systems. In this chapter we propose policy based architectures and test results catering to different needs of mobile computing

INTRODUCTION

Policies are rules that govern the overall functioning of the system. *Policy computing* is used in a variety of areas. *Mobile computing*, with its ever-expanding networks and ever-growing number of users, needs to effectively implement a policy-based approach to enhance data communication. This can result in increasing customer satisfaction as well as efficient mobile network management.

POLICY COMPUTING AND NEED FOR POLICY-BASED MOBILE COMPUTING

Policies in society and organizations are often captured and enforced as laws, rules, procedures, contracts, agreements, and memorandums. Policies are rules that govern the choices of system

behavior. A policy is defined as "a definite goal, course or method of action to guide and determine present and future decisions." Security policies define what actions are permitted or not permitted, for what or for whom, and under what conditions. Management policies define what actions need to be carried out when specific events occur within a system or what resources must be allocated under specific conditions. They are widely used for the mobile user whose requirements are dynamic.

Policy-based computing is the art of using policy-based approaches for effective and efficient computing; it is widely used because of its dynamicity. Hence in areas such as mobile computing, policy computing can be effectively used.

Much of the existing network systems' are configured statically (Fankhauser, Schweikert, & Plattner, 1999). In the present-day scenario, the number of mobile/wireless network users increases day by day. With the static systems being deployed, it is very difficult to achieve the needed dynamicity for mobile computing resulting from changing user base. In order to achieve efficient communication for fluctuating user base, policy-based systems need to be implemented in different areas of the existing wireless mobile network infrastructure.

POLICY IN MOBILE COMPUTING

Mobile computing is conducted by intermittently connected users who access network resources that need to escalate with increasing computing needs. Mobile computing has expanded the role of broadcast radio in data communication, and with increasing users, providing quality service becomes a challenging issue. The mobile users must be provided with the best possible service so that the service provider can stay in competition with peer service providers. In order for the best possible service to be provided to the mobile users, there are certain criteria that should be met. They are:

- The quality of service should be guaranteed.
- There should be effective service-level agreement (SLA) between the mobile user and the service provider.
- Security should be foolproof.

With the existing system (without a policy-based approach), it becomes very difficult to achieve the mentioned criteria. It is very difficult to provide a guaranteed quality of service (QoS), which is also dynamic (not statically configured). Moreover SLA is a very static procedure. Because of the mobility and dynamicity of mobile networks, SLAs also must be made very dynamic. Similarly, security should also be made very dynamic and efficient. To overcome all these shortcomings of the existing system, *a policy-based approach should be used in mobile networks.*

Policy computing can be effectively implemented in mobile networks using policy compilers. Policies can be written in different ways. There are different languages for writing policies that are used for different purposes of specifying policies. In order that the "security policies" be specified, languages such as Trust Policy Language (TPL), LaSCO, and so forth are used. In a similar way, for specifying management-related policies, languages such as Ponder, Policy Maker, and so forth are used. Thus for different scopes of application of policies, specific languages are used.

Policy validation checks a solution's conformance to the policy file. The actual process of policy validation has three primary stages. First, a node or hierarchy change event in Solution Explorer (such as add, drag, or delete event) begins the validation process. Then the validation process maps items discovered in the solution (such as files, references, classes, or interface definitions) to a corresponding Template Description Lan-

guage (TDL) policy ELEMENT node. Finally, for recognized ELEMENT nodes, the validation process checks the parent ELEMENT for policy compatibility with the child ELEMENT. When the policies are compatible, the validation process applies any ELEMENT-specific policy.

APPLICATIONS OF POLICY COMPUTING

Policy-based management is an over-arching technology for an automated management of networks (Lewis, 1996). Policy-based management is being adopted widely for different domains like quality of service, wireless networks, service-level agreement, virtual private networks (VPNs), network security, and IP address allocation. Therefore policy-based networking configures and controls the various operational characteristics of a network as a whole, providing the network operator with a simplified, logically centralized, and automated control over the entire network. In a wireless/mobile network, events are user and time based. These events are very dynamic in nature in order to provide the best service to the mobile users and also to maximize the profit of the service provider. But most of the existing systems are very static in nature. With the static systems being deployed, it is very difficult to achieve the needed dynamicity for mobile computing. In order to achieve this, policy-based systems need to be implemented in different areas of the existing wireless mobile network infrastructure such as QoS in wireless networks (especially differentiated networks), security, and SLAs.

Policy-Based Architecture for Security

Some of the key issues involved in providing services for wireless networks are (Sivanandam, Santosh Rao, Pradeep, & Rajeev, 2003):

1. **Bandwidth Cost:** Depending upon the number of users connected, the location (e.g., urban or remote), and the type of service (e.g., video, audio, etc.) being offered, dynamic allocation of bandwidth plays an important role.
2. **Limited Memory:** Today's wireless device places constraints on the amount of data that it can hold. Moreover, this limit depends upon the device being used and hence causes greater concern with low memory devices.
3. **Access Cost:** Optimizing the cost (Boertien, Janssen, & Middelkoop, 2001) of accessing and transferring data is more complex in wireless networks than in wired networks. If the number of servers used by a service or the number of services provided by an enterprise increases, then maintaining service consistency would turn out to be a cost factor in itself.
4. **Scalability Requirements:** These requirements force the service provider to think in terms of developing a solution that would support increasing and decreasing the number of services offered by the enterprise.

These constraints adversely affect the process of implementation of wireless/mobile services over the existing architecture. To overcome this, an identity management architecture for a wireless differentiated service schema that could be implemented using LDAP (lightweight directory access protocol) (Hodges & Morgan, 2002), directory structures are constructed.

Policy Warehouse

The concept of differentiated services entitles the maintenance of a large amount of information pertaining to the user (e.g., user names, passwords, services registered, premium amount, etc.) and an efficient quick access mechanism to retrieve the relevant details. This overhead increases when

it comes to wireless services. Here, the policy warehouse acts as the information backbone of the service provisioning system. The service provisioning engine contacts the policy warehouse whenever the service provider forwards to it a request from the user, after the right user has been authenticated.

1. **Id-Synch and P-Synch:** The synchronization of user identities and passwords pertaining to a single user is highly crucial in providing a hassle-free connection to services that require subscription to external back-end service providers. This architecture uses Id-Synch and P-Synch mechanisms for identity synchronization and password synchronization respectively.
2. **Meta Directories:** In general, service providers need to maintain a global user profile to uniquely identify a user over the various services provided to him. This information, which mainly comprises a collection of information pertaining to the user sign-on details of various services, is stored in meta directories. This profile makes it easy for the service provisions engine to authenticate the user. The architecture has a provision of compiling as well as retrieving meta directory information.
3. **LDAP Access Engine and Directory Structures:** Information pertaining to the user is stored in lightweight directory structures that can be retrieved using the LDAP. Directory structures are used to store user information because they provide a systematic mechanism for organizing data under a common head like user profiles, user services, user privileges, and so forth that are organized in a hierarchical manner on multiple workstations that are distributed over a network. This not only makes data retrieval fast, querying complexity less, and volume of data storage minimum, but it also makes easy implementation of policies that depend on the enterprise using the system.

The LDAP engine acts as an interface between the various servers like Id-Synch and P-Synch, and directories like meta directories and the underlying LDAP directory structures. They process the request for data from the higher layers and hand over the appropriate data to the requested application in the required format.

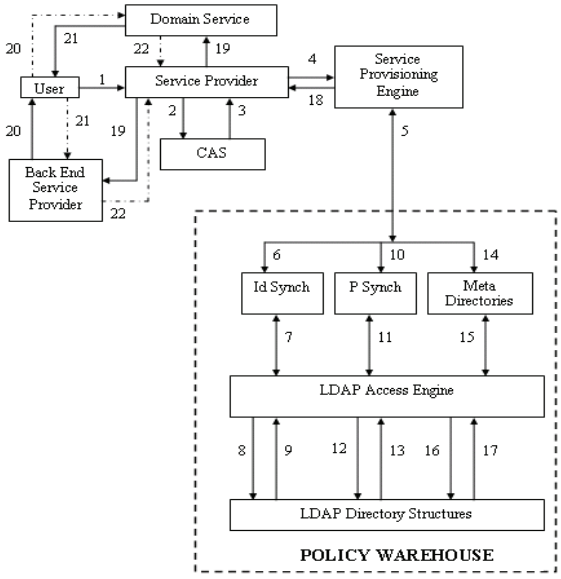
The LDAP directory services are part of the directory-enabled network services (DEN) that provide standard APIs for the access of network objects.

Service Provisioning Mechanism

The service provisioning system can be generally viewed in the following stages:

- **User Login:** In this stage, the user sends his details like User Name, Password, and Chip Index Number to the service provider for authentication.
- **Service Provider:** The Service Provider has to perform the following two actions on a service request:
  1. **Authentication:** The authentication is accomplished using the DSAP (distributed substring authentication protocol) (Sivanandam et al., 2003). This is done by fragmenting the user details into sub-strings and distributing them over a network which is monitored by a central authentication system. When the user is required to authenticate, the protocol fetches the appropriate sub-strings from the network and compares them to the user input. A match signifies a valid user. After this stage, the appropriate user policy is fetched from the policy warehouse using the service provisioning engine.

Figure 1. Policy-based provisioning



2. **Providing the Service:** This is the last step of the service provisioning. In this stage, the actual service that the user has requested is granted. The service could be from a back-end service provider or from the main service provider. This detail is an abstraction to the user who undertakes all transactions with the main service provider only. When the user disconnects from the service, intimation is sent to both the back-end service provider and to the main service provider. This has two implications: firstly, the main service provider's load is shared by the back-end service provider, and secondly, the intimation during the connection termination ensures that the main service provider gets the appropriate usage details. This can act as verification of the details that the back-end service provider will submit later.

Policy Based Architecture for QoS

The interface to the network device and the information models required for specifying policies are either standardized or being standardized in IETF and DMTF. An architecture for a policy-based QoS management system for Diffserv-based wireless networks, which are based on COPS for interfacing with the network device and on LDAP for interfacing with a directory server for storing policies, is constructed. The Diffserv policies are installed based on role combination assigned to the network device interfaces. The directory access could become a bottleneck in scaling the performance of the policy server, and it can be improved substantially by employing appropriate policy caching mechanisms. The framework considers various QoS parameters in the wireless network and proposes the policy-based architecture for QoS management in wireless networks.

Wireless Network QoS Parameters

The wireless/mobile network is affected by the following QoS parameters:

- **High Loss Rate:** Wireless/mobile networks are characterized by more frequent packet losses because of fading effects. The scheduler may think that a certain DSCP is being satisfied with the required number of packets scheduled, but the receiver is not receiving the packets at the required rate. It will be useful to have feedback from the receiver so that some compensation techniques can be employed. The base station (BS) can better handle compensation of lost bandwidth using this information.
- **Battery Power Constraints:** Current mobile battery technology does not allow more than a few hours of continuous mobile operation. Two of the major consumers of power in a mobile network are the network interface (14%) and the CPU/memory (21%). There-

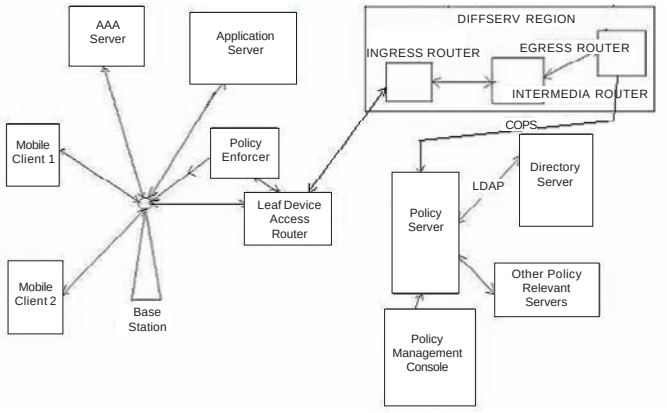
fore, network protocols should be designed to be more energy efficient (Agrawal, Chen, & Sivalingam, 1999). The mobile device can use the signaling mechanism to periodically send messages about its power level to the BS. The BS can then use this information to dynamically decide packet scheduling, packet dropping, and so forth.

- **Classification of Packets within a Flow:** Present Diffserv (Chan, Sahita, Hahn, & McCloghrie, 2003) mechanisms treat all packets within a flow identically. Even though a distinction can be made between packets as in-profile or out-of-profile, all in-profile packets are treated the same way. In many situations (e.g., while using layered video), it may be necessary to distinguish packets within a flow. This is because some packets from a flow level could be more important than the others, and a local condition like power level may lead to different treatments of these packets. Thus, the packets within a flow must be made distinguishable, and bits in the TOS field may be used for this purpose. To summarize, the various possible factors needed to make the Diffserv architecture suitable for wireless networks were discussed in this section.
- **Low Bandwidth:** Wireless networks available today are mostly low bandwidth systems. Most of the current LANs operate at 2 Mbps with migration up to 11 Mbps available. However, the available wireless LAN bandwidth is still an order of magnitude less than the typical wired LAN bandwidth of 100 Mbps. This leads to two decisions. First, the signaling protocol should be very simple and highly scalable. It is also better to modify an existing protocol for compatibility with other existing network protocols. Second, the mobile should not be swamped with too much data from a wired sender with higher network bandwidth. This can be handled to a large extent by transport protocol control, but the problem can be alleviated by handling it partially at the base station. Therefore, mechanisms may be used at the BS to send data to the mobile devices based on current conditions such as channel condition, bandwidth available, and so forth.

Policy-Based QoS

The IETF Resource Allocation Protocol (RAP) working group has defined, among other standards, the policy-based admission control framework, and the common open policy service (COPS) protocol and its extension—COPS for provisioning (COPS-PR). COPS is a simple query protocol that facilitates communication between the policy clients and remote policy server(s). Two policy control models have been defined: outsourcing and provisioning. While COPS supports the outsourcing model, its extension COPS-PR integrates both the outsourcing and provisioning models. The outsourcing model is tailored to signaling protocols such as the resource *r*eser*v*ation *p*rotocol (RSVP) (Braden, Zhang, Berson, Herzog, & Jamin, 1997), which requires traffic management on a per-flow basis. On the other hand, the provisioning or configuration model is used to control aggregate traffic-handling mechanisms such as the Differentiated Services (Diffserv) architecture. In the outsourcing model, when the PEP receives an event (e.g., RSVP reservation request) that requires a new policy decision, it sends a request (REQ) message to the remote policy decision point (PDP). The PDP then makes a decision and sends a decision (DEC) message (e.g., accept or reject) back to the PEP. The outsourcing model is thus PEP driven and involves a direct 1:1 relation between PEP events and PDP decisions. On the other hand, the provisioning or configurations model (Chan et al., 2001) makes no assumptions of such direct one-to-one correlation between PEP events and PDP decisions. The PDP may proactively provision the PEP reacting to external

Figure 2. Policy-based management system architecture



events, PEP events, and any combination thereof (N: M correlation). Provisioning thus tends to be PDP driven and may be performed in bulk (e.g., entire router QoS configuration) or in portions (e.g., updating a Diffserv marking filter).

Architecture of a Policy-Based Management System for a Diffserv-Based Wireless Network

Figure 2 illustrates the architecture of the policy-based management system for Diffserv-based wireless networks. The policy server is responsible for interpreting higher-level policies and translating them into device-specific commands for realizing those policies. For allocating resources on inter-domain links and for implementing SLAs, the policy server (especially the bandwidth broker component) has to communicate with the policy server in the provider.

The policy server is mainly responsible for the following:

- retrieving relevant policies created by the network administrator through the policy console after resolving any conflicts with existing policies;
- translating the policies relevant for each PEP into the corresponding policy information base (PIB) commands;
- arriving at policy decisions from relevant policies for policy decision requests, and maintaining those decision states; and
- taking appropriate actions such as deletion of existing decision states or modification of installed traffic control parameters in the PEP for any modifications to currently installed policies.

All the policies are stored in the LDAP server. The policy editor (PE) is the entity responsible for creating, modifying, or deleting policy rules or entries in the LDAP server. LDAP protocol provides access to directories supporting the X.500 models, while not incurring the resource requirements of the X.500 directory access protocol (DAP). It is specifically targeted at management applications and browser applications that provide read/write interactive access to directories. It does not have the mechanism to notify policy consumers of changes in the LDAP server. Therefore, it is the responsibility of the policy editor to indicate the changes in the LDAP server, as and when required, using an internal event messaging service. The

policy server, in addition to querying the LDAP server, queries other policy-relevant servers such as Certificate server, Time server, and so on.

The policy management client—also referred to as the policy editor—provides a high-level user interface, for operator input translates this input into the proper schema for storage in the directory server and pushes it out to the directory for storage. The authentication, authorization, and accounting (AAA) server is responsible for authentication, authorization, and accounting of the user after the relevant policies have been picked and enforced in the policy enforcers (routers). This AAA server is used by the base station to check if the user is authenticated and authorized for the resource he requests, and to check if he is accounted. The policy enforcer nearer to the base station enforces the policy decisions taken from the policy server. The base station then requests the nearest application server (after policies are enforced) and waits for the response from the application server.

The base station first sends the request to the leaf access router, which then sends it to the ingress router in the region. The ingress router then passes on the requests to the intermediate router. The request passes through the other intermediate routers and reaches the egress router, which sends the request to the policy server through COPS.

Policy-Based Architecture for SLA

Mobile ad hoc networks (MANETs) are autonomous networks operating either in isolation or as "stub networks" connecting to a fixed infrastructure. Depending on the nodes' geographical positions, transceiver coverage patterns, transmission power levels, and co-channel interference levels, a network can be formed and unformed on the fly. Ad hoc networks have found a growing number of applications: wearable computing, disaster management/relief and other emergency operations, rapidly deployable military battle-site networks, and sensor fields, to name a few. The main characteristics of ad hoc networks are:

- **Dynamic Topological Changes:** Nodes are free to move about arbitrarily. Thus, the network topology may change randomly and rapidly over unpredictable times.
- **Bandwidth Constraints:** Wireless links have significantly lower capacity than wired links. Due to the effects such as multiple accesses, multi-path fading, noise, and signal interference, the capacity of a wireless link can be degraded over time and the effective throughput may be less than the radio's maximum transmission capacity.
- **Multi-Hop Communications:** Due to signal propagation characteristics of wireless transceivers, ad hoc networks require the support of multi-hop communications; that is, mobile nodes that cannot reach the destination node directly will need to relay their messages through other nodes.
- **Limited Security:** Mobile wireless networks are generally more vulnerable to security threats than wired networks. The increased possibility of eavesdropping, spoofing, and denial-of-service (DoS) attacks should be carefully considered when an ad hoc wireless network system is designed.
- **Energy Constrained Nodes:** Mobile nodes rely on batteries for proper operation. As an ad hoc network consists of several nodes, depletion of batteries in these nodes will have a great influence on overall network performance. Therefore, one of the most important protocol design factors is related to device energy conservation.

To support mobile computing in ad hoc wireless networks, a mobile host must be able to communicate with other mobile hosts that may not lie within its radio transmission range. Therefore in order for one mobile host in the ad hoc network

to communicate with the other not lying in its transmission range, some other hosts in its transmission range should route the packets from the source to the destination host. The conventional routing protocols used in wired networks cannot be effectively used in ad hoc networks. Hence new routing mechanisms are suggested which may be used for routing in ad hoc networks. Routing issues in ad hoc networks are beyond the scope of this chapter and are not considered here.

Since many mobile hosts may be within transmission range of each other, there may be multiple routes for a packet to reach a destination. Therefore the source host should decide which route to use to send the packets to reach its destination. Obviously, the sending host has to decide on the best optimal route before sending its packets towards the destination. Thus, there should be a service level agreement between the source mobile host and the host which routes the packets to the destination host. Moreover there are certain constraints based on the characteristics of the ad hoc network which play a major role in deciding which route is optimal, given there are more routes to reach the destination.

Architectural Framework of the Policy-Based Mobile Ad Hoc Network

MANET is a collection of mobile hosts forming a temporary network without the aid of any centralized administration or standard support services. The architecture for the policy-based SLAs in ad hoc networks is given below. The architecture is designed where at least one host has connectivity with the wired network. In the architecture shown in Figure 3, the policy server is placed in the wired network. Polices are stored in the directory server. The ad hoc 'host1' is within the vicinity of both 'host2' and 'host3'. 'Host4' is not within the transmission region of 'host1'. So when 'host1' wants to send a packet to 'host4', intermediatory hosts, 'host2' and 'host3', help 'host1' with connection establishment.

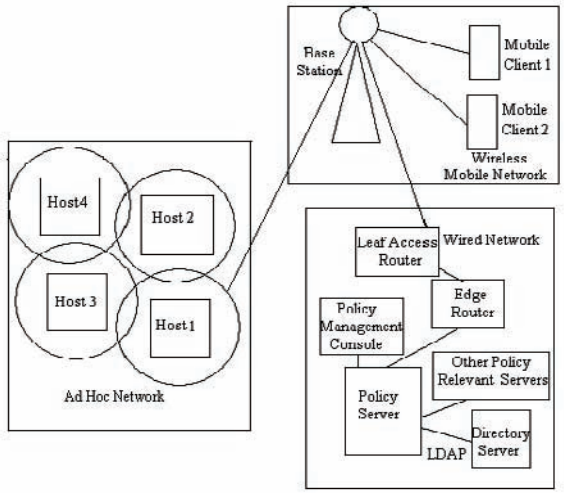
Assuming that both the host services satisfy the constraints of 'host1', 'host1' must choose a service level agreement among the two. In this case since 'host1' is connected to a base station, which in turn is connected to the wired network having the policy server, 'host1' can query the policy server through the base station and then the leaf access router and edge router. For simplicity we have shown the policy server being connected to the base station through only a few hops. But in practice it may be many hops away from it. Once the request reaches the policy server, it takes appropriate policies from the directory server through the LDAP.

The policy server also communicates with other relevant policy servers such as Time Servers, Certificate Servers, and AAA servers, and validates the host providing service by means of certificates and AAA. The policy server makes the decision on whether the host providing the service is an authenticated one, and his services are authorized with accountability and certificates. Then the policy server based on the higher level polices stored in the directory server chooses an agreement among the available agreements. The decision to choose an agreement from among the available ones may be done giving more weight to those performance metrics which affect the overall performance the most. Over a period of time, the history of the hosts providing the service will be stored; solutions based on a neural network model may be used for finding an optimal solution. Once the policy server chooses an agreement, it sends its reply back to 'host1', which agrees for the service with the appropriate host.

Policy Based E-Supply Chain Management Architecture

Internet-based e-purchases and e-supply chain management are now being widely used. This however has a major disadvantage of very limited mobility and the absence of a dynamic policy that will efficiently manage the entire supply chain.

Figure 3. Architecture of policy-based MANET



The activities that are required to provision mobile users comprise a surprisingly large number of steps that cross an entire enterprise. Policy setting and implementation, approval workflows, physical resource setup/teardown (provisioning), account maintenance, reconciliation of actual resource assignments with approved user lists, audit, and overall service management are some examples. They are together called policy-based provisioning. An extension to e-purchases through mobile phones using policy-based e-supply chain management is constructed.

Architecture of Policy-Based E-Supply Chain Management

The architectural framework of the e-purchase through policy based e-supply chain management is shown in Figure 4. Mobile customer, policy server, nodes (1-4), and suppliers (1 and 2) form the key elements of the proposed architecture. The mobile consumer requests the policy server of the service provider through the base station. The mobile user's authentication, authorization, and accounting rights are then verified by the AAA server. If the mobile user is found to be an authenticated one and his request for service is an authorized one, then the policy server fetches the corresponding policies for the user from the directory server through LDAP. The policy server of the service provider is connected to other nodes (Node 1 to Node 4) in the Internet. The nodes of the suppliers such as 'supplier1' and 'supplier2' are also connected to the nodes through the public network (Internet). When the relevant policies are fetched from the directory server, the user's request is sent to an efficient supplier who will supply the product to the mobile customer. The supply chain is made electronic as discussed earlier and is policy based. Once the user's request is sent to an efficient supplier relatively nearby to the customer, the ordered products will be delivered to the mobile customer through the shipping department. The billing of the products purchased is taken into the credit account of the mobile user and is charged along with the mobile phone bill, simplifying user billing and payment. The customer at the end of the month would pay the bill through the electronic account facility available with his existing bank account. Thus the whole process of placing the purchase order, delivering the product through the supply chain, and paying for the product is made electronically, thereby facilitating the customers, who are mostly travelers and tourists.

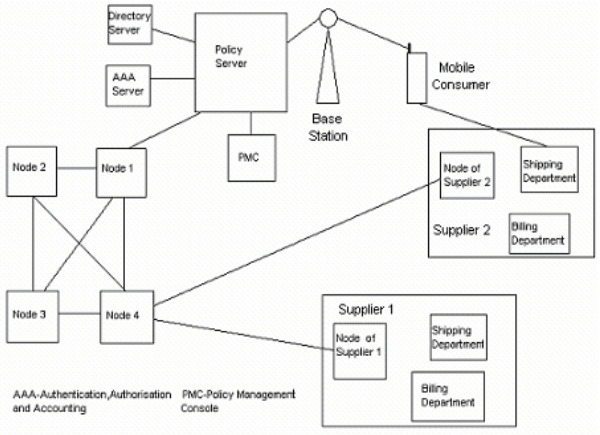
CASE STUDY

A simulation for the policy-based MANET shown in Figure 3 was performed using the QualNet Network Simulator, using the simplex method to solve the linear program model given below, and using the Ponder Toolkit. The mathematical model is given in the following section.

Mathematical Model

A mathematical model considered for policy-based MANET uses Linear Programming and Simplex Method to solve it. The following perfor-

Figure 4. Framework of e-purchase through policy-based e-supply chain management



mance metrics that are crucial for effective SLA trading and choice of route are considered in our model: (a) Bandwidth, (b) Delay, (c) Demand, (d) Packet Loss, (e) Congestion, (f) Queuing Delay, (g) Throughput, (h) Buffer Capacity, (i) Battery Consumption, and (j) Mobility. Let,

T_{ij} = Total (maximum) Bandwidth (channel capacity) available from host *i* to host *j*.
U_{ij} = Bandwidth being used for traffic flow between host *i* to host *j* at instant '*t*'.
R_{ij} = Reserved bandwidth from host *i* to host *j*.

Hence the bandwidth that can be leased to other hosts G_{ij} is given by $G_{ij} = T_{ij} - U_{ij} - R_{ij}$.

Let the required bandwidth—that is, the bandwidth consumed by the host *k* to reach host *j* through host *i*—be RB_{ij}.

And,

D_{ij} = Delay from host *i* to host *j*.
C_{ij} = Cost of reaching host *j* through host *i*.
F_{ij} = Fraction of bandwidth bought from host *i* to reach host *j*.

The objective here is to minimize the cost of reaching host *j* through other hosts.

$$Minimise \sum_{i,j} F_{ij} C_{ij} \quad (1)$$

As stated earlier, in the above equation represents the cost hostcharges to reach host through host.

Constraints

There are a set of constraints that define the model. The first constraint is the demand for bandwidth to reach host *j* through host *i*; De_{ij} should be less than or equal to the amount of bandwidth host *i* is ready to offer for cost to reach host *j*, G_{ij}.

$$DE_{ij} \leq \sum_{i,j} G_{ij} \quad (2)$$

The following constraints check if the service performance metrics in the service offered by the host *i* to reach host *j* fall within the predetermined and pre-calculated boundaries as expected by host *k* which needs the service. These boundary constants for the performance metrics can also be set dynamically and SLA negotiated accordingly.

Buffer Capacity B_{ij} should not be less than a bearable value given by the constant N=Number of packets that can be buffered.

$$B_{ij} \geq N \quad (3)$$

The time delay *D* should be set to a limit expressed by a constant '*p*1'as expected by the 'ISP k' which needs the service. The constant '*p*1' is arrived as derived as follows:

'*p*1' = Propagation Time + Transmission Time + Queuing Delay (+ Setup Time)
Propagation Time: Time for signal to travel length of network
= Distance/Speed of light
Transmission Time = Size/Bandwidth

Therefore, we have

$$D_{ij} \le p1 \quad (4)$$

Queuing Delay Q_{ij} should not exceed an allowable limit '$p2$' expressed as

$$p2 = \frac{D}{2} \times (N-1)$$

where, D = the time delay, N is the Buffer Capacity

$$Q_{ij} \le p2 \quad (5)$$

The Packet Loss P_{ij} for the service provided should not exceed a maximum limit set as constant '$p3$', and Congestion in the channel offered for service Co_{ij} should also be within the acceptable limits represented by the constant '$p4$', both of which are arrived at as shown as follows:

T_{min} = Minimum Inter-Arrival Time observed by the receiver.

P_0 = Out of order packet.

P_i = Last in-sequence packet received before P_0.

T_g = Time between arrival of packets P_0 and P_i.

n = Packets missing between P_i and P_0.

If $(n+1)\,T_{min} \le T_g < (n+2)T_{min}$, then n missing packets are lost due to transmission errors and hence '$p3$'='n'and

$$P_{ij} \le p3 \quad (6)$$

Else n missing packets are assumed to be lost due to congestion and hence '$p4$'='n' and

$$Co_{ij} \le p4 \quad (7)$$

Throughput TH_{ij} should be greater than or equal to '$p5$', which is given by

$$p5 = \{MSS / RTT\} \times C / (\sqrt{p})$$

where,

MSS = Maximum Segment size in bytes, typically 1460 bytes.

RTT = Round Trip Time in seconds, measured by TCP.

p = Packet loss.

C = Constant assumed to be 1.

$$TH_{ij} \ge p5 \quad (8)$$

The jitter J_{ij} should be within the acceptable limit '$p6$' given by

$$p6 = p6 + (|D(i-1,i)| - p6)/16$$

given

$$D(i,j) = (R_j - S_j) - (R_i - S_i)$$

where, S_i, S_j are sender timestamps for packets i, j and R_i, R_j are receiver timestamps for packets i, j.

Therefore

$$J_{ij} \le p6 \quad (9)$$

The Battery Consumption BC_{ij} for the offered service should be within the boundary constant '$p7$',

$$BC_{ij} \le p7 \quad (10)$$

The Mobility Factor M_{ij} which gives the idea of how long the host j will be in the transmission range of host for which packets need to be routed should not be smaller than a particular constant represented by '$p8$',

$$M_{ij} \ge p8 \quad (11)$$

This mobility factor M_{ij} plays a crucial role in ad hoc networks because the hosts are all mobile. It may be minutes or in any preferred time unit as the case may be. We generally assume that a mobile which has joined the ad hoc has more probability of staying in the network than the ones which came earlier than that. But the exact nature of the mobility of a host can be predicted only based on past performances of the mobile.

Non-Negativity Constraints

The following are the non-negativity constraints applied in the model:

Cost C_{ij} should always be positive,

$$C_{ij} \geq 0 \qquad (12)$$

Table 1. Performance metrics and other parameters of the hosts

Performance Metrics	Host 1	Host 2	Host 3
Total Bandwidth Allocated (MBps)	3	6	5
Bandwidth Used at Instant (MBps)	2	2	1
Reserve Bandwidth (MBps)	0	1	1
Remaining Bandwidth G_{ij} (MBps)	1	3	3
Demand for Bandwidth to Reach 'host4' (MBps)	1	0	0
Delay (x 10-3/sec)	7	8	10
Packet Loss Factor	7	5	6
Congestion Factor	30 2	0	25
Queuing Delay (x 10-4sec)	8	7	10
Throughput (x 103 Bits/sec)	100	100	90
Buffer Capacity (No. of Packets)	9	10 8	
Battery Consumption (mWh)	-	8	9
Mobility Factor? Minutes	-	25 1	8

Fraction of bandwidth bought from host *i* to reach host *j*, F_{ij} should also be positive,

$$F_{ij} \geq 0 \qquad (13)$$

The bandwidth that can be offered for cost to other hosts by host *i* should be positive,

$$G_{ij} \geq 0 \qquad (14)$$

Given the objective, for example, to minimize the agreement cost along with the performance metrics constraints, the proposed linear programming model solved using simplex method suffices for arriving at a suitable agreement for service with other hosts. There are always cases that the above model will fetch more than one solution if other solutions exist. Hence in such cases the decision of choosing the most appropriate of the available solutions should be taken which is described in the next section.

Table 2. Performance metrics and other parameters of 'host2' and 'host3'

Performance Metrics	Host 2	Host 3
Delay (x 10-3/sec)	2.9	3.1
Packet Loss Factor	0.2	0.3
Congestion Factor	0.3	0.3
Queuing Delay (x 10-4sec)	0.2	0.2
Throughput (x 103 Bits/sec)	4.2	4.2
Buffer Capacity (No. of Packets)	20 1	5
Battery Consumption (mWh)	8	9
Mobility Factor (minutes)	25 1	8
Jitter (x 10-4sec)	3.9	4.1
Fraction of Bandwidth that Can Be Given F_{ij} (MBps)	1	1
Cost C_{ij} ($)	2	6

The test environment has four ad hoc hosts from 'host1' to 'host4', as shown in Figure 3. The total bandwidth, used bandwidth, reserve bandwidth, battery consumption, mobility factor, and other performance metrics of the hosts are tabulated below.

In the simulation test environment, 'host1' needs to communicate with 'host4', which is not in its transmission range. So both 'host2' and 'host3' offer the service to 'host1'. Using the mathematical model proposed, 'host1' decides upon the suitable service among the offers using the SLA trading algorithm (Rajeev, Sivanandam, Sreenaath, & Bharathi Manivannan, 2005) and the mathematical model given previously. Since only the service offered by 'host2' adheres to the performance metric constraints, 'host1' chooses the service offered by 'host2'. All the simulation is done with respect to the packet flow from 'host1' to 'host4'.

The trade for the service is decided by using the simplex method to solve the linear programming model and SLA trading algorithm, by which a feasible solution is obtained. The performance constraints and other parameters of the hosts are given in Tables 1 and 2. According to the constraints given by 'host1' for the required service, the simplex method and SLA trading algorithm are used, and the best bid among the bids offered by the two hosts ('host2' and 'host3') is selected. Since only the bid for the service offered by 'host2' satisfies the constraints of 'host 1', SLA between 'host1' and 'host 2' takes place. The LP model is solved by using the simplex method. As only the trade provided by 'host2' satisfies all the constraints with the objective of minimum cost, the Service offered by 'host2' is agreed upon for trade.

From the performance metrics and the constraints on performance metrics, the objective of minimizing cost is arrived at (see Figure 5 and Table 3). Thus an effective SLA is traded between 'host1' and 'host2', satisfying the constraints on the performance metrics which affect the service.

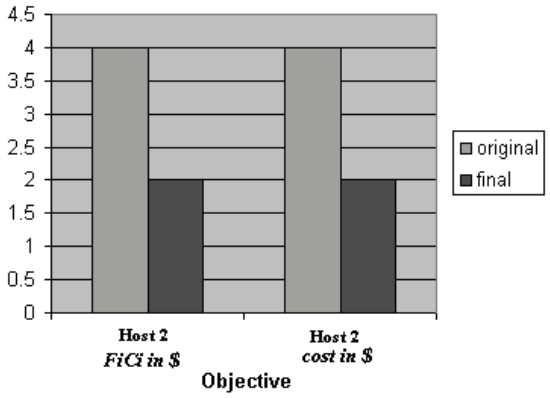
CONCLUSION AND FUTURE DIRECTIONS

Policy computing can be effectively used in mobile computing in various arenas such as QoS, security, SLA, and e-purchase. The architectural framework demonstrated in the case study gives insight as to how QoS, SLA, and security can be implemented in mobile networks. Policy-based architectures for billing in mobile networks are currently being constructed which could bring transparency in mobile billing with added dynamicity.

Table 3. Original and final value of the objective

Objective	Original Value	Final Value
$\sum_{i,j} F_{ij} G_{ij} C_{ij}$ of 'host2' ($)	12 6	
Cost of 'host2' ($)	4 2	

Figure 5. Objective



REFERENCES

Agrawal, P., Chen, J. C., & Sivalingam, K. M. (1999). *Energy efficient protocols for wireless networks*. Norwell, MA: Kluwer Academic Publishers.

Braden, R., Zhang, L., Berson, S., Herzog, S., & Jamin, S. (1997). *Resource ReSerVation Protocol (RSVP)—version 1 functional specification*. IETF RFC 2205.

Chan, K. et al. (2001). *COPS usage for policy provisioning (COPS-PR)*. IETF RFC 3084.

Chan, K., Sahita, R., Hahn, S., & McCloghrie, K. (2003). *Differentiated Services quality of service policy information base*. IETF RFC 3317.

Fankhauser, G., Schweikert, D., & Plattner, B. (1999). *Service level agreement trading for the Differentiated Services architecture*. Technical Report No. 59, Computer Engineering and Networks Lab, Swiss Federal Institute of Technology, Switzerland.

Hodges, J., & Morgan, R. (2002). *Lightweight Directory Access Protocol (v3): Technical specification*. IETF RFC 3377.

Lewis, L. (1996). Implementing policy in enterprise networks. *IEEE Communications Magazine, 34*(1), 50-55.

Rajeev, S., Sivanandam, S. N., Sreenaath, K. V., & Bharathi Manivannan, A. S. (2005). Policy-based SLA for wireless ad hoc networks. In *Proceedings of the International Conference on Services Management,* India.

Sivanandam, S. N., Santosh Rao, G., Pradeep, P., & Rajeev, S. (2003). Policy-based architecture for authentication in wireless Differentiated Services using Distributed Substring Authentication Protocol (DSAP). In *Proceedings of the International Conference on Advanced Computing,* India.

This work was previously published in Handbook of Research in Mobile Business, edited by B. Unkelkar, pp. 613-629, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.4
Field Evaluation of Collaborative Mobile Applications

Adrian Stoica
University of Patras, Greece

Georgios Fiotakis
University of Patras, Greece

Dimitrios Raptis
University of Patras, Greece

Ioanna Papadimitriou
University of Patras, Greece

Vassilis Komis
University of Patras, Greece

Nikolaos Avouris
University of Patras, Greece

ABSTRACT

This chapter presents a usability evaluation method for context aware mobile applications deployed in semi-public spaces that involve collaboration among groups of users. After reviewing the prominent techniques for collecting data and evaluating mobile applications, a methodology that includes a set of combined techniques for data collection and analysis, suitable for this kind of applications is proposed. To demonstrate its applicability, a case study is described where this methodology has been used. It is argued that the method presented here can be of great help both for researchers that study issues of mobile interaction as well as for practitioners and developers of mobile technology and applications.

INTRODUCTION

Mobile devices are part of many peoples' everyday life, enhancing communication, collaboration, and information access potential. Their vital charac-

teristics of mobility and anywhere connectivity can create new forms of interaction in particular contexts, new applications that cover new needs that emerge, and change the affordances of existing tools/applications.

A case of use of such devices, with particular interest, concerns *public places rich in information* for their visitors, in which mobile technology can provide new services. Examples of such places, are *museums* and other sites of culture (Raptis, Tselios, & Avouris, 2005), *public libraries* (Aittola, Parhi, Vieruaho, & Ojala, 2004; Aittola, Ryhänen, & Ojala, 2003), and *exhibition halls* and *trade fairs* (Fouskas, Pateli, Spinellis, & Virola, 2002). In these places, mobile devices can be used for information collection and exchange, for ad hoc communication with fellow visitors, and for supporting face-to-face interaction.

Usability evaluation of mobile applications is of high importance in order to discover, early enough, the main problems that users may encounter while they are immersed in these environments. Traditional usability evaluation methods used for desktop software cannot be directly applied in these cases since many new aspects need to be taken in consideration, related to mobility and group interaction. Therefore, there is a need either to adapt the existing methods in order to achieve effective usability evaluation of mobile applications or to create new ones. An important issue, that is discussed here, is the *process* and *media* used for recording user behaviour.

Data collection during usability studies is a particularly important issue as many different sources of data may be used. Among them, *video and audio* recordings are invaluable sources for capturing the context of the activity including the users' communication and interaction. It has been reported that in cases of studies that audio and video recordings were lacking, it was not possible to explain why certain behaviour was observed (Jambon, 2006). Recording user behaviour is a delicate process. Video and audio recording must be as unobtrusive as possible in order not to influence the behaviour of the subjects while, on the other hand, the consent of the users for their recording should be always obtained. In addition, questions related to the frame of the recorded scene, viewing angle, and movement of the camera are significant. It must be stressed that there is a trade off between capturing the interaction with a specific device and capturing the overall scene of the activity. For example, often, crucial details may be missing from a video if recording the scene from a distance. Therefore, this video has to be complemented by other sources of related information, like screen captures of the devices used.

In order to conduct a successful usability evaluation, apart from collecting activity data, techniques and tools are needed for analysis of the collected information. In the last years, new usability evaluation techniques have emerged, suitable for mobile applications. Many of these methods focus mainly on user interaction with the mobile device, missing interaction between users, and user interaction with the surrounding environment.

Taking into consideration these aspects, the aim of this chapter is to discuss techniques and tools used first, for collecting data during usability evaluation studies of mobile devices, and then for the analysis of these data. In the process, a combination of a screen capturing technique and some tools that can be used for analysis of data of usability studies are presented.

BACKGROUND

The usability of a product has been traditionally related with the ease of use and learn to use, as well as with supporting users during their interaction with the product (Dix, Finley, Abowd, & Beale, 2003; Schneiderman & Plaisant, 2004). There have been many attempts to decompose further the term and render it operational through attributes and apt metrics. According to ISO 9241-

11 standard, usability is defined as the "*extend to which a product can be used with effectiveness, efficiency and satisfaction in a specified context of use*" (ISO 9241). According to this view, a product's usability is directly related to the *user*, the *task*, and the *environment*. Consequently, usability cannot be studied without taking into consideration the goals and the characteristics of typical users, the tasks that can be accomplished by using the product, and the context in which it is going to be used. Making a step further on defining usability, the same standard suggests three potential ways in which the usability of a software product can be measured:

a. By analysis of the *features of the product* required for a particular context of use. Since ISO 9241 gives only partial guidance on the analysis process, in a specific problem there can be many potential design solutions, some more usable than others.
b. By analysis of the *process of interaction*. Usability can be measured by modeling the interaction with a product for typical tasks. However, current analytic approaches do not produce accurate estimates of usability since interaction is a dynamic process which is directly related to human behaviour that cannot be accurately predicted.
c. By analyzing the *effectiveness and efficiency*, which results from use of the product in a particular context, that is, measuring performance as well as the satisfaction of the users regarding the product.

Having in mind the three perspectives, there is a need for combining methods that capture the specific situation of use in a specific domain. Usability evaluation methods can be grouped in four categories (Nielsen, 1993): *Inspection, user testing, exploratory, and analytic methods.* Many techniques have been devised along these lines and have been extensively used in usability evaluation of desktop applications. Therefore the first approach in evaluating mobile applications was to apply these existing techniques. Such an approach can be found in Zhang. and Adipat's (2005) survey of usability attributes in mobile applications which identified nine attributes that are most often evaluated: learnability, efficiency, memorability, user errors, user satisfaction, effectiveness, simplicity, comprehensibility, and learning performance. Such an approach is, however, limited, given the special characteristics of mobile devices with respect to desktop environments (Kjeldskov & Graham, 2003).

The mobile applications introduce new aspects to evaluate. The evaluation cannot be limited only to the device (typical scenario in desktop applications) but it must be extended to include aspects of context. The context in which the application is used is highly relevant to usability issues and often bears dynamic and complex characteristics. There is the possibility that a single device is used in more than a single context, in different situations, serving different goals and tasks of a single or a group of users. Also, group interaction, a common characteristic in mobile settings, gives a more dynamic character to the interaction flow of a system and increases the complexity of the required analysis as well as the necessity of observational data.

Along these lines, a new breed of methods for usability evaluation has been proposed (Hagen, Robertson, Kan, & Sadler, 2005; Kjeldskov & Graham, 2003; Kjeldskov & Stage, 2004). The process of selecting appropriate usability attributes to evaluate a mobile application depends on the nature of the mobile application and the objective of the study. A variety of specific measures (e.g., task execution time, speed, number of button clicks, group interactions, seeking support, etc.) have been proposed to be used for evaluation of different usability attributes of specific mobile applications. In the next section problems of data collection during mobile usability studies will be discussed.

Data Collection Techniques

A significant step during a usability evaluation study is to collect appropriate observational data to be analyzed. Hagen, Robertson, Kan, and Sadler (2005) classify the data collection techniques for mobile human-computer interaction in three categories: (a) *Mediated data collection (MDC)*, access to data through participant and technology, *do it*—the user makes himself the data collection; *use it*—data is collected automatically through logs; *wear it*—user wears recording devices that collect the data. (b) *Simulations and enactments (SE)* where some form of pretending of actual use is involved and (c) *combinations* of the above techniques. A review of different techniques of data collection, according to Hagen, Robertson, & Kan (2005) is shown in Table 1.

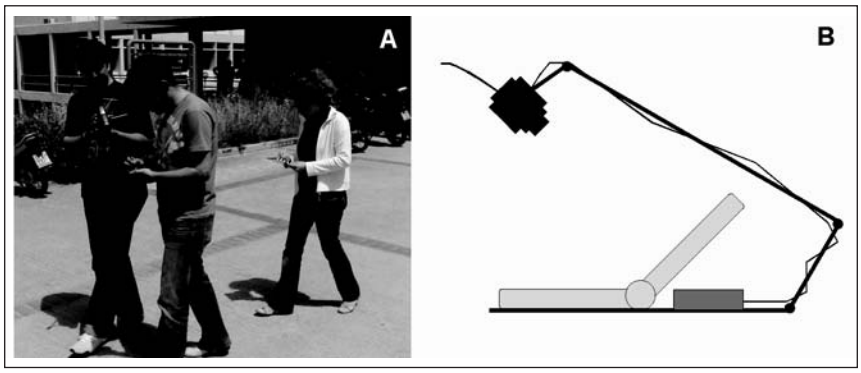
The data that are collected by these techniques come either directly from the user (through interviews, questionnaires, focus groups, diaries, etc.), by the evaluator (i.e., notes gathered during the experiment, observation of videos, etc.) or by raw data (log files, etc). All types of data need to be analyzed in order to become meaningful. Such data, in most cases, are in the following forms:

- *Log files* which contain click streams of user actions. These data can be derived by the application itself or by an external tool that hooks into the operating system message handler list. The latter case for mobile devices requires many system resources and therefore is not technologically feasible today, even in the most powerful mobile devices, like PDAs.
- *Audio/video recordings* of the users made through various means, like wearable mini cameras and/or audio recorders, static video cameras, operator or remote controlled cameras, from close or a far distance.
- *Screen recordings* by video cameras or by direct screen capturing through software (running on the device) the interaction

Table 1. Existing techniques for data collection used in studies of mobile technology. Adapted from Hagen, Robertson, and Kan (2005). F=Field, L=Laboratory, MDC=Mediated Data Collection, SE=Simulation and Enactments

Method	Description	Site*	Category
Artefacts (e.g. documents)	The use of objects or documents as sources for data collection. They may be objects (or photos of objects) from daily life or documents that users have created with devices being tested.	F	MDC
User Diaries	Users document information about their actions or thoughts, or impressions, often daily, for a period of time. Entries can be open and interpretive, or highly structured depending on the study.	F	MDC
Emulators	Emulators on desktop computers are used to simulate the interface of a potential mobile application.	L	SE
Focus Groups	Small groups of people are facilitated in unstructured discussion about an issue.	F, L	SE
Heuristics	Heuristics, often usability guidelines or design principles are applied by expert users to predict usability problems.	F, L	SE
Interviews	Interviews capture subject data from talking directly to participants. They can be open or structured and conducted in the field (including contextual interviews), online, over the phone and in labs.	F, L	SE
Log File analysis	Use logs are generated automatically (such as internet log files) or from systems specifically developed to capture content data and meta data.	F	MDC
NASA Task Load Analysis	Used in usability testing to determine work load.	F, L	SE
Observation/ Shadowing	Observation is used in field studies to capture use in context and can include, covert observation, participant observation, observing a place, or following a person. Data collection can include note taking, photography, and video.	F	MDC
Online data	Researchers gain access to information about the lives of users, and use practices from websites, forums and mailing lists.	F	MDC
Questionnaires	Quantitative or qualitative questionnaires are used to collect user opinions, feedback in evaluation, create user profiles or collect data about existing use practices. They can be done in person, or via phone or web.	F	SE
Role playing	Users and researchers play out different roles, or act out tasks or scenarios to explore existing and future use concepts	F, L	SE
Scenarios	Scenarios provide information about use situations giving examples of how technologies are used in practice.	F, L	SE
Think-Aloud	Participants describe out loud what they are thinking while they complete tasks using a device or prototype.	F, L	SE

Figure 1. A) Shadowing technique (see also Kjeldskov & Stage, 2004); B) Recording screen with wireless camera (see also Betiol & de Abreu Cybis, 2005)



flow in form of screen snapshots. This is a sequence of image representations of the user interface at certain instances that are usually taken at varying frequencies, usually a few snapshots per second. The screen snapshots can be stored either locally on the device (since it is feasible to store a large amount of data in memory cards) or on a central server over a wireless network connection

Screen recordings of mobile devices are invaluable resources that can greatly help evaluators identify usability problems. Various techniques can be used for capturing the screen of a mobile device: One is the recording of the screen by using a mini wireless camera (Figure 1B). It can be very helpful in cases of individual users but it is not suitable in the case of an application that involves beaming actions (e.g., Bluetooth, infrared) and/or interaction with the physical space because it can influence negatively the use of the device and can create obstacles in the infrared beams, sensors, or readers attached to the device (i.e., to an RFID reader). The main advantage of this technique is that the camera records, besides the screen, the movements of the users fingers or stylus, capturing valuable data identifying potential interaction problems (for example, the user hesitates to click something because the interface or the dialogs are confusing).

An alternative technique is the shadowing technique which can effectively work for individual users (Figure 1A). Again, this technique is not suitable for group activities, where the subjects often form groups and move continuously. Even in cases that it is considered possible to record properly, there could be many events missing because of the frequent movements of the subjects or the shielding of the screen by their body and hands.

The direct observation technique has also certain limitations (Cabrera et al., 2005; Stoica et al., 2005) because the observer must distribute his attention to many subjects. In case there are observers available for each user they will restrict the mobility of the users and they will distract their attention when being in so close range. Consequently, all these techniques impose the presence of the observer to the users, thus affecting their behavior.

Another significant issue that directly affects the usability evaluation is related to the location in which the study is conducted. There are many arguments in favour of *field usability studies* (Nardi, 1996; Kjeldskov, Skov, Als, & Hoegh, 2004; Zhang & Adipat, 2005; Kaikkonen, Kallio, Kekalaien, Kankainen, & Cankar, 2005). Comparative studies between laboratory and field evaluation studies have drawn, however, contradictory conclusions. In a recent survey of

evaluation studies of mobile technology (Kjeldskov & Graham, 2003), 71% of the studies were performed in the laboratory, which revealed a tendency towards building systems based on trial and error and evaluating systems in controlled environments at the expense of studying real use of them. So the question of what is useful and what is perceived problematic from a user perspective often is not adequately addressed.

In summary, in order to conduct a usability evaluation of a mobile application/system, there is a need to take into consideration the attributes that are going to be measured, the data collected for these measurements, the location in which the evaluation will take place, and finally, the appropriate tools to analyze them, having always in mind the user and the context of interaction.

Data Analysis

Usability evaluation of mobile applications is more complex than desktop software evaluation since new characteristics such as group activity and the interaction with the surrounding environment need to be taken into consideration. In order to acquire an understanding of group activity and performance, huge amounts of structured and unstructured data of the forms discussed in the previous section need to be collected. These data should capture the activity of subjects, including their movements, facial expressions, gestures, dialogues, interaction with the devices, and objects in the environment. Analysis of these data require special attention on details as well as the context of use, thus it can be a tedious process which can be facilitated by a suitable analysis tool (Benford et al., 2005).

Various tools have been developed to support usability evaluation studies and, in general, to record and annotate human activity. These tools often handle video and audio recordings and synchronize them with text files, containing hand-taken notes. This combination creates a dataset that is rich in information which is then annotated through an adequate annotation scheme, which creates quantitative and qualitative measures of the observed user-device interaction. Typical examples of such tools are: the *Observer XT* (Noldus, 2006), *HyperResearch* (Hesse-Biber, Dupuis, & Kinder, 1991; ResearchWare, 2006), *Transana* (Transana, 2006), *NVivo* (QSR, 2006; Rich & Patashnick, 2002; Welsh, 2002), and *Replayer* (Tennent & Chalmers, 2005). From them, only *Replayer* and *Observer XT* have special provisions for mobile settings. The extra characteristics in evaluation of mobile applications (group activity and interaction with the surrounding space) demand the extended use of multimedia files that thoroughly capture the activity. Thus, there is a need for a tool that combines and interrelates all of the observational data in a compact dataset and gives to the usability expert the ability to easily navigate them from multiple points of view (access in user—device interaction, access in user—space interaction).

All of the tools utilize video sources at a different extend, with the exception of *NVivo* that focuses more in textual sources. *NVivo* allows linking of evaluator's notes with video extracts, without permitting more fine grained handling of video content. On the other hand, *HyperResearch* and *Transana* do support flexible handling of video sources but they do not allow the integration and synchronous presentation of multiple video sources in the same study. Thus, *NVivo*, *HyperResearch*, and *Transana* cannot successfully respond to the extra characteristics of mobile applications. On the other hand, *Replayer* is a distributed, cross platform toolkit that allows the integration of multiple video sources and presents analysis data in various forms such as histograms and time series graphs. Although *Replayer* efficiently supports usability analysis of mobile applications, its failure to handle and to compare data that come from various studies makes it not suitable for cases of multiple studies

Table 2. Characteristics of usability evaluation tools

	Multiple multimedia sources	Aggregated results from multiple studies
Observer XT	☑	☑
HyperResearch		☑
Transana		
NVivo		☑
Replayer	☑	
ActivityLens	☑	☑

in which there is need to aggregate and generalise the findings. On the contrary, *Observer XT* is a powerful commercial tool, widely used in observation studies, that enables the synchronous presentation of multiple video files and also the derivation of overall results about the activity of multiple subjects. Although *Observer XT* meets the requirements of new characteristics of mobile applications, its use requires a prior lengthy training period.

A tool that has been especially adapted for analysis of data from mobile applications' evaluation studies is the *ActivityLens* which attempts to tackle some of the limitations of existing tools. Its main advantage is its ability to integrate multiple heterogeneous qualitative but also quantitative data. It allows the usability expert to directly access the collected data, thus to simultaneously focus on users' movements on the surrounding environment and user-device interaction. To sum up, *ActivityLens* supports analysis of collected data and produces results that cover the overall activity concerning all the participants.

Weitzman and Miles (as cited in Berkowitz, 1997) suggest that a criterion for the selection of an adequate analysis tool is related to the amount, types, and sources of data to be analyzed and the types of analyses that will be performed. In Table 2 a description is provided about how the tools support the extra characteristics of usability evaluation studies of mobile applications.

Data Analysis through ActivityLens

ActivityLens is a tool that embodies features especially designed for usability evaluation of mobile applications. *ActivityLens* is an evolution of the earlier Collaboration Analysis Tool (ColAT) (Avouris, Komis, Margaritis, & Fiotakis, 2004; Avouris, Komis, Fiotakis, Margaritis, & Voyiatzaki, 2005), originally designed for video analysis of collaborative learning activities. It was found particularly suitable for the proposed approach which involves multiple perspectives of the activity, based on different multimedia data.

In *ActivityLens*, all the collected data are organized into *Studies*. An example of a *Study* is the usability evaluation that was conducted in a Historical Cultural museum, described in the next sections. The tool allows *Projects* that belong to a specific *Study* to be defined. A *Project* is defined by the evaluator and can have different perspectives depending on the situation. For example, a *Project* can be defined as the set of data gathered from various groups over a set period of time, or it can be defined as a set of data of a specific group of users.

These data can be video and audio files, log files, images, and text files, including hand-taken notes of the observers. *ActivityLens* supports almost all the common video and audio file formats including file types that are produced by mobile devices such as .mp4 and .3gp. The

Figure 2. The usability evaluation tool—ActivityLens

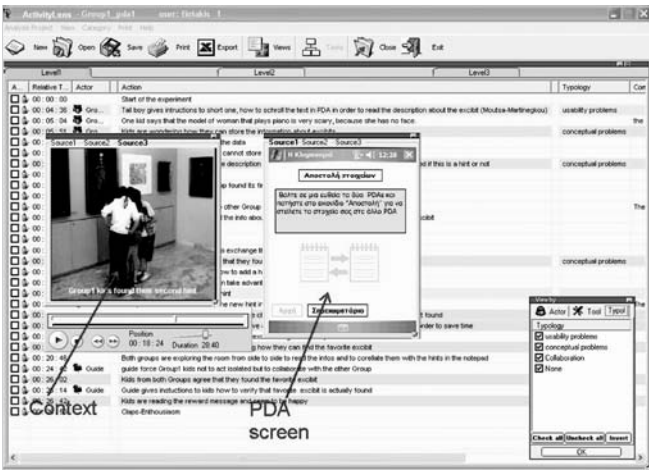
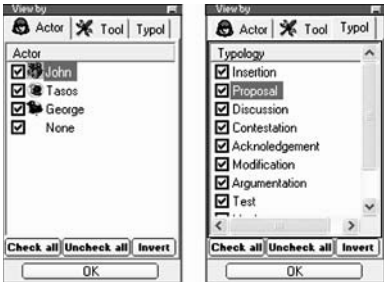


Figure 3. Event filtering tool through ActivityLens



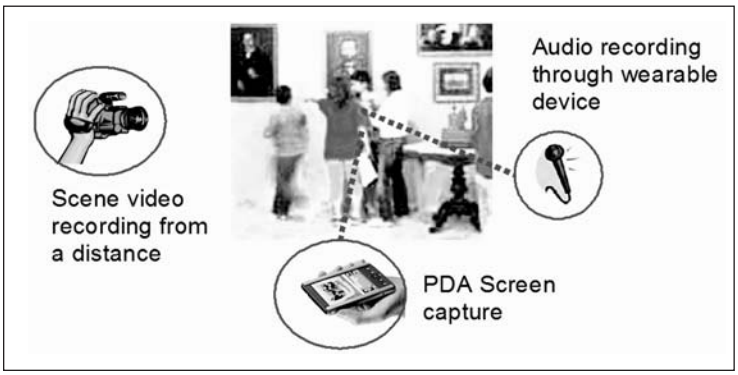
observed activity is reported in an XML log file. This file describes the activity as a set of events, reported in sequential order, following this typical structure:

<event id>, <time-stamp>, <actor>, <tool>, <event-description>, <type of event>, <comments of evaluator>

The log file events are presented via a simple spreadsheet view in order to be easily accessible for inspection and annotation. In addition, *ActivityLens* permits integration and synchronization of the collected multimedia files.

All the data can be reproduced and annotated on-the-fly in order to highlight interesting events. An example is shown in Figure 2, in which an overview video and a PDA screen are synchronized and annotated. The annotation of the observed events is based on a classification scheme defined by the evaluator. For example, an evaluator is analyzing videos that describe the activity of a group of students that try to solve a problem. During the activity some students propose ways to solve the problem and argue about it. Thus, one representative type of event could be defined as "Proposal." For usability studies, an evaluator can define typologies based on usability attributes, concerning for instance, user errors, comments expressing subjective view, and events marking successful completion of tasks.

Figure 4. Sources of observational data



ActivityLens provides the evaluator with the ability to reduce the huge amount of collected data through an event filtering mechanism. This feature is of high importance because it helps the evaluator to focus on interesting sequences of events and makes them emerge from the "noise." The evaluator is allowed to define criteria for specific Actors, tools used, and types of events or any combination between them. For example, the evaluator can choose to view all occurrences of "Proposals" made by "George" or "John." The criteria selection tool is shown in Figure 3.

PROPOSED METHODOLOGY

Based on the outlined data and analysis requirements, in this section a methodology suitable for usability evaluation of mobile applications is proposed. This method is proposed for applications deployed in places like museums, libraries, and so forth, in which groups of users interact among themselves and with the environment, in various ways. These semi-public spaces represent 'living organisms' that project, in a visible and tangible form, the various facets of information. For example, in a museum such applications assist the visitors in discovering and acquiring knowledge. A museum can be characterized as an ecology (Gay & Hebrooke, 2004) that is constituted by two main entities, the exhibits and the visitors, populating the same space. Items of the collection are exhibited to visitors, who react by discovering them in a way that is, at a large extent, influenced by the surrounding space. Also, visitors usually interact with each other, for example, because they comment the exhibits independently from the use of technology. This methodology involves, initially, the preparing study phase, the recording activity phase, and then the analysis of the activity.

Preparing the Study

Usually, activities that are expected to take place in semi-public spaces are desirable to be conducted in the field. For example, visitors inside a museum enjoy an experience that cannot be fully reproduced inside a laboratory. Therefore, the evaluator needs to conduct a study in a representative place, which should be adapted accordingly without disturbing its normal operation. Issues to be tackled are related with technological restrictions (e.g., wireless network infrastructure), recruitment of an adequate number of typical users, the extent of the study, and so forth. Consequently, it is evident that the preparation phase of the evaluation is a very important one, as it builds the foundation for a subsequently successful study.

Recording Activity Phase

A prerequisite in such environments is the low level of activity interference by the observers in order to minimize the behavioral change caused to the participants by the uncomfortable feeling of being observed and thus "disorienting the balance" in the ecology. The proposed recording activity includes an innovative combination of existing data gathering techniques in order to achieve the considered goal. The sources of data (Figure 4) include: (a) screen recordings of the mobile devices, (b) audio recordings using wearable recorders, (c) video recordings from the distance, where the camera is operated by an operator or preferably by remote control, and as complementary source (d) interviews and questionnaires to the users. A brief discussion of the process of collecting these data is included next.

(a) Screen Grabbing on the Mobile Device

In order to tackle problems related to the application nature (collaboration, interaction with the environment) it is proposed that the mobile device also be used as a screen recording device. The collected information can be in the form of screen-shots or aggregated in a low frame-rate video. The main requirement for a mobile device to become a screen recording device is that it must run a multitasking operating system in order to allow a background process to run in parallel with the main application. At the current technological status, this is the case for most mobile devices (PDAs and smart phones), as the main operating systems are multitasking: Symbian OS, Windows Mobile, Palm OS (version 6.0 onwards), Java OS, and so forth. Also, the needs of the market drove the mobile devices to handle large amounts of data that have to be consulted, edited, and updated by the user while speaking, browsing, watching TV, and so forth. As a result, mobile devices evolved from a single process, sequential to multitasking, and obtained increased storage capacities which permit the users to store a lot of information on them. Therefore, a mobile device can capture, by a parallel process, the screen and either save the pictures on their memory or send them directly to a server via a wireless connection.

A prototype application that is suitable for the Pocket PC/Windows Mobile environment has been developed and runs in parallel with the application which has to be evaluated. It captures screen snapshots and stores them on the device at a predefined time interval. In the tests, a compressed quarter VGA (240x320) screen shot was at most 32 KB that at a rate of 4 per second lead to a needed storage of about 450 MB/hour. It must be stressed that far better compression rates can be achieved by using video encoders.

The decision to grab the screen with a steady frequency and not per number of events, that would make sense in order to stop recording when the device is not used, was imposed by the technical current limitation: the scarce support for global system hooks on the Windows Mobile operating system. The lack of support is due to the fact that such hooks can critically affect the performance of the device.

(b) Audio Recording with Wearable Devices

Audio can capture dialogs between users that express difficulty in interacting with the application and the environment, or disagreement. Audio recordings can often reveal problems that users do not report during interviews or questionnaires.

The audio recordings from the inbuilt microphone of the video camera are sometimes not very useful due to the noise and to the fact that usually the dialogues are in a low voice. Also, the distance between the subject and the camera does not allow recording of good quality sound. The ideal solution would be that the mobile device itself could record both the screen and the

audio. Unfortunately, this is not feasible because of several reasons:

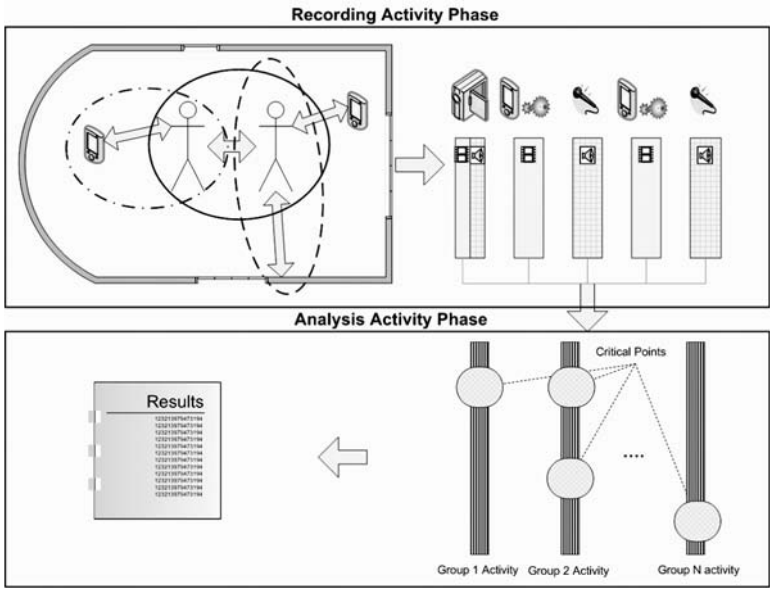
- The performance of the device degrades significantly by having two background processes running simultaneously, the one related to screen grabbing, discussed in (a) and the one to audio recording.
- The sounds that are produced by the device itself, in most cases, cover any other sound in the surrounding environment (i.e., a narration played back covers the dialogue.
- The storage might be a problem. Depending on the audio quality and compression used, 1 hour of recorded sound can take from 50 MB to 700 MB.

For these reasons, it seems that the most suitable solution is to use a wearable audio recorder that can store several hours of sound. These devices are very light; they weight less than 50 grams, including the battery. The user can wear it with the help of a neck strap or put it in a pocket and adjust a clip microphone. The wearable audio recorders guarantee that rich information concerning the dialogs between the subjects will not be lost, collaborating and interacting with the application and the environment.

(c) Discrete/Unobtrusive Video Recording

To complement the dialogs and the screen recordings, it is necessary to capture, in video, the ensemble. From this video, recording the context of the events, the social interactions between the group members (peers) and/or between groups can be depicted. In order to decrease as much as possible the level of obtrusion, the camera must be preferably maneuvered through remote control (allowing zoom and angle changes) or at least by a cameraman that will keep a large enough distance from the activity in order not to disturb the users. Often, many video recordings may need to be

Figure 5. Recording and analysis phases of proposed methodology: Interesting incidents are observed in the media files and are cross-checked for better understanding. These incidents are analyzed in terms of device and activity usability issues.



made from various angles, distance, or focusing in different aspects. These may be mixed in a single video stream if adequate equipment is used, or, more often, may be kept as separate sources of information. By studying these video recordings the evaluator can obtain a clear idea about the place in which the activity took place.

(d) Interviews and Questionnaires

Considering that the sources constitute the objective information, the users' subjective view through interviews and questionnaires also need to be obtained. Through these sources, which vary depending on the situation, someone can formulate results regarding *user's satisfaction, learning performance,* and so forth; attributes sometimes difficult to obtain simply through observation.

Analysing Activity Phase

The purpose of the analysis is to identify instances of use of the devices and the infrastructure, which identify usability problems of the technology used. Analysis of recorded activity of groups in semi-public spaces is not a simple process. Researchers have not only to focus just on the devices but to take into account more complicated issues concerning the interaction between groups, the interaction between peers in a certain group, and the interaction with the surrounding space. This analysis has to be meticulously performed in order to cover the above issues. During analysis all the collected sources that describe the group activity have to be combined and iteratively inspected. Initially, a quick inspection of recorded activity helps usability experts to isolate the segments that need thorough analysis. Then, detailed inspection of these segments is required to interpret the observed interaction and depict the usability problems. This process can help usability experts to detect certain critical points of interaction that can be further examined in order to measure their frequency and dispersion between groups and to be clear how they affect the use of mobile applications. The proposed methodology concerning the recording and analysis process can be seen in Figure 5.

EVALUATING USABILITY OF A COLLABORATIVE CONTEXT AWARE EDUCATIONAL GAME

An example of a study in which the proposed technique was applied was a usability evaluation of a collaborative mobile learning application supported by PDAs in a cultural-historical museum (Tselios et al., 2006). The study involved 17 students of the 5th-grade of an elementary school (11 years old) who were invited to visit the museum and use the prototype of an educational application that was temporarily installed there. All the students were familiar with the use of mobile phones but they had no former experience with PDAs. Furthermore, most of them described themselves in a pre-study questionnaire, as users of desktop computer systems on a daily basis.

The study took place in two of the museums' halls in which portraits and personal objects of important people of the local community were exhibited. First, a short introduction to the activity was provided by a member of the research team who undertook the role of the guide. The educational activity was designed in a way that students were motivated to read information about these important people and collaboratively search in order to locate a specific exhibit according to the activity scenario. The children were divided in two groups and each group consisted of two teams of 4 or 5 children each. Each group participated in a different session for approximately 1 hour.

In order to achieve the scenario's goal, each team was provided with a PDA equipped with a RFID tag reader. They used this equipment to locate hints that were hidden inside textual descriptions of the exhibits. These were obtained

by scanning the exhibit RFID tags. The students could store the hints it in a notepad of the PDA. After collecting all or most of the hints the teams were encouraged to share their hints, through beaming, to each other.

Then the students, using the found information, had to locate a specific-favorite exhibit which matched the description provided by the hints. When two teams agreed that they had found the favorite exhibit, they checked the correctness of their choice by scanning with both PDA's the RFID tag. A correct choice was indicated by the system with a verification message while a wrong one suggested a new search. When the study was over each student was requested to answer a set of questions related to the group activity in the museum.

Preparation of the Evaluation Study

During the preparation of the study the museum was contacted and the permission to run the evaluation study was obtained. The space of the museum was examined well in advance (e.g., for determining wireless network setup options) and afterwards a small scale pilot was run in a simulated environment in order to check the suitability of the technological infrastructure. In order to ensure the participation of subjects, a school in the vicinity of the museum was contacted and participation of a school party was requested for the study.

Collecting Data

In order not to miss important contextual information, three video cameras were used in this study. Two of them were steadily placed in positions overlooking the halls while the third one was handled by an operator who tenderly followed the students from a convenient distance. One student per team wore a small audio recorder in order to capture the dialogues between them, while interacting with the application and the environment. Furthermore, snapshots of the PDA screens were captured during the collaborative activity and stored in the PDA's memory. After the completion of the study, the guide, who was a member of the evaluator team, had an interview with the students, asking them to provide their opinion and experiences from the activity in the museum; while back at school a week later, their teacher asked them to write an essay describing their experience.

Figure 6. A) Instance of user—RFID tag interaction problem. B) and C) Photos from the collaboration activity inside the museum

Analysing Data with ActivityLens

In order to analyse all the collected data according to the proposed methodology, ActivityLens, that has already been effectively used in similar studies was used (Cabrera et al., 2005; Stoica et al., 2005).

The main reasons that ActivityLens was used among the discussed tools was its capacity of organizing observations into Studies (collection of projects) and its ability to present multiple perspectives of the whole activity (by integrating multiple media sources). Although Observer XT provides even more capabilities than ActivityLens, the choice of ActivityLens seemed to fit better the specific use case since its use did not require a long training time. In addition, ActivityLens permits easy access to the activities of the subjects recorded in different data sources.

Three usability experts, with different levels of experience, analysed the collected data in order to increase the reliability of the findings. Initially, a new *ActivityLens Study* including four projects (each project concerns the observations of a team) was created. The integrated multimedia files were extensively studied and the most interesting situations were annotated. It must be clarified that it was not wanted for the behaviour of each individual team member to be studied but wished that the performance of the whole team be evaluated. The performed analysis through ActivityLens revealed several problems related to the children's interaction with the device and the overall setting, given the surrounding physical space and groups.

Several problems were identified when the students interacted with the handheld devices.

The analysis indicated that almost all the groups could not successfully scan the Exhibits tags in their initial attempts and get information about the exhibits. The RFID tags were located underneath each exhibits label. Since the users had no clear indication of where to place the tag scanner, some of them experienced difficulties interacting with them. Also, there was an unexpected delay in the scanning process between tag and PDA (the PDA needed about 2 seconds to scan the tag). While from the scene, video recording, it seemed that the user was repeatedly scanning the same label, combining this with the PDA screen recording gave the real reason of this behavior—repeated unsuccessful tries to scan the tag. The users learned after a few frustrating attempts that they should target the center of the tags and hold the device for a couple of seconds.

A problem that troubled a specific group was the use of a scrollbar in the textual description of the exhibits. The users were not familiar with the procedure of scrolling on a PDA and they repeatedly discussed it amongst themselves. This problem was identified through the combined use of the audio and screen recording and was not visible from the scene video.

An unexpected problem was related to the content of some exhibits descriptions. They contained the word "hints" which confused the children and they were not sure if this was or was not a hint that they could add to the notepad. This was spotted from the complementary use of the overview video with the dialogue audio recordings. The problem was overcome by asking the help of the guide.

With the use of ActivityLens many problems that were related to the interaction with the physical space were managed to be detected. The most important one was that some of the exhibits tags were placed on the walls in such positions that they were not accessible by short students. In Figure 6 an instance of this problem is shown.

Another interesting element that was made clear through the students' dialogues and the videos was that in a certain area of the room an exhibit inspired fear to some of the children (e.g., a faceless piano player). Particularly, one student was clearly afraid to get near the puppet and said to the other members: "I am not going near her. She is very scary!!! Look at her, she has no face!" This situation made the team avoid that area,

which contained exhibits with useful information for the activity.

The children that participated in the study often expressed their concern about being delayed in their play due to the presence of other museum visitors (at a certain point an independent school party crowded the hall). Through the audio it was obvious that the kids expressed their frustration because they were delayed in playing the game and the visitors, because they were disturbed by the kids. These problems escape from the traditional usability analysis that focuses only on the device, because they contain the interaction between the user and the surrounding physical space.

The third dimension of the evaluation concerned investigation of the collaborative nature of the activity and the learning performance. An interesting observation was that by having two teams searching for hints at the same time, and the fact that one of the teams was more successful than the other, constituted a powerful motivation for the second team to search for hints. This was observed from the complementary scene video (pinpointing the event) and the dialog recordings (exclamations, etc.). Also, that some kids were too excited in using the PDAs and did not allow anyone else to use them was observed. Thus, disputes over use of the device influenced negatively the team spirit. From the audio streams, the disappointment of the kids that were not allowed to use the device were managed to be spotted.

Regarding the learning performance through the audio files and the PDA's screen it was found that one team was not reading the descriptions to locate the hints but they were searching for the parentheses that indicated the existence of a hint. It must be said that the solution with the parentheses and not colored text was adopted because it was wanted that those specific situations be avoided, but this did not actually work in all teams. In the future version, the hints will be visible only when the users click on them inside the description of the exhibits.

The results are also based on a study of questionnaires, independently of the ActivityLens analysis. In this point, the limitation of ActivityLens in analyzing user questionnaires has to be underlined. This weakness is a matter of further development and research.

In order to have a general view about the educational value of the activity when the children returned to their classrooms, they wrote an essay in which they reported on the museum experience. The teacher's view after going through these texts was that almost all the kids that participated in the activity learned something meaningful in a funny and enjoyable way. However, a more systematic study on these issues should involve a more quantitative experimental approach through a pre and post-test questionnaire and a control group.

CONCLUSION

This chapter has presented a brief overview of usability evaluation techniques for mobile applications, including collection of multiple observational data and their analysis. Due to the growing use of mobile devices, it is evident that there is a need for established techniques that support the collection and analysis of data while conducting usability evaluations. Since there are considerable differences between desktop and mobile environments, researchers are obliged to develop and fine tune these new techniques. Through this chapter a methodology for evaluating mobile applications focusing on collection and use of observational data was proposed. The proposed methodology was demonstrated through a usability study of an educational game in a Historical Museum.

The proposed recording activity technique can be characterized as unobtrusive regarding the users and allows evaluators to study the activity in conditions as close as possible to the typical conditions of use of the application, through various perspectives. The ActivityLens tool was used for analysis of the collected data which facilitates

interrelation and synchronization of various data sources and was found particularly useful since the collected data were of particularly high volume and often a finding was based on a combination of data sources. The methodology revealed usability problems of the application as well as issues about collaboration and interaction with the environment that would not be easy to discover in the laboratory and without the combined use of the multiple media data.

Studies that take place in semi-public spaces and involve groups of people have to tackle various problems. In most cases the willingness of people but also the availability of spaces is difficult to be guarantied for the long periods of time. Researchers that conduct such studies have to be as unobtrusive as possible to the users and pay special attention in order to minimize interference with the environment.

A limitation of the proposed approach is that it requires the users to carry light equipment (audio recorders) and also that a screen capturing software had to be installed in the mobile devices. However, these limitations did not inhibit the users to act naturally and recreate a realistic but controlled context of use. The typical studies of the proposed approach lasted a short time and thus, it is difficult to measure long term usability aspects like memorability and long term learning attitudes. It is still under investigation how to extend this technique to long term mobile usability studies involving different contexts of use.

What is however, missing from the story is an analysis scheme that can describe user interaction with the surrounding physical and information space and metrics that map usability attributes. Such a scheme would describe usability as a set of attributes that refer to interaction with the device, interaction with the space, and group interactions. This scheme could be supported by a tool like ActivityLens which facilitates easy navigation of the collected media data, allowing creation of pointers to incidents in the data, justifying the calculated values of the usability attributes. Definition of such a scheme should however, be the result of a wider research community process.

REFERENCES

Aittola, M., Parhi, P., Vieruaho, M., & Ojala, T. (2004). Comparison of mobile and fixed use of SmartLibrary. In S. Brewster & M. Dunlop (Eds.), *Proceedings of 6th International Symposium on Mobile Human-Computer Interaction (Mobile HCI 2004)* (pp 383-387). Berlin: Springer.

Aittola, M., Ryhänen, T., & Ojala, T. (2003). SmartLibrary—Location-aware mobile library service. In L. Chittaro (Ed.), *5th International Symposium on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI 2003)* (pp. 411-416). Berlin: Springer.

Avouris, N., Komis, V., Fiotakis, G., Margaritis, M., & Voyiatzaki, E. (2005). Logging of fingertip actions is not enough for analysis of learning activities. In *Proceedings of AIEDs Workshop on Usage Analysis in learning systems.* Retrieved February 27, 2007, from http://lium-dpuls.iut-laval.univ-lemans.fr/aied-ws/.

Avouris, N., Komis, V., Margaritis, M., & Fiotakis, G. (2004). An environment for studying collaborative learning activities. *Journal of International Forum of Educational Technology & Society, 7*(2), 34-41.

Benford, S., Rowland, D., Flintham, M., Drozd, A., Hull, R., Reid, J., Morrison, J., & Facer, K. (2005). Life on the edge: supporting collaboration in location-based experiences. *Proceedings of the SIGCHI conference on Human Factors in computing systems CHI 2005* (pp. 721-730). New York: ACM Press.

Berkowitz, S. (1997). Analyzing qualitative data. In J. Frechtling & L. Sharp Westat (Eds.), *User-friendly handbook for mixed method evaluations.*

Retrieved February 27, 2007, from http://www.ehr.nsf.gov/EHR/REC/pubs/NSF97-153/start.htm

Betiol, H. A., & de Abreu Cybis, W. (2005). Usability testing of mobile devices: A comparison of three approaches. In M. F. Costabile & F. Paterno (Eds.), *Proceedings of IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2005)* (pp. 470-481). Berlin: Springer.

Cabrera, J. S., Frutos, H. M., Stoica, A. G., Avouris, N., Dimitriadis, Y., Fiotakis, G., & Demeti, K. (2005). Mystery in the museum: Collaborative learning activities using handheld devices. In M. Tscheligi, R. Bernhaupt, & K. Mihalic (Eds.), *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services (Mobile HCI 2005)* (pp. 315-318). New York: ACM Press.

Dey, A. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal, 5*(1), 4-7.

Dix, A., Finley, J., Abowd, G., & Beale, R. (2003). *Human-computer interaction* (3rd ed.). Hertfordshire: Prentice Hall.

Fouskas, K., Pateli, A., Spinellis, D., & Virola, H. (2002). *Applying contextual inquiry for capturing end-users behaviour requirements for mobile exhibition services.* Paper presented at the 1st International Conference on Mobile Business. Athens, Greece.

Gay, G., & Hebrooke, H. (2004). *Activity-centered design. An ecological approach to designing smart tools and usable systems.* Cambridge, Massachusetts: MIT Press.

Hagen, P., Robertson, T. & Kan, M. (2005). *Methods for understanding use of mobile technologies.* Technical Report. Retrieved September 20, 2006, from http://research.it.uts.edu.au

Hagen, P., Robertson, T., Kan, M., & Sadler, K. (2005). Emerging research methods for understanding mobile technology use. In *Proceedings of the 19th Conference of the Computer-Human Interaction Special Interest Group (CHISIG) of Australia on Computer-human interaction: Citizens online: Considerations for today and the future OzCHI 2005* (pp. 1-10). New York: ACM Press.

Hesse-Biber, S., Dupuis, P., & Kinder, T. S. (1991). HyperRESEARCH, a computer program for the analysis of qualitative data with an emphasis on hypothesis testing and multimedia analysis. *Qualitative Sociology, 14,* 289-306.

Jambon, F. (2006). Reality testing of mobile devices: How to ensure analysis validity? In *Proceedings of CHI 2006 Workshop on Reality Testing: HCI Challenges in Non-Traditional Environments.* Retrieved February 27, 2007, from http://www.cs.indiana.edu/surg/CHI2006/WorkshopSchedule.html

Kaikkonen, A., Kallio, T., Kekalaien, A., Kankainen, A. & Cankar M. (2005). Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability Studies, 1*(1), 4-16.

Kjeldskov, J., & Graham, C. (2003). A review of Mobile HCI research methods. In L. Chittaro (Ed.), *5th International Symposium on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI 2003)* (pp. 317-335). Berlin: Springer.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human-Computer Studies, 60,* 599-620.

Kjeldskov, J., Skov, M. B., Als, B. S., & Hoegh, R. T. (2004). Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In S. Brewster & M. Dunlop (Eds), *6th International Symposium on Mobile Human-Computer Interaction (Mobile HCI 2004)* (pp 61-73). Berlin: Springer.

Nardi, B. (1996). Studying context: a comparison of activity theory, situated action models, and distributed cognition. In B. Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction* (pp. 69-102). Cambridge, Massachusetts: MIT Press.

Nielsen, J. (1993). *Usability engineering.* London: Academic Press.

Raptis, D., Tselios, N., & Avouris, N. (2005). Context-based design of mobile applications for museums: a survey of existing practices. In M. Tscheligi, R. Bernhaupt, & K. Mihalic (Eds.), *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services (Mobile HCI 2005)* (pp. 153-160). New York: ACM Press.

Rich, M., & Patashnick, J. (2002). Narrative research with audiovisual data: Video intervention/prevention assessment (VIA) and NVivo. *Int. Journal of Social Research Methodology 5*(3), 245-261.

Schneiderman, B., & Plaisant, K. (2004). *Designing the user interface: Strategies for effective human-computer interaction* (4th ed.). Boston: Addison Wesley.

Stoica, A., Fiotakis, G., Cabrera, J. S., Frutos, H. M., Avouris, N. & Dimitriadis, Y. (2005, November). *Usability evaluation of handheld devices: A case study for a museum application.* Paper presented at the 10th Panhellenic Conference on Informatics (PCI2005), Volos, Greece.

Tennent, P., & Chalmers, M. (2005). Recording and understanding mobile people and mobile technology. In *Proceedings of the1st International Conference on E-social science.* Retrieved February 27, 2007, from http://www.ncess.ac.uk/conference_05.htm/papers/

Tselios, N., Papadimitriou, I., Raptis, D., Yiannoutsou, N., Komis, V., & Avouris, N. (2006). *Design for mobile learning in museums.* To appear in J. Lumsden (Ed.), *Handbook of User interface design and evaluation for mobile technology.* Hershey, PA: IGI Global.

Welsh, E. (2002). Dealing with data: Using NVivo in the qualitative data analysis process. *Forum Qualitative Social Research Journal, 3*(2). Retrieved February 27, 2007, from http://www.qualitative-research.net/fqs-texte/2-02/2-02welsh-e.htm

Zhang, D., & Adipat, B. (2005). Challenges, methodologies, and issues in the usability testing of mobile applications. *International Journal of Human-Computer Interaction, 18*(3), 293-308.

KEY TERMS

ActivityLens: A usability analysis tool used to support usability studies for mobile and collaborative applications analyzing multiple media data.

Context: Any information that can be used to characterize the situation of an entity. An entity should be treated as anything relevant to the interaction between a user and an application, such as a person, a place, or an object, including the user and the application themselves. (Dey, 2001).

Context Aware: A device, a system, or an application that has the ability to sense aspects of context and change its behaviour accordingly.

Data analysis tool: A software package that supports extracting meaningful information and conclusions from collected data.

Data Collection: The process of gathering raw or primary specific data from a single source or from multiple sources.

Screen Recording: The operation of capturing the output of a devices' screen.

Semi-Public Space: A place which is public to people and imposes a set of common, and uni-

versally acceptable rules regarding their behaviour i.e. a museum, library, theatre.

Usability Evaluation: The process of assessing the usability of a given system or product.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 997-1013, copyright 2008 by Information Science Publishing (an imprint of IGI Global).

Chapter 8.5
Mobile Design for Older Adults

Katie A. Siek
University of Colorado at Boulder, USA

ABSTRACT

The global population of older people is steadily growing and challenging researchers in the human computer interaction community to design technologies to help them remain independent and preserve their quality of life. Researchers are addressing this challenge by creating assistive technology solutions using information appliances, such as personal digital assistants and mobile phones. Some have questioned whether older people can use information appliances because of age related problems. This chapter discusses work related to designing, implementing, and evaluating mobile applications for the aging. A discussion about what researchers should consider during the design process for information appliances shows the unique challenges posed by this population.

INTRODUCTION

Our world population is aging. The United States National Institute of Health estimates that the global older adult[1] population grows by 795,000 each month. They project that by 2030, the global older population will grow by 847,000 per month (Kinsella & Velkoff, 2001). In response to this increase, researchers in human computer interaction, social sciences, and ubiquitous computing communities are developing applications to help older people live independent and productive lives. Researchers use *information appliances* (Norman, 1999), such as personal digital assistants (PDAs) (Carmien, DePaula, Gorman, & Kintsch, 2004; Coroama & Rothenbacher, 2003) and mobile phones (Helal, Giraldo, Kaddoura, & Lee, 2003), to create *assistive technologies* for older people.

We contend that older adults can use information appliances if the physical and virtual interfaces are designed to meet their varying needs. Some may argue that older adults do not use information appliances and thus, researchers do not have to adjust designs for this population. However, a recent report in the United Kingdom revealed that 49% of older adults own a mobile phone and of that group, 82% make one or more calls per week (Office of Communications [OfCom], 2006). Thus, older adults are using information appliances, but they do encounter numerous problems, such as font and icon readability and interface complexity issues, discussed in greater detail in the background section.

Other people argue that since younger adults use information appliances now, they will not have a problem using similar technology in the future. Indeed, 82% of all United Kingdom residents own a mobile phone, whereas only 36% of people over 75 years old own a mobile phone (Office of Communications [OfCom], 2006). However, we know that (1) as people age their physical and cognitive abilities do not remain constant and (2) the *digital divide* is still present; factors such as age, socioeconomic status, and disabilities affect individuals' access to technology. Although *walk-up-and-use* systems are becoming more prevalent in our everyday lives, we cannot assume that by giving older people new technology, they will be able to easily interact with the device and application. We must work together now to create a set of guidelines to help inform the design and development of future technologies for older people to avoid problems associated with *technology determinism* (Warschauer, 2003).

In this chapter, we discuss issues that must be addressed when designing information appliance interfaces for older adults. We begin by highlighting design related work with older people and technology - traditional computers and information appliances. We then discuss best practices for conducting user studies with older populations and design issues to consider when developing applications and devices. We conclude the chapter with ideas for future work and challenges to the design, interaction, and technical communities.

BACKGROUND

We discuss how older people interact with traditional computers and information appliances in this section. The related work delves into design and interaction studies because interactions, physical and cognitive, have a major influence on design. Researchers have looked at how older populations interact with traditional desktop computers. Researchers are just beginning to look at how older populations interact with information appliances.

There has been a proliferation of information appliances designed for the general public, including PDAs, mobile phones, remote controls, digital cameras, digital music players, and game playing devices. The interfaces to these vary considerably, suggesting there may be variable age-related performance effects. Hence, when creating applications for older populations, designers must consider age-related abilities such as vision, dexterity, coordination, and cognition. Researchers have discovered that within older populations, there are noticeable differences in abilities, and that different design methodologies, such as universal design (Abascal & Civit, 2001) and user sensitive inclusive design (Newell & Gregor, 2001) should be used. Here we discuss some of the research that has been done to better understand older populations' interaction with technology.

Older People and Traditional Computers

Bernard, Liao, and Mills (2001) found that older people could read faster with a larger, more legible 14-point san serif font on websites. Researchers at

Georgia Tech studied how multimodal feedback (sound, touch, visual effect) could assist participants with varying vision problems perform basic mouse tasks (drag and drop). They found that all groups performed better when sound was added; however, groups performed the best when all three modal feedbacks were used (Jacko, Scott, Sainfort, Barnard, Edwards, Emery, et al., 2003).

A number of recent studies focused on the ability of older populations to use PC input devices (Chaparro, Bohan, Fernandez, & Choi, 1999; Charness, Bosman, & Elliott, 1995; Laursen, Jensen, & Ratkevicius, 2001; Smith, Sharit, & Czaja, 1999). The studies showed that older people completed tasks slower than younger groups. Charness et al. (1995) evaluated control key, mouse, and light-pen input devices and found older people preferred the light pen, followed by the mouse and control keys.

Smith et al. (1999) and Laursen et al. (2001) found older people made more mistakes than younger people and had difficulty with fine motor control tasks such as double clicking. Chaparro et al. (1999) found older people performed "point and click" and "click and drag" tasks slower than younger people, but with the same amount of accuracy. The researchers believed the reason that older people were slower was because of reduced fine motor control, muscle strength, and pincher strength associated with older age.

Older People and Information Appliances

Most of the human computer interaction studies on older adults and technology focus on the usability of traditional desktop computers. The usability of information appliances will be scrutinized more carefully as pervasive computing technology applications become more widespread. Researchers are already assessing the needs of older people with respect to mobile phones.

Maguire and Osman (2003) found that older people primarily considered mobile phones as a way to assist in emergencies, whereas younger people saw mobile phones as a way to interact socially. Older people were interested in small phones with large buttons and location aware systems. More specifically, older women were interested in finding the nearest retail shop that met their needs with location aware systems, whereas older men wanted to know how to get places with various forms of transportation. Abascal and Civit (2001) looked at the pros and cons of older adults using mobile phones. They found that older adults liked the safety and increased autonomy mobile phones gave them. But, they were primarily concerned about social isolation and loss of privacy by using a mobile phone. Sri Hastuti Kurniawan (2006) found that older women felt safer with a mobile phone. Unlike younger counterparts, older women wanted brightly colored, bulkier phones with an antenna so it would be easily identifiable in a cluttered purse.

Ziefle and Bay (2005) looked at the cognitive complexity of older adults using mobile phones. They found that older adults performed just as well as younger adults on less cognitively complex mobile phones. They also reported that as the mobile phone interaction became more complex, older participants' performance suffered. Irie, Matsunaga, and Nagano (2005) created a mobile phone for elders by relying heavily on speech input technologies to help decrease complexity and input methods.

Most of the findings in these studies for mobile phones can apply to PDAs as well; however, the needs assessments differ because PDAs have larger physical interfaces and different input mechanisms. The lack of research in the area of PDA technology use by the older adults prompted Darroch, Goodman, Brewster, and Gray (2005) to evaluate a suitable font size for older people who needed to read text on a PDA screen. They found older people preferred reading 12-point font on PDAs, but could read fonts as small as 10 points. The authors pointed out that the lower resolution of their PDA screen could account for

the smaller font size preferred by participants than what Bernard and colleagues had previously reported. We looked at how older adults physically interacted with PDAs. We found that older adults had no problem pushing buttons, identifying icons, voice recording, or barcode scanning. Similar to the Darroch study, we found that although older participants preferred to read icons 25-mm large, they could read icons less than 15-mm large (Siek, Rogers, & Connelly, 2005).

Researchers must take into consideration what drives older adults to adopt new technologies for assistive applications to help the target population. Melenhorst, Rogers, and Caylor (2001) found that older adults must understand the benefits of information appliances and alternative communication mediums before they will consider the necessary training to use new technology. In addition, researchers found that for older adults to adopt a new technology, they must feel the technology is useful, convenient, safe, and simple to use, especially in older adults with varying cognitive and physiological abilities (Smither & Braun, 1994).

The findings from this body of research suggest that older people can use information appliances; however, designers and researchers must look at these findings to help inform their designs. More specifically, researchers must look at the physical device capabilities, interface design, and interaction techniques.

MAIN FOCUS OF THE CHAPTER

In this chapter, we broadly define older adults as people over 65 years old. It is difficult to define an ideal older adult because of the variability in older populations' abilities affected by age, illness, and cognitive or physiological decline. Thus, when designing for older populations, it is important to carefully define the target population, recruit older adults who meet the defined criteria, conduct meaningful requirements gathering and user studies, and design *prototypes* with older adults in mind. Here we discuss each of these items in more detail from our experiences in developing assistive applications for older adults.

Recruiting Older Target Populations

The first thing designers and researchers must figure out is what type of older population they would like to target. Will the application or device be for older people with cognitive impairments? Will it be for older people with physical disabilities? Or will the design be for *all* older people? Eisma and colleagues (Eisma, Dickinson, Goodman, Mival, Syme, & Tiwari, 2003) recommend bringing in older people early on in the design process to assist with requirements gathering and prototype development. They found that the different backgrounds of older people and designers mutually inspired the group to create realistic aims for the project. Older people on the design team can help answer questions specific about the abilities of the targeted population. Researchers must keep in mind that if the design is for all older people, the target population will have to be large enough to test people with varying physical, mental, and social abilities.

Researchers typically post fliers, e-mail calls for participations on mailing lists, and recruit participants from their work or university. This may not be the best way to find a pool of older adult participants. Older participants may not have the same *social networks* as the researchers. Thus, researchers should branch out and connect with community centers, religious groups, veteran meetings, assisted living centers, disability support groups, alumni associations, or adult communities to recruit an older diverse population. Typically, researchers can set up a meeting with the activities coordinator, technology group, or outreach liaison to meet older adults.

I would have no need for one of these, so I don't have to touch it. [PDA handed to audience mem-

ber] But, what if I break it? [Grabs PDA more confidently after researcher says she does not have to worry about breaking it. Pushes a few buttons on the screen.] Well look at that – I could show pictures to my friends.

– Audience member speaking to presenter after recruiting presentation

Similar to any participant population, older adults want to know what is expected of them and what the researcher will do with the data. When recruiting older participants, it is easiest to volunteer to give a presentation about the intended study that includes why the research is being done, what type of person you are looking for (e.g., user profile data), what the participant will have to do, and how the data will be used. The researchers can field questions from the audience to assuage future participants' concerns. Presentations are also the perfect time to hand out preliminary questionnaires to audience members and schedule future meetings for *focus groups*, *interviews*, or user studies. If participants are expected to use technology that may be unfamiliar to them, bring along the information appliance and let audience members play with the technology after the presentation. Emphasize that you are not testing the participants, but the device or application, and that the device or application cannot be broken with simple interactions. Guided hands-on interactions can change a person's view of the technology as shown in the audience member quoted previously.

Meeting with Older Adults

Designers and researchers will inevitably have to meet with the older adults in their target population during requirements gathering and user studies. There has been quite a bit of research (Eisma et al., 2003; Kurniawan, 2006; Zajicek, 2004) that looks into the best way to meet with older adults. Focus groups and semistructured interview sessions are the most popular meeting methods for requirements gathering and user studies. In this section, we briefly summarize the pros and cons of each method and give tips for best practices.

I have my walking group at 9, craft group at 10:30, doctors at 11:30, lunch at 1…

– Participant and facilitator attempting to schedule another meeting time

A common misconception is that older adults have plenty of time to meet with designers and researchers because they may be retired or work fewer hours. However, researchers may quickly find that some older adults have equally busy schedules. Taking notes about what each person is interested in based on the person's schedule can give insight into how the information appliance would fit into the person's everyday life. The quote about scheduling a meeting shows the participant's varied activities. Would the information appliance always be with her/him during the study? If so, how would she/he carry it when attending each meeting? If not, how can we remind her/him to bring the information appliance to only certain activities? We found that older adults with lower social-economic status have busy schedules too because they were more likely to have chronic illness or responsible for caring for family members.

Participant 1: I do not understand what you are saying. I have to see your lips!

Participant 2: I cannot see the screen because of glare.

– Participants' comments during a focus group

Focus groups typically allow researchers to get peoples' opinions, test ideas for specifications, evaluate prototypes, and learn more from the group by spontaneous discussions. Researchers

are divided on how beneficial focus groups are when working with older adults. Zajicek (2004) found that focus groups with over three older people are challenging because of hearing impairments, visual impairments, cognitive abilities, and the ability to follow a conversation. Kurniawan (2006) reported no problems and found that focus groups with over three older people tended to work together and help with *cooperative learning* exercises.

For prototypes that run on information appliances, we found focus groups challenging because of screen glare problems and complex interactions. Information appliances are small; thus, when trying to show a feature or explain an interaction, it is difficult to show it to all participants at once. We have issued each participant an information appliance in focus groups, with multiple researchers on hand to help the facilitator explain concepts, interface components, and interactions. This method allows the participants to see the proposed application and associated small interface components. It also gives the participants the chance to interact with the device and see how input methods are different for information appliances than with traditional desktop computers. Participants typically talked with the facilitator or to the people next to them to compare what they saw and discuss what they thought. Unfortunately, this type of focus group requires more time, preparation, and coaching by researchers. In addition, time must be set-aside for the group to discuss their ideas about the information appliance or application.

Alternatively, we have projected the interface or device onto a larger viewing surface so everyone sees the screen and can discuss the issue at hand. Participants were more likely to talk openly and start new discussions about interface components. The latter method allowed us to guide the discussion more efficiently, but it did not give the participants' the same realistic feel for the interface with smaller buttons and less-controlled interactions as the former method. Designers will have to take into consideration the focus group interaction method to receive appropriate feedback from participants.

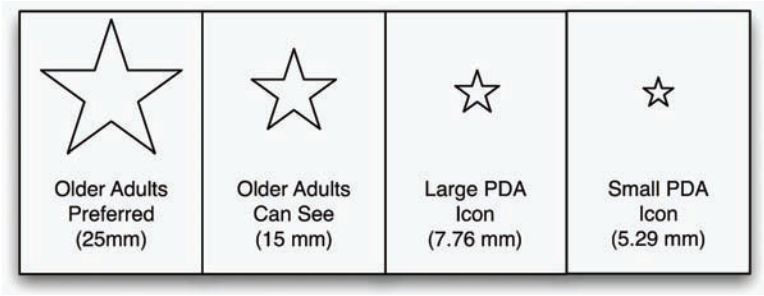
Interviews allow the facilitator to work one-on-one with a participant and ask more in-depth questions, or evaluate applications and devices more carefully. We found that we get the most detailed information about interface usability during semistructured interviews with accompanying task-centered user studies. The interview typically is quieter and has fewer distractions for the older user. In addition, the older user has a chance to interact with the device without worrying what others may think of him for not knowing how to do something on the information appliance.

My daughter thinks I am not smart because I cannot use a computer. But you know what—my daughter is not as smart as she thinks she is. One time when she was twelve, she came home from school and…

– Participant comment during interview

Our main problem with interviews, and sometimes with focus groups, is keeping on schedule. Older participants are more likely to share stories with the facilitator about their feelings towards technology when interacting in a one-on-one session. This rich data is useful, but there is a fine balance between keeping the conversation going and making sure the conversation does not diverge too much from the subject at hand, as shown in the previous quote. Another problem we encountered is that older people are more determined to finish each task than their younger counterparts, and will spend extra time to complete the tasks. We found that one-on-one interviews typically lasted one third longer than when working with younger participants.

Figure 1. Example of icon sizes older adults can view (preferred size and smallest viewable size) on a PDA



Physical Interfaces

In this section, we discuss some basic guidelines for the physical design of information appliances based on related work and our experiences. We found that older populations are interested in somewhat larger, more colorful physical devices and input components, although designers must find a balance between size and the perception of size.

When researchers conduct ethnographic studies or conduct studies where technology is discussed, but may not be necessarily used for data analysis, we find a persistent theme; most older adult populations want larger information appliances and input components (e.g., buttons, track wheels, etc.). A larger, bulky information appliance is easier to find, identify, and hold in one hand. Larger input components allow for quicker input. Indeed, a study found that older populations would prefer less overall functionality in exchange for larger buttons (Kurniawan, 2006). In terms of output, older adults would prefer to see a screen with more colors or contrast rather than have a larger screen.

In contrast, when studies have participants interact with the information appliance, they find that bigger is not always better. For example, participants in our studies were worried that their large fingers would press more than one button on an information appliance. The participants soon found that their perception of size was unfounded; they were able to interact with the smaller interface components (Siek et al., 2005). Another study found that older populations with specific physical ailments, such as paralysis, preferred smaller information appliances so they could be tucked into pockets easier (Eisma et al., 2003).

Since older populations are so diverse in abilities, it is difficult to create a strict guideline that specifies criteria of older adults who can use the information appliance. Instead, we have adopted an informal method of bringing information appliances to recruitment meetings and watching how older people interact with the devices. When we give an individual an information appliance, we collect her/his preliminary questionnaire and record comments about how she/he interacted with the device. After the recruitment meeting, the design team meets to discuss the interactions and questionnaire data to make correlations. Occasionally, we invite a clinician or an older adult to help us make conclusions about criteria needed to use the information appliance.

Virtual Interfaces

Similar to physical information appliance design, older adults are interested in the size of interface components and text. In addition, they prefer more common terminology to assist with interactions. Something that has not been studied with

information appliances is cognitive interactions and interactions with small widgets and interface components. In this section, we briefly discuss virtual interface guidelines that should be considered when designing information appliance applications for older people.

Older populations typically prefer larger fonts (e.g., 12-point font) (Darroch et al, 2005; Kurniawan, 2006) and icon sizes (e.g., 25mm) (Siek et al., 2005), as shown in Figure 1, but can read much smaller fonts (e.g., 10-point font) and icon sizes (e.g., 15mm). Design teams should take this information into consideration if they prefer to display interface information with text and icons. An application can be more appealing to older adults by using their font and icon size preferences; however, excess scrolling could make the application too complex. Indeed, we have found older populations have difficulty understanding the concept of scrolling on traditional computer Web browsers.

Besides the size of icons, older populations prefer realistic, picture-quality renderings to portray information in icons (Siek et al., 2005). Older participants prefer more detailed icons because the details helped them identify the function of the icon more efficiently.

Audience member 1: Why do I have to press Start to turn off my computer?

Audience member 2: Why do I have to press an apple to turn off my computer?

– Audience members' questions after recruiting presentation

Terminology used in virtual interfaces and user guides are often confusing to the general public. We found older users are more likely to voice their concerns and confusion about terminology. As the previous quote shows, audience members asked simple questions about the Windows and Apple desktop interfaces. At first it stumped the researcher; the reason why we press start and an apple symbol is because we always have. But just because we always have does not mean it is correct. If you would like to turn off your computer and in affect *end* all programs, why would one press start? These questions quickly prompted others to voice their concerns about e-mail and cell phone terminology. Researchers have documented older adults confusion about three-letter acronyms (e.g., SMS, MMS, etc.) and mobile phone terminology (e.g., What is a cell? What is roaming if I am always moving with a mobile phone?).

In this section, we described best practices that helped us and our fellow researchers develop successful information appliance applications for older adults. Since the older adults could use the applications and adopted them in their everyday lives, we assume these practices will help other researchers. We discuss in the next section future research directions for interface input components and interactions with these components. Research in this area will provide practitioners with guidelines to make consistently successful design decisions for information appliances.

FUTURE TRENDS

Information appliances are relatively new technologies, and mobile applications geared strictly towards older adult populations are only beginning to emerge. Researchers will continue to develop assistive applications for elders because of increases in the global older adult population. We must continue to address the issues proposed in this section to help further develop a guideline for information appliance development for older populations.

Researchers know how large common text and icons should be. We must look at how older populations use standard interface component widgets. Can they use standard size widgets with decreased fine motor skills? How large should the widget be?

I can only text people on my mobile phone if they text me first. I just push the reply button. I do not understand how to use the address book or how to enter people's names.

– Participant during interview

Cognitive interactions and interface complexity have been studied with traditional computers and Web sites. Currently, researchers have not delved into these issues for information appliances. Since information appliances have smaller screens and limited input capabilities, there will naturally be more interface screens and with it, increased complexity. As the participant noted in the quote, text messaging on mobile phones requires the user to input data from multiple sources (e.g., address book or alphanumeric key strokes) and send the message. However, once someone has sent a text message, it is easier to push one button and reply to the message. This interaction pattern could be a motivator for future research. How can we use this idea of one button interaction or precached contact data to increase communication mediums for older adults?

Interactions between the physical device and interface components are another area that must be researched for older adults to effectively use information appliances. For example, Charness and colleagues found that older adults had difficulties with traditional computer mouse and directional keyboard input because of varying fine motor control skills and the mapping between lateral movements with the mouse and the coordinate system on the screen. They found that light pens were optimal for older adults (Charness et al., 1995). In terms of information appliances, PDA screen input is similar to a light pen for optimal input. Despite this connection, designers must take into consideration that older adults may not have the fine motor control needed to select the standard, tiny interface components on PDA screens. In addition, current mobile phones pose an even bigger challenge, given the directional key presses needed to scroll and input information. It would be interesting to study if having these interactions close to the screen and on the same coordinate plane, such as in information appliances, will affect older peoples' perception of ease with information appliance input.

Along with standard interface development, researchers must strive to diversify the pool of older participants in their studies. Most studies summarized in this chapter worked with educated older populations. Indeed, recruitment from private assisted-living communities is fairly easy because the older adults who live in the community are educated and curious about technology. But a pessimistic view of the future may be that with such a large, ever growing population of older adults, the people who cannot afford private care will be monitored remotely by information appliances and *context aware* systems. If we create design guidelines and information appliance systems tested by people who are comfortable with technology, then we are leaving out the population who may need to use this technology one day. Researchers and designers must try to diversify their user pools by looking at education and socioeconomic status of their participants.

CONCLUSION

In this chapter, we looked at current research conducted with older adult populations using traditional desktop computers and information appliances. Research in the area of interface design for older adults is deficient because information appliances are relatively new, and design of assistive applications for older adults is just beginning to mature. We discussed issues and best practices that must be addressed when designing for information appliances. More specifically, we looked at the diversity of older adults, recruiting target populations, meeting with older adults in focus groups and interviews, and physical and virtual interface design considerations. We feel these best

practices are useful for researchers and the general practitioner because of our success with developing applications for older adults. Researchers and designers must strive to diversify their older adult target populations and consider people with different physical, cognitive, and emotional abilities. In addition, people from varying socioeconomic groups must be considered for the study to see how computing experience affects performance with information appliances.

REFERENCES

Abascal, J., & Civit, A. (2001). Universal access to mobile telephony as a way to enhance the autonomy of elderly people. In *Proceedings of the 2001 EC/NSF Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly* (pp. 93-99). New York, NY: ACM Press.

Bernard, M., Liao, C. H., & Mills, M. (2001). The effects of font type and size on the legibility and reading time of online text by older adults. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (pp. 175-176). New York, NY: ACM Press.

Carmien, S., DePaula, R., Gorman, A., & Kintsch, A. (2004). Increasing workplace independence for people with cognitive disabilities by leveraging distributed cognition among caregivers and clients. *Computer Supported Cooperative Work, 13*(5-6), 443-470.

Chaparro, A., Bohan, M., Fernandez, J., & Choi, S. (1999). The impact of age on computer input device - Psychophysical and psychological measures. *International Journal of Industrial Ergonomics, 24*(5), 503-513.

Charness, N., Bosman, E. A., & Elliott, R. G. (1995). *Senior-friendly input devices: Is the pen mightier than the mouse?* Paper presented at the 103rd Annual Convention of the American Psychological Association Meeting, New York.

Coroama, V., & Rothenbacher, F. (2003). The chatty environment - Providing everyday independence to the visually impaired. In *UbiHealth 2003.*

Darroch, I., Goodman, J., Brewster, S. A., & Gray, P. D. (2005). The effect of age and font size on reading text on handheld computers. In *Lecture Notes in Computer Science: Human-Computer Interaction - Interact 2005* (pp. 253-266). Berlin/Heidelberg, Germany: Springer.

Eisma, R., Dickinson, A., Goodman, J., Mival, O., Syme, A., & Tiwari, L. (2003). Mutual inspiration in the development of new technology for older people. In *Proceedings of INCLUDE 2003* (pp. 7:252-7:259). London, United Kingdom.

Helal, S., Giraldo, C., Kaddoura, Y., & Lee, C. (2003, October). Smart phone based cognitive assistant. In *UbiHealth 2003.*

Irie, T., Matsunaga, K., & Nagano, Y. (2005). University design activities for mobile phone: Raku Raku PHONE. *Fujitsu Sci. Tech. J., 41*(1), 78-85.

Jacko, J. A., Scott, I. U., Sainfort, F., Barnard, L., Edwards, P. J., Emery, V. K., et al. (2003). Older adults and visual impairment: What do exposure times and accuracy tell us about performance gains associated with multimodal feedback? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 33-40). New York, NY: ACM Press.

Kinsella, K., & Velkoff, V. A. (2001). *An aging world: 2001*(U.S. Census Bureau, Series P95/01-1). Washington, DC: U.S. Government Printing Office.

Kurniawan, S. (2006). An exploratory study of how older women use mobile phones. In *Proceedings of UbiComp 2006: Ubiquitous Computing* (pp. 105-122). New York, NY: ACM Press.

Laursen, B., Jensen, B. R., & Ratkevicius, A. (2001). Performance and muscle activity dur-

ing computer mouse tasks in young and elderly adults. *European Journal of Applied Physiology, 25*, 167-183.

Maguire, M., & Osman, Z. (2003). Designing for older inexperienced mobile phone users. In *Proceedings of HCI International 2003* (pp. 22-27), Mahwah, New Jersey: Lawrence Erlbaum Associates.

Melenhorst, A.-S., Rogers, W. A., & Caylor, E. C. (2001). The use of communication technologies by older adults: Exploring the benefits from the user's perspective. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (pp. 221-225).

Newell, A. F., & Gregor, P. (2001). Accessibility and interfaces for older people - A unique, but many faceted problem. In EC/NSF *Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly.*

Norman, D. (1999). *The invisible computer: Why good products can fail, the personal computer is so complex, and information appliances are the solution*. Boston: MIT Press.

Office of Communications. (2006). *Media literacy audit: Report on media literacy amongst older people*. London, United Kingdom: OfCom.

Siek, K. A., Rogers, Y., & Connelly, K. H. (2005). Fat finger worries: How older and younger users physically interact with PDAs. In *Lecture Notes in Computer Science: Human-Computer Interaction - Interact 2005* (pp. 267-280). Berlin/Heidelberg, Germany: Springer.

Smith, M. W., Sharit, J., & Czaja, S. J. (1999). Age, motor control, and the performance of computer mouse tasks. *Human Factors, 41*(3), 389-396.

Smither, J. A., & Braun, C. C. (1994). Technology and older adults: Factors affecting adoption of automatic teller machines. *The Journal of General Psychology, 121*(4), 381-389.

Warschauer, M. (2003). Demystifying the digital divide. *Scientific American*, 42-47.

Zajicek, M. (2004). Successful and available: Interface design exemplars for older users. *Interacting with Computers, 16*, 411-430.

Ziefle, M., & Bay, S. (2005). How older adults meet complexity: Aging effects on the usability of different mobile users. *Behaviour and Information Technology, 24*(5), 375-389.

KEY TERMS

Assistive Technologies: Applications and devices that pair human computer interaction techniques and technology to enhance the quality of life for people with various special needs.

Context Aware Systems: Technology embedded into our environments that communicates location, action, and other variables to help monitor the environment or individual.

Cooperative Learning: A method that allows individuals with different abilities to work together to improve their understanding of a subject.

Digital Divide: The gap between groups of people who do and do not have access to information technology.

Information Appliances: Electronic devices that allow people to send and receive various types of media (e.g., PDAs, mobile phones).

Focus Groups: A small group of selected participants who are asked questions about what they think about a specific topic or product (e.g., prototype); participants are free to discuss and build on what other participants say.

Interview: A participant is asked a series of questions by a facilitator to learn the participant's personal thoughts about a topic or product (e.g., prototype). Facilitators ask more open-ended ques-

tions in semistructured interviews and adapt future questions based on participants' feedback.

Prototype: A software or paper-based system that has a subset of the final application functionality; integral part of software development that allows researchers to get feedback from users before developing a fully functional system

Social Network: Connections between individuals with personal and professional relationships. Often the strength of the connections and influences of relationships are taken into account.

Technology Determinism: Idea that by introducing technology, people will understand and be able to use it.

Walk-Up-and-Use: Technologies that allow people to use the device or application without previous training or instruction (e.g., bank machines, self check-out kiosks at stores).

ENDNOTE

[1] Older people are defined here as 65 years old and over.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 624-634, copyright 2008 by Information Science Publishing (an imprint of IGI Global).

Chapter 8.6
Design for Mobile Learning in Museums

Nikolaos Tselios
University of Patras, Greece

Ioanna Papadimitriou
University of Patras, Greece

Dimitrios Raptis
University of Patras, Greece

Nikoletta Yiannoutsou
University of Patras, Greece

Vassilis Komis
University of Patras, Greece

Nikolaos Avouris
University of Patras, Greece

ABSTRACT

This chapter discusses the design challenges of mobile museum learning applications. Museums are undoubtedly rich in learning opportunities to be further enhanced with effective use of mobile technology. A visit supported and mediated by mobile devices can trigger the visitors' motivation by stimulating their imagination and engagement, giving opportunities to reorganize and conceptualise historical, cultural and technological facts in a constructive and meaningful way. In particular, context of use, social and constructivist aspects of learning and novel pedagogical approaches are important factors to be taken in consideration during the design process. A thorough study of existing systems is presented in the chapter in order to offer a background for extracting useful design approaches and guidelines. The chapter closes with a discussion on our experience in designing a collaborative learning activity for a cultural history museum.

INTRODUCTION

Use of mobile devices spreads in everyday human activities. These devices offer portability, wireless communication and connectivity to information resources and are primarily used as mobile digital assistants and communication mediators. Thus, it is no surprise that various attempts to use mobile appliances for learning purposes have been reported either inside or outside school (Roschelle, 2003). The term *mobile learning* or *m-learning* has been coined and concerns the use of wireless technologies, portable appliances and applications in the learning process without location or time restrictions. Practitioners' reports (Perry, 2003; Vahey & Crawford, 2002) and scientific findings (Norris & Soloway, 2004; Roschelle, 2003; Zurita & Nussbaum, 2004) communicate promising results in using these applications in various educational activities. The related bibliography proposes various uses of mobile appliances for learning. These Activities might concern access and management of information and communication and collaboration between users, under the frame of various learning situations.

A particular domain related to collaborative learning is defined as the support provided towards the educational goals through a coordinated and shared activity (Dillenbourg, 1999). In such cases, peer interactions involved as a result of the effort to build and support collaborative problem solving, are thought to be conducive to learning. On the other hand, traditional groupware environments are known to have various technological constraints which inflict on the learning process (Myers et al., 1998). Therefore, mobile collaborative learning systems (mCSCL) are recognized as a potential solution, as they support a more natural cooperative environment due to their wireless connectivity and portability (Danesh, Inkpen, Lau et al., 2001). While the mobility in physical space is of primary importance for establishing social interaction, this ability is reduced when interacting through a desktop system. It is evident that, by retaining the ability to move around it is easier to establish a social dialogue and two discrete communication channels may be simultaneously established through devices: one physical and one digital. Additionally, a mobile device can be treated as an information collector in a lab or in an information rich space (Rieger & Gay, 1997), as a book, as an organizing medium during transportation or even as a mediation of rich and stimulating interaction with the environment (i.e., in a museum). Effective usage of mobile appliances has been reported in language learning, mathematics, natural and social sciences (Luchini et al., 2002).

Furthermore, various technological constraints need to be taken in consideration during the design of activities which involve mobile devices. Such an example is the small screen, which cannot present all the information of interest while the lack of a full keyboard creates constraints in relation to data entry (Hayhoe, 2001). There is a need to provide the user with the possibility to 'go large' by getting information from both the virtual and physical world, while simultaneously 'going small,' by retrieving the useful and complementary information and getting involved into meaningful and easy to accomplish tasks (Luchini et al., 2002). In addition, despite the fact that technological solutions are proliferating and maturing, we still have a partial understanding of how users take effectively advantage of mobile devices. Specifically, in relation to communication and interaction, we need to investigate how mobile technology can be used for development of social networks and how it can provide richer ways for people to communicate and engage with others. In public spaces, like museums, a crucial question is if the serendipitous exchanges and interactions that often occur should be supported through mobile technology, how and where the interaction between people takes place and how is affected by this novel technology. Clearly, a better understanding of social activities and social interactions in public spaces should emerge to answer these questions.

A number of the aforementioned issues are discussed next in the context of a museum visit. First, we analyse how the context can affect any activity and application design. Then, we outline the most promising mobile learning applications and finally, we present our experiences of introducing collaborative learning activities using a novel approach based on the best practices surveyed previously, in a large scale project for a cultural history museum.

INTERACTION DESIGN FOR MOBILE APPLICATIONS

Interaction design is one of the main challenges of mobile applications design. Direct transfer of knowledge and practices from the user-desktop interaction metaphor, without taking in consideration the challenges of the new interaction paradigm is not effective. A new conceptualization of interaction is needed for ubiquitous computing. The traditional definition (Norman, 1986) of the user interface as a "means by which people and computers communicate with each other," becomes in ubiquitous computing, the means by which the people and the environment communicate with each other *facilitated* by mobile devices. As a result, interaction design is fundamentally different. In the traditional case, the user interacts with the computer with the intention to carry out a task. The reaction of the computer to user actions modifies its state and results in a dialogue between the human and the machine.

On the other hand, the user interaction with mobile devices is triadic, as the interaction is equally affected not only by the action of the user and the system's response, but by the context of use itself. The level of transparency of the environment, taking into account the presence of the mobile device and the degree of support to 'environmental' tasks meaningful to the user, are new issues to be considered. Consequently, new interaction design and evaluation criteria are required, since the design should not only focus on the user experience but pay also attention to the presence of other devices or objects of interest, including the level of awareness of the environment. By building the virtual information space into the real, the real is enhanced, but conversely, by drawing upon the physical, there is the opportunity to make the virtual space more tangible and intuitive and lower the overall cognitive load associated with each task.

To summarize, a number of design principles are proposed for mobile applications design:

a. Effective and efficient *context awareness* methods and models, with respect to the concept of context as defined by Dey (2001): 'Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.'
b. Presentation of useful information to the user *complementary* to the information communicated by the environment.
c. A*ccurate and timely update* of environmental data that affect the quality of interaction.
d. Contextualized and personalized information according to *personal needs.*
e. Information should be *presented to the user* rather than having the user searching for it.

Failure to look into these design issues can lead to erroneous interaction. For example, delays of the network, lack of synchronization between two artifacts of the environment or slight repositioning of the device can lead to misconceptions and illegal interaction states. In addition, information flow models should be aligned according to the information push requirement and relevant user modeling and adaptation techniques to support

this flow of information should be defined. Finally, new usability evaluation techniques, concerning mobile applications should emerge to shape a novel interaction paradigm.

Dix et al. (2000) present a framework to systematically address the discussed design issues and successive context awareness elements are inserted in the design process: (a) the *infrastructure* level (i.e., available network bandwidth, displays' resolution), (b) the *system* level (type and pace of feedback and feed through), (c) the *domain* level (the degree of adaptability that a system must provide to different users) and (d) the *physical* level (physical attributes of the device, location method and the environment). All these elements should be tackled independently and as a whole in order to study the effect of every design decision to each other.

We formulate the interaction design aspects discussed through the problem of designing a mobile learning application for a museum. During the visit a user has only a partial understanding of the available exhibits. This situation can be supported by complementary information included in the physical environment, for example alternative representations, concerning the historical role of the people or the artifacts presented the artistic value of a painting (Evans & Sterry, 1999), and so forth. This cognitive process of immersion into the cultural context, represented by the museum exhibits, could be supported by drawing upon the stimuli produced during the visit using context aware mobile devices. Therefore, these devices should be viewed as tools to enhance the involvement of a user in the cultural discovery process, tools that challenge the user to imagine the social, historical and cultural context, aligning her to a meaningful and worthy experience.

It is not argued here that the infusion of mobile technologies in museums will necessarily result to meaningful learning processes. Our analysis involves the potential use of the technology when integrated in educational activities (Hall & Bannon, 2006), which will offer a structured learning activity according to the characteristics of museums' content and the functionalities of the technologies used. To better illustrate this point (a) we briefly present a set of selected exemplary cases which demonstrate different ways of integrating mobile educational applications in museums and (b) we provide a more detailed account of such an application that we designed for a museum in Greece.

In the next section a number of approaches supporting such a visit are reviewed and examined using the design aspects as guiding paradigm and point of reference. Since the goal of the visitor is to see and learn more and not to explicitly use technology, a deep understanding of visitors' needs is important during the design phase, to avoid disturbances that can destruct her from her objective. Therefore, decisions made for the technology used and the styles of interaction, with the involved devices, have to deal with user's patterns of visit. Having the above requirements in hand, we use the framework proposed by Dix et al. (2000) to organize a coherent characteristics inspection of some representative examples of mobile museum guides.

MOBILE DEVICES AS MUSEUM GUIDES

In this section, some representative design approaches for mobile museum applications are discussed. An extended survey is included in Raptis, Tselios and Avouris (2005). The first system named "Electronic Guidebook," deployed in the Exploratorium science museum (Fleck et al., 2002), tries to involve the visitors to directly manipulate the exhibits and provides instructions as well as additional science explanations about the natural phenomena people are watching. The system of the Marble Museum of Carrara (Ciavarella & Paterno, 2004) stores the information locally in the PDA's memory, uses a map to guide the visitors around the museum and presents

content of different abstraction levels (i.e., room, section and exhibit). The "ImogI" system uses Bluetooth to establish communication between the PDAs and exhibits and presents the closest exhibits to the user, (Luyten & Coninx, 2004). The "Sotto Voce" system gives details about everyday things located in an old house (Grinter et al., 2002) by having pictures of the walls on the PDA's screen and asking from the user to select the exhibit she is interested in, by pressing it. The "Points of Departure" system (www.sfmoma.org) gives details in video and audio form by having 'thumbnails' of exhibits on the PDA's screen. It also uses 'Smart Tables' in order to enrich the interaction. A system, in the Lasar Segall Museum, Sao Paolo, Brazil (Dyan, 2004), automatically delivers information to the PDA, about more than 3,000 paintings. In the Tokyo University Digital Museum a system uses three different approaches to deliver content. The PDMA, in which the user holds the device above the exhibit she is interested in, the Point-it, in which the visitor uses laser-pointer to select specific exhibits and finally the Museum AR in which visitors wear glasses in order to get details about the exhibits (Koshizuka & Sakamura, 2000).

The system developed in the C-Map project, (Mase, Sumi, & Kadobayashi, 2000), uses active badges to simulate the location of the visitor, allowing tour planning and a VR system, controlled by the gestures of the visitor. In a Tour guide (Chou, Lee, Lee et al., 2004), the information about the exhibits is automatically presented and there is no variation in the form of the visit, but subjec-

Table 1. Design decisions affecting system context

	Location technology	**Storage of information**	**Flow of information**	**Additional functions**
"Antwerp project"	IrDA	In Server	Active	Cameras
C-Map	IrDA	In Server	Active, exhibit recommendations	Active Badges, Screens
Hippie	IrDA	In Server	Active, info presented based on the history of visit	
ImogI	Bluetooth	Info stored in Bluetooth transmitters	Active, proximity manager	
Lasar Segal Museum	IrDA	In Server	Passive	
Marble Museum	IrDA	Locally stored info, abstraction levels	Active, history of the visit	
PDMA, Point it, Museum AR	IrDA	In Server	Active	laser pointer, glasses
PEACH project	IrDA	In Server	Passive, task migration	Screens
Points of departure		Locally stored info	Active	Screens
Rememberer	RFID	In Server	Passive	Cameras
Sotto Voce		Locally stored info	Active	
Tour Guide System (Taiwan)	IrDA	In Server	Passive, subjective tour guides	

tive tour guides are used. A different approach is the one adopted in the Museum of Fine Arts in Antwerp (Van Gool, Tuytelaars, & Pollefeys, 1999), in which the user is equipped with a camera and selects exhibits, or details of an exhibit by taking pictures. A tour guide in the PEACH project, (Rocchi, Stock, Zancanaro et al., 2004), which migrates the interaction from the PDA to screens and uses a TV-like metaphor, using 'newscasters' to deliver content. Finally, a nomadic information system, the Hippie, developed in the framework of the HIPS project, (Oppermann & Specht, 1999), allows the user to access a personal virtual space during or after the visit. In the latter system, an electronic compass is used to identify the direction of a visitor.

The *infrastructure* context concerns the connections between the devices that comprise the system and influence the validity of the information that is provided through them to the users and needs not only to be addressed in problematic situations. It is also related with the validity and timely updates of available information. This can be clearly seen in collaboration activities where the user constantly needs to know the location of other users, the virtual space, the shared objects, and so forth. In the specific museum domain the results may not be so critical but can lead the user to various misunderstandings.

The mentioned systems use an indirect way of informing the user that her requests have been carried out: the user sees and hears the reflection of her requests on the PDA. There is no clear notification that the user's demands are executed successfully or not. Some of the systems use external factors, as signs of success, such as a led light ("Rememberer") and audio signals ("Marble Museum"). But in general terms, the user is on his or her own when problems occur and the systems leave it up to her to find it out, by observing that, there is no progress. We have to point that it could be very distracting and even annoying to have feedback messages in every state of interaction, but it is important for designers, to include a non-intrusive approach to inform that there is a problem and provide constructive feedback to overcome it.

Regarding the *system* context we can distinguish four different approaches as a means of awareness technology. In the first approach (Table 1), the PDA is the whole system. There are no other devices or awareness mechanisms involved and the information presented to the user is stored locally in the PDA. The second approach uses RFID tags to establish communication between the PDAs and the exhibits and the third which uses Bluetooth to establish communication with the exhibits and deliver content. The forth and most common approach uses IrDA technology to estimate the position of the visitor in space. Usually, IrDA tags are placed near every exhibit or in the entrance of each exhibition room and Wi-Fi derives the information to the PDA from a server Also, many different additional devices are built and integrated into these systems like screens (as a standalone devices or as interacting devices with the PDAs, where the user has the opportunity to transmit sequentially her interaction with the system from the PDA to a Screen). Regarding the *location*, all the studied systems use a topological approach to identify the position of a PDA, which informs approximately the system about the user's location. However, in the case of a museum with densely place exhibits, a more precise Cartesian approach can yield accurate user localization.

Domain context concerns aspects related to the situated interaction that takes place in the specific domain. Often in museum applications there is a lack of information about user profiles and characteristics. It is however important to consider that each visitor in a museum has different expectations, and is interested in different aspects regarding the exhibits. In the studied systems only in those that allow interaction of the users with servers there is a possibility for personalized interaction. Most of the systems require from the user to login, answer some specific questions, in

order to build a model of the user and present the information in her PDA according to her language, her expertise level and her physical needs (i.e., bigger fonts for those with sight impediments). When domain context is absent from the design process the system operates as a tool suited for the needs of a single hypothetical 'ideal' user. In such an environment this 'ideal' user will likely represent the needs and expectations of a small fraction of real visitors.

The system may push information to the user or it may wait until the user decides to pull it from the system. In the first case, special consideration should be taken to the user's specific activity and objective. Questions related to situated domain context are the following: Does the system propose any relevant information based on the history of users interaction? Does it adapt to actions repeatedly made by the user? Does it present content in different ways? For example, the "ImogI" system rearranges the order of the icons putting in front the mostly used ones. Also, in PEACH and in 'Points of Departure' the user can change the interaction medium from PDAs to Screens, in order to see more detailed information.

The *physical* context lays in the relation of the system with the physical environment and in problems concerning the physical nature of the devices. However, in the studied systems there is not a single mechanism of identifying the physical conditions. For example, in a room full with people, where a lot of noise exists, it would be appropriate if the system could automatically switch from an audio to a text presentation.

From the survey of the mobile guides applications presented here it seems that efficient design approaches could be achieved by augmenting physical space with information exchanges, by allowing collaboration and communication, by enhancing interactivity with the museum exhibits and by seamlessly integrating instantly available information delivered in various forms. However, the synergy between technology and pedagogy is not straightforward especially if we take into account the need to tackle issues such as efficient context integration, transparent usage of the PDA, and novel pedagogical approaches to exploit the capabilities of mobile devices. As a result, after discussing in detail usages of a mobile device as a mean of museum guidance, in the following, we attempt to discuss explicit educational activities mediated by mobile devices and a specific example of a new Mobile Learning environment.

DESIGNING MUSEUM MOBILE EDUCATIONAL ACTIVITIES

The level of exploitation of mobile devices in a museum setting is increasing and part of this use may have educational value. In this section we will focus on the added value of integrating educational mobile applications in museums. We will start our analysis by posing two questions that we consider central to this issue: (a) what is changing in the learning process taking place in a museum when mediated by mobile technology and (b) why these changes might be of educational or pedagogical interest? We will attempt to address these questions by focusing on three aspects related not only to the characteristics of mobile technology but also to the results of its integration in a museum. Specifically we will discuss: (a) the types of interaction between the visitor and the learning environment (e.g., the museum), (b) the learning activities that these interactions can support and (c) the role of context and motion in learning.

One facet of the learning process when mediated by mobile technology in museum visits involves the tangibility of museum artifacts: distant museum exhibits that were out there for the visitors allowing them just to observe now can be virtually touched, opened, turned and decomposed. In this case, technology provides to the user the key to open up the exhibit, explore it and construct an experience out of it. The tra-

ditional reading of information and observation of the exhibit is considered as one-dimensional "information flow" from the exhibit to the user. Mobile technology facilitates the transformation of the one dimensional relationship to a dialectic relationship between the user and the exhibit. Furthermore, this relationship can now include another important component (apart of the exhibits) of the museum environment: the other visitors. By providing a record of user–exhibit interaction for other visitors to see, reflect upon and transform technology can support social activities of communication, co-construction, and so forth between the visitors. To sum up, mobile technology mediates three types of interaction between the learner and the learning environment of a museum: (a) "exhibit–user" interaction (b) "user –exhibit" interaction and (c) "between the users" interaction about "a" and "b."

The enrichment of interaction between the learner and the museum might result in more or different learning opportunities (Cobb, 2002) the characteristics of which are outlined here. Specifically, the dialectic relationship between the user and the museum artifacts, mediated by mobile devices, might offer chances for analysis of the exhibit, experimentation with it, hypothesis formulation and testing, construction of interpretation, information processing and organization, reflection and many more, according to the educational activity designed. Collaboration and communication about the exhibits and information processing about them makes possible socio-constructive learning activities. By comparing these elements of the learning process to the reading or hearing of information about the exhibits (which is a the starting point for a non technology mediated museum visit) we realize that mobile technology has the potential to offer an active role to the learner: she can choose the information she wants to see, open up and de-construct an exhibit if she is interested in it, see how other visitors have interacted with a certain exhibit, discuss about it with them, exchange information, store information for further processing and use and so on.

Up until now, we described the role of mobile technology in learning with respect to two characteristics of the museum as learning environment: the exhibits and the other visitors. Another characteristic of the museum, which differentiates it from other learning environments (e.g., classroom) is that learning in a museum takes place while the learner moves. Learning while moving, quite often takes place very effectively without the support of technology. However there are cases that further processing with appropriate equipment is needed or some structuring of this "mobile learning experience" is proved to be useful. Mobile technologies can find in museums an important area of implementation not only because museum visits are structured around motion but because we have to support visitors *during* and not just after or before the visit (Patten, Arnedillo Sanchez et al., 2006). But why is it important to support learning during the visit? The answer here comes from the theory of situated learning (Lave & Wenger, 1991) which underlines the role of context in learning. Specifically, context facilitates knowledge construction by offering the practices, the tools, and the relevant background along with the objectives towards which learning is directed and has a specific meaning or a special function (knowledge is used for something). Finally, the use of mobile devices provides a new and very attractive way of interacting with the museum content especially for young children (Hall & Bannon, 2006).

As mentioned previously a large number of mobile applications have been developed during previous years for use in the museums (Raptis et al., 2005). All these mobile applications can add educational value to a museum visit in various ways. A survey of mobile educational applications for use inside the museum, led us to a categorization according to the educational approach followed in every occasion. The first category includes applications that mainly deliver

information to the visitor and concerns the vast majority of applications created for museums. Mobile devices take the place of the museums' docents and offer predetermined guided tours based upon certain thematic criteria. The aforementioned applications offer the museum visitor an enhanced experience which can support the learning process through a behaviorist approach. Enhancement is succeeded by supplying multimedia and context-related content.

The second category of applications, suitable for educational use in museums, consists of applications which provide tools that can support the learning process in a more profound way. Compared to the first category, they provide information about the exhibits of a museum but furthermore they include a series of functions that increase the interactivity with the user. Such an example is the Sotto Voce System (Grinter et al., 2002), which includes an electronic guide with audio content and the ability of synchronized sharing of this content between visitors. Thus, the users can either use individually the guide or "eavesdrop" to the information that another visitor listens.

Another example is the applications developed for the Exploratorium, a science museum in San Francisco (Fleck et al., 2002). In this museum, the visitor has the possibility to manipulate and experiment with the exhibits. Also, an electronic guidance was designed to provide information about the exhibits and the phenomena related with them, posing relative questions to provide deeper visitors' engagement. These applications are closer to social-cultural learning theories as they provide the user with tools to organize and control the provided information.

The third category of educational applications presents a specific educational scenario. Usually, game-based activities where the users, mostly children aged 5-15, are challenged to act a role and complete carefully designed pedagogical tasks. Such an example is the MUSEX application (Yatani, Sugimot & Kusunoki, 2004), deployed in the National Museum of Emerging Science and Innovation in Japan. MUSEX is a typical drill and practice educational system in which children work in pairs and are challenged to answer a number of questions. Children select an exhibit with their RFID reader equipped PDA and a question is presented in the screen with four possible answers. The activity is completed when each pair collects twelve correct answers. Children may collaborate and communicate either physically or via transceivers and monitor each group progress through a shared screen. After the completion of the activity the participators have the possibility to visit a Website and track their path inside the museum. The users can deeply interact with the exhibits, review the progress of her partner or ask for help (Yatani et al., 2004).

DinoHunter project includes several applications for the transmission of knowledge through game-based and mixed reality activities in the Senckenberg museum, Frankfurt, Germany. Three of these applications, namely DinoExplorer, DinoPick and DinoQuiz, are being supported by mobile technologies (Feix, Gobel, & Zumack, 2004). DinoExplorer delivers information to the users as an electronic guide, DinoPick allows the users to pick one part of the body of a dinosaur and get more multimedia information about this specific part and DinoQuiz provides a set of questions for further exploration of the exhibits of the museum.

Mystery at the Museum is another mobile, game-based, educational activity created for the Boston Museum of Science. It engages visitors in exploring and thinking in depth about the exhibits, thus making connections across them and encourages collaboration (Klopfer, Perry, Squire et al., 2005). High School students and their parents are called to solve a crime mystery where a band of thieves has stolen one of the exhibits. The users try to locate the criminals by using a PDA and a walkie-talkie. The participants must select upon the role of a technologist, a biologist or a detective. Depending on the chosen role they can interview

virtual characters, pick up and examine virtual objects by using virtual equipment (e.g., microscope), collect virtual samples via infrared tags or exchange objects and interviews through the walkie-talkies. A study confirmed deep engagement of the participants and extensive collaboration due to the roles set.

Another similar approach is presented through the Scavenger Hunt Game activity used in the Chicago Historical Society Museum (Kwak, 2004). In this case, the children are challenged to answer a series of questions related to the exhibits and the local history. They undertake the role of a historical researcher and they are called to answer 10 multiple choice questions while examining the exhibits. Each user is individually engaged into the activity and her progress is evaluated in a way similar to electronic games. The Cicero Project implemented in the Marble Museum of Carrara introduces a variety of games to the visitors (Laurillau & Paternò, 2004). The games vary from finding the missing parts of a puzzle to answering questions about the exhibits. Its main characteristic is the support it provides to the visitors to socially interact and collaboratively participate in activities concerning the exhibits of the museum, through peer-awareness mechanisms.

A series of mobile educational activities was also carried out in the frame of the Handscape Project in the Johnson Museum (Thom-Santelli, Toma, Boehner et al., 2005). The "Museum Detective" engage students in role-playing activities. Children working in pairs are called to locate an object described by one clue and learn as much as possible for it. A series of multiple-choice questions is presented for further exploration of the exhibit. Four types of interactive element are also provided for the exhibits: a painting, a drawing activity and a building activity and a multimedia narrative. The multiple-choice questions and the building activity were drill and practice activities and the rest were activities allowing children to make their own creations.

The systems of the latter category present coherent learning experiences comprised of planned and organized pedagogic activities, where an intervention has been purposefully designed to result a positive impact on children's cognitive and affective development. With respect to the contextual and interaction issues presented in the previous sections, we attempt to present in the next section an integrated application that involves children as role-playing characters by exploring the museum using a PDA.

AN EXAMPLE OF MOBILE ACTIVITY DESIGN FOR INDOOR MUSEUM VISIT

The "Inheritance" activity discussed here, is designed to support learning in the context of a cultural/historical museum visit. The application involves role-playing, information retrieval, data collecting and collaboration educational activities, suitable for children aged 10 or above working in teams of two or three members each. The activity scenario describes an imaginary story where the students are asked to help the Museum in finding the will of a deceased historian, worked for years in it. This will is hidden behind the historians' favorite exhibit. Clues to locate the document are scattered among the descriptions of some exhibits. If the children manage to find the will, all of the property of the historian will be inherited by the museum and not by his "greedy relatives." The scenario urges the students to read the description of the exhibits, find the clues and collaboratively locate the specific one.

During the design process of the activity we had to study the *museum context*, the *mobile technology* used and the *learning approach* to be followed in order to achieve the desired pedagogical outcome. The survey discussed in the previous section led us to adopt the following interaction design decisions. A PDA with wireless network capability is used and an RFID reader is attached

to it to 'scan' the RFID tags used to identify the exhibits. Wi-Fi infrastructure is being used to deliver data and establish communication between the visitors. When an exhibit is scanned, the PDA sends a request for information to the server which delivers the appropriate content presented in the form requested by the user. Data exchange between two users is accomplished through alignment of their devices while pointing one to the other, which mimics the exhibit scanning procedure. We also opted for small chunks of text since reading at low resolution screens reduces reading comprehension significantly.

The educational design of the activity was inspired by the social and cultural perspective of constructivism. It was structured around a set of learning objectives relevant to the thematic focus of the museum, to the exhibits' information, to the age and previous knowledge of the students, and to the fact that involves a school visit (as opposed to individual museum visits). The basic elements which shaped the activity were:

a. **Engagement of interest:** Engagement and interest hold an important role in the learning process. Student interest in a museum should not be taken for granted, especially because a visit arranged by the school is not usually based on the fact that some students might be interested to the theme of the museum. In the inheritance activity we considered to trigger student interest by engaging them in a game. The setting, the rules and the goal of the game were presented in the context of a story.

b. **Building on previous knowledge:** The focus of the activity was selected with respect to the history courses that students were taught in school. They had a general idea about the specific period of the Greek history and the activity offered complementary information about certain issues of this period. Building on previous knowledge was expected to support students in problem solving and hypothesis formulation and testing.

c. **Selecting–processing–combining pieces of information:** The scenario is structured around the idea that students read the offered information, select what is relevant to their inquiry and combine it with other pieces of information that have selected and stored earlier. Thus the students are expected to visit and re-visit the relevant exhibits, go through the information that involves them as many times as they think necessary and not just retrieve that information but combine it and use it in order to find the favorite exhibit which is the end point of the game

Figure 1. Screenshots of the "Inheritance" application: (a) Dialogue for RFID tags reading (b) information for a selected exhibit (c) clue selection (d) the notepad screen.

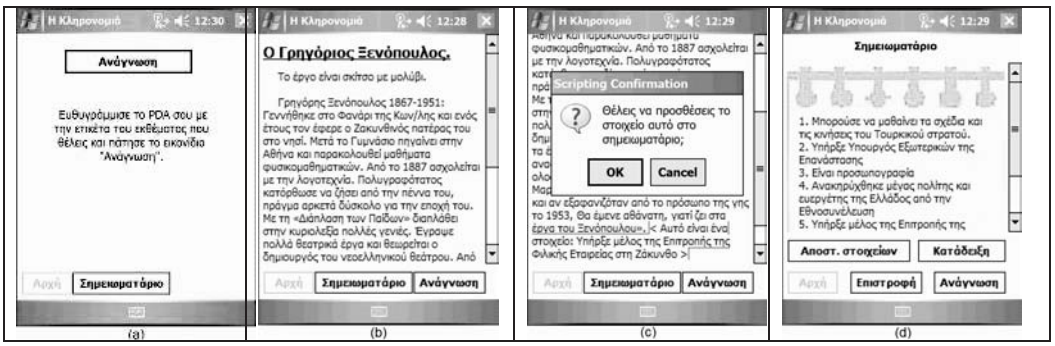


Figure 2. Children engaged in the activity

d. **Hypothesis formulation and testing:** When students have selected enough information from the exhibits around one room of the museum they can attempt to use some of the clues they have selected in order to find the favorite exhibit. If they fail they can go around the room to collect more information and try again.
e. **Communication and collaboration:** The activity is designed to facilitate inter and intra-group collaboration. Specifically, two groups of students are expected to collaborate to determine which exhibit they will interact with, to exchange clues using their PDA and to discuss their ideas about the favorite exhibit.

During the activity, the participating teams are free to explore any exhibit. Each team is provided with a PDA to extract information related to the exhibits by reading the tags attached to each of the exhibits (Figure 1). Only some of the exhibits contain 'clues,' which give information about the favourite exhibit to be found. Children must locate them, store them in the PDAs notepad and after collecting all or most of the clues the teams are able to beam their clues to each other. After collecting all six clues the students are challenged to locate the favorite exhibit. When both teams agree that one exhibit is the favorite one, they can check the correctness of their choice by reading with both PDAs the RFID tag of the chosen exhibit.

After the development of a prototype application, a case study was conducted inside the museum in order to validate the design choices. Seventeen students, aged 11, participated in the study (Figure 2). Data concerning all involving elements were collected to study the activity in depth. The activity was videotaped, PDA screen recording has been used and voice recorders were used to record dialogues among the participants.

The goal of the data analysis was twofold. First, to identify problems children encountered during the process in relation to each of the activity's elements. Then, to identify the nature of the interactions occurred during the procedure. Our analysis is based on the Activity Theory, concerning mainly human practices from the perspective of consciousness and personal development. It takes into account both individual and collaborative activities, the asymmetrical relation between people and things, and the role of artifacts in everyday life. The activity is seen as a system of human processes where a subject works on an object in order to obtain a desired outcome. In order to accomplish a goal, the subject employs tools, either conceptual or embodiments. Activity

Figure 3. Description according to the activity theory model

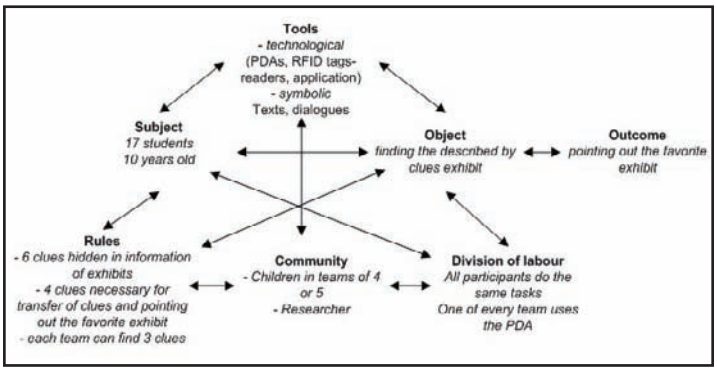


Table 2. An extract of the data analysis presenting action of the 'reading and searching for clues' class

Time	Actor	Tool	Events
00 : 08 : 50	Group1	PDA	Selection of "read"
00 : 08 : 52	Group1	texts	Information D. Stefanou
00 : 09 : 21	Group1	PDA	scrolling
00 : 09 : 33	Group1	PDA	scrolling
00 : 09 : 38	Group1	PDA	scrolling
00 : 09 : 42	Group1	PDA	scrolling
00 : 09 : 45	Group1	PDA	scrolling
00 : 09 : 49	Group1	PDA	scrolling
00 : 09 : 57	Group1	PDA	scrolling
00 : 10 : 02	Group1	PDA	scrolling
00 : 10 : 03	Group1	Dialogue	It doesn't have any (clues) here

is consisted by different components which are (Figure 3): (a) *subject*, (the persons engaged in the activity), (b) *object* (scope of the activity), (c) its *outcome* (c) *tools* used by the subjects (d) *rules-roles* that define the activity process, (e)*community* (context of the activity) and (f)*division of labour* (tasks division among the participants, Kuuti, 1995; Zurita & Nussbaum, 2004).

Activity Theory is of fundamental importance to deeper understand learning with mobile devices while visiting a museum, since in this case knowledge construction is mediated by cultural tools in a social context. The data collected were analyzed with the use of the Collaboration Analysis Tool (ColAT) environment which supports a multilevel description and interpretation of collaborative activities through fusion of multiple data (Avouris, Komis, Fiotakis et al., 2004).

In our analysis, an activity is a procedure during which objects become knowledge through three different levels-steps. Operation is the lower level where routine processes facilitate the completion of goal-oriented actions which in turn constitute the activity. Dialogues, user operations in the application and observations derived from the videos were transcribed in this first level of analysis. The actors, the operations and the mediating tools were noted in this level. Actors were the two participating teams and the researcher. The mediating tools were the dia-

logue among children and the researcher, texts of information (symbolic) and the application (technological). Some examples of operations in our case are text scrolling, RFID tag reading and transition from one screen to another. Analysis of these user operations led us to the identification of a problem in the use of the application. For example, due to data transfer delay from the server to the PDA, users in some occasions were frustrated and selected repeatedly an action due to lack of timely feedback.

In the second level, the different actions presented among the structural components of the activity are being studied. In order to identify and categorize the actions, a series of typologies were introduced. Typologies were set according to the goal and the mediating tool of each action. For example when children used the PDA to read the text information (mediating tools) their goal was not always the same. Three different typologies were used to describe the situation when the children read carefully the information provided ("Reading of information"), when they were reading the information and searched for clues also ("Reading and searching for clues"), and finally when they were "Searching for clues only." A "Reading and searching for clues" action example is presented in Table 2. Children scroll down the text and one of them states in the end of this action that they were unable to find a clue. When children search only for clues without paying attention to the information we observe rapid scrolling. A clear indication that they have already found all the clues is when children read only the information.

Other actions defined in our study were related to the dialogue between the children and the researcher aimed to overcome difficulties in using the application or understanding the rules-roles of the activity. Typologies where also introduced to describe the interaction between children related to the next step in the procedure (…"*Should we go there? …ok")* and the exchange of thoughts about the solution of the activity (…*"Well, tell me, the first clue is? …He could spy the Turkish army"…).* In the third level of analysis, patterns identified concerning the evolution of the procedure.

Clearly, the basic goals of the activity as described previously in this section have been fulfilled. Data analysis indicated that children were highly motivated by the activity and collaborated in order to achieve their goal. As derived from the analysis, the teams adopted different strategies to accomplish the task. Collaboration was observed mainly while making the choice of the next exhibit to be examined. After completing the task of finding the clues, the two teams collaborated more closely. They divided the work needed to find the exhibit described by the clues and looked in different parts of the room while collaborating and sharing their thoughts and suppositions. They used the clues as information filters thus eliminating the ones that did not match. Additionally, the learning result of this activity, as derived from subsequent students' essays describing the visit experience, was a deeper understanding of the historical role of the persons represented in the exhibits and their interrelations.

CONCLUSION AND FUTURE WORK

This chapter attempted to present current design approaches for mobile learning applications in the context of a museum visit. In addition, thorough study of similar approaches took place, which lead to useful design patterns and guidelines. As discussed, design of mobile learning systems, is not a straightforward process. In addition to the challenge of integrating the concept of context into the design process and independently from context conceptualization, a comprehension of pedagogical goals, desired learning transfers, user typical needs and objectives should take place. We argue that proper design decisions should take into account a solid theoretical cognitive framework, as well as the special characteristics of the mobile devices used and the challenges

of such an informal learning setting. A suitable activity should be properly supported by adequate interaction models, deeper understanding of the tasks involved to carry out the activity as a whole and their expectations while carrying out specific actions. For this reason, further validation of our proposed activity, took place in the actual museum. The activity was enjoyed by the students and enhanced their motivation to learn more about the cultural and historical context represented by the exhibits. The latter challenge has been better illustrated while discussing our experience of designing a collaborative learning activity in a cultural history museum and a case study validating its usefulness.

Clearly, the future of learning technology in museums lies in the blending, not the separation, of the virtual and the real world. That is because learning in a museum context could be conceived as the integration, over time, of personal, socio-cultural, and physical contexts. The physical setting of the museum in which learning takes place mediates the personal and socio-cultural setting. The so called 'interface transparency' should be treated as an effort to seamlessly integrate the computational device to our natural environment. This goal could be achieved by augmenting physical space with information exchanges, allowing collaboration and communication, enhancing interactivity with the museum exhibits and by seamless integrating instantly available information delivered in various forms. However, the synergy between technology and pedagogy is not straightforward, especially if we take into account the need to tackle issues such as efficient context integration, transparent usage of PDA, and novel pedagogical approaches to exploit the capabilities of the mobile devices. Therefore, further research effort should take place to experience established methods and practices.

REFERENCES

Avouris, N., Komis, V., Fiotakis, G., Dimitracopoulou, A., & Margaritis, M. (2004). Method and Tools for analysis of collaborative problem-solving activities. In *Proceedings of ATIT2004, First International Workshop on Activity Theory Based Practical Methods for IT Design* (pp. 5-16). Retrieved on February 28, 2007 from http://www.daimi.au.dk/publications/PB/574/PB-574.pdf

Chou, L., Lee, C., Lee, M., & Chang, C. (2004). A tour guide system for mobile learning in museums. In J. Roschelle, T.W. Chan, Kinshuk & S. J. H. Yang (Eds.), *Proceedings of 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education—WMTE'04* (pp. 195-196). Washington, DC: IEEE Computer Society.

Ciavarella, C., & Paterno, F. (2004). The design of a handheld, location-aware guide for indoor environments. *Personal Ubiquitous Computing, 8*, 82–91.

Cobb, P. (2002). Reasoning with tools and inscriptions. *The Journal of the Learning Sciences, 11*(2-3), 187-215.

Danesh, A., Inkpen, K.M., Lau, F., Shu, K., & Booth, K.S. (2001). Geney: Designing a collaborative activity for the palm handheld computer. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems—CHI 2001* (pp. 388-395). New York: ACM Press.

Dey, A. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal, 5*(1), 4-7.

Dillenbourg, P. (Ed.) (1999). *Collaborative learning: Cognitive and computational approaches.* Oxford, UK: Elsevier Science.

Dix, A., Rodden, T., Davies, N., Trevor, J., Friday, A., & Palfreyman, K. (2000). Exploiting space and location as a design framework for interactive mobile systems. *ACM Transactions on Computer-*

Human Interaction, 7(3), 285–321.

Dyan, M. (2004). *An Introduction to Art, the Wireless Way.* Retrieved on March 25, 2005 from http://www.cooltown.com/cooltown/mpulse/1002-lasarsegall.asp

Evans, J., & Sterry, P. (1999). Portable computers and interactive multimedia: A new paradigm for interpreting museum collections. *Journal Archives and Museum Informatics, 13*, 113-126.

Feix, A., Göbel, S., & Zumack, R. (2004). DinoHunter: Platform for mobile edutainment applications in museums. In S. Göbel, U. Spierling, A. Hoffmann, I. Iurgel, O. Schneider, J. Dechau & A. Feix (Eds.), *Proceedings of the Second International Conference on Technologies for Interactive Digital Storytelling and Entertainment: Conference Proceedings—TIDSE 2004* (pp. 264-269). Berlin: Springer.

Fleck, M., Frid, M., Kindberg, T., Rajani, R., O'Brien-Strain, E., & Spasojevic, M. (2002). From informing to remembering: Deploying a ubiquitous system in an interactive science museum. *Pervasive Computing, 1*(2), 13-21.

Grinter, R. E., Aoki, P. M., Szymanski, M. H., Thornton, J. D., Woodruff, A., & Hurst, A. (2002). Revisiting the visit: understanding how technology can shape the museum visit. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work—CSCW 2002,* (pp. 146-155). New York: ACM Press.

Hall, T., & Bannon, L. (2006). Designing ubiquitous computing to enhance children's learning in museums. *Journal of Computer Assisted Learning, 22*, 231–243.

Hayhoe, G. F. (2001). From desktop to palmtop: creating usable online documents for wireless and handheld devices. In *Proceedings of the IEEE International Conference on Professional Communication Conference–IPCC 2001* (pp. 1-11).

Klopfer, E., Perry, J., Squire, K., Jan, M., & Steinkuehler, C. (2005). Mystery at the museum: a collaborative game for museum education. In T. Koschmann, T. W. Chan & D. Suthers (Eds.), *Proceedings of the 2005 conference on Computer support for collaborative learning: the next 10 years!* (pp. 316-320). Mahwah, NJ: Lawrence Erlbaum.

Koshizuka, N., & Sakamura, K. (2000). The Tokyo University Museum. In *Kyoto International Conference on Digital Libraries: Research and Practice* (pp. 85-92).

Kuuti, K. (1995). Activity theory as a potential framework for human-computer interaction research. In B. Nardi (Ed.), *Context and consciousness: Activity theory and human computer interaction* (pp. 17-14). Cambridge: MIT Press.

Kwak, S.Y. (2004). *Designing a handheld interactive scavenger hunt game to enhance museum experience.* Unpublished diploma thesis, Michigan State University, Department of Telecommunication, Information Studies and Media.

Laurillau, Y., & Paternò, F. (2004). Supporting museum co-visits using mobile devices. In S. Brewster & M. Dunlop (Eds), *Proceedings of the 6th International Symposium on Mobile Human-Computer Interaction—Mobile HCI 2004* (pp 451-455). Berlin: Springer.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* New York: Cambridge University Press.

Luchini, K., Quintana, C., Krajcik, J., Farah, C., Nandihalli, N., Reese, et al. (2002). Scaffolding in the small: Designing educational supports for concept mapping on handheld computers. In *CHI 2002 Extended Abstracts on Human Factors in Computing Systems* (pp. 792-793). New York: ACM Press.

Luyten, K., & Coninx, K. (2004). ImogI: Take control over a context aware electronic mobile guide for museums. *In proceedings of the 3rd*

Workshop on HCI in Mobile Guides. Retrieved on February 24, 2007 from http://research.edm.luc.ac.be/~imogi/

Myers, B. A., Stiel, H., & Gargiulo, R. (1998). Collaboration using multiple PDAs connected to a PC. In *Proceedings of the ACM 1998 Conference on Computer Supported Cooperative Work — CSCW '98* (pp. 285-294). New York: ACM.

Mase, K., Sumi, Y., & Kadobayashi, R. (2000). The weaved reality: What context-aware interface agents bring about. In *Proceedings of the Fourth Asian Conference on Computer Vision - ACCV2000* (pp. 1120-1124).

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper (Eds.), *User centered systems design* (pp. 31-61). Mahwah, NJ: Lawrence Erlbaum.

Norris, C., & Soloway, E. (2004). Envisioning the handheld-centric classroom. *Journal of Educational Computing Research, 30*(4), 281-294.

Oppermann, R., Specht, M., & Jaceniak, I. (1999). Hippie: A nomadic information system. In H. W. Gellersen (Ed.), *Proceedings of the First International Symposium Handheld and Ubiquitous Computing - HUC'99* (pp. 330-333). Berlin: Springer.

Patten, B., Arnedillo Sanchez, I., & Tangney, B. (2006). Designing collaborative, constructionist and contextual applications for handheld devices. *Computers and Education, 46,* 294-308.

Perry, D. (2003). *Handheld Computers (PDAs) in Schools*. BECTA ICT Research Report. Retrieved on February 26, 2007 from http://www.becta.org.uk/ page_documents/research/handhelds.pdf

Raptis, D., Tselios, N., & Avouris, N. (2005). Context-based design of mobile applications for museums: a survey of existing practices. In M. Tscheligi, R Bernhaupt & K. Mihalic (Eds.), *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services- Mobile HCI 2005* (pp. 153-160). New York: ACM Press.

Rieger, R., & Gay, G. (1997). Using mobile computing to enhance field study. In R.P. Hall, N. Miyake & N. Enyedy (Eds.), *Proceedings of Computer Support for Collaborative Learning –CSCL 1997* (pp. 215–223). Mahwah, NJ: Lawrence Erlbaum.

Rocchi, C., Stock, O., Zancanaro, M., Kruppa, M., & Krüger, A. (2004). The museum visit: Generating seamless personalized presentations on multiple devices. In J. Vanderdonckt, N. J. Nunes & C. Rich (Eds.), *Proceedings of the Intelligent User Interfaces - IUI 2004* (pp. 316-318). New York: ACM.

Roschelle, J. (2003). Unlocking the learning value of wireless mobile devices. *Journal of Computer Assisted Learning, 19*(3), 260-272.

Thom-Santelli, J., Toma, C., Boehner, K., & Gay, G. (2005). Beyond just the facts: Museum detective guides. In *Proceedings from the International Workshop on Re-Thinking Technology in Museums: Towards a New Understanding of People's Experience in Museums* (pp. 99-107). Retrieved on February 25, 2007 from http://www.idc.ul.ie/museumworkshop/programme.html

Vahey, P., & Crawford, V. (2002). *Palm education pioneers program final evaluation report*. Menlo Park, CA: SRI International.

Van Gool, L., Tuytelaars, T., & Pollefeys, M. (1999). Adventurous tourism for couch potatoes. (Invited). In F. Solina & A. Leonardis (Eds.), *Proceedings of the 8th International Conference on Computer Analysis of Images and Patterns – CAIP 1999* (pp. 98-107). Berlin: Springer.

Yatani, K., Sugimoto, M., & Kusunoki, F. (2004). Musex: A System for Supporting Children's Collaborative Learning in a Museum with PDAs. In J. Roschelle, T.W. Chan, Kinshuk & S. J. H. Yang (Eds.), *Proceedings of 2nd IEEE International*

Workshop on Wireless and Mobile Technologies in Education – WMTE'04 (pp 109-113). Washington, DC: IEEE Computer Society.

Zurita, G., & Nussbaum, M. (2004). Computer supported collaborative learning using wirelessly interconnected handheld computers. *Computers and Education, 42*, 289-314.

KEY TERMS

Activity Theory: Is a psychological framework, with its roots in Vygotsky's cultural-historical psychology. Its goal is to explain the mental capabilities of a single human being. However, it rejects the isolated human being as an adequate unit of analysis, focusing instead on cultural and technical mediation of human activity.

Context: Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves (Dey, 2001).

Context-Aware: The ability to sense context.

Interaction Design: Interaction design is a sub-discipline of the design notion which aims to examine the role of embedded behaviors and intelligence in physical and virtual spaces as well as the convergence of physical and digital products. In particular, interaction design is concerned with a user experience flow through time and is typically informed by user research design with an emphasis on behavior as well as form. Interaction design is evaluated in terms of functionality, usability and emotional factors.

Mobile Device: A device which is typically characterized by mobility, small form factor and communication functionality and focuses on handling a particular type of information and related tasks. Typical devices could be a Smartphone or a PDA. Mobile devices may overlap in definition or are sometimes referred to as information appliances, wireless devices, handhelds or handheld devices.

Mobile Learning: Is the delivery of learning to students who are not keeping a fixed location or through the use of mobile or portable technology.

Museum Learning: A kind of informal learning which is not teacher mediated. It refers to how well a visit inspires and stimulates people into wanting to know more, as well as changing how they see themselves and their world both as an individual and as part of a community. It is a wide concept that can include not only the design and implementation of special events and teaching sessions, but also the planning and production of exhibitions and any other activity of the museum which can play an educational role.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 253-269, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.7
Component Agent Systems:
Building a Mobile Agent Architecture That You Can Reuse

Paulo Marques
University of Coimbra, Portugal

Luís Silva
University of Coimbra, Portugal

ABSTRACT

One central problem preventing widespread adoption of mobile agents as a code structuring primitive is that current mainstream middleware implementations do not convey it simply as such. In fact, they force all the development to be centered on mobile agents, which has serious consequences in terms of software structuring and, in fact, technology adoption. This chapter discusses the main limitations of the traditional platform-based approach, proposing an alternative: component-based mobile agent systems. Two case studies are discussed: the JAMES platform, a traditional mobile agent platform specially tailored for network management, and M&M, a component-based system for agent-enabling applications. Finally, a bird's eye perspective on the last 15 years of mobile agent systems research is presented along with an outlook on the future of the technology. The authors hope that this chapter brings some enlightenment on the pearls and pitfalls surrounding this interesting technology and ways for avoiding them in the future.

INTRODUCTION

A mobile agent (Chess et al., 1994; White, 1996) is a simple, natural and logical extension of the remote distributed object concept. It is an object with an active thread of execution that is capable of migrating between different hosts and applications. By using mobile agents, the programmer is no longer confined to have static objects and perform remote invocations but can program the objects to move directly between applications. In itself, a mobile agent is just a programming

abstraction: an active object that can move when needed. It is a structuring primitive, similar to the notion of class, remote object, or thread.

Two possible approaches for the deployment of mobile agents in distributed applications are:

a. To use a middleware platform that provides all the mechanisms and support for the execution of mobile agents. The basic characteristic of platform-based systems is that there is an infrastructure where all agents execute. This infrastructure typically corresponds to a daemon or service on top of which the agents are run. All agents coexist on the same infrastructure. When the programmer develops an application, he is in fact modeling different mobile agents, which execute on the platform. Typically, this is done by extending a `MobileAgent` class or a similar construct. In fact, some of the mobile agents may not even be mobile but just static service agents interfacing with other functionalities of the system. Examples include, among others, the SOMA platform (Bellavista, Corradi, & Stefanelli, 1999), D'Agents (Kotz et al., 1997), Ajanta (Tripathi et al., 2002), Aglets (Aglets Project Homepage, 2006; Lange & Oshima, 1998), and JAMES (Silva et al., 1999). This is by far the most common approach.
b. An alternative approach is to provide the support for mobile agents as software components that can be more easily integrated in the development of applications. This approach is followed by the M&M project (Marques, 2003), described in this chapter.

In this chapter, we present the results of two major projects that have been conducted in our research group: JAMES and M&M.

The **JAMES** platform was developed in collaboration with SIEMENS and consisted of a traditional mobile agent platform especially optimized for network management applications. Our industrial partners used this platform to develop some mobile agent-based applications that were integrated into commercial products. These applications used mobile agents to perform management tasks (accounting, performance management, system monitoring, and detailed user profiling) that deal with very large amounts of data distributed over the nodes of GSM networks. With this project, we learned that this technology, when appropriately used, provides significant competitive advantages to distributed management applications.

The main motivation for the second project, M&M, was to facilitate the development process and the integration of mobile objects within ordinary applications. M&M abandoned the classic concept of mobile agent platforms as extensions of the operating system. Instead, this middleware is able to provide for agent mobility within application boundaries, rather than within system boundaries. Its objective was to demonstrate that it is possible to create an infrastructure such that the mobile agent concept can be leveraged into existing object-oriented languages in a simple and transparent way, without interfering in the manner in which the applications are normally structured. In order to achieve this goal, a component-oriented framework was devised and implemented, allowing programmers to use mobile agents as needed. Applications can still be developed using current object-oriented techniques but, by including certain components, they gain the ability to send, receive, and interact with agents. The component palette was implemented using the JavaBeans technology and was, furthermore, integrated with ActiveX (Box, 1997; Denning, 1997), allowing programmers from any programming language that supports COM/ActiveX to take advantage of this paradigm. To validate the soundness of the approach, a large number of applications have been implemented using M&M. Two application domains were of particular interest: agent-enabling web servers (Marques, Fonseca, Simões, Silva,

& Silva, 2002a) and disconnected computing (Marques, Santos, Silva, & Silva, 2002b).

The rest of the chapter is organized as follows:

- In the **Background** section, a general introduction to platform-based systems for mobile agents is presented, followed by a case study: the JAMES platform;
- Then, in the **Component-Based Mobile Agent Systems** section, an alternative model is discussed, based on binary software components. In particular, it addresses some of limitations of the platform-based approach, presenting the M&M case study and its implications;
- The next section gives a **Bird's Eye Perspective** on the state of mobile agent technology;
- Finally, the last section presents the **Conclusion** and an outlook on the future of mobile agent technology.

BACKGROUND

Mobile Agent Platforms

The foundation for most platform-based systems is a server that sits on top of the operating system and where all agents execute. The platform is responsible for housing the agents and for providing every single feature needed by them and their surrounding environment (Delamaro, 2002; Marques, Simões, Silva, Boavida, & Gabriel, 2001). It provides functionalities like migration support, naming, security, inter-agent communication, agent tracking, persistence, external application gateways, platform management, and fault-tolerance. In fact, the agent platform plays the role of the "operating system of agents."

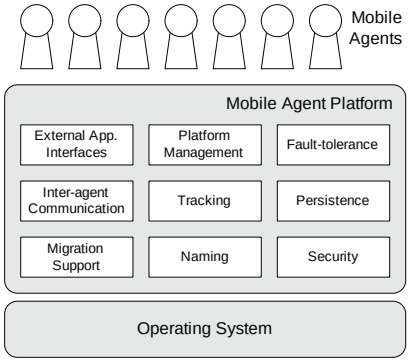
This list of supported features in an agent platform is by no means complete. Many application-specific domains have specific requirements. This leads to domain-specific implementations, having special features to address domain-specific requirements. Examples include: the JAMES platform (Silva et al., 1999), for telecommunication applications; aZIMAS (Arumugam, Helal, & Nalla, 2002) for supporting mobile agents in web servers; SOMA (Bellavista, 1999) for interoperability and integration with CORBA.

Figure 1 presents the typical architecture of such a system. As said, the operating system sits on the bottom. In the middle, there is the agent platform where all the agents execute. On the top, there are the agents belonging to all applications. This last point is especially important. Since all agents execute on top of the platform, usually all the agents from all applications can see each other. Although agents/applications can be divided into namespaces, as it happens in some platforms, and security permissions can be configured for proper access, in most cases the notion of application is quite weak or even inexistent.

In terms of programming model, in most platforms, the only support provided for application development is around the concept of mobile agent. Everything becomes a mobile agent, even entities that are conceptually services or gateways. Typically, inter-agent communication mechanisms are used to allow interactions between the different parts of the system. Some authors even refer to the concept of "whole application as a mobile agent" as the *fat-agent model* (Simões, Reis, Silva, & Boavida, 1999).

In Figure 2, this concept is illustrated. The developers have implemented two different applications, A and B, which in practice are just a set of mobile agents. True mobile agents are represented as white pawns. Black pawns represent static entities that are programmed as mobile agents due to the lack of proper infrastructure for application development. The interactions are fully based on inter-agent communication mechanisms. The concept of application is largely based on conventions about which agents communicate with what. Interestingly, many times there is even the

Figure 1. The mainstream mobile agent platform architecture



need to set up agents with special interfaces (and permissions) that are able to communicate with external applications via inter process communication (IPC) mechanisms. As strange as it seems, it is not uncommon for the only IPC mechanism available to provide integration with external entities to be a socket connection. It is also common for the agent platform to mediate all the access to operating system resources, especially if support for security is provided by the platform.

Developing distributed applications based on mobile agent systems has important advantages over the traditional client-server type of interactions. Mobile agents allow, among other things (Lange & Oshima, 1999), for: *reduced network traffic*, *easy software upgrading on-demand*, *easy introduction of new services in the network*, *higher robustness for the applications*, *support for disconnected computing*, *higher scalability*, *easy integration of vendor-proprietary systems*, and *higher responsiveness in the interactions with other systems.*

The JAMES Platform

In 1998, realizing the importance and advantages of using mobile agents in distributed systems, a consortium was setup for developing a mobile agent platform oriented for the development of applications in the field of telecommunication and network management. The project was called JAMES, and its partners were the University of Coimbra (Portugal), SIEMENS S.A. (Portugal), and SIEMENS AG (Germany); the being was implemented under the umbrella of a European Eureka Program (Σ!1921).

The JAMES platform provides the running environment for mobile agents. There is a distinc-

Figure 2. Applications on a standard agent platform are implemented as different mobile agents that map all the required functionality

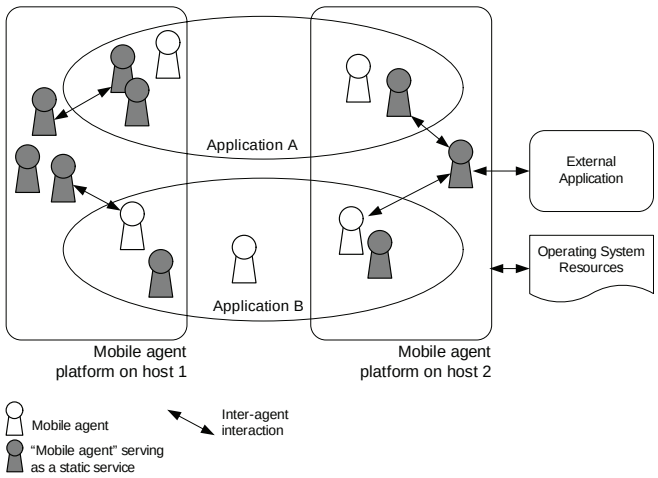
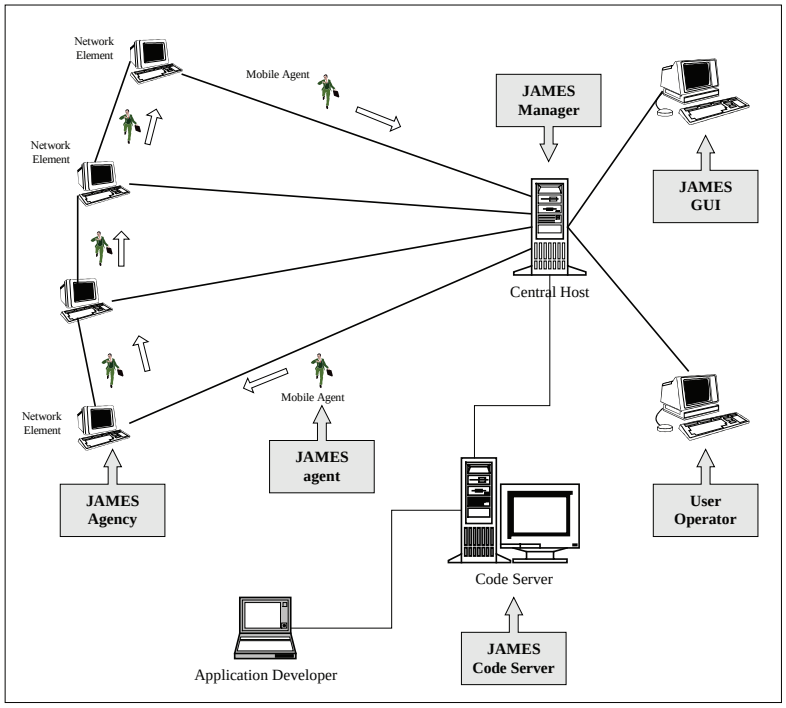


Figure 3. An overview of the JAMES platform



tion between the software environment that runs in the manager host and the software that executes in the network elements (NEs): the central host executes the JAMES manager while the nodes in the network run a JAMES agency. The agents are written by application programmers and execute on top of that platform. The JAMES system provides a programming interface that allows the full manipulation of mobile agents. Figure 3 presents a global snapshot of the system, with a special description of a possible scenario where the mobile agents will be used.

Every NE runs a Java virtual machine and executes a JAMES agency that enables the execution of the mobile agents. The JAMES agents will migrate through these machines of the network to access some data, execute some tasks, and produce reports that will be sent back to the JAMES manager. There is mechanism of authentication in the JAMES agencies to control the execution of agents and to avoid the intrusion of non-official agents. The communication between the different machines is done through stream sockets. A special protocol was developed to transfer the agents across the machines in a robust way and is atomic to the occurrence of failures.

The application developer writes the applications that are based on a set of mobile agents. These applications are written in Java and should use the JAMES API for the control of mobility. After writing an application, the programmer should create a JAR with all the classes that make part of the mobile agent. This JAR file is placed in a JAMES code server. This server can be a different machine or in the same machine where the JAMES manager is executing. In both cases, it maintains a code directory with all the JAR files

available and the mapping to the corresponding mobile agents.

The host machine that runs the JAMES manager is responsible for the whole management of the mobile agent system. It provides the interface to the end-user, together with a graphical user for the remote control and monitoring of agents, places, and applications. The JAMES GUI is the main tool for management and administration of the platform. With this interface, the user can manage all the agents and agencies in the system.

For lack of space we will not describe the inner details of the JAMES platform. However, in the following list we present the key features of our mobile agent system:

- Portability of the applications, through the use of the Java language;
- High-performance in mobility through the use of caching and prefetching techniques;
- Security mechanisms for code authentication;
- Resource control service to manage the use of underlying resources;
- An overview of the JAMES Platform (CPU, memory, disk and operating system resources);
- System monitoring;
- Fault-tolerance through the use of checkpointing and reconfiguration;
- Easy-to-use programming interface;
- Scalable execution of mobile agents, through the use of decentralized protocols;
- Easy customization of the software;
- "On-the-fly" software upgrading;
- Interoperation with classical network management protocols, like SNMP;
- Distributed management and easy configuration of the network;

The result of this platform was further exploited by the industrial partners that have developed some applications for performance management in telecommunications networks using mobile-agent technology and the JAMES platform.

COMPONENT-BASED MOBILE AGENT SYSTEMS

Introduction

Having completed the JAMES project, several important lessons were learned. Possibly the most important was that using a *mobile agent platform* as middleware seriously limits the acceptance of mobile agents as a structuring and programming paradigm. The problem is not the mobile agents by themselves but the use of monolithic platform for developing and deploying them. Platforms force the programmer to adopt a completely different development model from the one in mainstream use (object-oriented programming). When using an agent platform, the programmer is forced to center its development, its programming units, and its whole applications on the concept of agent. Although useful and relatively simple to implement, the use of platforms limit the acceptance of mobile agents as simple programming constructs (Kotz, Gray, & Rus, 2002).

In this section, we start by exploring how traditional mobile agent platforms limit the acceptance of mobile agents as a structuring primitive, and then we present M&M, an agent system specifically for overcoming those limitations.

Limitations of the Platform-Based Model

The reasons why the platform architecture limits the acceptance of the mobile agent paradigm can be seen from three perspectives: the programmer, the end-user, and the software infrastructure itself.

The Programmer

One fundamental aspect of the mobile agent paradigm is that, by itself, it does not provide more functionality than what is attainable with the traditional client/server model. One of the major advantages of the mobile agent paradigm is that *logically* (i.e., without considering its physical implementation), its functionalities as a whole and as a structuring primitive—*an active thread of execution that is able to migrate*—are particularly adapted for distributed computing (Papaioannou, 2000). Mobile agent technology, in itself, has no killer application (Chess et al., 1994; Kotz et al., 2002).

Taking this into consideration, a relevant question is: What strong reason can motivate a programmer to consider using mobile agents to develop its applications? After all, everything that can be attained by using mobile agents can be done using client/server. The most important reason, as discussed before, is that a mobile agent is a logical structuring primitive very adapted for distributed computing. Still, for it to be accepted, the price to be paid by the programmer cannot be too high. With traditional agent platforms, typically it is.

The problems include: the mobile agent concept is not readily available at the language level; the applications have to be centered on the mobile agents; and a complicated interface between the agents and the applications and operating system resources must be written. The programmers want to develop their applications as they currently do, by using object-oriented techniques, and by using mainstream APIs. Agents will typically play a small role on the application structuring. Current platforms force exactly the opposite. Instead of being middleware, agents are the *frontware*.

If one looks at history, it is easy to understand that the RPC success is due to its strong integration with structured programming environments. RPCs did not force programmers to abandon their programming languages, environments, and methodologies. On the contrary, RPCs embraced them. Understandably, if the RPC model had required completely different languages and methodologies, it would have failed. Programmers would have continued to use sockets. After all, everything that can be done using an RPC can be done using a socket, granted that it is so with different degrees of difficulty. The argument also applies to RMI and its integration with object-oriented languages. Remote method invocation did not require different languages or environments but instead blended into existing systems. In both cases, developers started to use the new technologies because: (a) they were readily integrated at the language level and into their programming environments; (b) the applications could continue to be developed using the existing methodologies and only use the new technologies as needed; c) no workarounds were necessary for integrating with existing applications.

The point is that mobile agent technology should not necessarily force complete agent-based software design. It should be a complement to traditional object-oriented software development and easily available to use in ordinary distributed applications, along with other technologies.

The End-User

From the viewpoint of the user, if an application is going to make use of mobile agents, it is first necessary to install an agent platform. The security permissions given to the incoming agents must also be configured, and the proper hooks necessary to allow the communication between the agents and the application must be setup. While some of these tasks can be automated by using installation scripts, this entire setup package is too much of a burden.

Usually, the user is not concerned with mobile agents nor wants to configure and manage mobile agent platforms. The user is much more concerned with the applications than with the middleware they are using in the background. In

the currently available mobile agent systems, the agents are central and widely visible. They are not the background middleware but the foreground applications.

Also, the term *mobile code* has very strong negative connotations, which makes the dissemination of mobile agent technology difficult. The user is afraid of installing a platform capable of receiving and executing code without its permission. This happens even though the existence of mobile code is present in technologies like Java, in particular in Applets, RMI, and JINI. The fundamental difference is that in those cases, the user is shielded from the middleware being used. In many cases, using mobile agents does not pose an increased security threat, especially if proper authentication and authorization mechanisms are in place. However, because the current agent platforms do not hide the middleware from the user, the risk associated with the technology is perceived as being higher than it is. This causes users to back away from applications that make use of mobile agents.

The Software Infrastructure

One final limitation of the platform-based approach lies in the architecture itself. Generally speaking, there are very few platforms that provide means for extensibility. Typically, the platform is a monolithic entity with a fixed set of functionalities. If it is necessary to provide new functionality, for instance, a new inter-agent communication mechanism, that functionality is directly coded into the platform. What this means is that if there are new requirements or features to be supported, it is necessary to recompile the whole platform and deploy it in the distributed infrastructure. This lack of support for system extensibility has several important consequences.

The first important consequence is management costs. As the name indicates, the *platform* is a software infrastructure that must be managed and attended at all times. Currently, when an operator deploys an agent-based application, it does not gain just one new application to administrate. Besides the application, it gains a full-blown agent platform to manage. This type of cost is not negligible. For instance, it is curious to observe how sensitive the network management and telecommunications communities are to management costs, even though they are amongst the ones that can most benefit from the use of this technology (Picco, 1998; Simões, Rodrigues, Silva, & Boavida, 2002).

Another facet of the monolithic structure of the agent platform problem has to do with changing requirements. It is well known in the software industry that the requirements of applications are constantly changing. In fact, this knowledge has motivated a whole series of software methodologies that take this into account, as rapid development (McConnell, 1996) and eXtreme programming (Beck, 1999). In most of the current agent platforms, each time a feature is added or an error corrected, the software is recompiled, originating a new version. Although it is quite simple to upgrade an application based on mobile agents—it may be as simple as to kill the agents and send out new ones—the same does not happen to the agent infrastructure. When there is a new version of the agent infrastructure, it is necessary to manually redeploy it across the distributed environment.

Even though it is easy to devise code-on-demand solutions for the complete upgrade of agent platforms or to use server-initiated upgrades, as in the case of JAMES (Silva et al., 1999), most existing platforms do not provide for it. In many cases, the cost of redeploying the software across the infrastructure may be unacceptable. This is a second type of management cost that must be paid. In terms of deployment, it would be much more appealing to have an infrastructure where parts could be plugged in, plugged out, and reconfigured as necessary. Voyager (Glass, 1997), MOA (Milojicic, Chauhan, & la Forge, 1998), and gypsy (Lugmayr, 1999) provided the first experi-

ments in this area, though, due to the monolithic architecture of most platforms, this type of solution is not readily available for them.

The M&M Agent System

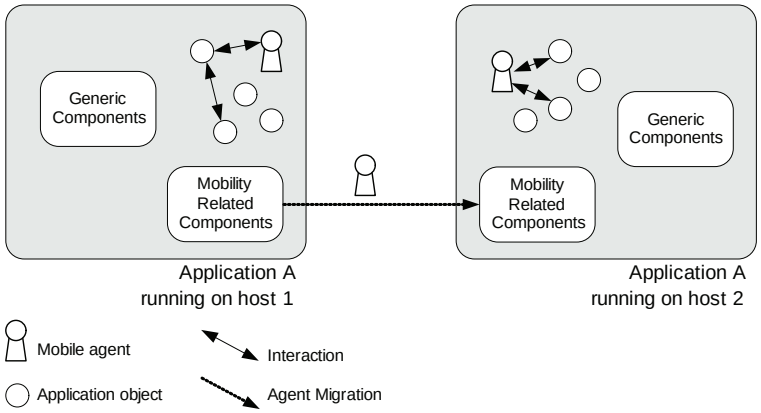
The most distinctive characteristic of M&M is that there are no agent platforms. Instead, the agents arrive and depart directly from the applications they are part of. The agents exist and interact with the applications from the inside, along with the other application objects.

The applications become agent-enabled by incorporating well-defined binary software components into their code. These components give them the capability of sending, receiving, and interacting with mobile agents. The applications themselves are developed using the current best-practice software methods and become agent-enabled by integrating these "mobility components," that is, M&M framework components. We call this approach ACMAS—*application centric mobile agent systems*—since the applications are central and mobile agents are just part of the system playing specific roles.

The key idea is that the different functionality typically found on a monolithic agent platform is factored out as independent pluggable components that can be added or removed from the applications. No underlying agent platform is involved. In Figure 4, the approach is shown. Here, an application is being run on two different hosts. This application was built by using object-oriented programming and by incorporating generic components, like the ones that provide easy database connectivity or graphics toolkits, and mobility-related components, as the ones that provide migration support and agent tracking. Agents are able to migrate between applications by using the mobility-related components.

Comparing this model with the platform-based architecture, it is quite clear that the applications are no longer a set of agents. In this approach, when inside an application, an agent is just another object that has an associated thread of execution. The agents are just like any other objects of the application. Different applications incorporate the specific components necessary for their operation, executing them side-by-side. Another advantage of this approach is that agents can be specific to their applications, not having all the agents from all the applications coexisting together.

Figure 4. The applications become agent-enabled by incorporating well-defined binary components



The M&M Component Palette

When developing an application by using the ACMAS approach, three different types of components are involved: generic third-party off-the-shelf components; application-specific components; and mobile agent-related components (see Figure 5):

- **Third-party off-the-shelf components** are components that are commercially available from software makers. Currently, there is a large variety of components available for the most different things, like accessing databases, designing graphical user interfaces, messaging, and others. All these components can be used for building the applications without having to re-implement the required functionalities.
- **Domain-specific components** are modules that must be written in the context of the application domain being considered, providing functionalities not readily available off-the-shelf. For instance, while implementing a particular application, it may be necessary to write special parsers for extracting information from files or to write supporting services for allowing agents to monitor the hardware of a machine. These modules can be coded as components and incorporated into the application.
- **Mobile agent-related components** provide the basic needs in terms of mobile-agent infrastructure. These components provide the functionalities typically found in agent platforms: mobility support, inter-agent communication mechanisms, agent tracking, security, and others. The M&M component palette fits into this category.

When writing an application, the programmer assembles the necessary components and interconnects them by using programmatic glue. The application logic that mandates the execution of the program can be a separate module or be embedded in the wiring of the components. Typi-

Figure 5. Applications are created by assembling the necessary components

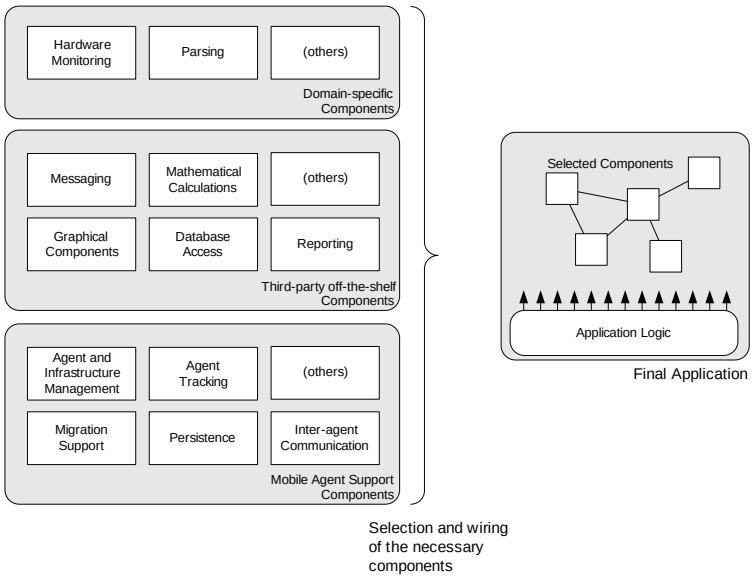


Table 1. Available components in the M&M component palette

Component(s)	Functionality
Mobility component	Provides the basic support for agent mobility, agent control, and monitoring. It incorporates an extensibility mechanism that allows other components to interact with the mobile agents.
Management components	Allows agents and the instantiated components to be monitored and controlled locally and remotely by applications and by administrative agents.
Agent tracking components	Allows the agents, local, and external applications to know the location of each agent in the distributed application.
Security component	Allows agents to safely execute inside the applications and for the applications to safely execute the agents. It is responsible for the provision of authentication and authorization services, and of monitoring and controlling what operations each agent is allowed to perform.
Local communication components	Supports message exchange between agents and applications or other agents, in the context of a single running application, using several paradigms (message passing and publisher-subscriber, both synchronously and asynchronously).
Global communication components	Allows the agents and the applications to exchange messages using several paradigms (message passing and publisher-subscriber, both synchronously and asynchronously), in the global context of a distributed application.
Disconnected computing components	Provides support for disconnected computing, allowing agents to be saved into persistent storage if they are not able to migrate to a disconnected device, and to migrate when the device comes back online. Persistent storage is also implemented as a separate component.
Web publishing components	Allows agents that migrate to a Web server to publish information and act as Web resources.

cal examples of the former are backend database and transaction processing applications; examples of the latter are form-based applications for data entering and querying. The final application is the combination of components, wiring glue, and backend application logic.

When considering a component palette for supporting mobile agents, two different aspects must be considered. On the one hand, there are components that are integrated into the application, giving it special abilities in terms of interacting with mobile agents. One example of this is a component that gives the application the ability to track the agents whenever they migrate.

On the other hand, when programming the mobile agents themselves, there are components that can be used for building them, for instance, a component that when included in a mobile agent gives it the ability to communicate with other mobile agents. Currently, the M&M framework supports both types.

Thus, when discussing the component palette of M&M, presented in Table 1, it is important to realize that there are components for including

into the applications and components for including into agents. In fact, some of the components can even be used for both purposes (e.g., the client components of inter-agent communication).

Another important point is that sometimes the components do not implement the full functionality of a service and serve only as access points to certain functionalities. For instance, the client component for agent tracking connects to a network server that holds and updates the location of the agents. It should be noted that the M&M component list is by no means static or fixed. The M&M framework has well-defined interfaces that allow third-party developers to create new components and add them to the system and to upgrade existing components. In fact, M&M somewhat resembles a Lego system where new pieces can be created and fitted into the existing ones.

Figure 6 shows an application being created in a visual development tool and being agent-enabled by using M&M. On the left it is possible to see that the mobility component has been included; on the right, the properties of this component are shown; on the top, the M&M component palette is visible.

A detailed account of the M&M system and its implementation can be found in Marques (2003).

Consequences of Using a Component-Based Approach

Using a component-based approach for developing applications that use mobile agents has several important consequences. Some of these consequences are directly related to the characteristics of component technology, others are a product of the way M&M was designed. Some of the most important aspects are:

- **The users do not see agents or manage agent platforms:** As agents are sent back into the middleware instead of being "frontware," what the users ultimately see are applications and their interface. The adoption of a certain application is once again based on its perceived added value, not on the technology that it is using. Also, because users do not have to manage agents nor platforms, applications become conceptually simpler. No middleware infrastructure is explicitly visible to install and maintain. Although there are still distributed applications to install and manage, this is much easier than managing a separate infrastructure shared by a large number of distributed applications with different policies and requirements.
- **Tight integration with the end applications:** Since the agents can migrate from end-applications to end-applications, interacting with them from the inside, development becomes much more straightforward. There is no need to set up service agents, configure their policies, and devise workarounds based on IPCs. Once inside an application, a mobile agent is just a thread of execution that can access all the objects of that application, under proper security concerns. This contributes to better performance and scalability since the interaction with the applications does not have to go through the middlemen—the service agents.
- **Tight integration with existing programming environments:** For supporting mobile agents in an application, all the developer has to do is to include some of the components into the application and interconnect them. The applications can continue to be developed using object-oriented methodologies, having the necessary support for active migrating objects. By using M&M components in the applications, developers do not have to center all the development on the agents. Agents become just "one more" powerful distributed programming construct, as once advocated by Papaioannou (Papaioannou, 2000). What this means is that the develop-

ment model is well integrated with current software development practices and environments. The path to using mobile agents in the applications becomes as smooth as using, for instance, remote method invocation.

- **Possibility of using visual development tools:** Software components are normally designed so that they can be manipulated visually. Instead of focusing on an API approach, components emphasize on well-defined visual entities with properties and interconnections that the programmer can configure visually. This can result in a large increase in productivity, a smoother learning curve, and a wider acceptance of a technology. By using a component-based approach, M&M takes benefit of all these characteristics.
- **Support for any programming language:** It is possible to use a JavaBeans/ActiveX bridge to encapsulate JavaBeans components as ActiveX ones. Thus, in the Windows platform it is possible to use a JavaBeans component from any programming environment, as long as it supports COM/ActiveX. This was the approach taken in the M&M framework. By using such a bridge, it was possible to have applications written in several languages—Visual Basic, Visual C++, Java, and so forth—sending and receiving agents between them. Even though, from a research point of view, this might not seem so important, from an integration and acceptance point of view it is quite important: It is directly related to whether a technology gets to be used or not.
- **Security can be integrated into the application:** One of the valuable lessons learned when designing the Internet was the importance of the end-to-end argument in system design (Saltzer, Reed, & Clark, 1984). There is a fragile balance between what can be offered generically by an infrastructure and what should be left to the application design on the endpoints. In terms of security, this is especially important. In many cases, it is only at the end applications, and integrated with the application security policies and its enforcement, that it is possible to take sensible security decisions. By using components directly integrated into the applications, the security of the mobile agents of an application becomes integrated with the security policies of the application itself.
- **Only the necessary components are included in each application:** Because the developer is implementing its application and only using the necessary features of mobile code, only the required components that implement such features need to be included. What this means is that it is not necessary to have a gigantic platform that implements every possibly conceivable feature, because in the end many of these features are not used in most applications. By using specific components it is possible to build a software infrastructure that is much more adapted to the needs of each specific application.
- **Extensibility and constant evolution:** In the last section it was discussed how "changing requirements" are an important problem and how monolithic platforms are ill-equipped to deal with new features, new releases, and redeployment. By using a component-based approach, it is possible to continually implement new components, with new features and characteristics, without implying the rebuilding of a new version of "the platform." The component palette can always be augmented.
- **Reusability and robustness:** One of the acclaimed benefits of using components is their tendency to become more robust over time. As a component is more and more reused in different applications, more errors are discovered and corrected, leading

to new and more robust versions. Because components are black boxes with well defined interfaces and functionalities, they are also easier to debug and correct. By using a component-based approach to implement a mobile agent framework, it is also possible to benefit from these advantages.

Overall many applications have been implemented using M&M accessing its usefulness, easy of use, and close integration with programming environments and software. In this context, two application domains were of particular interest: agent-enabling web servers and supporting disconnected computing.

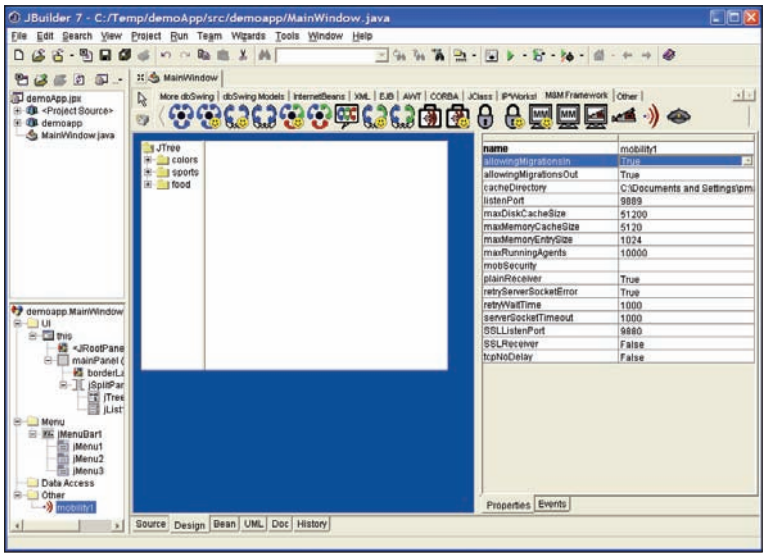
- The M&M component framework allows any Web server that supports the *servlet specification* to send, receive, and use mobile agents. M&M provides a clean integration, requiring neither a custom-made Web server nor a special purpose agent platform. A full account of this exploration can be found in Marques et al. (2002a).
- Finally, supporting disconnected computing has allowed us to understand the current limitations of the framework in supporting very low-level operations. This account can be found in Marques et al. (2002b).

A BIRD'S EYE PERSPECTIVE

Over the years, there has been a huge proliferation of mobile agent platforms. In fact, already in 1999 there were 72 known platforms. Nevertheless, the technology is still far from common use by mainstream programmer and, oddly as it seems, there now seems to exist more agent platforms than mobile agent-based applications.

In the previous sections, we have discussed how the agent platform architecture seriously impairs the actual leverage of the mobile agent paradigm into real applications, by real programmers, to real users. The paradigm does not bring a "one order of magnitude" increase in functionality, and at the same time, the agent platform architecture imposes a too-high price

Figure 6. Screen capture of the M&M component palette being used for developing a demonstration application



to developers, users, systems administrators, and institutions that could make use of the technology. The road imposed by the platform architecture is not one of integration with existing tools and environments but one of complete conversion to a new architecture. This conversion is asked for without giving anything substantial enough in return, and imposing many limitations on what can be implemented and how.

About the Proliferation of Mobile Agent Platforms

In our view, the huge proliferation of mobile agent platforms is directly connected with two factors: (a) the monolithic nature of agent platforms; (b) the advent of the Java language.

When considering different application fields, each one has its specific requirements. For instance, in the network management area, an important requirement is integration with SNMP. In other domains, there are others. Because in most cases it is not possible to extend a generic agent platform to take into account the needs of a certain application domain, what has happened is that researchers, and in fact the industry, have developed many new agent platforms that try to address specific domains. The Java language made it extremely easy to develop basic infrastructures for object and thread mobility, and thus to experiment in this area. In most cases, these platforms were developed from scratch.

The problems with these platforms are: they are not reusable across domains (many times not even across problems in a single domain); they do not take into account results in research (e.g., how to properly implement agent tracking or fault-tolerance); and, because in many cases they are quite experimental, they lack the robustness needed to be usable in the real world. The result is the current huge number of platforms that are not used by anyone and a large disappointment with the mobile agent paradigm.

The Mobile Agent Community

Another important problem is that the mobile agent community is too biased on the platform architecture, and little concerned with its integration with existing systems and infrastructures. This strong bias is easy to observe in practice. For instance, when looking at standards for mobile agents systems such as MASIF (OMG, 2000), from the Object Management Group, and FIPA (FIPA, 2000; FIPA, 2002), from the Foundation for Intelligent Physical Agents, it can be seen that the standardization effort was done around the concept of "agent platform", the platform on top of which all agents *should* execute. Neither alternative execution models nor any provisioning for integration with existing programming environments were considered.

Curiously, two of the most respected researchers in the area of intelligent agents—Nick Jennings and Michael Wooldridge—have long been arguing about the dangers of trying to leverage agents into the market. In the classic article *Software Engineering with Agents: Pitfalls and Pratfalls* (Wooldridge & Jennings, 1999), they examine what has been happening in the intelligent agent community. But, in fact, the observations also apply to the mobile agent community. Some of the key lessons from their experience are: "*You oversell agents*," "*You ignore related technology*," "*You ignore legacy*," "*You obsess on infrastructure*," and "*You see agents everywhere*". These are all relevant points if one wants to bring the mobile agent paradigm into real-world programming environments.

This view is somewhat supported by the success of Voyager (Glass, 1997) when compared with other agent infrastructures. Voyager is not a mobile agent platform but an ORB that, among other things, is also able to send and receive agents. Voyager does not force everything to be centered on agents, nor forces the applications to be agents. Applications are developed using ordinary methods and can use Voyager as an ordinary ORB. In

practice, what has happened is that programmers who were using Voyager as a commercial ORB understood that they also had support for sending and receiving "active threads." Not having any prejudice, they started using this facility. Users continued to see applications as they always had, and system administrators did not gain complex agent platforms to manage.

Security

Security is traditionally singled out as the biggest problem facing mobile agents, and that prevented their dissemination as a paradigm. The idea of having arbitrary code that can execute on a host can be a scary one. There is the danger of resource stealing, denial-of-service attacks, data spying, and many other problems (Farmer, Guttman, & Swarup, 1996; Greenberg, Byington, & Harper, 1998).

Even so, although there are still some hard problems to be solved in a practical way, like the malicious host problem (Hohl, 1998) and proper resource control, security is probably one of the most active areas of research in mobile agent systems (e.g., Loureiro, 2001, contains an extensive survey on existing research on the topic).

When developing mobile agent applications, two different scenarios have to be considered. The first scenario consists of deploying an application that uses mobile agents in a closed environment. What this means is that it is possible to identify a central authority that controls all the nodes of the network where the agents are deployed. For example, a network operator may deploy an agent platform on its network, for being able to deliver new services in an easy and flexible way. Although different users, with different permissions, may create agents, the key point is that there is a central authority that is able to say who has what permissions and guarantee that nodes do not attack agents.

A completely different scenario is to have agents deployed in an open environment. In this picture, agents migrate to different nodes controlled by different authorities, possibly having very different goals. One classic example of this situation is an e-commerce application on the Internet. Different sites may deploy an agent platform, allowing agents to migrate to their nodes and query about products and prices, and even perform transactions on behalf of their owners. Here the sites will be competing against each other for having the agents making the transactions on their places. There is no central authority, and each site may attack the agents, stealing information from them, or making them do operations that they were not supposed to do.

Although it may appear that deploying applications on closed environments is too restrictive, there is a large number of applications that are worth deploying in such setting. Examples include network management, telecommunication applications, software distribution and upgrading, parallel and distributed processing, and groupware. For such environments, the currently available solutions are well adapted and sufficient. In many cases, it is a matter of proper authentication and authorization mechanisms, associated with a public key infrastructure.

The key argument is that although a lot of research is still necessary for securely deploying agents in open environments, there is a multitude of applications that can be developed securely for existing computing infrastructures in closed environments. On closed environments, besides the psychological effect of having code that is able to migrate between hosts, there is no additional risk when compared with existing technologies. The risks are similar to having *rexec* daemons running on machines, or using code servers in Java RMI.

In our view, the argument that it is security that is preventing the deployment of mobile agent technology is a misguided one. Many applications can be developed securely; and considering distributed systems, there are many technologies that operate without any special security considerations. That lack of security has not prevented

them from being developed or having a large user base. A classical example is SNMP.

Agent Languages: Could They Be the Solution?

Over the years, many researchers came up with new languages that express mobile processes and mobile computations directly (e.g., Visual Obliq, Cardelli, 1995; Jocaml Conchon, & Fessant, 1999; Nomadic Pict & Wojciechowski, 1999). Although these languages integrate the mobile agent paradigm at the language level and are interesting in terms of the lessons learned in expressing new abstractions, they present the same generic problem as the platform architecture.

These languages force the programmers to use completely different programming paradigms and software structuring methodologies. At the same time, because using mobile agents does not bring any large increase in productivity nor enables anything important that cannot be achieved by classical means, programmers are not compelled to try out these new languages. In fact, it does not make much sense to ask developers to abandon proven development techniques and environments in favor of new languages that do not allow anything new or powerful, have not proven themselves, and force all the development to be centered on different abstractions.

The mobile agent technology should be available at the language level, but this means that the middleware should allow the creation, management, and integration with mobile entities directly, probably through an API at the same level than other programming libraries. It means that the programmer should be able to continue to use its current programming environments, languages, and development methodologies. When necessary, and only then, it should be able to create an active thread of execution that would be able to migrate and interact with the objects on different applications. This does not mean that all the development should be centered on the agents or that new languages are really necessary.

CONCLUSION

Mobile agent research is now almost 15 years old. Many valuable lessons have been learned in areas so diverse as persistence, resource allocation and control, tracking, state capture, security, communication, coordination, and languages. Nevertheless, no killer application has emerged, and only a few commercial applications have been developed. Two standardization efforts were made and failed.

Although the mobile agent paradigm has not entered the realm of mainstream programming, the fact is that *mobile code* and *mobile state* are now mainstream programming techniques. This has happened not as most researchers would expect it to have, namely as mobile agents, but in many different and more subtle ways. Java RMI code servers are a standard mechanism in use. Java object serialization and .NET's Remoting mechanism, which are for everyday use, offer state mobility and also code mobility in certain circumstances. Remotely upgradeable software, ActiveX-enabled pages, and other forms of code and state mobility are now so common that we do not even think about them. Even mainstream undergraduate books like Coulouris, Dollimore, and Kindberg (2000) and Tanenbaum and Steen (2002) discuss code and state mobility, and even mobile agents, without any prejudice. Books that are not mobile agent-specific but are related to code mobility are available (Nelson, 1999). Mobile code and mobile state have entered the realm of distributed programming.

It is always hard and error prone to make predictions. But, quoting Kotz et. al., it is also our view that "*The future of mobile agents is not specifically as mobile agents*" (Kotz, 2002). We also do not believe that the future of mobile agents is connected to the use of agent platforms as we

know them today. That belief arises from our experiences with the JAMES and M&M systems, as presented in this chapter.

In our view, the future of mobile agents will be a progressive integration of mobile agent concepts into existing development environments. This integration will be as readily available at the API level as object serialization and remote method invocation have become. The programmer will be able to derive from a base class and with no effort have an object that is able to move between applications. That class and that object will be ordinary ones among the hundreds or thousands used in any particular application. This evolution will probably occur as the development of object serialization APIs becomes more complete, powerful, and easy to use.

ACKNOWLEDGMENT

We would like to thank to all the students and researchers that over the years worked on the JAMES and M&M systems. Also, this investigation was partially supported by the Portuguese Research Agency – FCT, through the CISUC Research Center (R&D Unit 326/97).

REFERENCES

Aglets Project Homepage (2006). Retrieved April 27, 2006, from http://sourceforge.net/projects/aglets

Arumugam, S., Helal, A., & Nalla, A. (2002). aZIMAs: Web mobile agent system. *Proceedings of 6th International Conference on Mobile Agents (MA'02)* (LNCS 2535). Barcelona, Spain: Springer-Verlag.

Beck, K. (1999). *eXtreme programming explained: Embrace change*. Addison-Wesley.

Bellavista, P., Corradi, A., & Stefanelli, C. (1999). A secure and open mobile agent programming environment. *Proceedings of the 4th International Symposium on Autonomous Decentralized Systems (ISADS'99)*, Tokyo, Japan.

Box, D. (1997). *Essential COM*. Addison-Wesley.

Cardelli, L. (1995). A language with distributed scope. *Computing Systems Journal, 8*(1).

Chess, D., Grossof, B., Harrison, C., Levine, D., Parris, C., & Tsudik, G. (1994). *Mobile Agents: Are they are good idea? (RC19887)*. IBM Research.

Conchon S., & Fessant, F. (1999). Jocaml: Mobile agents for objective-caml. *Proceedings of the Joint Symposium on Agent Systems and Applications/Mobile Agents (ASA/MA'99)*, Palm Springs, CA.

Coulouris, G., Dollimore, J., & Kindberg, T. (2000). *Distributed systems: Concepts and design* (3rd ed.). Addison-Wesley.

Delamaro, M., & Picco, G. (2002). Mobile code in .NET: A porting experience. *Proceedings of 6th International Conference on Mobile Agents (MA'02)* (LNCS 2535). Barcelona, Spain: Springer-Verlag.

Denning, A. (1997). *ActiveX controls inside out* (2nd ed.). Redmond, WA: Microsoft Press.

FIPA. (2000). *FIPA agent management support for mobility specification*. DC000087C. Geneva, Switzerland: Foundation for Intelligent Physical Agents.

FIPA. (2002). *FIPA abstract architecture specification*. SC00001L. Geneva, Switzerland: Foundation for Intelligent Physical Agents.

Farmer, W., Guttman, J., & Swarup, V. (1996). Security for mobile agents: Issues and requirements. *Proceedings of the 19th National Information Systems Security Conference (NISSC'96)*, Baltimore.

Glass, G. (1997). *ObjectSpace voyager core package technical overview.* ObjectSpace.

Greenberg, M., Byington, J., & Harper, D. (1998). Mobile agents and security. *IEEE Communications Magazine, 36*(7).

Gschwind, T., Feridun, M., & Pleisch, S. (1999). ADK: Building mobile agents for network and systems management from reusable components. *Proceedings of the Joint Symposium on Agent Systems and Applications/Mobile Agents (ASA/MA'99),* Palm Springs, CA.

Hohl, F. (1998). A model of attack of malicious hosts against mobile agents. In *Object-Oriented Technology, ECOOP'98 Workshop Reader / Proceedings of the 4th Workshop on Mobile Object Systems (MOS'98): Secure Internet Mobile Computations* (LNCS 1543). Brussels, Belgium: Springer-Verlag.

Kotz, D., Gray, R., Nog, S., Rus, D., Chawla, S., & Cybenko, G. (1997). AGENT TCL: Targeting the needs of mobile computers. *IEEE Internet Computing, 1*(4).

Kotz, D., Gray, R., & Rus, D. (2002). Future directions for mobile agent research. *IEEE Distributed Systems Online, 3*(8).

Lange, D., & Oshima, M. (1998). Mobile agents with Java: The Aglet API. *World Wide Web Journal,* (3).

Lange, D., & Oshima, M. (1999). Seven good reasons for mobile agents. *Communications of the ACM, 42*(3).

Loureiro, S. (2001). *Mobile code protection.* Unpublished doctoral dissertation, Institut Eurecom, ENST, Paris.

Lugmayr, W. (1999). *Gypsy: A component-oriented mobile agent system.* Unpublished doctoral dissertation, Technical University of Vienna, Austria.

Marques, P. (2003). *Component-based development of mobile agent systems.* Unpublished doctoral dissertation, Faculty of Sciences and Technology of the University of Coimbra, Portugal.

Marques, P., Fonseca, R., Simões, P., Silva, L., & Silva, J. (2002a). A component-based approach for integrating mobile agents into the existing Web infrastructure. In *Proceedings of the 2002 IEEE International Symposium on Applications and the Internet (SAINT'2002).* Nara, Japan: IEEE Press.

Marques, P., Santos, P., Silva, L., & Silva, J. G. (2002b). Supporting disconnected computing in mobile agent systems. *Proceedings of the 14th International Conference on Parallel and Distributed Computing and Systems (PDCS2002).* Cambridge, MA.

Marques, P., Simões, P., Silva, L., Boavida, F., & Gabriel, J. (2001). Providing applications with mobile agent technology. *Proceedings of the 4th IEEE International Conference on Open Architectures and Network Programming (OpenArch'01),* Anchorage, AK.

McConnell, S. (1996). *Rapid development: Taming wild software schedules.* Redmond, WA: Microsoft Press.

Milojicic, D., Chauhan, D., & la Forge, W. (1998). Mobile objects and agents (MOA), design, implementation and lessons learned. *Proceedings of the 4th USENIX Conference on Object-Oriented Technologies (COOTS'98),* Santa Fe, NM.

Nelson, J. (1999). *Programming mobile objects with Java.* John Wiley & Sons.

OMG. (2000). *Mobile agent facility, version 1.0.* Formal/00-01-02: Object Management Group.

Papaioannou, T. (2000). *On the structuring of distributed systems: The argument for mobility.* Unpublished doctoral dissertation, Loughborough University, Leicestershire, UK.

Picco, G. (1998). *Understanding, evaluating,*

formalizing, and exploiting code mobility. Unpublished doctoral dissertation, Politecnico di Torino, Italy.

Saltzer, J., Reed, D., & Clark, D. (1984). End-to-end arguments in system design. *ACM Transactions in Computer Systems, 2*(4).

Silva, L., Simões, P., Soares, G., Martins, P., Batista, V., Renato, C., et al. (1999). James: A platform of mobile agents for the management of telecommunication networks. *Proceedings of the 3rd International Workshop on Intelligent Agents for Telecommunication Applications (IATA'99)* (LNCS 1699). Stockholm, Sweden: Springer-Verlag .

Simões, P., Reis, R., Silva, L., & Boavida, F. (1999). Enabling mobile agent technology for legacy network management frameworks. *Proceedings of the 1999 International Conference on Software, Telecommunications and Computer Networks (SoftCOM1999),* FESB-Split, Split/Rijeka Croatia, Trieste/Venice, Italy.

Simões, P., Rodrigues, J., Silva, L., & Boavida, F. (2002). Distributed retrieval of management information: Is it about mobility, locality or distribution? *Proceedings of the 2002 IEEE/IFIP Network Operations and Management Symposium (NOMS2002),* Florence, Italy.

Tanenbaum, A., & Steen, M. (2002). *Distributed systems: Principles and paradigms.* Prentice Hall.

Tripathi, A., Karnik, N., Ahmed, T., Singh, R., Prakash, A., Kakani, V., et al. (2002). Design of the Ajanta system for mobile agent programming. *Journal of Systems and Software, 62*(2).

White, J. (1996). Telescript technology: Mobile agents. In J. Bradshaw (Ed.), *Software agents.* AAI/MIT Press.

Wojciechowski, P., & Sewell, P. (1999). Nomadic pict: Language and infrastructure design for mobile agents. *Proceedings of the Joint Symposium on Agent Systems and Applications/Mobile Agents (ASA/MA'99),* Palm Springs, CA.

Wooldridge, M., & Jennings, N. (1999). Software engineering with agents: Pitfalls and pratfalls. *IEEE Internet Computing, 3*(3).

This work was previously published in Architectural Design of Multi-Agent Systems: Technologies and Techniques, edited by H. Lin, pp. 95-114, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.8
Building Applications to Establish Location Awareness:
New Approaches to Design, Implementation, and Evaluation of Mobile and Ubiquitous Interfaces

D. Scott McCrickard
Virginia Polytechnic Institute and State University (Virginia Tech), USA

Miten Sampat
Feeva Technology, Inc., USA

Jason Chong Lee
Virginia Polytechnic Institute and State University (Virginia Tech), USA

ABSTRACT

An emerging challenge in the design of interfaces for mobile devices is the appropriate use of information about the location of the user. This chapter considers tradeoffs in privacy, computing power, memory capacity, and wireless signal availability that accompany the obtaining and use of location information and other contextual information in the design of interfaces. The increasing ability to integrate location knowledge in our mobile, ubiquitous applications and their accompanying tradeoffs requires that we consider their impact on the development of user interfaces, leading to an agile usability approach to design borne from agile software development and usability engineering. The chapter concludes with three development efforts that make use of location knowledge in mobile interfaces.

INTRODUCTION

A key challenge in the emerging field of ubiquitous computing is in understanding the unique user problems that new mobile, wearable, and embedded technology can address. This chapter focuses on problems related to location determination—different ways to determine location at low cost with off-the-shelf devices and emerging computing environments, and novel methods for integrating location knowledge in the design of

applications. For example, many Web sites use location knowledge from IP addresses to automatically provide the user with relevant weather and traffic information for the current location. There is significant opportunity in the use of location awareness for human-computer interaction (HCI) researchers to explore information-interaction paradigms for the uncertainty and unpredictability that is inherent to many location detection systems—particularly indoor systems that use Wifi signals which can be blocked by roofs, walls, shelves, and even people!

The prior knowledge of location to make such decisions in the presentation of information affords it to be categorized as *context awareness*, the use information that can be used to identify the situation of an entity to appropriately tailor the presentation of and interaction with information to the current situation (Dey, 2001). While context awareness can include a wide variety of information—including knowledge of who is in your surrounding area, events that are happening, and other people in your vicinity—this chapter focuses on the identification and use of location information, perhaps the most cheaply and readily available type of context information. This chapter considers the tradeoffs in privacy, computing power, memory capacity, and wireless (Wifi) signal availability in building interfaces that help users in their everyday tasks. We discuss our own SeeVT system, which uses Wifi signals in location determination (Sampat, Kumar, Prakash, & McCrickard, 2005). The SeeVT system provides the backbone for supplying location information to mobile devices on a university campus. Numerous interfaces built on SeeVT provide timely and appropriate location information to visitors in key areas of the campus.

The increasing ability to integrate location knowledge in our mobile, ubiquitous applications requires that we consider its impact on the development of user interfaces. This chapter describes the merging of agile software development methods from software engineering with the scenario-based design (SBD) methodology from usability engineering to create a rapid iteration design approach that is heavy in client feedback and significant in its level of reusability. Also presented are three interfaces developed using our Agile Usability methodology, focusing on the benefits found in using the Agile Usability approach and the tradeoffs made in establishing location awareness.

BACKGROUND

From the early days, navigation has been central to progress. Explorers who set sail to explore the oceans relied on measurements with respect to the positions of celestial bodies. Mathematical and astronomical techniques were used to locate oneself with respect to relatively stationery objects. The use of radio signals proved to be fairly robust and more accurate, leading to the development of one of the first modern methods of navigation during World War II, called long range navigation (LORAN). LORAN laid the foundation of what we know as the Global Positioning System or GPS (Pace et al., 1995). Primarily commissioned by the United States Department of Defense for military purposes, GPS relies on 24 satellites that revolve around the Earth to provide precision location information in three dimensions. By relying on signals simultaneously received by four satellites, GPS provides much higher precision than previous techniques. GPS navigation is used in a wide range of applications from in-car navigation, to geographic information system (GIS)-mapping, to GPS-guided bombs.

GPS has become the standard for outdoor location-awareness as it provides feedback in a familiar measurement metric. Information systems like in-car navigators have adopted GPS as the standard for obtaining location, since it requires little or no additional infrastructure deployments and operates worldwide. However, GPS has great difficulty in predicting location

in dense urban areas, and indoors, as the signals can be lost when they travel through buildings and other such structures. With an accuracy of about 100 meters (Pace et al., 1995), using GPS for indoor location determination does not carry much value. Along with poor lateral accuracy, GPS cannot make altitude distinctions of three to four meters—the average height of a story in a building—thus making it hard to determine, for example, whether a device is on the first floor or on the second floor. Despite continued progress through technological enhancements, GPS has not yet evolved sufficiently to accommodate the consumer information-technology space. This chapter primarily focuses on technologies making inroads for indoor location determination.

While GPS has clear advantages in outdoor location determination, there have been other efforts focused around the use of sensors and sensing equipment to determine location within buildings and in urban areas. Active Badges was one of the earliest efforts at indoor location determination (Want, Hopper, Falcao, & Gibbons, 1992). Active Badges rely on users carrying badges which actively emit infrared signals that are then picked up by a network of embedded sensors in and around the building. Despite concerns about badge size and sensor deployment costs, this and other early efforts inspired designers to think about the possibilities of information systems that could utilize location-information to infer the context of the user, or simply the context of use. One notable related project is MIT's Cricket location system, which involved easy-to-install motes that acted as beepers instead of as a sensor network (Priyantha, Chakraborty, & Balakrishnan, 2000). The user device would identify location based on the signals received from the motes rather than requiring a broadcast from a personal device. Cricket was meant to be easy to deploy, pervasive, and privacy observant. However, solutions like Cricket require deployment of a dense sensor network—reasonable for some situations, but lacking the ubiquity necessary to be an inexpensive, widely available, easy-to-implement solution.

To provide a more ubiquitous solution, it is necessary to consider the use of existing signals—many of which were created for other purposes but can be used to determine location and context. For example, mobile phone towers, IEEE 802.11 wireless access points (Wifi), and fixed Bluetooth devices (as well as the previously mentioned GPS) all broadcast signals that have identification information associated with them. By using that information, combined with the same sort of triangulation algorithms used with GPS, the location of a device can be estimated. The accuracy of the estimation is relative to the number and strength of the signals that are detected, and since one would expect that more "interesting" places would have more signals, accuracy would be greatest at these places—hence providing best accuracy at the most important places. Place Lab is perhaps the most widespread solution that embraces the use of pre-existing signals to obtain location information (LaMarca et al., 2005). Using the broadcasted signals discussed previously, Place Lab allows the designer to establish location information indoors or outdoors, with the initiative of allowing the user community to contribute to the overall effort by collecting radio environment signatures from around the world to build a central repository of signal vectors. Any client device using Place Lab can download and share the signal vectors for its relevant geography—requiring little or no infrastructure deployment. Place Lab provides a location awareness accuracy of approximately 20 meters.

Our work focuses specifically on the use of Wifi access networks, seeking to categorize the benefits according to the level of access and the amount of information available in the physical space. We propose three categories of indoor location determination techniques: *sniffing* of signals in the environment, *Web-services access* to obtain information specific to the area, and *smart algorithms* that take advantage of other

information available on mobile devices. In the remainder of this chapter, we describe these techniques in more detail, and we discuss how these techniques have been implemented and used in our framework, called SeeVT using our Agile Usability development process.

CATEGORIZING INDOOR LOCATION DETERMINATION TECHNIQUES

When analyzing location awareness, it is clear that the goal is not just to obtain the location itself, but information associated with the location—eventually leading to full context awareness to include people and events in the space, as described in Dey (2001). For example, indoor location awareness attributes such as the name of the building, the floor, surrounding environments, and other specific information attributed with the space are of particular interest to designers. Designers of systems intended to support location awareness benefit not only from location accuracy, but also from the metadata (tailored to the current level of location accuracy) that affords several types of cross-interpretations and interpolations of location and other context as well.

Access to this information can be stored with the program, given sufficient computing power and memory. This approach is reasonable for small areas that change infrequently—a library or a nature walk could be examples. Information about the area can be made accessible within the application with low memory requirements and rapid information lookup. However, changes to the information require updates to the data, a potentially intolerable cost for areas where location-related changes occur frequently. For example, a reconfigurable office building where the purpose and even the structure of cubicles change frequently would not be well served by a standalone application. Instead, some sort of Web-based repository of information would best meet its needs. Taking this model another step, a mobile system could request and gather information from a wide range of sources, integrating it for the user into a complete picture of the location. As an example, a university campus or networked city would benefit from a smart algorithm that integrated indoor and outdoor signals of various types to communicate a maximally complete picture of the user's location.

Of course, each added layer of access comes with additional costs as well. Simple algorithms may sense known signals from the environment (for example, GPS and wireless signals) to determine location without broadcasting presence. However, other solutions described previously might require requesting or broadcasting of information, revealing the location to a server, information source, or rogue presence—potentially resulting in serious violations of privacy and security. The remainder of this section describes the costs and benefits for three types of indoor location determination approaches: sniffing, Web services, and smart algorithms.

Sniffing

As the name suggests, sniffing algorithms sense multiple points of a broadcast environment, using the points to interpret the location of a device. The radio environment is generally comprised of one or more standard protocols that could be used to interpret location—modern environments include radio signals including Wifi, Bluetooth, microwaves, and a host of other mediums—creating interesting possibilities for location interpolation. Sniffing is also desirable because all location interpolation and calculations are performed on the client device, eliminating the need for a third-party service to perform the analysis and produce results. As mentioned previously, there are some benefits and disadvantages to this approach.

Performing the location determination on the client device eliminates the need for potentially slow information exchange over a network. This approach gives designers the flexibility they need

in order to perform quick and responsive changes to the interfaces as well as decision matrices within their applications. For example, a mobile device with a slow processor and limited memory will need a highly efficient implementation to achieve a speedy analysis. A limiting factor for this approach is the caching of previously known radio vectors. Since most analysis algorithms require a large pool of previously recorded radio-signal vectors to interpolate location, it translates into large volumes of data being precached on the client device. A partial solution for this exists already, precaching only for regions that the user is most likely to encounter or visit. Though this is not a complete solution to the resource crunch, it is a reasonable approach for certain situations with periodic updates or fetches when radio-vectors are upgraded or the system encounters an unknown location.

Herecast is an example of a system using the sniffing model (Paciga & Lutfiyya, 2005). It maintains a central database of known radio vectors, which are then published to client devices on a periodic basis. The clients are programmed to cache only a few known locations that the user has encountered, and relies largely on user participation to enter accurate location information when they enter new areas that the system has not encountered before. The accuracy for these systems is generally acceptable, but there is always the worry of not having a cache of an area that the application is about to encounter. The lack of linking to a service also means that other contextual information associated with the location is hard to integrate with this approach due to device caching constraints and metadata volatility.

Web-Services Model

Keeping with the fundamental idea of mobile devices facing a resource crunch, this approach has client devices and applications use a central service for location determination. This means that the client device simply measures or "sees" the radio environment and reports it to the central service. The service then performs the necessary computation to interpolate the user location (potentially including other timely information) and communicates it back to the client. This also allows the client to store a minimal amount of data locally and to perform only the simplest of operations—important for mobile devices that often trade off their small size for minimal resources.

The approach is elegant in many ways, but faces several challenges in its simplistic approach such as the problem of network latency leading to lengthy times to perform the transactions. However, as the speed and pervasiveness of mobile networks is on the rise, as is the capability of silicon integration technologies for mobile platforms, designing large-scale centralized systems based on the Web-services model will be a reasonable approach for many situations. Mobile online applications such as Friend Finders and child tracking services for parents are classic examples of tools that require central services to allow beneficial functionality to the end user.

Our own SeeVT system uses the Web services model by allowing its clients to perform Simple Object Access Protocol (SOAP) Web service calls to a standard Web interface, and submit a radio vector for analysis. It then performs the necessary location determination using a probabilistic algorithm and returns the location to the client. SeeVT provides the interface designer access to functionality on the service end as well. It allows the designers to control sessions and monitor the progress of clients by using a logging feature, and provides handles to integrate other widgets as well. For example, if an application wants to perform a search based on the user's current location, SeeVT allows the designer to add functionality to its modules to perform further server-side computations.

Smart Algorithms

Looking ahead, algorithms that span large and diverse geographic areas will require the integration of many signals, information requests, and additional inputs. Place Lab attempts to address this issue for all radio signals (LaMarca et al., 2005). Currently it can compute location using mobile phone tower signals, Wifi, fixed Bluetooth devices, and GPS. However, we expect that other information will be used for location determination in the near future. For example, the ARDEX project at Virginia Tech seeks to use cameras—quickly becoming commonplace on mobile devices—to create a real-time fiducially-based system for location determination based on augmented reality algorithms (Jacobs, Velez, & Gabbard, 2006). The goal of the system is to integrate it with SeeVT such that anyone at defined hot spots can take a picture of their surrounding area and obtain information about their location. In an interesting twist on this approach, the GumSpots positioning system allows users to take a picture of the gum spots on the ground in urban areas and performs image recognition on them to return user location (Kaufman & Sears, 2006). Other information recording devices could be used in similar ways to help determine or enhance the understanding of our current location.

BUILDING INTERFACES FOR LOCATION-KNOWLEDGEABLE DEVICES

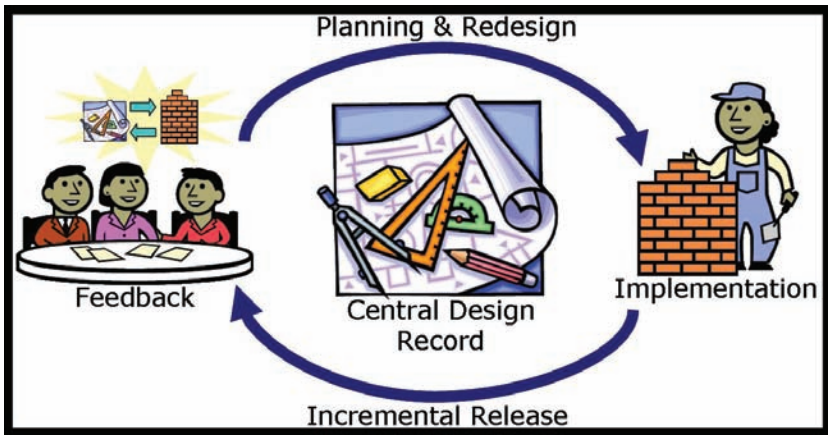
This section begins with a discussion on possible application scenarios that can leverage location knowledge in mobile devices. This section first describes *Agile Usability*, an extension of agile software development methodologies to include key aspects of usability engineering—resulting in an interface building technique that is well suited to ubiquitous and location-knowledgeable computing devices, both from the standpoint of interaction as well as development processes. Next, three case studies illustrate real world applications that have been built using these processes. Each case study describes key aspects of the application, illustrating one of the indoor location determination techniques and highlighting key lessons learned from the use of Agile Usability.

Agile Software Development, Usability Engineering, and Agile Usability

Ubiquitous and pervasive systems are often introduced to augment and support everyday tasks in novel ways using newly developed technology or by using existing technology in different ways. Since end-user needs are often ill-defined for ubiquitous systems, development needs to quickly incorporate stakeholder feedback so the systems can be iteratively improved to address new and changing requirements. This section discusses the use of an agile development methodology to build ubiquitous systems. Based on our own work (Lee et al., 2004; Lee, Chewar, & McCrickard, 2005; Lee & McCrickard, 2007) and on prior investigation of agile development methods (Beck, 1999; Constantine, 2001; Koch, 2004), we present a usability engineering approach for the construction of interfaces for mobile and ubiquitous devices.

Agile software development methodologies have been developed to address continuous requirements and system changes that can occur during the development process. They focus on quick delivery of working software, incremental releases, team communication, collaboration, nd the ability to respond to change (Beck et al., 2001). One stated benefit of agile methods is a flattening of the cost of change curve throughout the development process. This makes agile methods ideally suited to handle the iterative and incremental development process needed to effectively

Figure 1. The agile usability process. The central design record bridges interface design with implementation issues. This enables incremental improvement incorporating feedback from project stakeholders and usability evaluations.

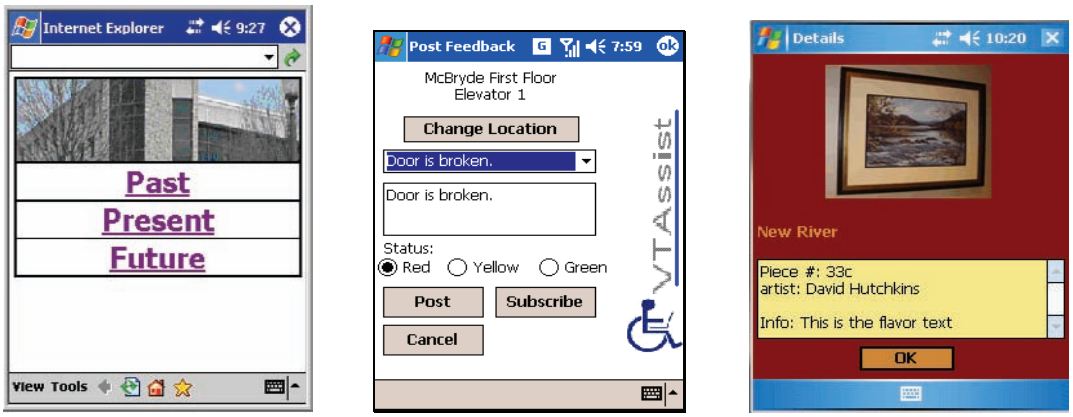


engineer ubiquitous systems. One shortcoming of many agile methods is a lack of consideration for the needs of end users (Constantine, 2001). Current agile development methodologies have on-site clients to help guide the development process and ensure that all required functionality is included. However, many ubiquitous and pervasive systems require continuous usability evaluations involving end-users to ensure that such systems adequately address their needs and explore how they are incorporated in people's daily tasks and affect their behavior. Researchers, including Miller (2005), Constantine (2001), and Beyer, Holtzblatt, and Baker (2004) have developed ways to integrate system and software engineering with usability engineering. We present our approach to agile usability engineering, henceforth referred to as Agile Usability, with the added benefit of usability knowledge capture and reuse.

Our approach combines the software development practices of extreme programming (XP) with the interaction design practices of scenario-based design (SBD) (Beck, 1999; Rosson & Carroll, 2002). The key features of this process are an incremental development process supported by continual light-weight usability evaluations, close contact with project stakeholders, an agile interface architecture model, known as a central design record (CDR), that bridges interface design and system implementation issues, and proactive knowledge capture and reuse of interface design knowledge (Figure 1).

Running a large-scale requirements analysis process for developing ubiquitous systems is not as beneficial as when designing other types of systems as it can be very difficult to envision how a ubiquitous system will be used in a specific situation or how the introduction of that system will affect how people behave or use it. In this type of development process, portions of the system are developed and evaluated by end users on a continual basis. This helps developers in uncovering new requirements and dealing with changing user needs as development proceeds. This type of development process requires some amount of discipline and rigor in terms of the types of development practices to follow. Specific details of these XP programming practices are detailed

Figure 2. Screenshots from three applications built on SeeVT. From the left, the alumni tour guide, VTAssist, and SeeVT-ART



in Beck's book on the subject (Beck, 1999). Our use of these practices are elaborated in a technical report (Lee, Wahid, McCrickard, Chewar & Congleton, 2007).

An incremental development process necessitates close collaboration with customers and end users to provide guidance on what features are needed and whether the system is usable. Ideally, representatives from these groups will be onsite with the developers working in the same team. Our customers were not strictly on site, although they were in the same general location as the developers. Regularly scheduled meetings and continual contact through e-mail and IM were essential to maintaining project velocity.

The key design representation is the central design record (CDR), which draws on and makes connections between design artifacts from XP and SBD. Stories that describe individual system features are developed and maintained by the customer with the help of the developers. They are prioritized by the customer and developed incrementally in that order. These include all features needed to develop the system including underlying infrastructure such as databases, networking software, or hardware drivers. Scenarios, which are narratives describing the system in use, are used to communicate interface design features and behaviors between project stakeholders. Claims, which describe the positive and negative psychological effects of interface features in a design, are developed from the scenarios to highlight critical interaction design features. Story identification and development may lead to changes to the scenarios and claims. The reverse may also be true. This coupling between interface design and system implementation is critical for ubiquitous systems as developers must deal with both interactional and technological issues when deploying a system to the population.

In addition to acting as a communication point between stakeholders and highlighting connections between interface design and implementation, the CDR is important as a record of design decisions. As developers iterate on their designs, they often need to revisit previous design decisions. The explicit tradeoffs highlighted in the claims can be used by developers and clients to determine how best to resolve design issues that come up. Perhaps most important, Agile Usability drives developers to explore key development techniques in the development of location-based

interfaces—the techniques used advance the field and can be reused in other situations.

Agile Usability has been applied in numerous situations, three of which are highlighted in this chapter as case studies. Each case study describes how the user tasks were identified, how stakeholder feedback was included, how our agile methodology was employed, and how appropriate location detection technologies were integrated. The discussion portion of this section will compare and contrast the lessons learned in the different case studies—highlighting specific usability engineering lessons and advancements that can be used by others.

Case Study 1: Alumni Tour Guide

The alumni tour guide application was built for visitors to the Virginia Tech (VT) campus. The system notifies users about points-of-interest in the vicinity as and when they move about the VT campus (Nair et al., 2006). This image-intensive system provides easy-to-understand views of the prior and current layout of buildings in the current area. By focusing on an almost exclusively image-based presentation, users spend little time reading text and more time reflecting on their surroundings and reminiscing about past times in the area. See Figure 2 for a screenshot of the guide.

The earliest prototypes of the tour guide proposed a complex set of operations, but task analyses and client discussions performed in the Agile Usability stages indicated that many alumni—particularly those less familiar with handheld and mobile technology—would be unlikely to want to seek out solutions using the technology. Instead, later prototypes and the final product focused on the presentation and contrast of historical and modern images of the current user location. For example, alumni can use the tour guide to note how an area that once housed some administrative offices in old homes has been rebuilt as a multistory technology center for the campus. This pictorial comparison, available at any time with only a few clicks, was well received by our client as an important step in connecting the campus of the past with the exciting innovations of the present and future.

As the target users are alumni returning to campus, most are without access to the wireless network, and the logistics are significant in providing access to the thousands of people who return for reunions, sporting events, and graduations. As such, the Alumni Tour Guide uses the sniffing location detection method to identify current location. This method fits well with the nature of the tasks of interest to alumni: they care most about the general space usage and the historical perspectives of a location that change little over time.

Case Study 2: VTAssist

Building interfaces is often difficult when the target audience has needs and skills different than those of the developer: for example, users with mobility impairments. It often takes many iterations to focus on the most appropriate solutions—a perfect candidate for Agile Usability. A pair of developers used our methodology to build VTAssist, a location-aware application to enable users with mobility impairments, specifically users in wheelchairs, to navigate a campus environment (Bhatia, Dahn, Lee, Sampat, & McCrickard, 2006). VTAssist helps people in wheelchairs navigate in an environment more conducive to those who are not restricted in movement. Unlike typical handhelds and Tablet PC applications (the two platforms for which VTAssist was created), the VTAssist system must attract the user's attention at times of need or danger, guide them to alternate paths, and provide them with a means to obtain personal assistance when necessary. Perhaps most importantly, VTAssist allows users to quickly and easily supply feedback on issues and difficulties at their current location—both helping future visitors and building a sense of community

among those who traverse the campus. See Figure 2 for a screenshot of the VTAssist.

In developing VTAssist using Agile Usability, we found that needs and requirements changed over time requiring that the methodology account for those changes. For example, the original design was intended to help wheelchair users find location accessible resources and locations, but later the need was identified to keep that information constantly updated, resulting in the addition of the collaborative feedback feature. It was this feature that was deemed most important to the system—the feature that would keep the information in VTAssist current, and would enable users to take an active role in maintaining the information, helping others, and helping themselves.

Due to the importance of the feedback feature in maintaining up-to-date information for those in wheelchairs, VTAssist uses the Web services model. Certainly it would be possible to obtain some benefit from the sniffing model, but the client reaction indicated the importance of user feedback in maintaining an accurate database of problems and in providing feedback channels to frustrated users looking for an outlet for their comments. In addition, the server-side computations of location and location information (including comments from users and from facility administrators) results in faster, more up-to-date reports about the facilities.

Case Study 3: Conference Center Guide

The conference center guide, known as SeeVT-ART, addresses the desires of visitors and alumni to our area in coming to, and generally in returning to, our university campus—specifically the campus alumni and conference center (Kelly, Hood, Lee, Sampat, & McCrickard, 2006). SeeVT-ART provides multimodal information through images, text, and audio descriptions of the artwork featured in the center. Users can obtain alerts about interesting regional and university-specific features within the center and they can be guided to related art by the same artist or on the same topic. The alerts were designed to be minimally intrusive, allowing users to obtain more information if they desired it or to maintain their traversal through the center if preferred. See Figure 2 for a screenshot of SeeVT-ART.

Agile Usability was particularly effective in this situation because of the large amount of input from the client, who generated a lot of ideas that, given unlimited time and resources, would have contributed to the interface. Agile Usability forced the developers to prioritize—addressing the most important changes first while creating placeholders illustrating where additional functionality would be added. Prioritization of changes through Agile Usability also highlighted the technological limitations of the underlying SeeVT system, specifically those related to the low accuracy of location detection, and how that influenced the system design. For example, when a user enters certain areas densely populated with artistically interesting objects, SeeVT-ART requires the user to select from a list of the art pieces, as it is impossible to determine with accurate precision where the user is standing or (with any precision) what direction the user is facing. These limitations suggested the need for smart algorithms that use information about the area and that integrate additional location determination methods.

Smart algorithms that store location data over time and use it to improve location detection can be useful in determining data such as the speed at which a user is walking and the direction a user is facing. SeeVT-ART can use this data to identify the piece of art at which a user most likely is looking. Our ongoing work is looking at integrating not only the widely accessible broadcast signals from GPS, cellular technology, and fixed Bluetooth, but also RFID, vision algorithms, and augmented reality (AR) solutions. Our early investigation into a camera-based AR solution combines information about the current location with image processing by a camera mounted on the handheld to identify

the artwork and augment the user's understanding of it with information about the artist, provenance, and so forth. These types of solutions promise a richer and more complete understanding of the importance of a location than any one method could accomplish alone.

CONCLUSION AND FUTURE DIRECTIONS

The three location-knowledgeable SeeVT applications described in this document offer a glimpse into the possibilities for location-knowledgeable mobile devices. The increasing presence of wireless networks, improvements in the power and utility of GPS, and development of other technologies that can be used to determine location portends the ubiquity of location-knowledgeable applications in the not-too-distant future. Delivery of location-appropriate information in a timely and useful manner with minimal unwanted interruption will be the goal of such systems. Our ongoing development efforts seek to meet this goal.

In support of our development efforts, we explore new usability engineering approaches particularly appropriate for location-knowledgeable applications. The use of stories and the knowledge capturing structures of Agile Usability combined with its rapid multiple iterations enable convergence on solutions to the most important issues faced by emerging application areas. We repeatedly found that designers are able to identify issues of importance to the target users, while keeping in perspective the design as a whole. Our ongoing work seeks ways to capture and share the knowledge produced from designing these applications not only within a given design but across designs, leading to the systematic scientific advancement of the field.

In the future, these developing Agile Usability techniques will be supported by specific tools and toolkits for leveraging the location-awareness needs of on-the-go users. An early contribution that can be drawn from this work is the novel methods for supporting location awareness in users—browseable historical images of the current location, rapid feedback methods for reporting problems, new map presentation techniques—all methods that should be captured in a toolkit and reused in other location awareness situations.

REFERENCES

Bahl, P., & Padmanabhan, V. N. (2000). RADAR: An in-building RF-based user location and tracking system. In *Proceedings of IEEE INFOCOM,* Tel Aviv, Israel, (Vol. 2, pp. 775-784).

Beck, K. (1999, October). Embracing change with extreme programming. *IEEE Computer, 32*(10), 70-77.

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M. et al. (2001). *The Agile Manifesto.* Retrieved January 25, 2008, from http://agilemanifesto.org

Beyer, H. R., Holtzblatt, K., & Baker, L. (2004). An agile customer-centered method: Rapid contextual design. In *Proceedings of Extreme Programming and Agile Methods 2004 (XP/Agile Universe),* Calgary, Canada (pp. 50-59).

Bhatia, S., Dahn, C., Lee, J. C., Sampat, M., & McCrickard, D. S. (2006). VTAssist-a location-based feedback notification system for the disabled. In *Proceedings of the ACM Southeast Conference (ACMSE '06),* Melbourne, FL (pp. 512-517).

Constantine, L. L. (2001). Process agility and software usability: Toward lightweight usage-centered design. *Information Age, 8*(2). In L. Constantine (Ed.), *Beyond chaos: The expert edge in managing software development.* Boston: Addison-Wesley.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing, 5*(1), 4-7.

Jacobs, J., Velez, M, & Gabbard, J. (2006). ARDEX: An integrated framework for handheld augmented reality. In *Proceedings of the First Annual Virginia Tech Center for Human-Computer Interaction Research Experience for Undergrads Symposium*, Blacksburg, VA (p. 6).

Kaufman, J., & Sears, J. (2006). GSPS: GumSpots positioning system. In *IPT 2006 Spring Show.* Retrieved January 25, 2008, from http://itp.nyu.edu/show/detail.php?project_id=539

Kelly, S., Hood, B., Lee, J. C., Sampat, M., & McCrickard, D. S. (2006). *Enabling opportunistic navigation through location-aware notification systems.* Pending paper submission.

Koch, A. S. (2004). *Agile software development: Evaluating the methods for your organization.* Artech House Publishers.

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., et al. (2005). Place lab: Device positioning using radio beacons in the wild. In *Proceedings of the 3rd International Conference on Pervasive Computing (Pervasive 2005),* Munich, Germany (pp. 134-151).

Lee, J. C., Chewar, C. M., & McCrickard, D. S. (2005). Image is everything: Advancing HCI knowledge and interface design using the system image. In *Proceedings of the ACM Southeast Conference (ACMSE '05),* Kennesaw, GA (pp. 2-376-2-381).

Lee, J. C., Lin, S., Chewar, C. M., McCrickard, D. S., Fabian, A., & Jackson, A. (2004). From chaos to cooperation: Teaching analytic evaluation with LINK-UP. In *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn '04),* Washington, D.C. (pp. 2755-2762).

Lee, J. C., & McCrickard, D. S. (2007). Towards extreme(ly) usable software: Exploring tensions between usability and agile software development. In *Proceedings of the 2007 Conference on Agile Software Development (Agile '07),* Washington DC, (pp. 59-70).

Lee, J. C., Wahid, S., McCrickard, D. S., Chewar, C. M., & Congleton, B. (2007). Understanding Usability: Investigating an Integrated Design Environment and Management System. *International Journal of Information Technology and Smart Education (ITSE), 2*(3), 161-175.

Miller, L. (2005). Case study of customer input for a successful product. In *Proceedings of the Agile 2005 Conference,* Denver, CO (pp. 225-234).

Nair, S., Kumar, A., Sampat, M., Lee, J. C., & McCrickard, D. S. (2006). Alumni campus tour: Capturing the fourth dimension in location based notification systems. In *Proceedings of the ACM Southeast Conference (ACMSE '06),* Melbourne, FL (pp. 500-505).

Pace, S., Frost, G. P., Lachow, I., Frelinger, D., Fossum, D., Wassem, D. et al. (1995). *The global positioning system: Assessing national policies* (Ref. No. MR-614-OSTP). Rand Corporation.

Paciga, M., & Lutfiyya, H. (2005). Herecast: An open infrastructure for location-based services using WiFi. In *Proceedings of Wireless and Mobile Computing, Networking, and Communications (WiMoB 2005),* Montreal, Canada (pp. 21-28).

Priyantha, N. B., Chakraborty, A., & Balakrishnan, H. (2000). The cricket location-support system. In *Proceedings of the Sixth Annual International Conference on Mobile Computing and Networking (MOBICOM 2000),* Boston, MA (32-43).

Rosson, M .B., & Carroll, J. M. (2002). *Usability engineering: Scenario-based development of human-computer interaction.* New York: Morgan Kaufman.

Sampat, M., Kumar, A., Prakash, A., & McCrickard, D. S. (2005). Increasing understanding of a new environment using location-based notification systems. In *Proceedings of 11th Internation-*

alConference on Human-Computer Interaction (HCII '05). Las Vegas, NV.

Sciacchitano, B., Cerwinski, C., Brown, I., Sampat, M., Lee, J. C., & McCrickard, D. S. (2006). Intelligent library navigation using location-aware systems. In *Proceedings of the ACM Southeast Conference (ACMSE '06),* Melbourne, FL (pp. 371-376).

Tom, H. (1994). The geographic information systems (GIS) standards infrastructure. *StandardView, 2*(3), 33-142.

Want, R., Hopper, A., Falcao, V., & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems, 40*(1), 91-102.

Youssef, M., & Agrawala, A. K. (2004). Handling samples correlation in the Horus system. In *Proceedings of IEEE INFOCOM*. Hong Kong, China.

KEY TERMS

Agile Usability: Design methodologies that incorporate practices from agile software development methods and usability engineering methods to enable the efficient development of usable software.

Extreme Programming: An agile software development methodology centered on the values of simplicity, communication, feedback, courage, and respect.

Location Awareness: Functionality in mobile devices that allows them to calculate their current geographic location.

Mobile Devices: Handheld, portable computing devices such as smart phones and personal digital assistants.

Scenario-Based Design: Usability engineering methodology that uses descriptions of how people accomplish tasks—scenarios—as the primary design representation to drive the development and analysis of systems.

SeeVT: Location aware system that uses Wifi signals to calculate the position of wireless-enabled mobile devices.

Ubiquitous Computing: Technology embedded in the environment that becomes implicit and tightly integrated into peoples' day to day tasks.

Wifi: Wireless local area networking technology and standards developed to improve the interoperatility of wireless communication devices.

This work was previously published in Ubiquitous Computing: Design, Implementation, and Usability, edited by Y. Theng and H. Duh, pp. 253-265, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.9
From Ethnography to Interface Design

Jeni Paay
Aalborg University, Denmark

Benjamin E. Erlandson
Arizona State University, USA

ABSTRACT

This chapter proposes a way of informing creative design of mobile information systems by acknowledging the value of ethnography in HCI and tackling the challenge of transferring that knowledge to interface design. The proposed approach bridges the gap between ethnography and interface design by introducing the activities of field-data informed design sketching, on a high level of abstraction, followed by iterative development of paper-based mock-ups. The outcomes of these two activities can then be used as a starting point for iterative prototype development—in paper or in code. This is particularly useful in situations where mobile HCI designers are faced with challenges of innovation rather than solving well-defined problems and where design must facilitate future rather than current practice. The use of this approach is illustrated through a design case study of a context-aware mobile information system facilitating people socialising in the city.

INTRODUCTION

This chapter looks at the mobile technology design problem of taking an ethnographic-based approach to gathering field data and making this data available to the design process in a form that is easily assimilated by designers to inform user-centred design of mobile technology. Interface design for mobile technologies presents unique and difficult challenges that sometimes render traditional systems design methods inadequate. Ethnography is particularly well-suited to design for mobile technology. Mobile usability is often highly contextual and ethnographic approaches can facilitate richer understandings of mobile use contexts providing insight into the user's perspective of the world. Exploring the huge po-

tential of mobile devices presents designers with a unique opportunity for creativity. In thinking about mobile technology design for *future*, rather than *current* practice, the challenge becomes even greater.

Before this discussion proceeds further it is worth clarifying the use of the term *ethnography*. Traditionally, ethnographic studies within sociology are conducted from a particular theoretical viewpoint and for the purpose of contributing to theory. However, ethnography, as it is understood in HCI research, generally refers to a collection of techniques used for gathering and organizing field materials from observational studies (Dourish, 2006). By its very definition, ethnography is primarily a form of reportage. It provides both empirical observational data, and makes an analytical contribution in the organization of that data. The virtue of ethnography is that it takes place in real-world settings and provides access to the ways people perceive, understand, and do things (Hughes et al., 1997). Ethnographically-oriented field methods can be used in HCI to provide a deeper understanding of an application domain, a holistic understanding of users, their work, and their context, which can then be drawn into the design process at the earliest stages (Millen, 2000). Ethnographic studies involve detailed observations of activities within their natural setting, providing rich descriptions of people, environments and interactions, and acknowledging the situated character of technology use (Millen, 2000). These observations can provide valuable insights into the processes needed for systems requirements specifications (Sommerville et al., 1993).

In the literature, the terms ethnography and ethnomethodology are both used to refer to field studies using ethnographic methods to understand how people perceive their social worlds. Other terms such as technomethodology (Button & Dourish, 1996), rapid ethnography (Millen, 2000) and design ethnography (Diggins & Tolmie, 2003) are also used to distinguish different aspects of the use of ethnography in the design of technology. For the sake of simplicity, this chapter uses the term ethnography to encompass these understandings as being important to the discussion of the relationship between their outputs and the inputs they provide to the design process.

For ethnography to make a worthwhile contribution to the design of mobile technologies, we need to find ways for translating ethnographic findings into forms that are suitable for informing design processes. In the following sections, the historical relationship between ethnography and HCI is discussed, including how it has been incorporated into the process of interface design. The theoretical and methodological background for how to gather and interpret ethnographic data and use this for informing design is described. A design case study is then presented in which an ethnographic approach has been applied to mobile technology design in a real world research project through a structured series of activities. The overall process is described, and the two steps of developing *design sketches* and *paper-based mock-ups* are introduced as a way of bridging the gap between ethnography and interface design. Finally, lessons learned from using design sketches and paper-based mock-ups in the development process are outlined.

BACKGROUND

Ethnography and HCI

The issue of bridging the gap between ethnography and interface design has been a topic of discussion in HCI research for over a decade. Ethnography is now regarded as a common approach to HCI research and design (Dourish, 2006). Yet there is still no overall consensus on how best to incorporate the results of ethnographic fieldwork into the design processes (Diggins & Tolmie, 2003). In the early 90s seminal work by sociologists, such as Suchman, Hughes, Harper, Heath and Luff, inspired the use of ethnography for under-

standing the social aspects of work processes and informing user interface design (Hughes et al., 1995). However, researchers struggled with the challenge of utilizing insights provided by ethnography into the activity of designing. By the mid 90s, ethnography was hailed as a new approach to requirements elicitation for interactive system design, particularly through its application in the development of computer-supported cooperative work (CSCW) systems (Hughes et al., 1995). Even so, some researchers still held reservations about the ability of ethnographic methods to inform design (Hughes et al., 1997) and ethnography was regarded as a relatively untried approach to systems development, despite the fact that it was increasingly being used to inform and critique actual systems (Button & Dourish, 1996). Toward the end of the 90s, researchers were beginning to develop systematic approaches to social analyses for the purpose of influencing design (e.g., Viller & Sommerville, 1999). However, despite many research efforts, bridging the gap between ethnography and design still remains a matter of concern to HCI researchers today (Diggins & Tolmie, 2003).

The turn towards ethnography within HCI was motivated by a growing need to design for complex real world situations. This began with the belief that methods from the social sciences, such as ethnography, could provide means for understanding these contextual issues of technology use better. In the light of today's ubiquitous and mobile networked computing environments, the need to understand contexts of technology use, such as peoples' dynamic work and social practices, is challenging HCI researchers and designers more than ever. Supporting innovation in a world of emerging technologies can be done by submerging designers, who understand emerging technical possibilities, into rich ethnographic field data about potential users' lives and current practices (Holtzblatt, 2005). In this way technology design drives an understanding of the user's situation, which in turn, propels innovation.

Ethnography and Interface Design

The process of transition from field data to prototype design is a difficult one (Cheverst et al., 2005; Ciolfi & Bannon, 2003). A design process involving ethnography generally starts with observations and interviews collected through ethnographic methods. Key findings are then summarized and design ideas are drawn out with a set of features that can be tied back to the findings. The next step involves, "design suggestions" or "design implications," which may evolve into requirements through the development of a low-fidelity prototype. This prototype is then iterated with feedback from users and evolves into the operational system. The data collected by ethnographic methods reflects the richness of the user's situation in a way that is difficult to derive from a limited set of questions or measures as employed in traditional analysis methods (Wixon, 1995). In contrast to traditional systems analysis that looks at data, structures, and processing, ethnography is concerned with participants and interactions (Sommerville et al., 1993). This provides the designer with a rich understanding of the context of use for the artifacts that are being designed (Millen, 2000). In looking at a situation through the user's eyes rather than the designers, ethnography provides a view of the situation that is independent of design preconceptions (Hughes et al., 1997).

Ethnography has much to contribute to interface design—particularly in mobile device design due to the highly contextual nature of mobile usability and use. However, one of the main problems is finding a suitable mechanism for the transference of knowledge between these two fundamentally different disciplines. Ethnographic findings need to be understood and communicated to designers (Hughes et al., 1995). And yet, current mechanisms for incorporating ethnographic findings into the design process still

fail to capture the value of these investigations (Dourish, 2006).

Ethnography deals in "the particular," and software design in "the abstract" (Viller & Sommerville, 1999). While willing to listen to each other, both disciplines speak different languages and use different methodologies. Ethnographers deal in text, notes, reports, and transcriptions, and produce detailed results giving a rich and concrete portrayal of the particulars of everyday practical action in context, presented in a discursive form; software designers and engineers deal in the creation and manipulation of more formal graphical abstractions, notations and description techniques to simplify the complexity of the situation and extract critical features. Ethnographers avoid judgements; designers make them. Where ethnographers take an analytic role, including gathering and interpreting data, software designers have a synthesis role, designing from abstract models of situations (Button & Dourish, 1996; Hughes et al., 1995). In addition to the problems of communication there are also problems of timing. Ethnography is generally conducted over a long period of time; in fact, it is difficult to define an end point for gathering understanding. On the other hand, software designers are often under restricted time pressure to deliver a product.

The problem has been in finding a timely method and a suitable form to present field findings that can be assimilated by and are readily usable for designers (Hughes et al., 1995; Viller & Sommerville, 1999). The needs of the software designer have to be aligned with a representation of the essential "real world" practices of users in context. Simply describing the social events being observed is not sufficient, designers need to be able to model and use this understanding in design.

USING ETHNOGRAPHY IN THE DESIGN PROCESS

Gathering Data

From HCI research it can be seen that using ethnography as a data gathering method requires the development of more structured approaches to conducting and reporting from ethnographic studies that better support the development of design requirements.

One approach is to conduct ethnography concurrently with design and bring ethnographic results into the design process in a more systematic way throughout the development process. This can, for example, be achieved through meetings between ethnographers and the design team (Hughes et al., 1995). This approach results in a change in the way that ethnography is conducted. Rather than extended periods in the field, ethnographers working in cooperation with software designers to create a system design, making short and focused field studies, reporting back to designers, and often taking design questions back into the field to focus their observations and questions to users. To structure the process, the communication of fieldwork to designers can be supported by dedicated software packages (Diggins & Tolmie, 2003; Sommerville et al., 1993). In this situation, the ethnographic record becomes a joint resource with ethnographers regularly reporting their findings in an electronic form, and designers using this content to develop structured design requirements. Constructing these records in a connected manner preserves backward and forward traceability between ethnographic findings and evolving system requirements.

Another approach is to lead into the design process through *rapid ethnography* (Millen, 2000). Rapid ethnography provides the field worker with a broad understanding of the situation which can then be used to sensitize designers to the use situation rather than identifying specific design issues. It is aimed at gaining a reasonable

understanding of users and their activities in the short time available for this in a software development process. Rapid ethnography provides a more structured approach to ethnographic field studies by limiting the scope of the research focus before entering the field. It focuses time spent in the field by using key informants in the real situation and interactive observation techniques. Rapid ethnography also uses multiple observers in the field to ensure several views of the same events and to create a richer representation and understanding of the situation (Millen, 2000).

Interpreting Data

Ethnography is not simply about the collection of data in the field, it is also about reflection on and interpretation of that field data. Effective communication between ethnography and design is at the heart of the matter of bridging the gap between the two disciplines (Hughes et al., 1997). By recognizing the different natures and input and output requirements of ethnography and interface design, integration between the two disciplines can be achieved through enhancing and structuring the communication between them during the interpretation phase.

One approach to interpreting the data collected is to have a cross-discipline team participating in the fieldwork. In this situation designers go into the field with ethnographers to experience themselves how users work. They also contribute to the representation of the gathered data, shaping it into a form that is easier for designers to use (Diggins & Tolmie, 2003). Representing ethnographic findings through pictorial stories, drawings, data models, analogies and metaphors are ways to communicate field learning to cross-discipline teams (Millen, 2000). Videotapes of field observations and design documentaries play a similar role using a more designer-accessible communication mode than a written report (Raijmakers et al., 2006).

Another approach to interpretation is to have both ethnographers and designers involved in the conceptual design process. In this situation, the ethnographer is an ongoing member of the design team, providing grounded insights and interpretations into the abstracted requirements as they evolve and the design emerges. The ethnographer acts as a substitute user during the design process (Viller & Sommerville, 1999). Through their knowledge of the actual situation, they can participate in discussions with the designers, providing insights and access to instances of specific relevant situations.

A third approach is for the designer to play the part of a pseudo-ethnographer. This involves designers going "into the wild" and being exposed to users by watching real work while it is being done, and hence truly experiencing the richness of work (Wixon, 1995). Structured methods such as rapid ethnography and *contextual design* (Beyer & Holtzblatt, 1998) make this possible. In contextual design, the user and the designer explore the design space together using *contextual interview* or *facilitated enactment* of their practices in context (Holtzblatt, 2005). *Affinity diagramming*, from the contextual design method, provides a synthesis of the data into hierarchical classifications where the meaning contained in the data elements can be reflected on in relation to the design question, facilitating understanding and innovation for designers.

Informing Design

After the ethnographically gathered field data has been interpreted, abstracted findings are used to derive design opportunities and design requirements. The designer uses the outputs from the interpretation of the field data as input into the design process. Sometimes the ethnographers are involved in this design process bringing their intimate knowledge of the users and the situation of use, and their deep relationship to the data, to the team (Cheverst et al., 2005). They participate

in the identification of design incentives by drawing attention to general design opportunities, and relevant topics and concerns. Otherwise, the designers must draw understanding entirely from the reports, discussions, diagrams and models, which represent the ethnographic record.

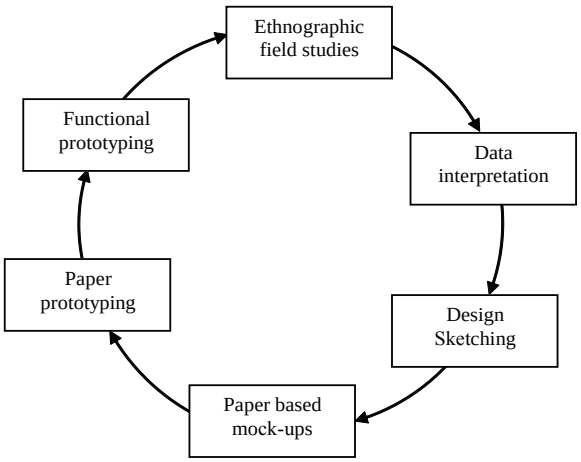
Design is a matter of making, and is used to create and give form to new ideas and new things (Fallman, 2003). A recent approach to informing design and achieving a close connection between the design team and the field data is the use of field observation videos or design documentaries. These videos mediate between ethnographic and design perspectives. As the design team watches them they incorporate interpretation of data into the design process on the fly through discussions drawing design sensitivities and identifying design concerns. Designers become sensitized to relevant issues visible in the real world interactions depicted in the video (e.g., Ciolfi & Bannon, 2003; Raijmakers et al., 2006). This method requires a high level of design experience, and in bridging the gap between ethnography and design, these designers work in an inspirational, ephemeral and creative way. For others this creative leap across the divide is very difficult, and more structured methods are needed to guide the process of envisioning design from ethnographic outputs. In response to new interface design challenges, including mobile technology, HCI researchers are investigating new techniques for guiding designers through this difficult transition – of particular interest to this chapter are the techniques of *design sketching* (Buxton, 2007), *paper-based mock-ups* (Ehn & Kyng, 1991) and *paper prototyping* (Snyder, 2003).

Design sketching is fundamental to the process of design, and can be used by information system designers to bring about the realization of an idea in the way designers think (Fallman, 2003). Sketching is the art of giving form to the unknown; it makes it possible to "see" ideas or envision whole new systems, and is especially critical in the early ideation phase of design (Buxton, 2007). According to Buxton, sketches should be rapid, timely, inexpensive, disposable, plentiful, clear, un-detailed, light, informal representations that practitioners can produce and interact with to suggest and explore ideas. Sketching is not only a way to visualize existing ideas, but it is about shaping new ideas. In making a sketch of something, the visualization talks back to the designer with a new perspective on that idea, providing a link between vision and realization of new ideas.

Paper-based mock-ups are closely related to the notion of design sketching. In this technique from the participatory design tradition, representational artifacts are constructed from paper, cardboard and materials at hand. Informed by studies of practice, mock-ups can play an important mediating role in connecting use requirements and design possibilities in a form recognizable to multi-disciplinary design teams (Ehn & Kyng, 1991). These mock-ups can be used to incorporate materials from the ethnographic study, embody envisioned new technological possibilities, convey design ideas in relation to existing practices and reveal requirements for new practices (Blomberg & Burrell, 2003).

Paper prototyping is a widely used technique for designing, testing and refining user interfaces (Snyder, 2003). This technique helps with the development of interfaces that are useful, intuitive, and efficient, by initiating testing of the interface at a stage when the design is in its formative stages and therefore still open to the input of new ideas. Paper prototyping can be used to reflect on field study findings while developing and refining the design (Holtzblatt, 2005). A collection of interface designs, drawn from ideas generated through design sketching and paper-based mock-ups are given functional and navigational connections through the process of paper prototyping. A paper prototype is a useful vehicle for giving visual form to identified design requirements. It forms the focus for design refinement discussions and cognitive walkthroughs by the design team, and

Figure 1. The overall process of designing the Just-for-Us mobile information system



is in itself part of the design specification for implementation of the system.

A DESIGN CASE STUDY

The project used as a design case study in this chapter involved the development of a context-aware mobile information system, *Just-for-Us*, designed to facilitate people socialising in the city by providing information about people, places, and activities in the user's immediate surroundings. The case study location was a specific city precinct covering an entire city block, Federation Square, Melbourne, Australia. This location was chosen because it is a new, award-winning architectural space providing a variety of activities through restaurants, cafes, bars, a museum, art galleries, cinemas, retail shops, and several public forums spanning an entire city block. The design intention for the civic space was to incorporate digital technologies into the building fabric creating a combination of virtual information space and physical building space for people to experience. Thus, this particular place provided a unique setting for studying people's situated social interactions in a "hybrid" space and for inquiring into the user experience of mobile technology designed to augment such a physical space with a digital layer.

Process

The Just-for-Us mobile information system was designed specifically for Federation Square on the basis of an ethnographic study of people socialising there. The development process involved seven major activities:

- Ethnographic field studies
- Field data interpretation
- Design sketching on a high level of abstraction
- Paper-based mock-up development
- Iterative paper prototyping
- Implementation of a functional prototype
- Field studies of prototype use in-situ

The specific content and outcome of these activities are described in the following subsections.

Figure 2. Ethnographic observations and contextual interviews at Federation Square

Details of the implemented system and findings from the field study of its use are not covered here, but can be found in Kjeldskov and Paay (2006).

As illustrated in Figure 1, data from ethnographic field studies of situated social interactions in public were subjected to data interpretation, using the *grounded theory* approach (Strauss & Corbin, 1990) and affinity diagramming (Beyer & Holtzblatt, 1998). In trying to bridge the gap between our ethnographic data and actual mobile device interface design, outcomes from the interpretation of field data were used to inform a systematic activity of design sketching (Buxton, 2007). The purpose of this activity was to generate design ideas on a high level of abstraction inspired by ethnographic findings but without getting into too much detail about specific look, feel and functionality. On the basis of selected design sketches, we developed a number of paper-based mock-ups (Ehn & Kyng, 2991) of potential design solutions. This forced us to become more specific, but still allowed us to focus on overall functionality and interaction rather than on technical details. After this, we engaged in a number of paper prototyping (Snyder, 2003) iterations with the purpose of developing a detailed set of system requirements and a coherent interface concept prior to writing any program code. Finally, these specifications were implemented in a functional prototype allowing us to introduce new technology into the field and revisit peoples' socialising behavior in the city while using the operational Just-for-Us context-aware mobile information system.

Gathering and Interpreting Data

The aim of our ethnographic field study was to inquire into peoples' social interactions at Federation Square. The field study was guided by a subset of McCullough's typology of everyday situations (McCullough, 2004) for classifying peoples' social activities when out on the town: eating, drinking, talking, gathering, cruising, belonging, shopping, and attending. The study applied a rapid ethnography approach and consisted of a series of contextual interviews (Beyer & Holtzblatt, 1998) and ethnographic field observations (Blomberg & Burrell, 2003) with the designers acting as pseudo-ethnographers and gathering the field data (Figure 2). Three different established social groups participated in the study. Each group consisted of three young urban people, mixed gender, between the ages of 20 and 35, with a shared history of socialising at Federation Square. The groups determined the activities undertaken and the social interactions that they engaged in. Prior to the field visits, each group received a 10-minute introduction to the study followed by a 20-minute interview about their socialising experiences and preferences. This introduction occurred at a place familiar to the group, where they might meet before socialising in the city. This encouraged them to reflect on past social interactions, to relax about the visit, and gave the interviewer insight into the situated interactions that the group typically participated in. One of the members of the group was then taken to Federation Square and asked to arrange to meet up with the other members of the group. The group was then asked to do what they would usually do as a group when socialising out on the town—while "thinking aloud" as they moved around the space, and responding to

Figure 3. Graphical image of inhabited social context at Federation Square

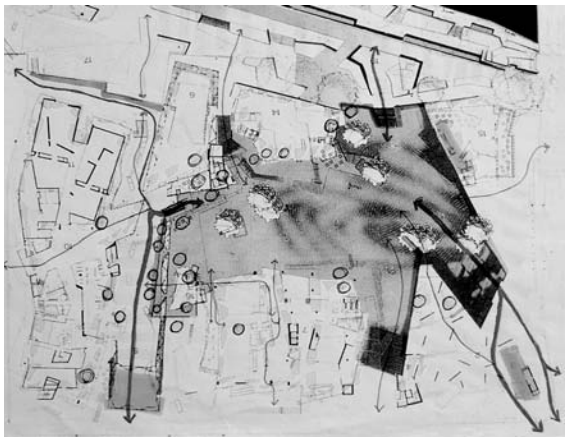
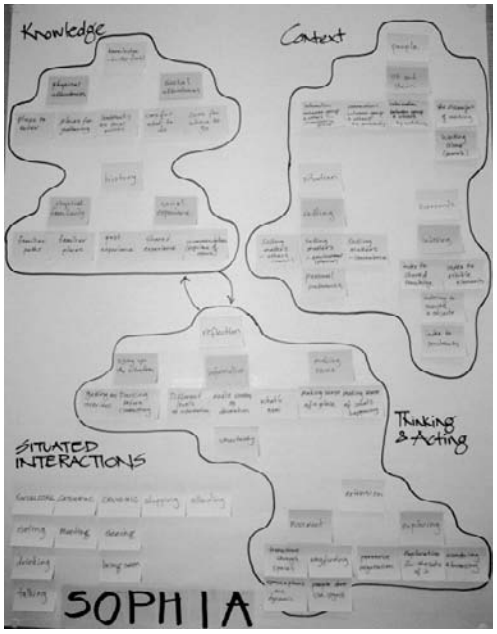


Figure 4. Affinity diagram of situated social interactions at Federation Square



questions from the interviewer. Two researchers were present in the field, providing multiple views on the data collected.

Each field visits lasted approximately three hours and allowed the groups to engage in a number of social activities. The outcome of the ethnographic field studies amounted to eight hours of video and approximately 30 pages of written notes.

In addition to the observational studies of people socialising at Federation Square an architecturally trained observer carried out a single *expert audit* (Lynch, 1960) focusing on the physical space of Federation Square. The expert audit documented architectural elements and their relationships to surrounding context, including the people inhabiting the space through 124 digital photographs and corresponding field notes.

Interpreting data gathered from the ethnographic study involved two phases. Firstly, photographic data and written notes from the expert audit were analyzed using *content analysis* (Millen, 2000) and affinity diagramming (Beyer & Holtzblatt, 1998). Concepts and themes describing the physical space of Federation Square were overlaid onto a map of the precinct to produce a color-coded multi-layered abstraction of the space (Figure 3). This provided an overview of the spatial properties of Federation Square highlighting constraints and enablers for situated social interactions there with traceable links back to specific observations.

Secondly, video data from the contextual interview and observational field study of people socialising at Federation Square was transcribed and then analyzed using open and axial coding adapted from *grounded theory* analysis (Strauss & Corbin, 1990). Identifying key words or events in the transcript, and analyzing the underlying phenomenon created the initial open codes. Analysis of these codes resulted in a collection of categories relating to actions and interactions. After the codes were grouped into categories, higher-level themes were extracted using axial

Figure 5. Design sketching informed by interpreted ethnographic field data. The delineated area corresponds to the paper-based mock-up produced later and highlighted on Figure 6.

coding. Affinity diagramming was then used to draw successively higher levels of abstraction from the data by grouping and sorting the themes until a set of high-level concepts, representing the essence of the data and encompassing all lower level themes, had been formed. The process of affinity diagramming produced a hierarchical conceptual framework containing three overall clusters of themes abstracted from the transcripts (Figure 4). This provided a rich story about how people interact with each other while socialising in public, with traceable links back to specific observations in the field study sessions.

As illustrated in Figures 3 and 4, outcomes from the interpretation of our ethnographic field data were primarily on an abstract level, providing a deeper understanding of peoples' situated social interactions in the physical space of Federation Square. While this is an important part of the foundation for good design, in their current form these outputs did not point towards any particular design ideas. As an example, the analytical outcomes from interpreting the field data included a series of qualitative statements similar to those in the following list (For a detailed account of findings from the ethnographic field studies see Paay and Kjeldskov (2005)).

- Federation Square has four key districts with distinctly different characteristics, each with an associated landmark.
- Federation Square has visible surroundings, general paths, general entrances, focal structures and no clear paths, so people need to use the structures and surrounds in finding their way around the space.
- People socialising at Federation Square like getting an overview of what is happening around them, and want to know about the presence and activities of other people.
- People's past experience with places and people at Federation Square play an important factor in choosing places and activities for socialising.
- People give directions at Federation Square by referring to shared experiences and visible elements, and use their history and physical familiarity with a place to find their way around using familiar paths.

In order to move forward from data interpretation toward an overall design concept as well as actual interface design and system requirements for a context-aware mobile information system for people socialising at Federation Square, the design team engaged in two steps of developing

Figure 6. One of the paper-based mock-ups of possible mobile device screens

design sketches and paper-based mock-ups (as described earlier). Each of these techniques produced interface design artifacts on different levels of detail and abstraction. These two "bridging" steps between ethnography and interface design are described in the following sections.

Design Sketching

The first step in the design of the Just-for-Us mobile information system was to develop a series of conceptual design ideas based on the insight from our data analysis. For this purpose, the design team spent two days generating, discussing, sketching, and refining design ideas on the basis of the abstract models of the architectural space of Federation Square and the clustering of themes in the affinity diagram from the analysis of people socialising there.

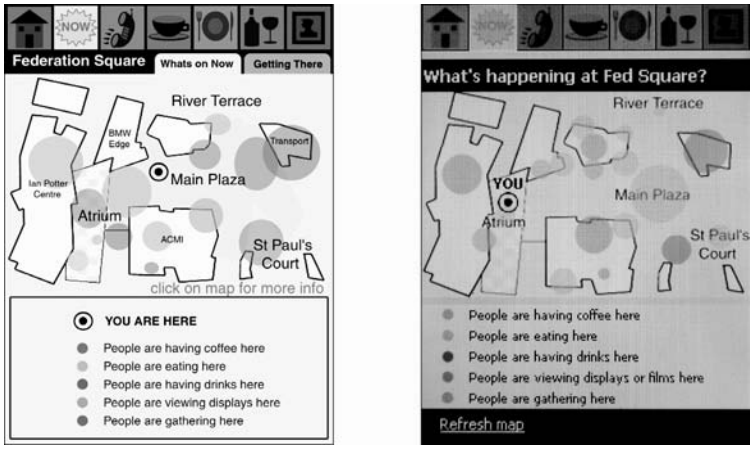
The design sketching activity was done in a dedicated design workspace with sheets of A1 paper lining the walls on which we could sketch and refine design ideas. Each sketch took its origin in a specific finding or observation from the interpreted field data. This field finding would firstly be discussed in more detail to ensure shared understanding among the design team. Secondly, we would start sketching possible design ideas, for example, how to facilitate an observed practice. Hence, we were, in a sense, using collaborative data analysis, as described in the rapid ethnography method, to drive the generation of design ideas.

During the process of sketching, the conceptual outcomes from the data interpretation phase were continually revisited and, in turn, the sketches were continuously annotated with post-it notes referring to the data. For example, a section of the affinity diagram included the themes of "social experience," encompassing "past experience" and "shared experience." A diagram was then sketched to explore the intersections between past and shared experiences in groups of friends. In this way, we ensured a strong link between data and design, and maintained clear traceability between the two. This activity was about sketching the social concepts that came out of the data models, not about generating solutions. In doing this, we were able to explore the field data findings in a graphical form, and to explore derivations from these concepts by generating multiple understandings of them. Design sketching was used as a mechanism to understand the field outcomes, to generate graphical overviews of the design space, and create graphical representations of design opportunities within that space.

The outcome from the two-day design workshop was a collection of design sketches on A1 paper (Figure 5), each describing conceptually a potential design idea or design opportunity, for parts of the Just-for-Us mobile information system, including envisioned general functionality, general ideas for graphical design and user interaction, with clear references back to the empirical data.

The design sketches provided a new visual abstraction to the ethnographically interpreted field data, translating understanding encapsulated in the abstract findings into design parlance.

Figure 7. Detailed paper prototype screen (left) and the corresponding final functional prototype screen (right), designed from the paper-based mock-up highlighted in Figure 6



Engaging in the process of design sketching rather than jumping straight to specifying system requirements, enabled us to see the ethnographic findings from a new perspective and to play with design ideas on a high level of abstraction. This allowed us to distance ourselves from the role of "problem solvers" and to explore instead, on a conceptual level, design ideas facilitating potential future practice in technology use.

Paper-Based Mock-Ups

While useful for generating and working with overall design ideas, conceptual design sketches are far too abstract for informing specific system requirements. Hence, moving directly on to detailed prototype design and implementation is likely to commit designers to specific solutions too early and impede their flexibility to try out new ideas. In an attempt to overcome this problem, the next step of our process from ethnography to interface design was to produce a series of paper-based mock-ups of possible specific design solutions (Figure 6).

The production of paper-based mock-ups took place over several days and facilitated a series of long discussions within the design team leading to an overall concept for the Just-for-Us mobile information system providing functionality such as: an augmentation of the user's physical surroundings; chat capability with friends out on the town; content indexed to the user's physical and social context and history of interactions in the city; a graphical representation of places, people and activities within the user's vicinity; and way-finding information based on indexes to landmarks and familiar places. These design ideas were screen-based solutions to design opportunities identified during design sketching.

Working with each of these ideas in more detail, the paper-based mock-ups gave the design team a medium for trying out and modifying specific design ideas for what the system should be able to do and what it should look like—long before any actual coding was done. Consequently, the mock-ups coming out of this activity had already undergone several iterations of redesign and refinements.

Discussions during the mock-up phase took place on different levels of abstraction: from

screen design, system functionality, privacy issues, problems designing for small screens, what aspects of the user's context to capture in the system, and how to do this. We also had several discussions about whether or not the implementation of the produced mock-ups would be feasible within current mobile technologies, and if not, which enabling technologies would have to be developed. Through these discussions and continued refinements and redesigns, a set of specific design requirements slowly began to take shape—gradually taking us into the "safer ground" of interface design.

Prototyping

Having completed the paper-based mock-up phase, the final steps of our development process were much more straightforward. On the basis of the mock-ups, more detailed paper prototypes were produced using Adobe Photoshop (Figure 7 left). This forced the design team to work within the graphical limitations of the target device and to use the specific graphical user interface elements available in the target browser, for this web based application. Also, the detailed paper prototypes allowed the designers to discuss some of the more dynamic interaction issues such as navigation structure and handling of pushed information. While most design changes were done in Adobe Photoshop at this time, some of the more serious issues, such as how to fit the Internet chat screen(s) into the limited design space, forced the design team back to working with paper-based mock-ups for a short time. After several cognitive walkthroughs, a full paper prototype with a detailed set of requirements was agreed upon and implemented as an operational mobile web site providing context-aware information to users, with very few modifications (Figure 7 right).

The design specified by the paper prototype was implemented as a functional Web-based system accessible through the Web browser of a PDA (personal digital assistant) providing context-related information, dynamic maps and location specific annotated graphics to the user. It also keeps a history of the user's visits to places around the city. The functional prototype uses WLAN or GPRS for wireless Internet access and resolves the user's location and the presence of friends in vicinity by means of Bluetooth beacons potentially embedded into the environment. The implementation of a functional prototype allowed us to close the circle depicted in Figure 1 by returning to Federation Square to do an ethnographic field study of people socialising there—this time facilitated by the Just-for-Us system. For details on this use study see Kjeldskov and Paay (2006).

FUTURE TRENDS

The future trends for bridging between ethnography and interface design for mobile technologies are many. As a part of a drive toward more user centered innovative design for both current and future practice, new techniques are emerging, which respond to the specific challenges of mobile technology design and use. These include, for example, cultural probes, digital ethnography, video diaries, film documentaries, facilitated enactment, acting-out in context, role-playing and body storming. Through these new techniques, the roles of ethnographers, designers, and future users are becoming more interwoven, facilitating a smoother and more effortless transition from ethnography to interface design. Techniques such as these reflect the fact that mobile technology design is not only about designing for existing work practices but also about designing for future practices in peoples' private and social lives and responding to the challenge of innovating for non-work in as yet non-existing use situations. They also respond to issues raised by many researchers that mobile technologies are often used in dynamic and continually changing contexts, offering information directly related to those contexts, and that it can be very difficult

to predict what future user-adaptations of mobile technology might evolve.

The techniques of sketching and mocking-up introduced in this chapter are not new. Both have a long tradition in other design disciplines. However, like many of the above emerging approaches, we have combined existing techniques in a new way that provides designers with a more structured path to follow when making the difficult transition of transferring knowledge from the field into the design process.

CONCLUSION

This chapter addresses the issue of ethnography informing interface design for mobile technologies. It has described how ethnographic studies can be used in HCI design and how such studies can be useful for understanding current practice as well as providing a backdrop for envisioning potential future practice. However, as confirmed in the literature, bridging between ethnography and design is difficult, and techniques are needed that enable designers to better use ethnographic findings in the design process. In response to this, the two steps of conceptual design sketching and creating paper-based mock-ups have been proposed as bridging activities between ethnographic data interpretation and iterative prototype development.

Illustrating how this can be done in practice, this chapter has described a recent project involving the design of a context-aware mobile information system on the basis of a rapid ethnographic field study. In this project, the process of design sketching from analytical data made a useful link between interpretation and design. It provided a means of communicating a conceptual understanding of current practice into the early stages of interface design, and helped "translate" findings from the field data into design parlance. Working with sketches allowed the design team to play with design ideas on a conceptual level rather than moving straight to specifying system requirements. It also allowed them to distance themselves from the role of "problem solvers" and to explore instead potential future practice of technology use.

The process of creating and refining paper-based mock-ups on the basis of selected design sketches gave the design team a medium for being a bit more specific while still maintaining a high level of flexibility. It allowed for drilling down into some specific design ideas and the exploration and modification of ideas for interface design and functionality before doing any coding. It also allowed the team to engage in discussions about possible screen designs, different functionality, privacy, small screens, etc., and to rapidly implement, evaluate, and refine design ideas. By working with paper-based mock-ups, it was possible to generate a strong set of specific design requirements, which provided a solid foundation for subsequent activities of paper and functional prototyping.

Innovative interface design for mobile technologies is both an art and a science. It requires us to be creative and inspired as well as structured and focused. Facilitating creativity and inspiration provides the art. Grounding interface design in empirically informed understanding of people and current practice provides the science. The challenge we are faced with is not just how to perform the art and science of design better individually, but more so how to support a fruitful interplay between the two. For this purpose, techniques such as conceptual design sketching and creation of paper-based mock-ups are valuable tools for researchers and designers on their journey from ethnography to interface design.

REFERENCES

Beyer, H., & Holtzblatt, K. (1998). *Contextual design—Defining customer centred systems*. San Francisco: Morgan Kaufmann.

Blomberg, J., & Burrell, M. (2003). An ethnographic approach to design. In J. Jacko & A. Sears (Eds.), *Handbook of human-computer interaction* (pp. 964-986). Mahwah, NJ, USA: Lawrence Erlbaum Associates Inc.

Button, G., & Dourish, P. (1996). Technomethodology: Paradoxes and possibilities. In *Proceedings of CHI 96*, (pp. 19-26). Vancouver, Canada: ACM.

Buxton, B. (2007). *Sketching user experiences: Getting the design right and the right design*. San Francisco, Morgan Kaufman Publishers.

Cheverst, K., Gibbs, M., Graham, C., Randall, D., & Rouncefield, M. (2005). Fieldwork and interdisciplinary design. *Notes for tutorial at OZCHI 2005.* Retrieved October 24, 2007, from http://www.comp.lancs.ac.uk/rouncefi/Tutout.html

Ciolfi, L. & Bannon, L. (2003). Learning from museum visits: Shaping design sensitivities. In *Proceedings of HCI International 2003* (pp. 63-67). Crete, Greece: Lawrence Erlbaum.

Diggins, T., & Tolmie, P. (2003). The 'adequate' design of ethnographic outputs for practice: some explorations of the characteristics of design resources. *Personal and Ubiquitous Computing, 7*, 147-158.

Dourish, P. (2006). Implications for Design. In *Proceedings of CHI 2006* (pp. 541-550). Montreal, Canada: ACM.

Ehn, P., & Kyng, M. (1991). Cardboard computers: Mocking-it-up or hands-on the future. In J. Greenbaum & M. Kyng (Eds.), *Design at work: Cooperative design of computer systems* (pp. 167-195). Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Publishers.

Fallman, D. (2003). Design-oriented human-computer interaction. In *Proceedings of CHI 2003* (pp. 225-232). Florida, USA: ACM.

Holtzblatt, K. (2005). Customer-centred design for mobile applications. *Personal and Ubiquitous Computing, 9,* 227-237.

Hughes, J., King, V., Rodden, T., & Andersen, H. (1995). The role of ethnography in interactive systems design. *Interactions, 2*(2), 56-65.

Hughes, J., O'Brien, J., Rodden, T., & Rouncefield, M. (1997). Designing with ethnography: A presentation framework for design. *In Proceedings of DIS '97* (pp. 147-158). Amsterdam, Holland: ACM.

Kjeldskov, J., & Paay, J. (2006). Public pervasive computing in the city: Making the invisible visible. *IEEE Computer, 39*(9), 30-35.

Lynch, K. (1960). *The image of the city.* Cambridge, MA, USA: The MIT Press.

McCullough, M. (2004). *Digital ground—Architecture, pervasive computing and environmental knowing.* Cambridge, MA, USA: The MIT Press.

Millen, D. R. (2000). Rapid ethnography: Time deepening strategies for HCI field research. In *Proceedings of DIS´00* (pp. 280-286). Brooklyn, NY: ACM.

Paay, J., & Kjeldskov, J. (2005). Understanding situated social interactions in public places. In *Proceedings of Interact 2005* (pp. 496-509). Rome, Italy: Springer-Verlag.

Raijmakers, B., Gaver, W., & Bishay, J. (2006). Design documentaries: Inspiring design research through documentary film. In *Proceedings of DIS 2006* (pp. 229-238). Pennsylvania, USA: ACM.

Snyder, C. (2003). *Paper prototyping.* San Francisco: Morgan Kaufmann Publishers.

Sommerville, I., Rodden, T., Sawyer, P., Bentley, R., & Twidale, M. (1993). Integrating ethnography into the requirements engineering process. In *Proceedings of IEEE International Symposium on Requirements Engineering* (pp. 165-181). San Diego, CA, USA: IEEE Computer Society Press.

Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research*. Newbury Park, CA, USA: Sage Publications.

Viller, S., & Sommerville, I. (1999). Coherence: An approach to representing ethnographic analyses in systems design. *Human-Computer Interaction, 14*, 9-41.

Wixon, D. (1995). Qualitative research methods in design and development. *Interactions, 2*(4), 19-24.

KEY TERMS

Affinity Diagramming: One of the techniques of the contextual design process, used during data interpretation sessions to group related individual points together, creating a hierarchical diagram showing the scope of issues in the work domain being studied.

Content Analysis: A qualitative research technique for gathering and analyzing the content of text, where content can be words, meanings, pictures, symbols, ideas, themes, or any message that can be communicated, to reveal messages in the text that are difficult to see through casual observation.

Contextual Design: A collection of techniques supporting a customer-centered design process, created by Beyer and Holtzblatt (1998), for finding out how people work to guide designers to find the optimal redesign for work practices.

Design Sketch: A graphical representation of a concept or design idea on a high level of abstraction. It should be quick, timely, open, disposable, un-detailed, and informal, and is usually hand-drawn on paper.

Expert Audit: A field reconnaissance done by an architecturally trained observer maping the presence of various elements of the physical environment and making subjective categorizations based on the immediate appearance of these elements in the field and their visible contribution to the image of the city.

Ethnography: A collection of techniques used for gathering and organizing field materials from observational studies, involving detailed observations of activities within their natural setting, to providing rich descriptions of people, environments and interactions.

Grounded Theory: A theory based analytical approach, which takes a set of data collected using ethnographic methods and provides a set of specific procedures for generating theory from this data.

Paper Prototype: A paper representation of a system design, able to simulate operation of that system, which is independent of platform and implementation, and can be used for brainstorming, designing, testing and communication of user interface designs and for identifying usability problems at an early stage of the design process.

Paper-Based Mock-Up: A representation of a specific design idea that is built from simple materials such as paper and cardboard, keeping it cheap and understandable, but making it a physical representation of a design idea for a final system, good for envisioning future products in the very early stages of the design process.

Rapid Ethnography: A collection of field methods to provide designers with a reasonable understanding of users and their activities given a limited amount of time spent in the field gathering data.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 1-15, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.10
Mobile e-Learning for Next Generation Communication Environment

Tin-Yu Wu
I-Shou University, Taiwan

Han-Chieh Chao
National Dong Hwa University, Taiwan

ABSTRACT

This article develops an environment for mobile e-learning that includes an interactive course, virtual online labs, an interactive online test, and lab-exercise training platform on the fourth generation mobile communication system. The Next Generation Learning Environment (NeGL) promotes the term "knowledge economy." Internetworking has become one of the most popular technologies in mobile e-learning for the next generation communication environment. This system uses a variety of computer embedded devices to ubiquitously access multimedia information, such as smart phones and PDAs. The most important feature is greater available bandwidth. The learning mode in the future will be an international, immediate, virtual, and interactive classroom that enables learners to learn and interact.

INTRODUCTION

The development of new approaches and technologies to support distance learning are undergoing now. In particular Web-based and mobile asynchronous learning environments and virtual classrooms via the Internet have been adopted widely. Static information as an instructional delivery method is the current trend in e-learning. Learners using these kinds of conventional learning methods are only able to browse through the mass static information. This is passive learning by reading online.

In the last decade, technologies enabling e-learning have increased learning location flexibility. Wireless communication technologies further increase the options for learning location. Advances in wireless communication technologies have provided the opportunity for educators

to create new educational models. With the aid of wireless communication technology, educational practice can be embedded into mobile life without wired-based communication. With the trend in educational media becoming more mobile, portable, and individualized, the learning form is being modified in spectacular ways (Gang & Zongkai, 2005).

In the third generation cellular system (3G) environment (such as Universal Mobile Telecommunications System, UMTS), the data rate reaches 2Mbps while the user is standing and 384Kbps while the user is moving slowly. Multimedia streaming, video conferencing, and online interactive 3D games are expected to attract increasing numbers of users. Such bandwidth is not sufficient for these increasingly popular applications and would be the major challenge for wireless networks. The 3G bandwidth has great problems with interactive teaching (Bos & Leroy, 2001).

In the future, wireless network traffic is expected to be a mix of real-time traffic such as voice, music, multimedia teleconferencing, online games, and data traffic such as Web page browsing, instant messaging, and file transfers. All of these applications will require widely varying and very diverse quality of service (QoS) guarantees for the different types of offered traffic (Dixit, 2001).

For these reasons, a fourth generation improved mobile communication system is necessary. The 4G system can support more bandwidth than other systems. It has advantages like authentication, mobile management, and quality of service (QoS). How to implement future distance learning environments for the fourth generation mobile communication system is the question. In this article, we distinguish four kinds of interactive courses, virtual online labs, interactive online tests, and lab-exercises training platform to deliver over the fourth generation mobile communication system. The fourth generation mobile communication system can use a variety of computer embedded devices to ubiquitously access multimedia information, such as smart phones and PDAs. Most important is that have more bandwidth. Hence, it supply ubiquitous learning environment (Girish & Dennett, 2000).

These new functions can improve the latency and location limits during transmission. Our proposed Next Generation Learning Environment offers learners the opportunities to use all kinds of mobile nodes that can connect to an Internet learning equipment system for access using All-IP communication networks. The Sharable Content Object Reference Model (SCORM) is used to compose information. Hence, as you can imagine, the condition of the learning mode in the future will be an international, immediate, and virtual interactive classroom that enables learners to learn and interact.

The article is organized as follows. We first describe the environments for mobile learning, followed by the virtual online classroom. The 4G testbed system design analyses are dealt with, and then the mobile e-learning results are discussed. The last section concludes the article.

ENVIRONMENTS FOR MOBILE LEARNING

Several investigations have focused on how to support great service for mobile e-learning. How many services will be able to fill the bill? In this session, we are introducing that mobile e-learning environment possesses many unique characteristics as follows (Tony, Sharples, Giasemi, & Lonsdale, 2004).

- Better adaptation to individual needs
- Ubiquitous and responds to urgent learning need
- Flexibility of location and time to learn
- Interactive knowledge acquisition
- Efficiency due both to re-use and feedback
- Situational instructional activities

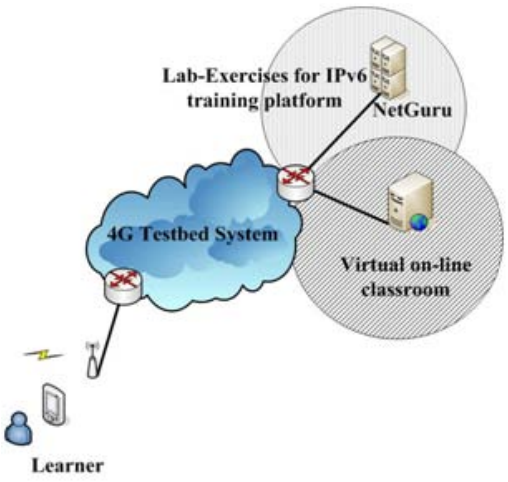
- Integrated instructional context (Chao, Wu, & Kao, 2004)

The mobile e-learning system includes interactive courses, virtual online labs, interactive online tests, and lab-exercises training platform on the fourth generation mobile communication system. As shown in Figure 1 the following sessions will present each part:

- **Interactive course system**: In learning history learners could experience interactive learning only in the classroom. The e-learning systems support only a single way for learning. These ways cannot support learning anytime and anywhere. We therefore developed an interactive course system to do that. Learners can choose which chapter they want to learn in this system. This learning method is not limited by the environment.
- **Virtual online labs system:** Generally speaking, experiments must be conducted in a laboratory. Learners are thereby limited to a specific learning area. To solve this problem special equipment is required. How can this problem be overcome? We simulated an experiment on laboratory all the time by Flash program. This virtual online lab platform supports step-by-step experimentation. Learners are therefore not restricted in the laboratory.
- **Interactive online test system:** An online interactive testing system is used to examine the teaching effect on students. The instructors can know how many learners were impacted via the testing system. Learners can obtain the learning effect on themselves.
- **Lab-exercises training platform:** The learners have more items for experimentation. NetSmooth Inc. developed a complete solution called NetGuru platform to tackle this issue. The learners can access the lab-exercises training platform via pre-arranged authorization.

A communication system is required to transfer the learning data. The most common communication platform used by students is the third generation cellular system. The data rate reaches

Figure 1. The virtual classroom on 4G system



2Mbps while the user is standing and 384Kbps while the user is moving slowly. This kind of system does not have enough bandwidth and no All-IP core network. Therefore, we developed a 4G testbed system that can support high transfer bandwidth.

VIRTUAL ONLINE CLASSROOM

Today there is much work going on in the field of virtual online classrooms around the world. The Web-based virtual classroom via the Internet as an instructional delivery method is a popular trend. Traditional learning methods only allowed the student to browse through mass static information. This is passive learning. In this session, we will introduce an interactive virtual classroom that includes the interactive course, virtual online labs, an interactive online test, and a lab-exercise training platform. For more information, see the virtual online classroom interactive Web site: http://6book.niu.edu.tw/ (6BOOK).

Setting up the Interactive Learning Course Web Site Platform

The Internet has uni-location and unlimited time features. Early online teaching materials included video lessons captured by DV, e-books, poster messages, and so on. These materials were used via the Internet. However, these approaches are single direction learning. These approaches are not good approaches for learning. These ways cannot attain learning anytime and anywhere. Therefore, we developed the interactive course system to do that (see Figure 3). The learners can choose which chapter they want to learn in this system. It can repeat whatever the learners want. The course collocates the interactive online test with interactive capability. The learners can learn anytime and anywhere unlimited by the environment.

Learners can select the chapters that they want to learn or review rapidly. They can study the chapters in order or preview or review any chapters, in any order. They can save all of their previous study processes. During learning, the system supports sliders with hints and oral explanations. Learners can control their learning speed and repeat it at will. Learners can see clearly just like taking the classes live (see Figure 4).

Setting up Virtual Online Lab Exercises

Generally, learner lab exercises must be conducted in a laboratory. They cannot perform experiments without a laboratory. This reduces a lot of opportunity to learn. Therefore, we used FLASH to produce a series of online lab exercises, explaining the lab exercises from the beginning and performing the exercises with detailed background voice and subtitles. There are explanations in great detail for each exercise. These explanations include the experiment goals, steps, and approaches that can help learners understand the background.

Most important, the learners can control the speed at which the lab-exercises proceed by themselves. Relying on online lab exercises, learners can perform lab exercises an unlimited number of times. They can perform experiments anytime from anywhere. The instructors do not have to spent time to prepare lab exercises or setup equipment. If learners have any questions about the exercises, they can use hyperlink to text to the Web site for answers. This teaching platform covers both theory and lab exercises interactively.

Setting up On-Line Exercises

The Virtual online lab exercises and interactive learning Web site platform help learners study efficiently. This system is able to identify the learning effect. We developed online interactive exercises for each chapter. These exercises identify

Figure 2. The interactive virtual online classroom Web site

Figure 3. Interactive learning course Web site platform

the comprehension of each learner for the instructor that uses these teaching materials. The system tutors learners that do not exhibit complete lesson understanding. Learners can also know clearly what areas should be enhanced and the content of each chapter by practicing the exercises.

Lab-Exercises Training Platform

The lab-exercises training platform is set-up using the NetSmooth Inc. test platform. This platform supports another solution with lab exercises for learners. The proposed NetGuru platform helps instructors to conduct network courses easily with Web-based tutorial courseware. It also assists students to strengthen the concepts of network with hands-on lab experiences (Chiang, Liang, Wu, & Chao, 2005).

The pragmatic lab exercises for the IPv6 training platform use a small-sized personal computer. There are some characters as follows (shown in Figure 7):

- All necessary lab hardware equipment is bundled together. No PC is required.
- Large-scale training labs are supported with multiple Netguru sets.
- The default setting is easily restored to initiate another lab work.
- Built-in 3 hosts and 3 hubs. (Each host has 3 NICs)
- Each set of NetGuru supports 3 groups to do Lab work.
- Simply connect monitor, mouse and keyboard with NetGuru to start Lab work instantly.

NetGuru integrates hardware, lab software, and training media into one complementary training set. We equipped the system with common use software to easily implement network services, such as routing, DNS, VPN, DHCP, NAT, Firewall, and so on. With build-in Ethereal tools for packet analyzing, learners will reinforce their conception about the packet structure. Based on the online commands, the environment will restore and default setting to initiate another lab work easily. In past days, while establishing a network environment, we not only needed computers, but also the heavy and complicated equipment configura-

Figure 4. The teaching slides with voice on the platform

Figure 5. The virtual online lab exercises

Figure 6. The online interactive exercises platform

Figure 7. NetGuru platform framework interface

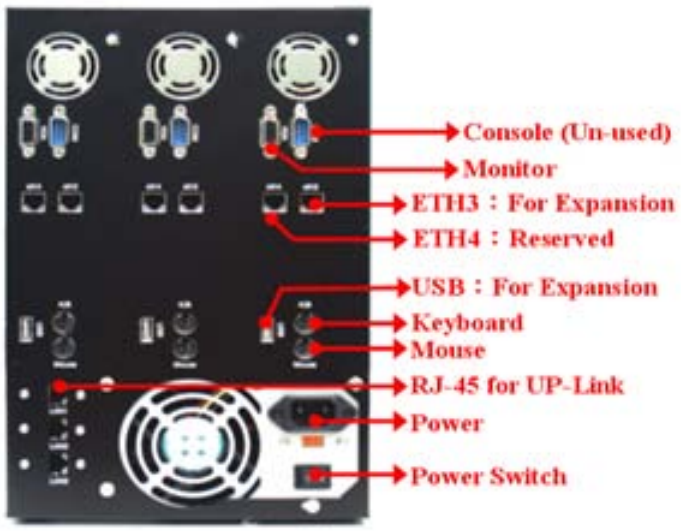
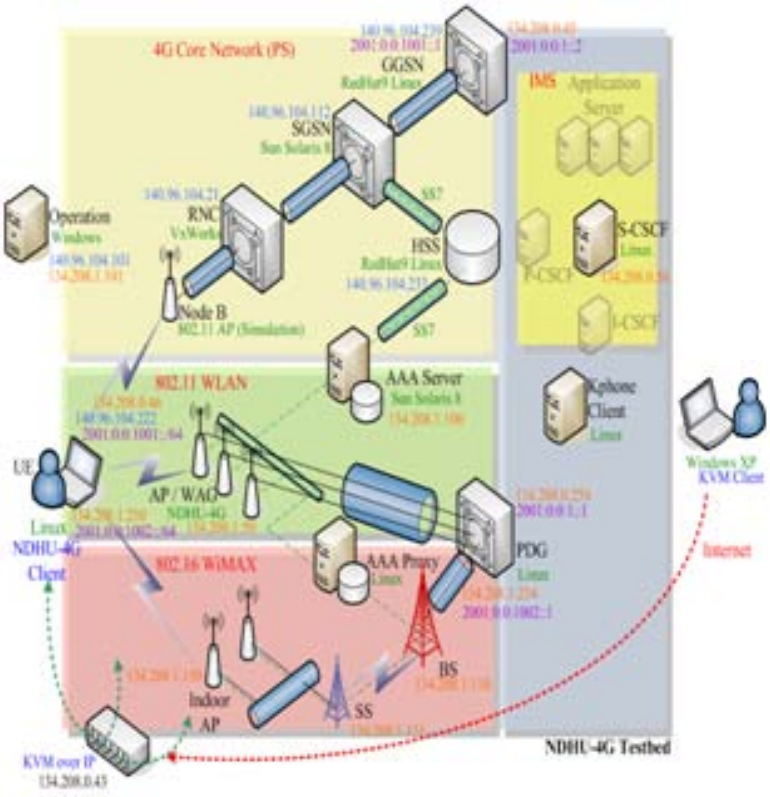


Figure 8. The cross-layer coordination plane



tion. With NetGuru, the TCP/IP lab environment can be easily set-up requiring no PC. We scaled down the size, and the small footprint allows easy relocation. Thus, we can set-up the TCP/IP lab environment anytime and anywhere. As we mentioned before, with multiple sets of NetGuru, large-scale training labs can easily be supported. NetGuru also supports extended network devices. Instructors can design other advanced lab work for use with this system.

4G Testbed System Design Analyses

We propose a fourth generation mobile communication testbed system. The system can support greater bandwidth than other systems. It has advantages like authentication, mobile management, and quality of service (QoS).

This session will introduce our fourth generation communication testbed system. We followed the specification defined in 3GPP to design our system. This system is composed of two main components: RAN (Radio Access Network) and Core-Network. RAN includes RNC and Node B. The Core-Network then includes SGSN (Serving GPRS Support Node), GGSN (Gateway GPRS Support Node), and HSS (Home Subscriber Server), as shown in Figure 8.

At RAN, Node B works like the access point of wireless network, providing the ability for UE (User Equipment) to connect to the core network through radio interface, each RNC can work with single or multiple Node B to form a RNA. RAN is then constituted by these RNS.

At the core network, SGSN is responsible for tasks such as connecting to the core network with single or multiple RAN, access control, location management, routing management etc. GGSN is an interface responsible of connecting core network and outer network, also routing traveling packets. It is also responsible for mobility management (Uskela, 2001).

HSS is a data center responsible for recording the operations of the entire network. HLR is its main component. Its function is to store the user's identity, location, and registered services that are allowed to the user.

Figure 9. The fourth generation mobile communication testbed system

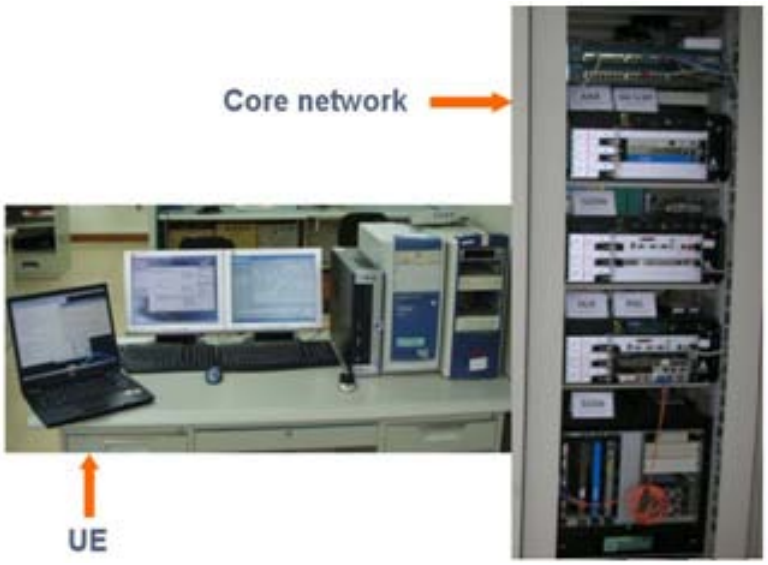


Figure 10. To learn in virtual classroom via PDA

Since the radio frequency used by a 3GPP cell phone is a licensed band, a legal license must be acquired. Therefore we used 802.11g which belongs to the ISM band instead. Through broadcasting UDP packets to simulate the radio network, and because the protocol stack of the simulation program is executed in UE according to the 3GPP standard, all generated packets are identical to packets generated by an actual 3GPP cell phone. UE enables us to acquire the flow chart of packets generated through the data exchange process between UE and the network. Figure 9 shows the entire system (3GPP, 3GPP TS 23.228, 3GPP TS 23.234).

Measurement Results with Mobile e-Learning

Wireless networks and mobile systems will continue to have explosive growth in the future. The traffic is expected to be a mix of real-time traffic such as voice, music and multimedia, and data traffic such as Web page browsing, instant messaging, and file transfers. All of these applications will require widely varying and very diverse quality of service (QoS) guarantees for the different types of offered traffic. Therefore, the mobile e-learning environment will be replacing traditional e-learning. We proposed a fourth generation mobile communication testbed system with advantages such as high transmission rate,

Figure 11. End-to-end delay

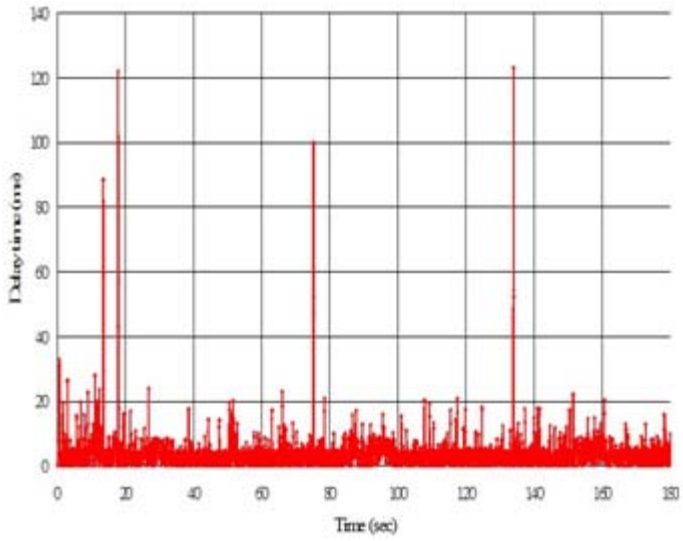
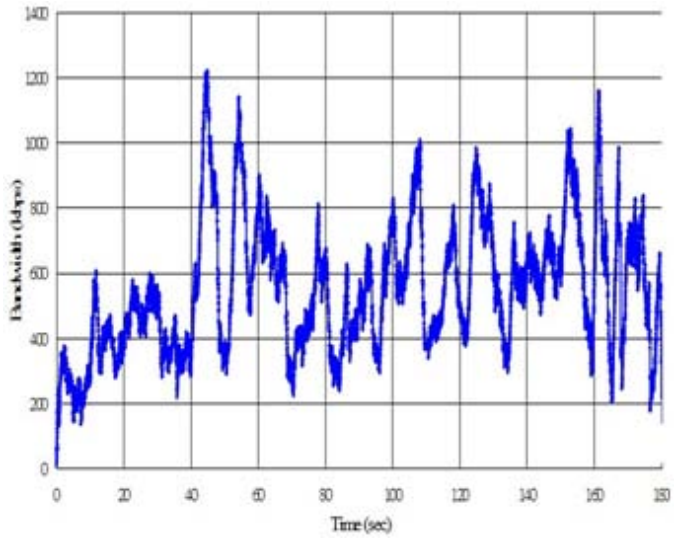


Figure 12. End-to-end bandwidth



robust wireless QoS control, wide cover area and supply IMS technology. Next generation communication technology can supply a variety of portable devices to ubiquitously access multimedia information, such as smart phones and PDAs. All Learners can use portable devices to log-on to the virtual classroom. Figure 10 shows a portable device surfing the virtual classroom via 4G.

In this session, we measure the mobile e-Learning system results via 4th generation core network. The scenarios are the UEs ability to connect to WLAN. Two scenarios are used for this measurement.

The first scenario is that the UE connects to the core network through WLAN. We measure the end-to-end delay of voice packet delivery. We captured the packets using the Wireshark protocol analyzer. We installed the protocol analyzer at the end of the core network client to capture the packet and decode the voice stream.

End-to-end delay refers to the time cost used for a packet to be transported across a network from the source to destination. For voice packet transmission, we calculate the end-to-end delay according to RFC 3550. Figure 11 shows the end-to-end delay for voice packet delivery in WLAN.

The second scenario is that the UE connects to core network through WLAN. We measure the throughput of video packet delivery. These results show the transported bandwidth in the 4G testbed system, as shown in Figure 12.

CONCLUSION

The explosive development of the Internet and wireless communications has made personal communication more convenient. Mobile computing uses the Next Generation Learning Environment (NeGL) to set up learning systems. We proposed a mobile e-learning system that includes interactive courses, virtual online labs, interactive online testing, and a lab exercise training platform via the fourth generation mobile communication system. It offers learners opportunities to use all kinds of mobile nodes or anything that can

connect to an Internet learning equipment system to be accessed using All-IP communication networks. In order for Content Object Reference Model (SCORM) to compose information, the 4G can use a variety of computer embedded devices to ubiquitously access multimedia information, such as smart phones and PDA. Most important is that more bandwidth is available. As you can imagine, the condition of the learning mode in the future will be an international, immediate and virtual interactive classroom that enables learners to learn and interact.

REFERENCES

3GPP. Third Generation Partnership Project, http://www.3gpp.org

3GPP TS 23.228 V6.10.0 (2005-06). IP multimedia subsystem

3GPP TS 23.234 V6.5.0 (2005-06). 3GPP system to Wireless Local Area Network (WLAN) interworking

6BOOK: http://6book.niu.edu.tw

Bos, L., & Leroy, S. (2001). Toward an all-IP-based UMTS system architecture. *IEEE Network, 15*(1), 36-45.

Chao, H.-C., Wu, T.-Y., & Kao, T. C.M. (2005). Environments for mobile learning: Pervasive and ubiquitous computing using IPv6. Chapter of *Encyclopedia of Online Learning and Technology, Information Science.*

Chiang, F.-Y., Liang, M.-H., Wu, T. Y., & Chao, H.-C. (2005). Pragmatic lab exercises for IPv6 training. *Proceedings of iCEER-2005,* Tainan, Taiwan, March 1-5.

Dixit, S.S. (2001). Evolving to seamless all-IP wireless/mobile networks. *IEEE Communications Magazine, 39*(12), 31-32.

Gang, Z. & Zongkai, Y. (2005). Learning resource adaptation and delivery framework for mobile learning. *Proceedings of Frontiers in Education, 2005 (FIE '05).*

Girish, P. & Dennett, S. (2000). The 3GPP and 3GPP2 movements toward an all-IP mobile network. *IEEE Personal Communications, 7*(4), 62-64.

Tony, C., Sharples, M., Giasemi, V., & Lonsdale, P. (2004). Educational metadata for mobile learning. *Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education 2004*, pp. 197-198.

Uskela, S. (2001). All IP architectures for cellular networks. *Proceedings of the Second International Conference on 3G Mobile Communication Technologies*, pp.180-185.

This work was previously published in International Journal of Distance Education Technologies, Vol. 6, Issue 4, edited by S. Chang and T. Shih, pp. 1-13, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 8.11
An Interactive Wireless Morse Code Learning System

Cheng-Huei Yang
National Kaohsiung Marine University, Taiwan

Li-Yeh Chuang
I-Shou University, Taiwan

Cheng-Hong Yang
National Kaohsiung University of Applied Sciences, Taiwan

Jun-Yang Chang
National Kaohsiung University of Applied Sciences, Taiwan

INTRODUCTION

Morse code has been shown to be a valuable tool in assistive technology, augmentative and alternative communication, and rehabilitation for some people with various conditions, such as spinal cord injuries, non-vocal quadriplegics, and visual or hearing impairments. In this article, a mobile phone human-interface system using Morse code input device is designed and implemented for the person with disabilities to send/receive SMS (simple message service) messages or make/respond to a phone call. The proposed system is divided into three parts: input module, control module, and display module. The data format of the signal transmission between the proposed system and the communication devices is the PDU (protocol description unit) mode. Experimental results revealed that three participants with disabilities were able to operate the mobile phone through this human interface after four weeks' practice.

BACKGROUND

A current trend in high technology production is to develop adaptive tools for persons with disabilities to assist them with self-learning and personal development, and lead more independent lives. Among the various technological adaptive tools available, many are based on the adaptation of computer hardware and software. The areas of application for computers and these tools include training, teaching, learning, rehabilitation, communication, and adaptive design (Enders, 1990;

McCormick, 1994; Bower et al., 1998; King, 1999).

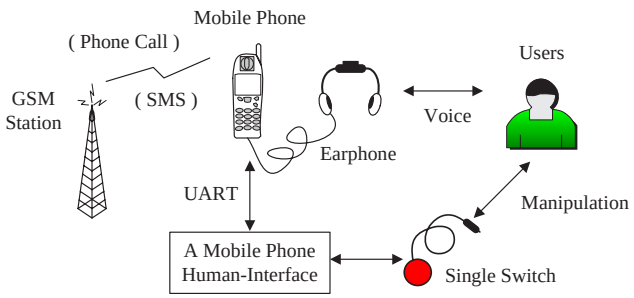
Many adapted and alternative input methods now have been developed to allow users with physical disabilities to use a computer. These include modified direct selections (via mouth stick, head stick, splinted hand, etc.), scanning methods (row-column, linear, circular) and other ways of controlling a sequentially stepping selection cursor in an organized information matrix via a single switch (Anson, 1997). However, they were not designed for mobile phone devices. Computer input systems, which use Morse code via special software programs, hardware devices, and switches, are invaluable assets in assistive technology (AT), augmentative-alternative communication (AAC), rehabilitation, and education (Caves, 2000; Leonard et al., 1995; Shannon et al., 1981; Thomas, 1981; French et al., 1986; Russel & Rego, 1998; Wyler & Ray, 1994). To date, more than 30 manufactures/developers of Morse code input hardware or software for use in AAC and AT have been identified (Anson, 1997; http://www.uwec.edu/Academic/Outreach/Mores2000/morse2000.html; Yang, 2000; Yang, 2001; Yang et al., 2002; Yang et al., 2003a; Yang et al., 2003b). In this article, we adopt Morse code to be the communication method and present a human interface for persons with physical disabilities.

The technology employed in assistive devices has often lagged behind mainstream products. This is partly because the shelf life of an assistive device is considerably longer then mainstream products such as mobile phones. In this study, we designed and implemented an easily operated mobile phone human interface device by using Morse code as a communication adaptive device for users with physical disabilities. Experimental results showed that three participants with disabilities were able to operate the mobile phone through this human interface after four weeks' practice.

SYSTEM DESIGN

Morse code is a simple, fast, and low-cost communication method composed of a series of dots, dashes, and intervals in which each character entered can be translated into a predefined sequence of dots and dashes (the elements of Morse code). A dot is represented as a period ".", while a dash is represented as a hyphen, or minus sign, "-". Each element, dot or dash, is transmitted by sending a signal for a standard length of time. According to the definition of Morse code, the tone ratio for dot to dash must be 1:3. That means that if the duration of a dot is taken to be one unit, then that of a dash must be three units. In addition, the silent ratio for dot-dash space to character-space also has to be 1:3. In other words, the space between the elements of one character is one unit while

Figure 1. System schematics of the mobile phone human-interface



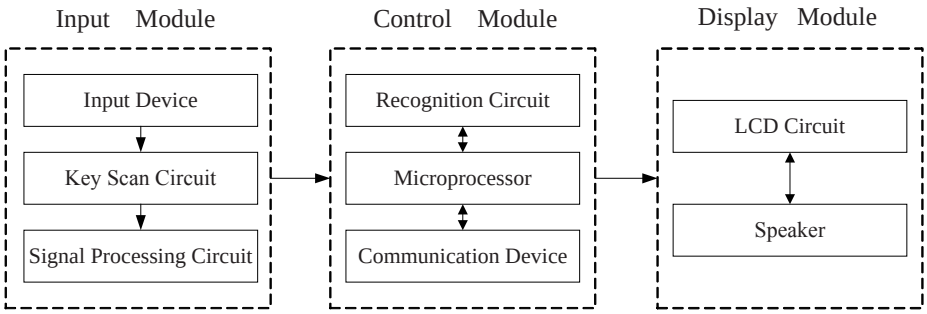
the space between characters is three units (Yang et al., 2002).

In this article, the mobile phone human interface system using Morse code input device is schematically shown in Figure 1. When a user presses the Morse code input device, the signal is transmitted to the key scan circuit, which translates the incoming analog data into digital data. The digital data are then sent into the microprocessor, an 8051 single chip, for further processing. In this study, an ATMEL series 89C51 single chip has been adopted to handle the communication between the press-button processing and the communication devices. Even though the I/O memory capacity of the chip is small compared to a typical PC, it is sufficient to control the device. The 89C51 chip's internal serial communication function is used for data transmission and reception (Mackenzie, 1998). To achieve the data communication at both ends, the two pins, TxD and RxD, are connected to the TxD and RxD pins of a RS-232 connector. Then the two pins are connected to the RxD and TxD of an UART (Universal Asynchronous Receiver Transmitter) controller on the mobile phone device. Then, persons with physical disabilities can use this proposed communication aid system to connect their mobile communication equipment, such as mobile phones or GSM (global system for mobile communications) modems, and receive or send their messages (SMS, simple message service). If they wear an earphone, they might be able to dial or answer the phone. SMS is a protocol (GSM 03.40 and GSM 03.38), which was established by the ETSI (the European Telecommunications Standards Institute) organization. The transmission model is divided into two models: text and PDU (protocol description unit). In this system, we use the PDU model to transmit and receive SMS information through the AT command of the application program (Pettersson, 2000). Structurally the mobile phone human-interface system is divided into three modules: the input module, the control module, and the display module. The interface framework is graphically shown in Figure 2. A detailed explanation is given below.

INPUT MODULE

A user's input will be digitized first, and then the converted results will be sent to the micro controller. From the signal processing circuit can monitor all input from the input device, the Morse code. The results will be entered into the input data stream. When the user presses the input key, the micro-operating system detects new input data in the data stream, and then sends the corresponding characters to the display module. Some commands and/or keys, such as *OK, Cancel, Answer, Response, Send, Receive, Menu, Exit,*

Figure 2. Interface framework of mobile phone for persons with physical disablities



and so forth, have been customized and perform several new functions in order to accommodate the Morse code system. These key modifications facilitate the human interface use for a person with disabilities.

CONTROL MODULE

The proposed recognition method is divided into three modules (see Figure 3): space recognition, adjustment processing, and character translation. Initially, the input data stream is sent individually to separate tone code buffer and space recognition processes, which are based on key-press (Morse code element) or key-release (space element). In the space recognition module, the space element value is recognized as a dot-dash space or a character space. The dot-dash space and character space represent the spaces existing between individual characters and within isolated elements of a character respectively. If a character space is identified, then the value(s) in the code buffer is (are) sent to character translation. To account for varying release speeds, the space element value has to be adjusted. The silent element value is sent into the silent base adjustment process. Afterwards, the character is identified in the character translation process.

A Morse code character, x_i, is represented as follows:

$$m_1(x_i), b_1(x_i), \ldots, m_j(x_i), b_j(x_i), \ldots, m_n(x_i), b_n(x_i)$$

where

$b_j(x_i)$: jth silent duration in the character x_i.

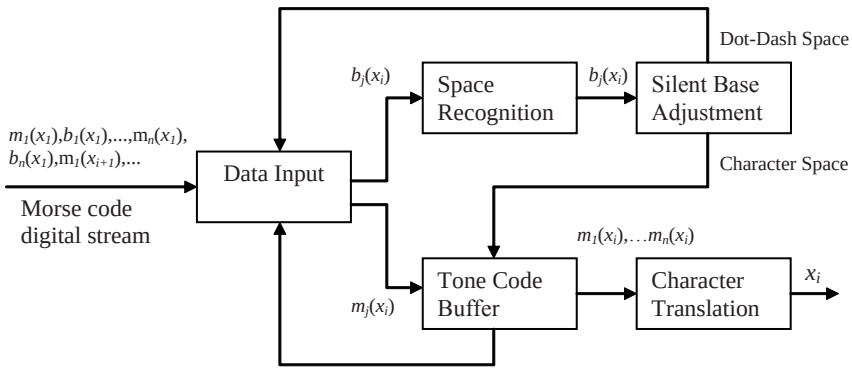
n: the total number of Morse code elements in the character x_i.

$m_j(x_i)$: the jth Morse code element of the input character x_i.

DISPLAY MODULE

Since users with disabilities have, in order to increase the convenience of user operations, more requirements for system interfaces than a normal person, the developed system shows selected items and system condition information on an electronic circuit platform, which is based on LCD (liquid crystal display). The characteristics

Figure 3. Block diagram of the Morse code recognition system



of the proposed system can be summarized as follows: (1) easy operation for users with physical disabilities with Morse code input system, (2) multiple operations due to the selection of different modes, (3) highly tolerant capability from adaptive algorithm recognition, and (4) system extension for customized functions.

RESULTS AND DISCUSSIONS

This system provides two easily operated modes, the phone panel and LCD panel control mode, which allow a user with disabilities easy manipulation. The following shows how the proposed system sends/receives simple message service (SMS) message or make /respond to a phone call.

SMS Receiving Operation

First, when users receive a message notification and want to look at the content, this system will provide "phone panel" and "LCD panel" control modes to choose from. In the phone panel mode, users can directly key-in Morse code "…" (as character 'S'). The interface system will go through the message recognition process, then exchange the message into AT command "AT+CKPD='S', 1", to execute the "confirm" action of the mobile phone. The purpose of this process is the same as users keying-in "yes" on the mobile phone keyboard, then keying-in Morse code ". - - ." (as key '↓'). The system will recognize the message, then automatically send the "AT+CKPD='↓', 1" instruction. The message cursor of the mobile phone is moved to the next line, or key-in Morse code ". - . -" (as key '↑') for moving it to the previous data line. Finally, if users want to exit and return to the previous screen, they only need to key-in Morse code ". . - ." (as character 'F'), and start the c key function on the mobile phone keyboard. If LCD panel mode is selected, one can directly follow the selected items on the LCD crystal, to execute the reception and message reading process.

SMS Transmitting Operation

Message transmission services are provided in two modes: phone panel and LCD panel. In the phone panel mode, continually type two times the Morse code ". - . - ." (as key '→'). The system will be converted into AT Command and transferred into mobile phone to show the selection screen of the message functions. Then continuing to key-in three times the Morse code ". . ." (as character 'S'), one can get into the editing screen of message content, and wait for users to input the message text data and receiver's phone number. The phone book function can be used to directly save the receiver's phone number. After the input, press the "yes" key to confirm that the message sending process has been completed. In addition, if the LCD panel mode is selected, one can follow the LCD selection prompt input the service selection of all the action integrated in the LCD panel. Then go through the interface and translate to a series of AT command orders, and batch transfer these into the mobile phone to achieve the control purpose.

The selection command "Answer a phone," displays on the menu of the LCD screen, and can be constructed using Morse code. The participants could press and release the switch, and input the number code ". - - - -" (as character '1') or hot key ". -" (as character 'A'). The mobile phone is then answered automatically. Problems with this training, according to participants, are that the end result is limited typing speed and users must remember all the Morse code set of commands.

Three test participants were chosen to investigate the efficiency of the proposed system after practicing on this system for four weeks. Participant 1 (P1) was a 14-year-old male adolescent who has been diagnosed with cerebral palsy. Participant 2 (P2) was a 14-year-old female adolescent with cerebral palsy, athetoid type, who experiences in-

voluntary movements of all her limbs. Participant 3 (P3) was a 40-year-old male adult, with a spinal cord injury and incomplete quadriparalysis due to an accident. These three test participants with physical impairments were able to make/respond to phone calls or send/receive SMS messages after practice with the proposed system.

FUTURE TRENDS

In the future, Morse code input device could be adapted to several environmental control devices, which would facilitate the use of everyday appliances for people with physical disabilities considerably.

CONCLUSION

To help some persons with disabilities such as amyotrophic lateral sclerosis, multiple sclerosis, muscular dystrophy, and other conditions that worsen with time and cause the user's abilities to write, type, and speak to be progressively lost, it requires an assistive tool for purposes of augmentative and alternative communication in their daily lives. This article presents a human interface for mobile phone devices using Morse code as an adapted access communication tool. This system provides phone panel and LCD panel control modes to help users with a disability with operation. Experimental results revealed that three physically impaired users were able to make/respond to phone calls or send/receive SMS messages after only four weeks' practice with the proposed system.

ACKNOWLEDGMENTS

This research was supported by the National Science Council, R.O.C., under grant NSC 91-2213-E-151-016.

REFERENCES

Anson, D. (1997). *Alternative computer access: A guide to selection.* Philadelphia, PA: F. A. Davis.

Bower, R. et al. (Eds.) (1998). *The Trace resource book: Assistive technology for communication, control, and computer access.* Madison, WI: Trace Research & Development Center, Universities of Wisconsin-Madison, Waisman Center.

Caves, K. (2000). *Morse code on a computer—really?* Keynote presentation at the First Morse 2000 World Conference, Minneapolis, MN.

Enders, A., & Hall, M. (Ed.) (1990). *Assistive technology sourcebook*. Arlington, VA: RESNA Press,.

French, J. J., Silverstein, F., & Siebens, A. A. (1986). An inexpensive computer based Morse code system. In *Proceedings of the RESNA 9th Annual Conference, Minneapolis* (pp. 259-261). Retrieved from http://www.uwec.edu/Academic/Outreach/Mores2000/morse2000.html.

King, T. W. (1999). *Modern Morse code in rehabilitation and education.* MA: Allyn and Bacon.

Lars Pettersson. (n.d.). *Dreamfabric.* Retrieved from http://www.dreamfabric.com/sms

Leonard, S., Romanowski, J., & Carroll, C. (1995). Morse code as a writing method for school students. *Morsels, University of Wisconsin-Eau Claire, 1*(2), 1.

Mackenzie, I. S. (1998). *The 89C51 Microcontroller* (3rd ed.). Prentice Hall.

McCormick, J. A. (1994). Computers and the Americans with disabilities act: A manager's guide. Blue Ridge Summit, PA: Wincrest/McGraw Hill.

Russel, M., & Rego, R. (1998). A Morse code

communication device for the deaf-blind individual. In *Proceedings of the ICAART, Montreal* (pp. 52-53).

Shannon, D. A., Staewen, W. S., Miller, J. T., & Cohen, B. S. (1981). Morse code controlled computer aid for the nonvocal quadriplegic. *Medical Instrumentation, 15*(5), 341-343.

Thomas, A. (1981). Communication devices for the non-vocal disabled. *Computer, 14*, 25-30.

Wyler, A. R., & Ray, M. W. (1994). Aphasia for Morse code. *Brain and Language, 27*(2), 195-198.

Yang, C.-H. (2000), Adaptive Morse code communication system for severely disabled individuals. *Medical Engineering & Physics, 22*(1), 59-66.

Yang, C.-H. (2001). Morse code recognition using learning vector quantization for persons with physical disabilities. *IEICE Transactions on Fundamentals of Electronics, Communication and Computer Sciences, E84-A*(1), 356-362.

Yang, C.-H., Chuang, L.-Y. Yang, C.-H., & Luo, C.-H. (2002). An Internet access device for physically impaired users of Chanjei Morse code. *Journal of Chinese Institute of Engineers, 25*(3), 363-369.

Yang, C.-H. (2003a). An interactive Morse code emulation management system. *Computer & Mathematics with Applications, 46*, 479-492.

Yang, C.-H., Chuang, L.-Y., Yang, C.-H., & Luo, C.-H. (2003b, December). Morse code application for wireless environmental control system for severely disabled individuals. *IEEE Transactions on Neural System and Rehabilitation Engineering, 11*(4), 463-469.

KEY TERMS

Adaptive Signal Processing: Adaptive signal processing is the processing, amplification and interpretation of signals that change over time through a process that adapts to a change in the input signal.

Assistive Technology (AT): A generic term for a device that helps a person accomplishes a task. It includes assistive, adaptive and rehabilitative devices, and grants a greater degree of independence people with disabilities by letting them perform tasks they would otherwise be unable of performing.

Augmentative and Alternative Communication (AAC): Support for and/or replacement of natural speaking, writing, typing, and telecommunications capabilities that do not fully meet communicator's needs. AAC, a subset of AT (see below), is a field of academic study and clinical practice, combining the expertise of many professions. AAC may include unaided and aided approaches.

Global System for Mobile Communications (GSM): GSM is the most popular standard for global mobile phone communication. Both its signal and speech channels are digital and it is therefore considered a 2nd generation mobile phone system.

Morse Code: Morse code is a transmission method, implemented by using just a single switch. The tone ratio (dot to dash) in Morse code has to be 1:3 per definition. This means that the duration of a dash is required to be three times that of a dot. In addition, the silent ratio (dot-space to character-space) also has to be 1:3.

Simple Message Service (SMS): A service available on digital mobile phones, which permits the sending of simple messages between mobile phones.

This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 352-356, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.12
A Mobile Computing Framework for Passive RFID Detection System in Health Care

Masoud Mohammadian
University of Canberra, Australia

Ric Jentzsch
Compucat Research Pty Limited, Australia

INTRODUCTION

The cost of health care continues to be a world wide issue. Research continues into ways and how the utilization of evolving technologies can be applied to reduce costs and improve patient care, while maintaining patient's lives. To achieve these needs requires accurate, near real time data acquisition and analysis. At the same time there exists a need to acquire a profile on a patient and update that profile as fast and as possible. All types of confidentiality need to be addressed no matter which technology and application is used. One possible way to achieve this is to use a passive detection system that employs wireless radio frequency identification (RFID) technology. This detection system can integrate wireless networks for fast data acquisition and transmission, while maintaining the privacy issue. Once this data is obtained, then up to date profiling can be integrated into the patient care system. This article discussed the use and need for a passive RFID system for patient data acquisition in health care facilities such as a hospital. The development of profile data is assisted by a profiling intelligent software agent that is responsible for processing the raw data obtained through RFID and database and invoking the creation and update of the patient profile.

BACKGROUND

Health is on everyone's agenda whether they are old or young. Millions of hours of lost time is recorded each week by employers' whose staff are in need of health care. It is and has been known that more research into applications and

innovative architectures is needed. To this end the use of Radio Frequency Identification (RFID), a relatively new technology and is showing itself to be a viable and promising technology as an aid to health care (Finkenzeller, 1999; Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster, Allen, & Brock, 2007; Shepard, 2005). This technology has the capability to penetrate and add value to nearly every area of health care. It can be used to lower the cost of some services as well as improving service to individuals and the health care provider. Although many organizations are developing and testing the possible use of RFIDs, the real value of RFID is achieved in conjunction with the use of intelligent software agents. Thus the issue becomes the integration of these two great technologies for the benefit of assisting health care services.

To begin with, let us look at data collection. In health care, we can collect data on the patients, doctors, nurses, institution itself, drugs and prescriptions, diagnosis, and many other areas. It would not be feasible to do all of these nor would all of these be able to effectively use RFID. Thus for our perspective we will concentrate on a subset with the understanding that all areas could, directly or indirectly, benefit from the use of RFID and intelligent software agents in a health care and hospital environment.

In this research, we begin to look at the architecture of integrating intelligent software agents technology with RFID technology, in particular in managing patients' health care data in a hospital environment.

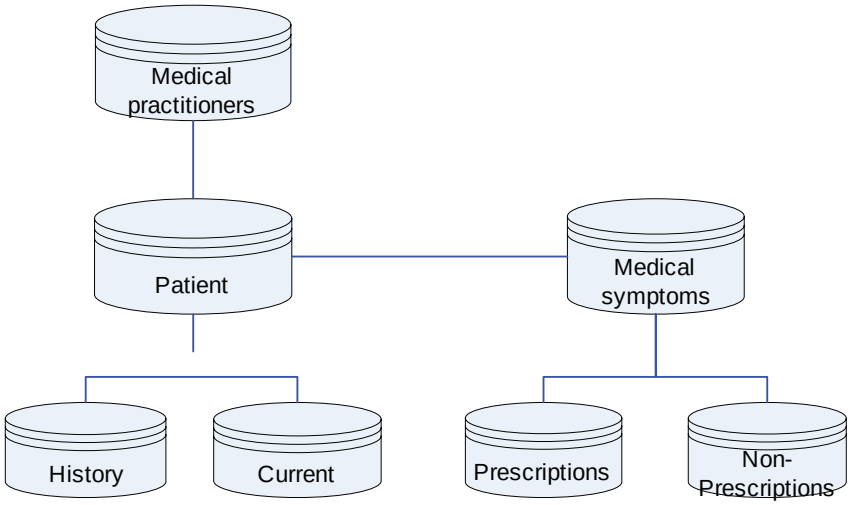
An intelligent software agent can continuously profile a patient based on their medical history, current illness, and on going diagnostics. The RFID provides the passive vehicle to obtain the data via its monitoring capabilities. The intelligent software agent provides the active vehicle in the interpretation profiling of the data and reporting capacity. There are certain data that is stored about each patient in a hospital. The investigation of this data provides an analysis that describes the patient's condition, is able to monitor their status, and cross reflect on why the patient was admitted to the hospital. Using this information an evolving profile of each patient can be constructed and analysed.

Using the data and analysis this will allow us to assist in deciding what kind of care he/she requires, the effects of ongoing care, and how to best care for this patient using available resources (doctors, nurses, beds, etc.) for the patient. The software agent is used to build a profile of each patient as they are admitted to the health care institution. Although not shown in the illustration, an additional profile for each doctor can be developed that practices in the hospital can be developed. If this is done, then the patient and doctor profile can be correlated to obtained the availability of the best doctor to suit the patient. However, this will require an additional data repository, as shown in Figure 1.

The patient profiling is useful in a variety of situations:

- The profile provides a personalized service based on the patient and not on symptoms or illness. to a particular patient. For example, by identifying the services that the patient requires this will allow us to target that which will be directed to speeding up their recovery progress;
- A good profile will assist the medical facilities in trying to prevent the need for the patient to return to the hospital any sooner than necessary;
- Disambiguating patient's diagnostic based on patient profile may help in assisting in matching a doctor's specialization to the right patient;
- When a patient needs to re-enter a hospital, a past profile can make it easier to match the patient's needs to a relevant available doctor;
- Presenting information about the patient on an on-going, continuous basis for the

Figure 1. Data repositories for patient and doctors



doctors means that current up to date information is available rather than information that needs to be searched for and compiled before it is useful; and

- Providing tailored and appropriate care to reduce health care costs.

Profiling is being done in many business operations today. Often profiling is combined with personalization, and user modeling for many e-commerce applications such as those by IBM, ATG Dynamo, BroadVision, Amazon, and Garden (18). However, there is very little in the way of the use of such systems in hospital and very little in health care in this perspective has been reported. It must be remember that there are different definitions of personalization, user modeling, and profiling. In e-commerce the practice of tracking information about consumers' interests by monitoring their movements online is considered profiling or user modeling. This can be done without using any personal information, but simply by analyzing the content, URL's, and other information about a user's browsing path/ click-stream. Many user models try to predict the user's preference in a narrow and specific domain. This works well as long as that domain remains relatively static and, as such, the results of such work may be limited.

In this research, profiling is a technique whereby a set of characteristics of a particular class of person, patient, is inferred from their past and data-holdings are then searched for individuals with a close fit to symptom characteristics. One of the main aims of profiling and user modeling is to provide information recipients with correct and timely response for their needs. This entails an evolving profile to ensure that as the dynamics of that which is being monitored change, the profile and model reflects these updates as appropriate.

There are several ways in which a patient's visit to a hospital can be recorded. A patient's visit may simply be classified as a regular visit.

This may be for a check up, for tests, or at the request of a doctor. A patient might be at the hospital because of an emergency or an ad hoc appointment due to lack of other facilities being available. Of course there are a whole set of patients that visit the hospital for reasons that are less well defined. In each situation, the needs of the patients are different.

The patient's profile can assist the attending doctor in being aware of the particular patient's situation. This provides the attending doctor with information that is needed without waiting for the patient's regular doctor. The regular doctor may be unavailable and therefore the profile of the patient can be matched with the available doctor suitable to the needs of the patient. The patient to doctor assignment is a type of scheduling issue and is not going to be discussed in this article.

However, in an emergency visit, there is no assigned doctor for such a patient. The doctor in emergency section of the hospital will provide information about a patient after examination and a patient profile then can be created. In this case, the intelligent agent can assist the patient by matching the profile of the patient with the doctors suitable to the needs of the patient. Also the doctors can be contacted in a speedier manner as they are identified and their availability is known.

An appointment visit is very similar to a regular visit but it may happen only once and therefore the advantages mentioned for regular patients applies here.

We will endeavor to describe several of these, but will expand on one particular potential use of RFIDs in managing patient health data. First let us provide some background on RFIDs and present some definitions. We will discuss the environment that RFIDs operate in and their relationship to other available wireless technologies such as the IEEE 802.11b, IEEE 802.11g, IEEE 802.11n, and so forth, in order to fulfill their requirements effectively and efficiently.

This research is divided into four main sections. Section two is based on the patient to doctor profiling and intelligent software agents. The third section is a RFID background; this will provide a good description of RFIDs and their components. This section discusses several practical cases of RFID technology in and around hospitals. It will also list three possible applicable cases assisting in managing patients' medical data. The final section discusses the important issue of maintaining patients' data security and integrity and relates that to RFIDs.

PATIENT TO DOCTOR PROFILING

A profile represents the extent to which something exhibits various characteristics. These characteristics are used to develop a linear model based on the consensus of multiple sets of data, generally over some period of time. A patient or doctor profile is a collection of information about a person based on the characteristics of that person. This information can be used in a decision analyze situation between the doctor, domain environment, and patient. The model can be used to provide meaningful information for useful and strategic actions. The profile can be static or dynamic. The static profile is kept in prefixed data fields where the period between data field updates is long such as months or years. The dynamic profile is constantly updated as per evaluation of the situation. The updates may be performed manually or automated. The automated user profile building is especially important in real time decision-making systems. Real time systems are dynamic. These systems often contain data that is critical to the user's decision making process. Manually updated profiles are at the need and discretion of the relevant decision maker.

The profiling of patient doctor model is base on the patient/doctor information. These are:

- The categories and subcategories of doctor specialization and categorization. These categories will assist in information processing and patient/doctor matching.
- Part of the patients profile based on symptoms (past history problems, dietary restrictions, etc.) can assist in prediction of the patient's needs specifically.
- The patients profile can be matched with the available doctor profiles to provide doctors with information about the arrival of patients as well as presentation of the patients profile to a suitable, available doctor.

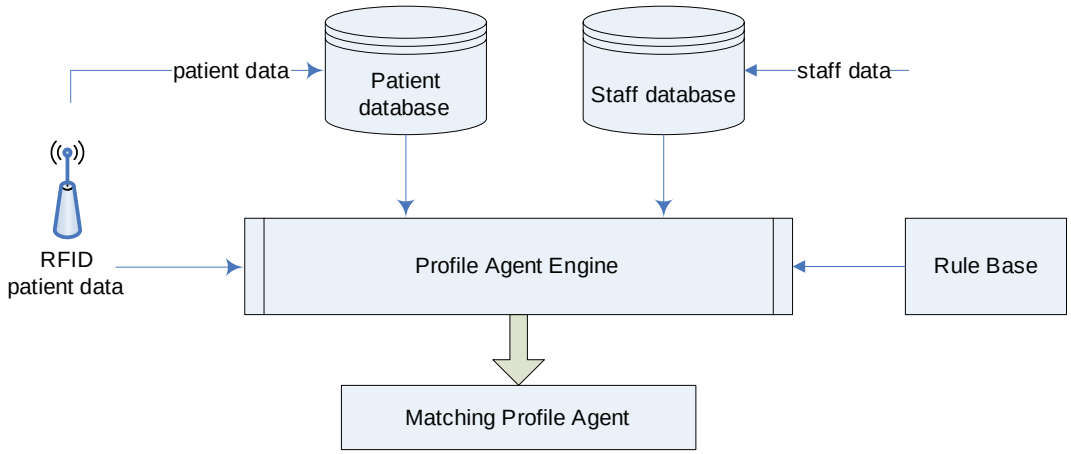
A value denoting the degree of association can be created form the above evaluation of the doctor to patient's profile. The intelligent agent based on the denoting degrees and appropriate, available doctors can be identified and allocated to the patient.

In the patient/doctor profiling, the agent will make distinctions in attribute values of the profiles and match the profiles with highest value. It should be noted that the agent creates the patient and doctor profiles based on data obtained from the doctors and patient namely:

- Explicit profiling occurs based on the data entered by hospital staff about a patient.
- Implicit profiling can fill that gap for the missing data by acquiring knowledge about the patient from its past visit or other relevant databases if any and then combining all these data to fill the missing data. Using legacy data for complementing and updating the user profile seems to be a better choice than implicit profiling. This approach capitalizes on user's personal history (previous data from previous visit to doctor or hospital).

The proposed agent architecture allows user profiling and matching in such a time intensive important application. The architecture of the agent profiling systems using RFID is given in Figure 2.

Figure 2. Agent profiling model using RFID



PROFILE MATCHING

Profile matching done is based on a vector of weighted attributes. To get this vector, a rule based systems can be used to match the patient's attributes (stored in patient's profile) against doctor's attributes (stored in doctor's profile). If there is a partial or full match between them, then the doctor will be informed (based on their availability from the hospital doctor database).

INTEGRATION OF INTELLIGENT SOFTWARE AGENTS AND RFID TECHNOLOGIES

Intelligent software agent technology has been used in order to provide the needed transformation of RIFD passive data collection into an active organizational knowledge assistant (Finkenzeller, 1999; Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster et al., 2007; Shepard, 2005). Intelligent agent should be able to act on new data and already stored profile/knowledge and thereafter to examine its current actions based on certain assumptions, and inferentially plans its activities. Furthermore, intelligent software agents must be able to *interact* with other agents using symbolic language (Bigus & Bigus, 1998; Wooldridge & Jennings, 1995) and able to substitute for a range of human activities in a situated context. (In our case the activities are medical/patient assignment and the context is a hospital environment)

Context driven Intelligent software agents' activities are also dynamic and under continuous development in an historical time related environment (Bigus & Bigus, 1998; Wooldridge & Jennings, 1995).

Medical and hospital patient applied ontology's describing the applied domain are necessary for the semantic communication and data understanding between RFID inputs and knowledge bases inference engines so that profiling of both patients and doctors can be achieved (Gruber, 1993; Guarino, Carrara, & Giaretta, 1994).

The integration of RFID capabilities and intelligent agent techniques provides promising development in the areas of performance improvements in RFID data collection, inference, knowledge acquisition, and profiling operations.

By using mediated activity theories, an RIFD agent architecture could be modeled according to the following characteristics:

- The ability to use patient/doctor profile in natural language, ACL, or symbolic form as communicative tools mediating agents cooperative activities.
- The ability to use subjective and objective properties required by intelligent software agents to perform bidirectional multiple communication activities.
- The ability to internalize representations of medical/patient profile patterns from agents or humans.
- The ability to externalize internally stored representations of medical assignment patterns to other agents or humans.

The Agent Language Mediated Activity Model (ALMA) agent architecture currently under research is based on the mediated activity framework described and is able to provide RFID with the necessary framework to profile a range of internal and external medical/patient profiling communication activities performed by wireless multi-agents.

RFID DESCRIPTION

RFID or Radio Frequency Identification is a progressive technology that has been said to be easy to use and well suited for collaboration with intelligent software agents. Basically an RFID can:

- Be read-only;
- Volatile read/write; or
- Write once/read many times
- RFID are:
    - Noncontact and
    - Non line-of-sight operations.

Being noncontact and non line-of-sight will make RFIDs able to function under a variety of environmental conditions and while still providing a high level of data integrity (Finkenzeller, 1999; Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster et al., 2007; Shepard, 2005).

MAIN COMPONENTS

A basic RFID system consists of four components:

1. The RFID tag (sometimes referred to as the transponder);
2. A coiled antenna;
3. A radio frequency transceiver; and
4. Some type of reader for the data collection.

Basically there are three components as often components are combined such as the transponder or transceiver or the antenna.

Transponders

The reader emits radio waves in ranges of anywhere from 2.54 centimeters to 33 meters. Depending upon the reader's power output and the radio frequency used and if a booster is added that distance can be somewhat increased. When RFID tags (transponders) pass through a specifically created electromagnetic zone, they detect the reader's activation signal. Transponders can be online or off-line and electronically programmed with unique information for a specific application or purpose. A reader decodes the data encoded on the tag's integrated circuit and passes the data to a server for data storage or further processing.

Coiled Antenna

The coiled antenna is used to emit radio signals to activate the tag and read or write data to it. Antennas are the conduits between the tag and the transceiver that controls the system's data acquisition and communication. RFID antennas are available in many shapes and sizes. They can be built into a doorframe, book binding, DVD case, mounted on a tollbooth, embedded into a manufactured item such as a shaver or software case (just about anything) so that the receiver tags the data from things passing through its zone (Finkenzeller, 1999; Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster et al., 2007; Shepard, 2005).

Transceiver

Often the antenna is packaged with the transceiver and decoder to become a reader. The decoder device can be configured either as a handheld or a fixed-mounted device. In large complex, often chaotic environments, portable or handheld transceivers would prove valuable.

TYPES OF RFID TRANSPONDERS

RFID tags can be categorized as active, semi-active, or passive. Each has and is being used in a variety of inventory management and data collection applications today. The condition of the application, place and use determines the required tag type.

Active Tags

Active RFID tags are powered by an internal battery and are typically read / write. Tag data can be rewritten and/or modified as the need dictates. An active tag's memory size varies according to manufacturing specifications and application requirements; some tags operate with up to 5 megabyte of memory. For a typical read/write RFID work-in-process system, a tag might give a machine a set of instructions, and the machine would then report its performance to the tag. This encoded data would then become part of the tagged part's history. The battery-supplied power of an active tag generally gives it a longer read range. The trade off is greater size, greater cost, and a limited operational life that has been estimated to be a maximum of 10 years, depending upon operating temperatures and battery type (Finkenzeller, 1999; Glover & Bhatt, 2006).

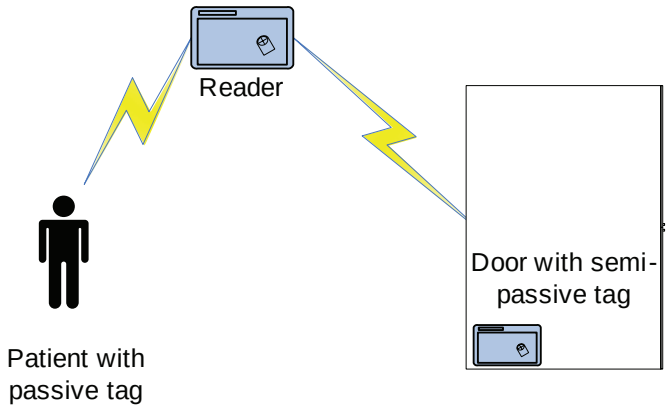
Semi-Active Tags

The semi-active tag comes with a battery. The battery is used to power the tags circuitry and not to communicate with the reader. This makes the semi-active tag more independent than the passive tag, and it can operate in more adverse conditions. The semi-active tag also has a longer range and more capabilities than a passive tag (Shepard, 2005). Linear barcodes that reference a database to get product specifications and pricing are also data devices that act is a very similar way. Semi-passive tags are preprogrammed, but can allow for slight modifications of their instructions via the reader/interrogator. However, it is bigger, weighs more, and is more complete than a passive tag. A reader is still needed for data collection.

Passive Tags

Passive RFID tags operate without a separate external power source and obtain operating power generated from the reader. Passive tags,

Figure 3. Semi-passive tag



since they have no power source embedded, are consequently much lighter than active tags, less expensive, and offer a virtually unlimited operational lifetime. However, the trade off is that they have shorter read ranges, than active tags, and require a higher-powered reader.

Read-only tags are typically passive and are programmed with a unique set of data (usually 32 to 128 bits) that cannot be modified. Read-only tags most often operate as a data device that utilizes a database for all data storage (Finkenzeller, 1999; Shepard, 2005).

Range

RFID systems can be distinguished by their deployment and frequency range. RFID tags generally operate in two different types of frequencies that make them adaptable for nearly any application. These frequency ranges are:

Low Frequency Range (Short Range)

Low-frequency (30 KHz to 500 KHz) systems have short reading ranges and lower system costs. They are most commonly used in security access, asset tracking, and animal identification applications (Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005).

High Frequency Range (Long Range)

High-frequency (850 MHz to 950 MHz and Industry, Science and Medical - 2.4 GHz to 2.5 GHz) systems, offer longer reading ranges (greater than 33 meters) and high reading speeds. These systems are generally used for such applications as railroad car tracking, container dock and transport management, and automated toll collection. However, the higher performance of high-frequency RFID systems incurs higher system operating costs (Glover & Bhatt, 2006; Schuster et al., 2007).

Hospital Environment

In hospitals, systems need to use rules and domain knowledge that is appropriate to the situation. One of the more promising capabilities of intelligent software agents is their ability to coordinate information between the various resources.

In a hospital environment, in order to manage patient medical data we need both types; fixed and handheld transceivers. Also, transceivers can be assembled in ceilings, walls, or doorframes to collect and disseminate data. Hospitals have become large complex environments.

In a hospital, nurses and physicians can retrieve the patient's medical data stored in transponders (RFID tags) before they stand beside a patient's bed or as they are entering a ward.

Given the descriptions of the two types and their potential use in hospital patient data management we suggest that:

- It would be most useful to embed a passive RFID transponder into a patient's hospital wrist band;
- It would be most useful to embed a passive RFID transponder into a patient's medical file (there are several versions and perspectives that we can take no this).

Doctors should have PDAs equipped with RFID or some type of personal area network device. Either would enable them to retrieve some patient's information whenever they are near the patient, instead of waiting until the medical data is pushed to them through the hospital server (there are several versions and perspectives that we can take no this):

- *Active RFID tags* are more appropriate for the continuous collection of the patient's medical data. Since the patient's medical data needs to be continuously recorded to an active RFID tag and an associated

reader needs to be employed. Using an active RFID means that the tag will be a bit bulky because of the needed battery for the write process and there is a concern with radio frequency admissions. Thus, it is felt that an active tag would not be a good candidate for the patient wrist band. However, if the patient's condition is to be continuously monitored, the collection of the data at the source is essential. The inclusion of the tag in the wrist band is the only way to recorder the medical data on a real-time base using the RFID technology. As more organizations get into the business of manufacturing RFIDs and the life and size of batteries decrease, the tag size will decrease and this may be a real possible use.

- *Passive RFID tags* can be also used as well. These passive tags can be embedded in the doctors PDA, which is needed for determining their locations whenever the medical staff requires them. Also, passive tags can be used in patients' wrist bands for storage of limited amount of data- on off-line bases, for example, date of hospital admission, medical record number, and so forth.

After examining both ranges, we can suggest the following:

- *Low frequency range tags* are suitable for the patients' band wrist RFID tags. Since we expect that the patients' bed will not be too far from a RFID reader. The reader might be fixed over the patient's bed, in the bed itself, or over the door-frame. The doctor using his/her PDA would be aiming to read the patient's data directly and within a relatively short distance.
- *High frequency range tags* are suitable for the physician's tag implanted in their PDAs. As physicians use to move from one location to another in the hospital, data on their patients could be continuously being updated.

One final point in regards to the range of RFIDs: until 2002, the permissible radio frequency range was not regulated, that is, it still operated in some low frequency ranges (30- 500 KHz) and in the free 2.45 GHz ISM band of frequency. The IEEE's 802.11b and IEEE's 802.11g (WiFi) wireless networks also operate in the same range (actually there are many other wireless application that operate in that range). This band of frequency is crowded. Where equipment in a hospital is often in the ISM band of frequency, there may be some speed of transmission degradation. The IEEE 802.11n builds upon previous 802.11 standards by adding MIMO (multiple-input multiple-output). MIMO uses multiple transmitter and receiver antennas to allow for increased data throughput via spatial multiplexing and increased range by exploiting the spatial diversity. Note that 802.11n draft 2.0 has been released but the certification of products is still in progress. What this means is that even though 802.11n has greater benefits then previous standards, it is still a draft. The full version is not expected until 2008; thus; products may take several years to be compliant and incorporate that into RFIDs (Hedgepeth, 2007).

Shapes of RFID Tags

RFID tags come in a wide variety of shapes and sizes. Animal tracking tags that are inserted beneath the animal's skin can be as small as a pencil lead in diameter and about one centimeter in length. Tags can be screw-shaped to identify trees or wooden items, or credit card shaped for use in access applications. The antitheft hard plastic tags attached to merchandise in stores are RFID tags (Glover & Bhatt, 2006). Manufacturers can create the shape that is best for the application, including flexible shaped tags that

act like and resemble human skin. RFID tags can be flexible and do not have to be rigid.

Transceivers

The transceivers/interrogators can differ quite considerably in complexity, depending upon the type of tags being supported and the application. T\he overall function of the application is to provide the means of communicating with the tags and facilitating data transfer. Functions performed by the reader may include quite sophisticated signal conditioning, parity error checking, and correction. Once the signal from a transponder has been correctly received and decoded, algorithms may be applied to decide whether the signal is a repeat transmission, and may then instruct the transponder to cease transmitting or temporarily cease asking for data from the transponder. This is known as the "Command Response Protocol" and is used to circumvent the problem of reading multiple tags over a short time frame. Using interrogators in this way is sometimes referred to as "Hands Down Polling." An alternative, more secure, but slower tag polling technique is called "Hands Up Polling." This involves the transceiver looking for tags with specific identities, and interrogating them in turn. A further approach may use multiple transceivers, multiplexed into one interrogator, but with attendant increases in costs (Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster et al., 2007).

Hospital patient data management deals with sensitive and critical information (patient's medical data). *Hands Down polling* techniques in conjunction with multiple transceivers that are multiplexed with each other, form a wireless network. The reason behind this choice is that, we need high speed for transferring medical data from medical equipment to or from the RFID wrist band tag to the nearest RFID reader, and then through a wireless network or a network of RFID transceivers or LANs to the hospital server. From there it is a short distance to be transmitted to the doctor's PDA, a laptop, or desktop through a WLAN IEEE 802.11b, 802.11g, or 802.11n, or wired LAN which operates at the 5.2 GHz band with a maximum data transfer rate exceeding 104 Mbps.

The "Hand Down Polling" techniques, as previously described, provides the ability to detect all detectable RFID tags at once (i.e., in parallel). Preventing any unwanted delay in transmitting medical data corresponding to each RF tagged patient.

RFID TRANSPONDER PROGRAMMERS

Transponder programmers are the means, by which data is delivered to write once, read many (WORM) and read/write tags. Programming can be carried out off-line or online. For some systems re-programming may be carried out online, particularly if it is being used as an interactive portable data file within a production environment, for example. Data may need to be recorded during each process. Removing the transponder at the end of each process to read the previous process data, and to program the new data, would naturally increase process time and would detract substantially from the intended flexibility of the application. By combining the functions of a transceiver and a programmer, data may be appended or altered in the transponder as required, without compromising the production line.

We conclude from this section that RFID systems differ in type, shape, and range; depending on the type of application, the RFID components shall be chosen. Low frequency range tags are suitable for the patients' band wrist RFID tags. Since we expect that a patients' bed is not too far from the RFID reader, which might be fixed on the room ceiling or door-frame. High frequency range tags are suitable for the physician's PDA

tag. As physicians use to move from on location to another in the hospital, long read ranges are required. On the other hand, transceivers which deal with sensitive and critical information (patient's medical data) need the Hands Down polling techniques. These multiple transceivers should be multiplexed with each other forming a wireless network.

PRACTICAL CASES USING RFID TECHNOLOGY

This section explains in details three possible applications of the RFID technology in three applicable cases. Each case is discussed step-by-step then represented by a flowchart. Those cases cover issues as acquisition of Patient's Medical Data, locating the nearest available doctor to the patients location, and how doctors stimulate the patient's active RFID tag using their PDAs in order to acquire the medical data stored in it.

Case I: Acquisition of Patient's Medical Data

Case one will represent the method of acquisition and transmission of medical data. This process can be described in the following points as follows:

1. A biomedical device equipped with an embedded RFID transceiver and programmer will detect and measure the biological state of a patient. This medical data can be an ECG, EEG, BP, sugar level, temperature or any other biomedical reading.
   After the acquisition of the required medical data, the biomedical device will write-burn this data to the RFID transceiver's EEPROM using the built in RFID programmer. Then the RFID transceiver with its antenna will be used to transmit the stored medical data in the EEPROM to the EEPROM in the patient's transponder (tag) which is around his/her wrist. The data received will be updated periodically once new fresh

Figure 4. Acquisition of patient data

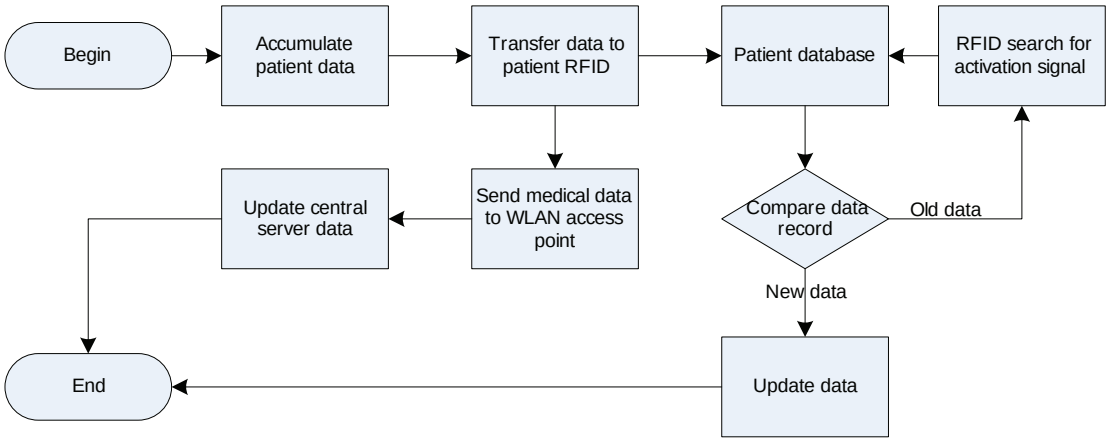
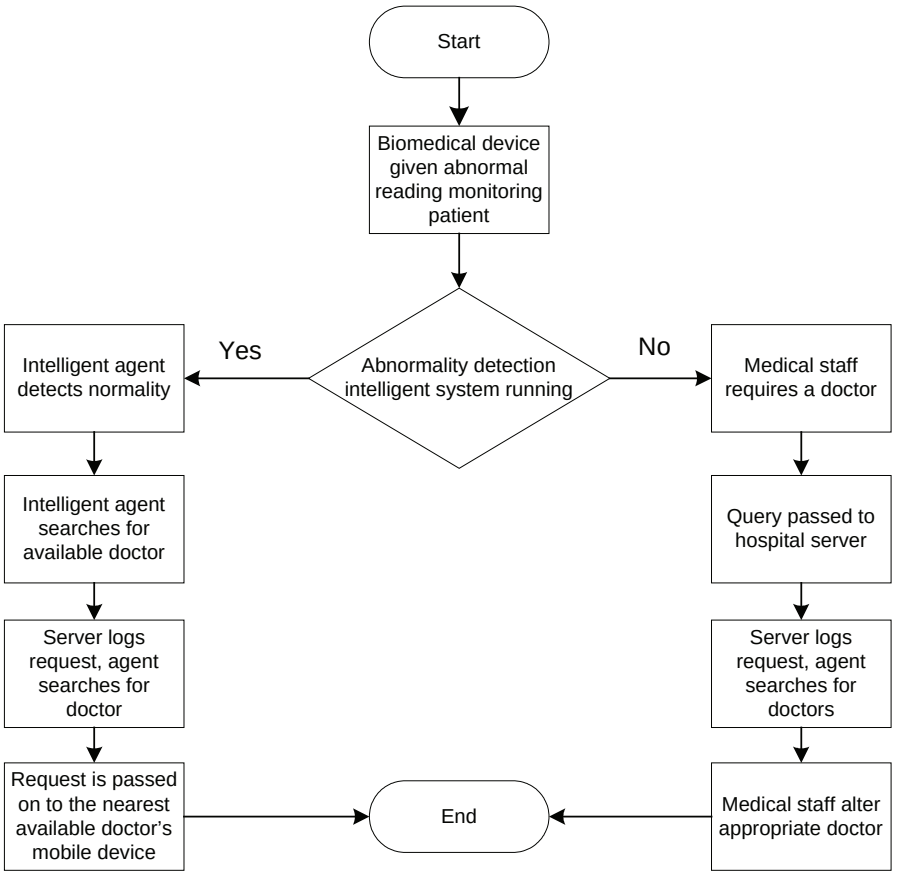


Figure 5. Locating nearest doctor



readings are available by the biomedical device. Hence, the newly sent data by the RFID transceiver will be accumulated to the old data in the tag. The purpose of the data stored in the patient's tag is to make it easy for the doctor to obtain medical information regarding the patient directly via the doctor's PDA, tablet PC, or laptop.

2. Similarly, the biomedical device will also transfer the measured medical data wirelessly to the nearest WLAN access point. Since high data rate transfer rate is crucial in transferring medical data, IEEE 802.11b or g is recommended for the transmission purpose.
3. The wirelessly sent data will be routed to the hospitals main server; to be then sent (pushed) to:
   i. Other doctors available throughout the hospital so they can be notified of any newly received medical data.
   ii. To an online patient monitoring unit or a nurse's workstation within the hospital.
   iii. Or the acquired patients' medical data can be fed into an expert (intel-

ligent) software system running on the hospital server. To be then compared with other previously stored abnormal patterns of medical data, and to raise an alarm if any abnormality is discovered.

4. Another option could be using the in-built-embedded RFID transceiver in the biomedical device to send the acquired medical data wirelessly to the nearest RFID transceiver in the room. Then the data will travel simultaneously in a network of RFID transceivers until reaching the hospital server.

Case II: Locating the Nearest Available Doctor to the Patients Location

This case will explain how to locate the nearest doctor who is needed urgently to attend an emergency medical situation. This case can be explained as follows:

1. If a specific surgeon or physician is needed in a specific hospital department, the medical staff in the monitoring unit (e.g., nurses) can query the hospital server for the nearest available doctor to the patient's location. In our framework an intelligent agent can perform this task.
2. The hospital server traces all doctors' locations in the hospital through detecting the presences of their wireless mobile device; for example, PDA, tablet PC, or laptop in the WLAN range.
3. Another method that the hospital's server can use to locate the physicians is making use of the RFID transceivers built-in the doctor's wireless mobile device. Similarly to the access points used in WLAN, RFID transceivers can assist in serving a similar role of locating doctor's location. This can be described in three steps, which are:
   i. The fixed RFID transceivers throughout the hospital will send a stimulation signal to detect other free RFID transceivers—which are in the doctors PDAs, tablets , r laptops, and so forth.
   ii. All free RFID transceivers will receive the stimulation signal and reply back with an acknowledgement signal to the nearest fixed RFID transceiver.
   iii. Finally, each free RFID transceiver cell position would be determined by locating to which fixed RFID transceiver range it belongs to or currently operating in.
4. After the hospital server located positions of all available doctors, it determines the nearest requested physician (pediatrics, neurologist, and so forth) to the patient's location.
5. Once the required physician is located, an alert message will be sent to his\her PDA, tablet PC or laptop indicating the location to be reached immediately. This alert message could show:
   i. The building, floor, and room of the patient (e.g., 3C109).
   ii. Patient's case (e.g., heart stroke, arrhythmia, etc.)
   iii. A brief description of the patient's case.
6. If the hospital is running an intelligent agent as described in the proposed framework on its server, the process of locating and sending an alert message can be automated. This is done through comparing the collected medical data with previously stored abnormal patterns of medical data, then sending an automated message describing the situation. This system could be used instead of the staff in the patient monitoring unit or the nurse's workstation where nurses observe and then send an alert message manually.

Case III: Doctors Stimulate the Patient's Active RFID Tag Using their PDAs in Order to Acquire the Medical Data Stored in it

This method can be used in order to get rid of medical files and records placed in front of the patient's bed. Additionally, it could help in preventing medical errors- reading the wrong file for the wrong patient and could be considered as an important step towards a paperless hospital.

This case can be described in the following steps:

1. The doctor enters into the patient's room or ward. The doctor wants to check the medical status of a certain patient. So instead of picking up the "hard" paper medical file, the doctor interrogates the patient's RFID wrist tag with his RFID transceiver equipped in his\her PDA, tablet PC, or laptop, and so forth.
2. The patient's RFID wrist tag detects the signal of the doctor's RFID transceiver coming from his\her wireless mobile device and replies back with the patient's information and medical data.
3. If there was more than one patient in the ward possessing RFID wrist tags, all tags can respond in parallel using Hands Down polling techniques back to the doctor's wireless mobile device.
4. Another option could be that the doctor retrieving only the patient's number from the *passive* RFID wrist tag. Then through the WLAN the doctor could access the patient's medical record from the hospital's main server.

RFID technology has many potential important applications in hospitals, and the discussed three cases are a real practical example. Two important issues can be concluded from this section: WLAN is preferred for data transfer; given that IEEE's wireless networks have much faster speed and coverage area as compared to RFID transceivers\transponders technology. Yet, RFID technology is the best for data storage and locating positions of medical staff and patients as well.

The other point is that we need a RFID transceiver & programmer embedded in a biomedical device for data acquisition and dissemination, and only a RFID transceiver embedded in the doctor's wireless mobile device for obtaining the medical data. With the progress the RFID technology is currently gaining, it could become a standard as other wireless technologies (Bluetooth, for example), and eventually manufacturers building them in electronic devices; biomedical devices for our case.

MAINTAINING PATIENTS' DATA SECURITY AND INTEGRITY

Once data is transmitted wirelessly, security becomes a crucial issue. Unlike wired transmission, wirelessly transmitted data can be easily sniffed out leaving the transmitted data vulnerable to many types of attacks. For example, wireless data could be easily eavesdropped on using any mobile device equipped with a wireless card. In worst cases wirelessly transmitted data could be intercepted and then possibly tampered with, or in best cases, the patient's security and privacy would be compromised. Hence emerges the need for data to be initially encrypted from the source.

In this section, a discussion on how to apply encryption for the designed wireless framework for hospital is considered. Suggesting exactly where data needs to be encrypted and\or decrypted depending on the case that is being examined does this.

First a definition of the type of encryption that would be used in the design of the security (encryption\decryption) framework is discussed,

followed by a flowchart demonstrating the framework in a step-by-step process.

Layers of Encryption

Two main layers of encryption are recommended. They are:

Physical (Hardware) Layer Encryption

This means encrypting all collected medical data at the source or hardware level before transmitting it. Thus, we insure that the patient's medical data would not be compromised once exposed to the outer world on its way to its destination. So even if a person with a malicious intent and also possessing a wireless mobile device steps into the coverage range of the hospitals' WLAN, this intruder will gain actually nothing since all medical data is encrypted, making all intercepted data worthless.

Application (Software) Layer Encryption

This means encrypting all collected medical data at the destination or application level once receiving it. Application level encryption runs on the doctor's wireless mobile device (e.g., PDA, tablet PC, or laptop) and on the hospital server. Once the medical data is received, it will be protected by a secret pass-phrase (encryption\decryption key) created by the doctor who possesses this device. This type of encryption would prevent any person from accessing patient's medical data if the doctor's wireless mobile device gets lost, or even if a hacker hacks into the hospital server via the Internet, intranet or some other mean.

Framework of Encrypting Patient's Medical Data

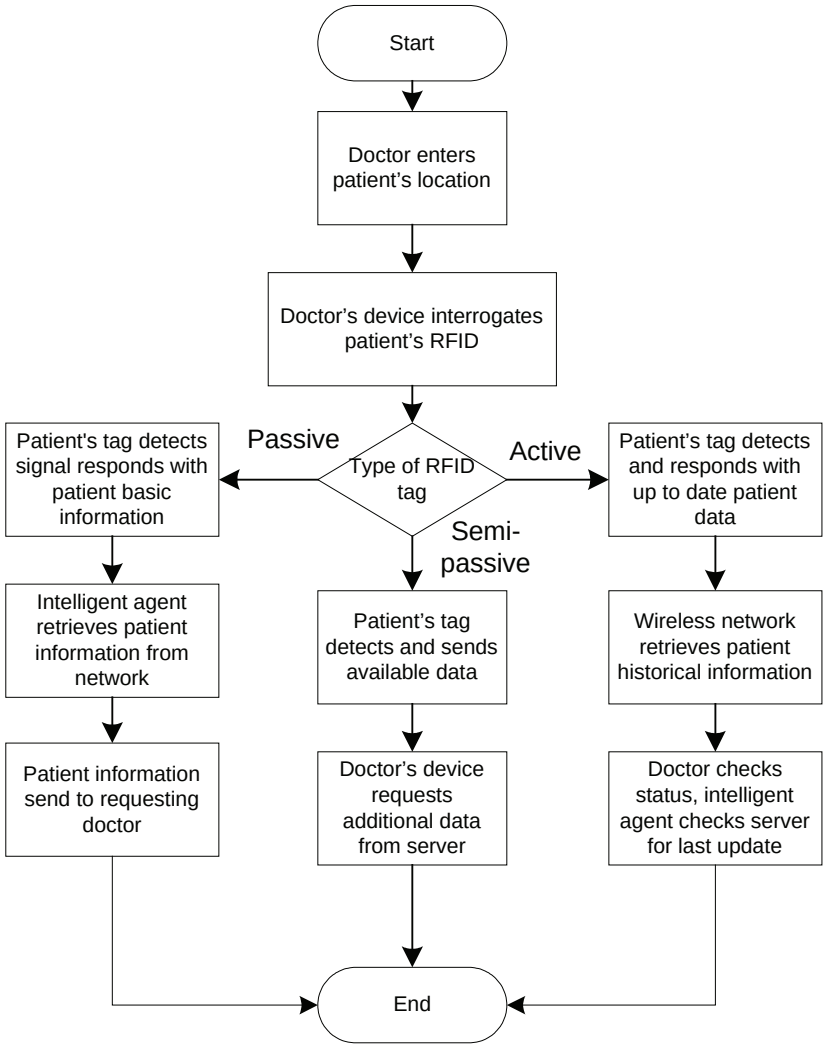
The previous section (Practical Cases using RFID Technology) focused on how to design a wireless framework to reflect how patients' medical data can be managed efficiently and effectively leading to the elimination of errors, delays, and even paperwork. Similarly, this section will focus on the previously discussed framework from a security perspective, attempting to increase security and data integrity.

i. Acquisition of Patients' Medical Data
ii. Doctors stimulating the patient's active RFID tag using their wireless mobile devices in order to acquire the medical data stored in it.

While the third case which was about locating the nearest available doctor to the patients location, is more concerned about locating doctors than transferring patient's data, so it is not discussed here.

The lower part of Figure 6 represents the physical (hardware) encryption layer. This part is divided into two sides. The left side demonstrates the case of a doctor acquiring patient's medical data via a passive RFID tag located in a band around the patient's wrist. The passive RFID tag contains only a very limited amount of information such as the patients name, date of admission to the hospital and above all his/her medical record number (MRN), which will grant access to the medical record containing the acquired medical data and other information regarding the patient's medical condition. This process is implemented in six steps, and involves two pairs of encryption and decryption. The first encryption occurs after the doctor stimulates the RFID passive tag to acquire the patient's MRN, so the tag will encrypt and reply back the MRN to the doctors PDA for example. Then the doctor will decrypt the MRN and use it to access the patient's medical record from the hospital's server. Finally the hospital server will encrypt and reply back the medical record, which will be decrypted once received by the doctors' PDA.

Figure 6. Functional flow



The right side of Figure 6 represents a similar case but this time using an active RF tag. This process involves only one encryption and decryption. The encryption happens after the doctor stimulates the active RFID tag using his PDA which has an in-equipped RFID transceiver, so the tag replies with the medical data encrypted. Then the received data is decrypted through the doctors' PDA.

The upper part of Figure 6 represents the application encryption layer, requiring the doctor to enter a pass-phrase to decrypt and then access the stored medical data. Whenever the doctor wants to access patient's medical data, the doctor simply enters a certain pass-phrase to grant access to either wireless mobile device or a hospital server depend where the medical data actually resides.

CHOOSING LEVEL OF SECURITY FOR THE WIRELESSLY-TRANSMITTED MEDICAL DATA

Securing medial data seems to be uncomplicated, yet the main danger of compromising such data comes from the people managing it, for example, doctors, nurses, and other medical staff. For that, we have seen that even though the transmitted medical data is initially encrypted from the source, doctors have to run application level encryption on their wireless mobile devices in order to protect this important data if the devices gets lost, left behind, robbed, and so forth. Nevertheless, there is a compromise. Increasing security through using multiple layers, and increasing length of encryption keys decreases the encryption\decryption speed and causes unwanted time delays, whether we were using application or hardware level of encryption. As a result, this could delay medical data sent to doctors or online monitoring units.

Figure 6 represents the case of high and low level of security in a flowchart applied to the previously discussed two cases in the last report.

At the end of this section, we conclude that there are two possible levels of encryption, software level (application layer) or hardware level (physical layer), depending on the level of security required. Both physical (hardware) layer and application layer encryption are needed in maintaining collected medical data on hospital servers and doctors wireless mobile devices.

Encrypting medical data makes the process of data transmission slower while sending data unencrypted is faster. We have to have a compromise between speed and security. For our case, medical data has to be sent as fast as possible to medical staff, yet the security issue has the priority.

CONCLUSION

Managing patients' data wirelessly (paperless) can prevent errors, enforce standards, make staff more efficient, simplify record keeping, and improve patient care. In this research report, both passive and active RFID tags were used in acquiring and storage of medical data, and then linked to the hospitals' server via a wireless network Moreover, three practical applicable RFID cases discussed how the RFID technology can be put in use in hospitals, while at the same time maintaining the acquired patients' data security and integrity.

This research in the wireless medical environment introduces some new ideas in conjunction to what is already available in RFID technology and wireless networks. Linking both technologies to achieve the research main goal, delivering patients medical data as fast and secure as possible, to pave the way for future paperless hospitals.

Finally, as reported by Frost and Sullivan, the high cost of radio frequency identification (RFID) technology is a deterrent for health care providers, though RFID has great benefits to hospitals in tracking patients, monitoring patients, assisting in health care administration, and reducing medical costs. With the reduction in cost of radio frequency identification (RFID) technology, increased use of RFID technology in health care in monitoring patients and assisting in health care administration is expected.

REFERENCES

Bigus, J. P., & Bigus, J. (1998). *Constructing intelligent software agents with Java – a programmers guide to smarter applications.* Wiley. ISBN: 0-471-19135-3.

Finkenzeller, K. (1999). *RFID handbook*. John Wiley and Sons Ltd.

Glover, B., & Bhatt, H. (2006). *RFID essentials.* O'Reilly Media, Inc. ISBN: 10-0596009445.

Gruber, T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5,* 199-220.

Guarino, N., Carrara, M., & Giaretta, P. (1994). An ontology of meta-level categories. Journal of knowledge representation and reasoning. In *Proceedings of the Fourth International Conference (KR94)*, Morgan Kaufmann, San Mateo, CA.

Hedgepeth, W. O. (2007). *RFID metrics: Decision making tools for today's supply chains.* Boca Raton, FL: CRC Press. ISBN: 9780849379796.

Lahiri, S. (2005). *RFID sourcebook*. IBM Press. ISBN: 10-0131851373.

Odell, J. (Ed.). (2000, September). *Agent technology.* OMG Document 00-09-01, OMG Agents interest Group.

RFID Australia. (2003). *Why use RFID.* Retrieved February 15, 2008, from http://www.rfid-australia.com/files/htm/rfid%20brochure/page4.html

Schuster, E. W., Allen, S. J., & Brock D. L. (2007). *Global RFID: The value of the EPCglobal network for supply chain management*. Berlin; New York: Springer. ISBN: 9783540356547.

Shepard, S. (2005). *RFID: Radio frequency identification*. New York: McGraw-Hill. ISBN: 0071442995.

Wooldridge, M., & Jennings, N. (1995). Intelligent software agents: Theory and Practice. *The Knowledge Engineering Review,* 10(2), 115-152.

This work was previously published in Encyclopedia of Healthcare Information Systems, edited by N. Wickramasinghe & E. Geisler, pp. 890-905, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.13
Widely Usable User Interfaces on Mobile Devices with RFID

Francesco Bellotti
University of Genoa, Italy

Riccardo Berta
University of Genoa, Italy

Alessandro De Gloria
University of Genoa, Italy

Massimiliano Margarone
University of Genoa, Italy

ABSTRACT

Diffusion of radio frequency identification (RFID) promises to boost the added value of assistive technologies for mobile users. Visually impaired people may benefit from RFID-based applications that support users in maintaining "spatial orientation" (Mann, 2004) through provision of information on where they are, and a description of what lies in their surroundings. To investigate this issue, we have integrated our development tool for mobile device, (namely: MADE, Bellotti, Berta, De Gloria, & Margarone, 2003), with a complete support for RFID tag detection, and implemented an RFID-enabled location-aware tour-guide. We have evaluated the guide in an ecological context (fully operational application, real users, real context of use (Abowd & Mynatt, 2000)) during the EuroFlora 2006 international exhibition (EuroFlora). In this chapter, we describe the MADE enhancement to support RFID-based applications, present the main concepts of the interaction modalities we have designed in order to support visually impaired users, and discuss results from our field experience.

INTRODUCTION

Starting from the European Union cofounded E-Tour project, we designed the tourist digital assistant (TDA) concept and developed multimedia

tour guides on mobile devices (PocketPC and Smartphone devices) for a number of European tourist sites, such as the Costa Aquarium of Genoa, "Strada Nuova" architectonical area and the city of Genoa, the Castellon region in Spain, and the city of Uddevalla in Sweden (Bellotti, Berta, De Gloria, & Margarone,, 2002).

The tour guide provides multimedia contents, added-value information, and location-based services to the tourists. Added-value services are implemented by integrating the mobile devices with additional hardware and software tools such as GPS, electronic compasses, wireless connectivity, digital cameras, written text input, databases, and so forth.

See Figure 1 for snapshots of tourist guide applications.

Relying on the argument that "play is a powerful mediator for learning throughout a person's life," we developed the "educational territorial-gaming" concept in VeGame (Bellotti, Berta, De Gloria, Ferretti, & Margarone, 2003), a computer-supported educational wireless team-game played along Venice's narrow streets to discover the art and the history of the city (see Figure 2), and in ScienceGame (Bellotti, Berta, De Gloria, Ferretti, & Margarone, 2004), a sort of treasure-hunt game inviting players to discover the mysteries and the marvels of the science (see Figure 3) during the "Festival della Scienza" exhibition held in Genoa every year.

These applications were developed from scratch. From these first experiences, we identified common needs and came up with a system to support design of multimedia applications

Figure 1. Snapshots from the Aquarium and Strada Nuova tour guides on PocketPC device

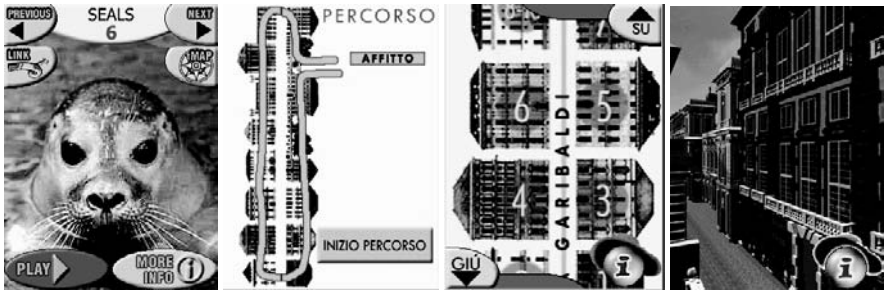
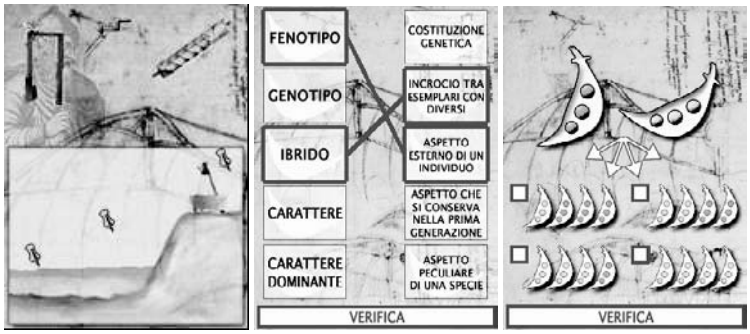


Figure 2. Snapshots from VeGame

Figure 3. Snapshots from ScienceGame



for mobile devices, called Mobile Applications Development Environment (MADE) (Bellotti et al., 2002).

MADE includes M3P (MicroMultiMedia Player), a network-enabled multimedia player easily programmable through the micromultimedia services language (MSL). MSL provides high-level components encapsulating advanced services (e.g., positioning, database query, path search, etc.) that can be easily integrated in multimedia applications. This allows building modular software programs that provide information-rich services to the general public through a coherent and homogeneous HCI that can be learned with low mental workload. On the other hand, MADE hides the low-level aspects of multimedia and service management, allowing designers to focus on the modalities of presentation of information and on user interaction, reducing learning, development, and code maintenance time.

In this chapter, we describe the latest MADE enhancement: we have integrated it with a complete support for RFID detection to allow development of multimedia mobile applications directly connected with the physical world (Want, Fishkin, Gujar, & Harrison, 1999). All low-level aspects of the hardware tag-detection system that are necessary to identify and locate physical objects with attached small RF tags (Want, 2004) are hidden to MSL programmer by the MADE system.

This chapter will also show the use of MADE with the RFID support in a real context such as EuroFlora 2006 international exhibition. This guide differs from others because it has been ad-hoc developed in order to meet strict usability needs. In particular, the novel interface design assists visually impaired people in maintaining "spatial orientation" (Mann, 2004) through provision of information on where they are, hazards that might be in the way, and a description of what lies in their surroundings.

MADE SUPPORT OF RFID TECHNOLOGY

Location-Aware Computing

Recent research has developed several systems, to determinate physical location, that differ by accuracy, cost, and coverage (Boriello, Chalmers, La Marca, & Nixon, 2005). The global positioning system (GPS), which uses signal from satellite to estimate position (Djuknic & Richton, 2001), is

the most used system, but only for applications in outdoor areas. In indoor areas and urban areas with poor sky visibility, the system does not work properly. Moreover, it has a long start-up time.

To overcome these limitations, the first indoor positioning system was the active badge system (Want, Hopper, Falcão, & Gibbons, 1992), which is based on sensors that receive infrared ID broadcast from tags worn by people. This system gives a poor (room-grained) localization precision. After the active badge system, typical indoor location systems are based on radio frequency and on the estimation of position computed from the measured signal strength. Various technologies can be used: Wi-Fi (Howard, Siddiqi, & Sukhatme, 2003), Bluetooth (Bruno & Delmastro, 2003) and nowadays RFID (Liu, Corner, & Shenoy, 2006).

The first two solutions can give an accuracy of around some meters, but require expensive fixed base stations. RFID tags, instead, are very inexpensive and have the same performance. The literature reports also of many location estimation algorithms based on cellular radio networks (Xu & Jacobsen, 2005). However, there is not a generally agreed solution today, and each algorithm has pros and cons, depending on environmental issues. Finally, some vision-based algorithms (López de Ipiña, Mendonça, & Hopper, 2002) are promising because they do not require infrastructure (like tags, satellite, or base station). However, it is difficult to set up a system to locate a user with a 1-meter precision. In the selection of the best methodology for our system, we have taken into account three major issues: the possibility to have a system for outdoor/indoor spaces (like the EuroFlora 2006 exhibition area), a technology with a low cost for the deployment of the infrastructure, and a likely pervasive availability of the system in the near future. All these requirements are satisfied by the RFID technology.

RFID Application Fields

Major RFID application domains include monitoring physical parameters, such as temperature or acceleration, during fragile or sensitive products delivery, monitoring product integrity from factory to retail locations (Siegemund & Floerkemeier, 2003), utilities for home and office automation (Langheinrich, Mattern, Romer & Vogt,, 2000). Nowadays we have passive or active inexpensive RFID (approaching 35 cents today, with a goal of 5 cents (Quaadgras, 2005)) that makes these kinds of sensors practical for tourist applications. For example, a museum exposition can place tags attached to each point of interest so that tourists can receive information about exposition in the right moment at the right place; when near to the object. The research community has actively explored this possibility at the Exploratorium, the interactive science museum in San Francisco. The HP Laboratories researchers have implemented a system that uses three types of identification technology: infrared beacon, barcodes, and RFIDs (Fleck, Frid, Kindberg, O'Brian-Strain, Rajani, & Spasojevic, 2002). In Goker et al. (Goker, Watt, Myrhaug Whitehead, Yakici, Bierig, et al., 2004), a special tag that can work with mobile devices to provide ambient information to users on the move is described. In the Cooltown project (Kindberg & Barton, 2001), RFIDs are used to attach pointers from everyday objects to entities in the computational world. A full exploitation of RFID potentials requires study and implementation of human–computer interaction (HCI) modalities able to support usability of the enhanced mobile tool by the general public. This implies the necessity to resort to programming methodologies and tools specifically dedicated to support the RFID technology. Thus, we have extended the MADE toolkit to support a link between applications and physical world through RFID sensors.

MADE Architecture

A typical MADE application consists of a set of pages containing multimedia and service objects. The micromultimedia services language (MSL) script specifies pages' layout and objects' appearance, synchronization, and user-interaction modalities. MSL scripts are interpreted at runtime by the M3P player that manages presentation of contents and user interaction according to the instructions specified in the input MSL script.

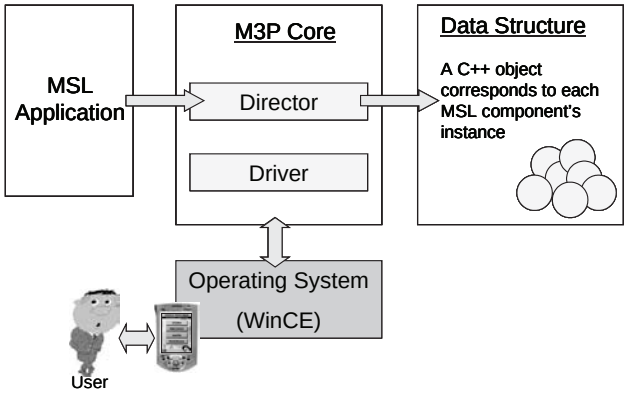
M3P player relies on two-layer architecture (see Figure 4) involving a high-level, platform-independent director and a low-level driver. The director is responsible for creating, initializing, and managing the objects that implement the language functionalities. In order to support incremental development of the player, M3P is composed by a set of modules. In particular, the director has been designed to be independent of the components it manages. According to the object-oriented methodology, this has been achieved by encapsulating the functions of the components in the code of their class, and by structuring the hierarchy so that the director can simply keep a reference to the presentation's pages and convey them events.

According to the instructions specified by the MSL developer in the script, events (either from the system or from user interaction) are conveyed to the director that simply redirects them to the components of the page currently on show, which is the higher-priority choice or, with lower priority, to the other pages of the presentation.

Events are implemented as string messages that are to be interpreted by the target objects. This design choice allows the director's code to be independent of the components and the components to be independent of each other. The basic assumption of this schema is that each component exports a well-defined interface (i.e., a set of messages to which it is able to react) and implements this interface (i.e., implements the reaction to such messages).

Thus, components can be seamlessly added and interchanged (in this last case, as long as they comply with the same interface). Adding a new component (i.e., a new functionality) does not involve any change either in the director's code, or in the other components' code.

Figure 4. MADE architecture



Such a design choice supports easy incremental development, allowing seamless integration of services within a single application framework. This implies that a homogeneous HCI can be applied to an application that hosts several different services that have been developed independently of each other (e.g., intelligent tour planning, interactive maps, positioning, and database access).

MSL relies on a component-based data structure. That is, an MSL file specifies creation of components, attributes of components, and their reaction as a consequence of user interaction. Components are organized in three main libraries: multimedia (e.g., audio, image, video, button), synchronization (utilities like timers that can be used to implement synchronization and scheduling of contents), and services (objects that encapsulate services such as positioning, shortest path search, tour planning, database query, etc).

Every different type of component has its own kind of attributes (fields). The fields record data for specifying the appearance (such as position) and the behaviour (i.e., reactions to events) of the component. In general, components are contained in a special container component, called CARD, that can be thought of as an empty page on which the developer can add components.

The core of M3P involves a platform-independent director and a platform-dependent driver. The director manages the multimedia objects that implement the presentation. Objects are compounded in hierarchical structures. For instance, a CARD (i.e., a multimedia page) may include several images, buttons, and mpeg players. The driver implements the functions to access the hardware, while the director deals with the logic of the multimedia presentation.

Integration of RFID Subsystem

A major feature of MADE consists in the possibility of incrementally adding new hardware and software modules, that are integrated into the HCI framework with no need for modifying the M3P core, since every component's class is responsible for interpreting its receivable messages, independent of the others. MADE can integrate, into a common framework, various hardware modules independently developed to augment the mobile device potentiality. M3P driver's classes, which have to be developed to integrate every new hardware subsystem, manage low-level aspects of the hardware modules, while the MSL interface to the application developer abstracts the services at high level. This implies that a homogeneous HCI can be applied to an application that hosts several different services that have been developed independently of each other (e.g., automatic positioning, intelligent tour planning, and database access can be integrated in an interactive map), and the MSL developer can simply exploit the service modules focusing on the integration of the HCI. Examples of hardware modules already integrated in MADE are a positioning and orientation module that an MSL developer can exploit to get geographical position from a GPS receiver and direction from a digital compass, and the remote communication module able to exploit the hardware available for connection with the external world (e.g., wired/wireless LAN, Bluetooth, GSM/GPRS cellular networks).

In order to enable applications to react to objects in the physical world, our new M3P module, called RFID sensing module (RfidSM), detects presence of RFID tags in the surrounding space and notifies the M3P run-time objects with events implemented as string messages (see Figure 5).

The script interface of the RfidSM is a new MSL component, called RFID, that exposes the fields shown in Table 1.

When the RfidSM component is started, and until it receives a stop event, it scans the surrounding environment to check the presence of tags every "period" of time. The list of detected tags is then sent with the MADE message-exchange modalities to the components specified in the "target" field. In addition, the component has an

Figure 5. Integration (within MADE) of the RFID sensing module (RfidSM)

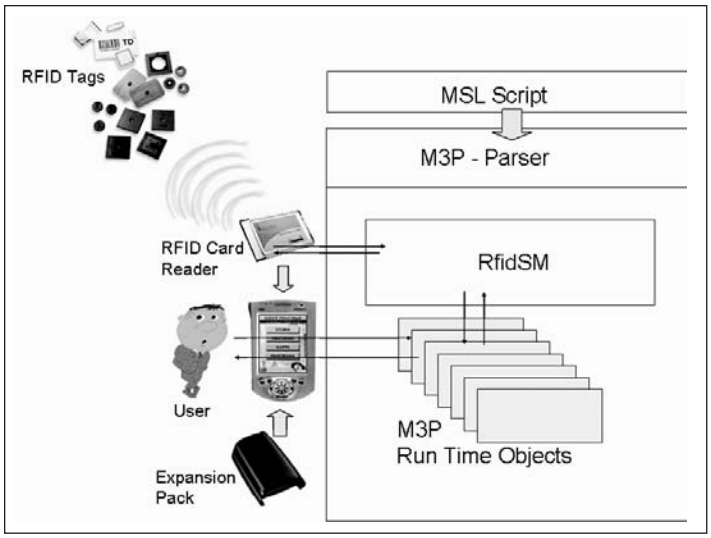


Table 1. RfidSM fields description

Component Field	Description
Target	List of identifiers of the components to which information about identified tags are sent
Period	A time period in milliseconds between two consecutive environmental scans to detect tags
Repetition	A number of tag detection operations executed consecutively on each scanning action
Id	A list of RFID tags that are of interest for the component
Delay	A list of time frames, one for each interesting tag, in which tags are not identified again
dBm	A list of signal strength values, one for each interesting tag, that specify thresholds for tag identifications
onFound	A list of events, one for each interesting tag, that RFID component launch when a tag is identified
Start	If a component launch this event on a RFID component starts the scanning of tags
Stop	If a component launch this event on a RFID component stops the scanning of tags

"id" field to allow programmer expressing interest in a set of tags, and defining (through the field "onFound") the corresponding events list that should be executed. Each interesting tag is also featured with a signal strength threshold (through the field "dBm") that specifies a limit under which the tag is considered in range.

There is the problem of collisions, since the scan results are typically imperfect due to not all tags are detected in every scan. To solve this problem, a tag typically awaits a random number of time slots before it answers the RF pulse sent by the reader. However, the problem still remains and grows as the number of tags in the surrounding

environment grows. The MADE RFID sensing module tackles this issue, allowing the programmer to specify, through the field "repetition," a number of times that the reader should repeat the scanning before returning to the founded tags. The list of founded tags is the collection of all tags observed in each scan. A small value of repetition results in a fast scan with high risk of collision, whereas large repetition value results in a slow scan with few collisions. This trade-off should be resolved by the programmer basing his decision on application constrains: long delays can result in human-computer interaction problems if the application allows a user expectation for immediate reaction to tags. This is the case of applications in which user a voluntarily accosts the mobile device to tagged objects to obtain information. Similar problems arise if the application has a short time frame to detect tags, for example, in applications where the user moves at relatively high speed in the environment, like in territorial games. Instead, others type of applications can gain advantage from precise but slow detections. It is the case of a tourist mobile guide for a naturalistic park in which the user moves along a path with some points of interest largely spaced, like tree species or rare flowers.

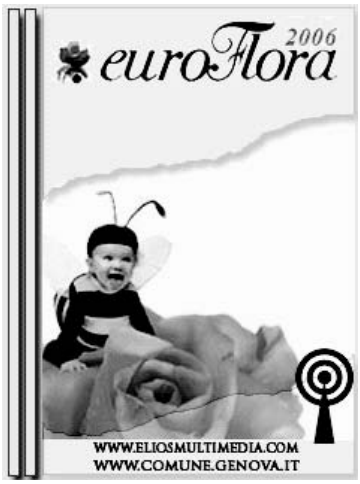
The other problem affecting the RFID technology is the "tag detection flickering" (Römer, Schoch, & Mattern, & Dubendorfer, 2003): due to the collision problem, some tags can appear and disappear from sequential scanning, generating a fast list of tag identifications. The MADE RFID sensing module allows the programmer to decide how to convert scan results into applications events handling this problem. Programmer can specify, through the "delay" field, a time period (for each interesting tag) starting from the detection. During this time, subsequent detection events of the same tag are discarded; also, the exact definition of this delay is application dependent. Applications with events that occur only one time, like tourist guide for museums with linear path, can have delay values set to infinite. Instead, in applications with events generated multiple times closer each other, like territorial games, the delay should be short or zero.

Currently, we have implemented the low-level driver support for the iCARD Identec reader in a PCMCIA card format (IDENTEC). This card can be integrated in handheld, portable, or laptop computers to communicate with the iQ and iD active RFID tags at a distance of up to 100 meters. The RF signal is in the UHF radio band (915 MHz or 868 MHz), providing long-range communication and high-speed transmission rates for reliable data exchange.

THE EUROFLORA GUIDE

In order to assess the possibility of using RFID technology to develop widely usable interfaces, we present a real-world application (see Figure 6) developed through the MADE toolkit and deployed in an ecological environment (fully op-

Figure 6. Snapshot of the cover page of EuroFlora Guide application



erational, reliable, and robust application, used by real users and in a real context of use) at EuroFlora 2006 (the international flower exhibition that is held in Genoa every 5 years). With over 500,000 visitors in 10 days, EuroFlora is one of the most important exhibitions of Europe.

The developed application concerns the research area of assistive technologies for visually impaired people. Such assistive applications have the potential to improve the quality of life of a large portion of population (by 2020, there will be approximately 54 million of blind persons over age 60 worldwide (WHO, 1997)).

In this field, maintaining spatial orientation is a major challenge for people with visual impairment. There is the need of systems in providing blind people with information on where they are, hazards that might be in the way, and a description of what lies in their surroundings (Mann, 2004). The notion of "spatial orientation" refers to the ability to establish awareness of space position relative to landmarks in the surrounding environment (Guth & Rieser, 1997). The goal of our application is to support functional independence to visually impaired people, providing support to indoor awareness of elements in the surroundings (Ross, 2004).

The EuroFlora guide is organized in two parts. One part provides general information about the exhibition, the guide, and their services. The other part provides the description of the selected interest points. While first part is directly accessible by the user at any moment, the second one is event driven. More precisely, every interest point description is associated to an RFID tag, and when a user enters that area (i.e., her/his handheld device recognizes the RFID tag), the software asks the user whether to launch the corresponding description.

We placed 99 RFID sensors on an area of 30,000 mq of exhibition, covering 99 points of interest, services, and major areas (see Figure 7). RFID sensors were IP65 compliant in order to resist to water and dust, and self-powered. Power level of sensors could be set in two levels, low and high.

Design Methodology

The necessity for combining the flexibility and multimedia potential of a mobile device with the extreme simplicity of interaction, required for use by a wide audience (also visually impaired people), involves facing three main HCI issues:

Figure 7. a) The packaging of the multimedia guide in a leather case; b) Snapshots from the tests: users visit EuroFlora 2006 supported by the guide and touch some dedicated plants

- **Usability by general users:** The tourist has little time and willingness to learn how to use the new technological tool, since she or he is there to visit the exhibition and not to learn a tool. Most of the tourists use such a tool for the first time and just for a short time (typically, from 30 to 90 minutes). Thus, the complexity of the platform should be hidden from visitors, making the guide immediately usable, with no effort by users. This implies that the interface is to be as simple and intuitive as possible.
- **Usability by visually impaired people:** Visiting an exhibition is a critical task for the blind, mainly for the combination of several reasons: the site is often crowded and unfamiliar to the visitor, it may be noisy, it is difficult to orientate in a highly dynamic place. In this context, the guide should be not intrusive, with few and very recognizable input interface elements (also with tactile feedback), and should give information in a proactive modality when needed by the user.
- **Presentation of information:** Added-value information (e.g., how the various specimens live in their natural environment) should be synergistic with the direct experience of the visitor at the exhibition. Provision of information has to be structured in order to enhance the direct perception of the visitor, leading to a better and more pleasant comprehension of her/his surrounding environment. For example, the guide should make use of environmental sound (e.g., waterfall) and scent (e.g., flower smell) to connect content information and the objects in the space.

We have tackled such issues resorting to the methodologies of the user-centric design (Carroll, 1997), in an iterative development of the guide involving participatory design (Beck, 1993), definition of usability specifications, and contextual design, as shown in the following:

- Participatory design consisted of the participation of botanists at the design decisions, authors skilled in writing for blind people and visually impaired end-users, together with technical developers. The most significant contribution of the first three categories consisted in the definition of the targets and in the concrete perspective they brought into the project.
- Usability specifications provide explicit and measurable targets to verify the suitability of the work done. Examples of such goals are "90% of the users should be able to operate the guide without asking questions to the personnel," "90% of the users should be able to use the interface with no errors," "90% of the users should be able to understand the meaning of all the touchable controls within 120 seconds." All these objectives were verified in early lab and field tests in order to take the appropriate corrective actions.
- Contextual design involved early field tests with experts and users at the exhibition in the preopening days when the infrastructure of EuroFlora was being built. Field tests have been helpful to highlight problems and shortcomings that had been overlooked or ignored in lab.

Structure of the Interface

The interface of the EuroFlora guide has been designed to support immediate use by the general public, also by visually impaired people. To this end, we used general design principles (we already described them in the introduction) such as overall simplicity, low intrusiveness, and support for natural interaction and knowledge acquisition. Moreover, we added further features in order to meet the specific needs of visually impaired people:

- Tactile feedback in the control interface
- Tutorial stage
- Event-driven interface
- Facilities to support orientation

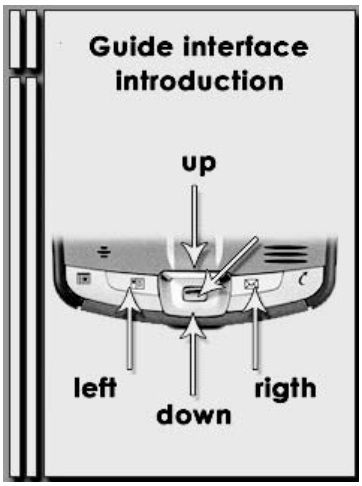
The basic element of the interface is the multimedia card. A multimedia card corresponds to each subject of a presentation (e.g., a flower species). Each multimedia card provides, in an audio format, texts specifically written for visually impaired people (i.e., highlighting olfactive and tactile sensorial information, providing detailed ambient descriptions).

The tactile feedback is necessary to allow impaired people to easily understand the position of the controls and give her/him feedback. Our previous multimedia guides had the interface embedded in the graphic contents, exploiting the touch screen of a pocket-pc device. During the early field tests, visually impaired people pointed out some important shortcomings in these solutions. They felt that the screen was too large and their fingers were lost in a space without roughness. Since most of such users are well acquainted with the common cell phones' relief keyboard, we studied a new solution exclusively based on the hardware buttons (Figure 8).

The hardware buttons of the interface are highlighted. The choice of this area as navigation control allows visually impaired people to have a tactile feedback. The meaning of the button is "up" to accept a content description (which is automatically triggered when the user enters a new cell), "down" to reject it, "right" to exit from a section of the guide, and back to the main menu, "left" to have contextual information about user's current position.

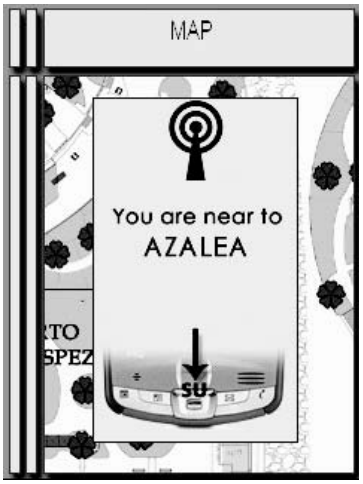
The tutorial stage is an initial guide section in which users could freely experiment withthe interface of the tool in order to allow people to use the guide in an independent way. In this stage, users are invited to freely press buttons. A speech description briefly explains the meaning of each pressed button. This tutorial stage prevents the necessity for providing papers or long explanations when users rent the guide.

Figure 8. Snapshot of the guide instruction



The event-driven interface allows a user to get information about points of interest (POIs) and orientation when they are in the proximity of a POI. For example, in Figure 9, a user near the azalea stand is told about the presence of this flower by a pop-up window (the guide is to be usable by everybody) and a corresponding jingle sound. If she/he wants to listen to this content, she/he can press the "up" hardware button. By default, the system skips the presentation. This operational mode has low intrusiveness (users are asked whether to listen to a content), but it also provides a certain degree of proactivity. Information is not only botanical, as in the example of azalea, but also concerns the positioning of the user. Many tags are placed in the proximity of facilities, such as lavatories, cafés, exits, and intersections. This localization system lowers the mental workload necessary for the tourist to synchronize the physical space of the exhibition with the virtual space of the program.

Figure 9. Snapshot of the event-driven interface



This message (accompanied by a jingle) is shown to the user when she/he has just entered a POI area. The combination of audio and graphics is due to the fact that the guide may be used also by not visually impaired people. In the example in this figure, the user is near to the azalea flower, and if she/he is interested in the description she/he can press the "up" hardware button to access the related content.

One of the main tasks of the guide is to assist the visitor in her/his exploration of the exhibition space. A facility to support orientation (not useful for visually impaired people) is a section with the map that helps the tourist to orient herself /himself in the exhibition. The map (see Figure 10) shows the structure of the EuroFlora, including lavatories, café, exits, and so forth, and the location of the points of interests. In order to enhance user's orientation, the map is centered on the position of the user, as determined by the currently perceived RFID tags.

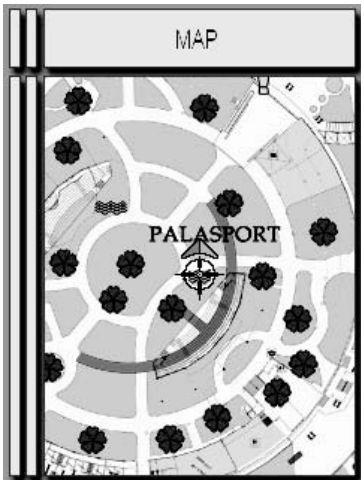
FIELD EVALUATION

Experimental Framework

Real evaluation of advanced mobile device applications and of the impact on their intended population is difficult and costly. Evaluation requires analysis of real users, in a real context of use. In order to adequately evaluate interaction with computation resources, test-people should use a fully operational, reliable, and robust tool, not just a demonstration prototype (Abowd & Mynatt, 2000). Hence, it is important to perform early tests in the authentic context of use in order to verify end-user acceptance and overall usefulness of the system, and to receive feedback to inform future design.

In this chapter, we describe the early stage analysis of acceptance and usefulness of the developed multimedia guide. The tests were performed at EuroFlora 2006, the international flower exhibition that is held in Genoa every 5

Figure 10. Snapshot of the map section of EuroFlora Guide



years. With over 500,000 visitors in 10 days, EuroFlora is one of the most important exhibitions of Europe.

The exhibition area (around 90,000 squared metres) was equipped with an infrastructure of 99 RFID tags. The experimentation involved 120 visually impaired tourists who used the tour guide, and were observed and interviewed by expert of disabilities and HCI designers. Subjects were visually impaired (25%) or blind persons (75%) representing a wide range of age (28% age <30; 32% age between 30 and 50; and age >50 40%). Moreover, the tests involve 64 females and 56 males.

The tour guide consisted of a PocketPC equipped with a special leather package with a lace dropping from the neck for a more comfortable use. Headphones were also used in order to isolate the user from the highly noisy surrounding environment (see Figure 1).

Preexhibition Tests

In an early test session-performed 2 days before the official opening of the exhibition, when some stands were already readyenabling a realistic test-we prepared a prototype software version that was used by five selected visually impaired users visiting 30% of the total exhibition area. We followed and interviewed the users in this phase, in order to understand shortcomings, defects and weaknesses, and strong points of the product. In this phase, we understood and solved some problems on user interface and contents, such as the most suited assignment of buttons to presentation control functionalities and the length of the descriptions. Some test-users found the long silence time between a presentation activation and the next one (i.e., the period of time in which the user is walking through areas not covered by RFID tags) frustrating. We partially tackled this issue by periodically providing a message saying that the user is currently in an area not close to a POI.

Ecological Tests

One hundred and twenty blind people used the guide during the exhibition. Sixty of them (aged from 12 to 78 years old) participated in an ecological test conducted by the authors. We interviewed the users at the return of the guide. We evaluated three main performance factors: usability (including effectiveness, efficiency and pleasantness of use), usefulness, and capability to support spatial orientation (in particular the approach to the POIs). We asked users to give a general comment on the guide and a 1-5 grade for each factor (which was carefully explained and defined by the interviewers). An overall survey of results is reported in Table 2; it clearly shows the high acceptance by the users.

Analyzing the variables' correlations based on the chi-square test, we observed that usability is correlated with the perceived support for spatial orientation (χ=25.3, df (degree of freedom) = 16, 90% confidence), and that perceived utility of the tools is strictly correlated with perceived support for spatial orientation (χ=30.2, df=16, 99.9% confidence). This suggests the importance of our design choice to use mobile technology to support orientation of visually impaired people. Moreover, test results also show that the tool is perceived as useful and usable.

Considering the free comments, the guide was judged as an excellent tool for users to orientate themselves inside the exhibition. Several people expressed a similar concept, which we can synthesize with the words of one visitor: "after always having been guided, for the first time I myself have been able to guide my wife and to explain the exhibition!" Such positive comments were also confirmed by the blind assistance experts, who highlighted the significant degree of independence the blind could reach through the guide.

Shortcomings in the interface were reported by some elderly users, while some people asked for more extended descriptions, though each point of interest included at least one. The high

Table 2. Overall survey results

Issue	Average	Standard Deviation
Usability	4.00	0.64
Usefulness	4.25	0.75
Support for spatial orientation	4.20	0.66
Session length time	201 minutes	30 minutes

performance and reliability of hardware, software, and batteries assured long sessions of use with no troubles for the user.

FUTURE TRENDS AND VISION

The research community is envisaging a new model of a "tagged world" as an intelligent environment that allows providing visually impaired people with information about architectonical barriers, safe paths, points of interest, potential danger areas, and other useful information. A sample scenario description may give an idea of this likely future.

Maria is visually impaired. She is in a foreign city on a business trip. Maria owns a mobile device with a mobility-assistance system (MAS: it is similar to the EuroFlora Guide, but with a much larger action range). The MAS accompanies her in her path to her destination office, and signals pedestrian crossings, traffic lights, safe paths in work-in-progress areas, and so forth. All objects in the world send their signals, but Maria's wearable device has an intelligent reasoning algorithm (based on user preferences and interpretation of the user's current activity) and a suitable human-computer interaction (HCI) in order to provide her only with the needed information. This information is extracted from a mass of data that are continuously received from the close-by RFID tags. Thus, the wearable device notifies Maria about a pedestrian crossing only if it knows that this is useful for her current activity (i.e., going to office). Not useful information will not be provided, in order not to distract Maria. Along her path to her destination, Maria passes by a newsagent. The World Guide scans all the magazines and identifies today's issue of Maria's favourite magazine. It queries Maria's database, which replies that Maria has not purchased this issue yet; so, it notifies her about the opportunity to buy the magazine.

CONCLUSION

The ubiquitous presence of smart tags will offer, in the near future, a critical mass of information, embedded in the world, that will be exploitable to rethink the relationships between people involved in their daily-life activities and the surrounding world.

With MADE we have designed a system that continuously scans the tagged world, interprets the large amount of information coming from the surrounding objects, and provides it to the users through multimedia human-computer interaction. Moreover, the application in the future will filter the raw data coming from the environment (with

artificial intelligence behaviour) taking into account the user needs, preferences, and profile.

The field test at EuroFlora 2006 has demonstrated the feasibility of our vision, by deploying the system in a real-world setting (an exhibition area with indoor and outdoor instrumented environments), and performing extensive field tests with real users. In a longer-term view, with such an application, we intend to investigate the future scenarios that will be enabled by a massive presence of RFID tags in our environments. This "early prototyping" has allowed us to understand, as early as possible, costs, limits, strengths, and benefits of the new technology. We have also obtained a significant positive feedback on user acceptance. Usability results show that the guide is perceived as highly usable and useful, in particular because of its ability to support spatial orientation.

The next step towards a "tagged world" will require integration of data and services, and capability of interpreting a variety of sources according to the specific and dynamic user needs. Achieving these goals will involve a huge research effort that will be successful only if it will lead to the deployment of compelling applications that will be perceived as useful by the users. In a user-centered design view, this implies a rapid prototyping of applications and extensive user testing in the real context of use, which was our inspiring principle in the EuroFlora project.

REFERENCES

Abowd, G. D., & Mynatt, E. D. (2000). Charting past, present, and future research in ubiquitous computing. *ACM Transaction in Computer-Human Interaction, 7*(1), 29-58.

Beck, A. (1993). User participation in system design: Results of a field study. In M. J. Smith, *Human-computer interaction: Applications and case studies* (pp. 534-539). Amsterdam: Elsevier.

Bellotti, F., Berta, R., De Gloria, A., Ferretti, E., & Margarone, M. (2003). VeGame: Field exploration of art and history in Venice. *IEEE Computer, 26*(9), 48-55.

Bellotti, F., Berta, R., De Gloria, A., Ferretti, E., & Margarone, M. (2004). Science game: Mobile gaming in a scientific exhibition. eChallenges e2004. *Fourteenth International Conference on eBusiness, eGovernment, eWork, eEurope 2005 and ICT.* Vienna.

Bellotti, F., Berta, R., De Gloria, A., & Margarone, M. (2002). User testing a hypermedia tour guide. *IEEE Pervasive Computing, 1*(2), 33-41.

Bellotti, F., Berta, R., De Gloria, A., & Margarone, M. (2003). MADE: Developing edutainment applications on mobile computers. *Computer and Graphics, 27*(4), 617-634.

Borriello, G., Chalmers, M., La Marca, A., & Nixon, P. (2005). Delivering real-world ubiquitous location systems. *Communications of the ACM, Special Issue on The Disappearing Computer, 48*(3) 36–41.

Bruno, R., & Delmastro F., (2003). Design and analysis of a bluetooth-based indoor localization system. *Personal Wireless Communications, 2775*, 711-725.

Carroll, J. M. (1997). Human-computer interaction: Psychology as a science of design. *International Journal of Human-Computer Studies, 46*(4) 501-522.

Djuknic, G. M., & Richton, R.E. (2001). Geolocation and Assisted GPS. *IEEE Computer, 34*(2), 123-125.

EuroFlora. (s.d.). EuroFlora 2006 Home Page. Retrieved from http://www.fiera.ge.it/euroflora2006/index_eng.asp

Fleck, M., Frid, M., Kindberg, T., O'Brian-Strain, E., Rajani, R., & Spasojevic, M. (2002). From Informing to Remembering: Ubiquitous Systems

in Interactive Museums. *IEEE Pervasive Computing, 1*(2), 11-19.

Goker, A., Watt, S., Myrhaug, H. I., Whitehead, N., Yakici, M., & Bierig, R. (2004). An ambient, personalised, and context-sensitive information system for mobile users. *Second European Union Symposium on Ambient Intelligence.* Eindhoven, The Netherlands.

Guth, D. A., & Rieser, J. J. (1997). Perception and the control of locomotion by blind and visually impaired pedestrians. In *Foundation of orientation and mobility* (2nd ed.) (pp. 9-38). AFB Press.

Howard, A., Siddiqi, S., & Sukhatme, G. (2003). *An experimental study of localization using wireless Ethernet.* International Conference on Field and Service Robotics

IDENTEC. IDENTEC Solution Ltd Home Page. Retrieved from http://www.identec.com

Kindberg, T., & Barton, J. (2001). A Web-based nomadic computing system. *Computer Networks, 35*(4), 443-456.

Langheinrich, M., Mattern, F., Romer, K., & Vogt, H. (2000). *First steps towards an event-based infrastructure for smart things.* PACT 2000. Philadelphia.

Liu, X., Corner, M.D., & Shenoy, P. (2006, September). *RFID* Localization for pervasive multimedia. In *Procedings of the 8th International Conference on Ubiquitous Computing (UbiComp '06).* California, USA.

López de Ipiña, D., Mendonça, P., & Hopper A. (2002). TRIP: A low-cost vision-based location system for ubiquitous computing. *Personal and ubiquitous computing, 6*(3), 206-219

Mann, W. C. (2004). The aging population and its needs. *IEEE Pervasive Computing, 3*(2), 12-14.

Quaadgras, A. (2005). Who joins the platform? The case of the RFID business ecosystem. *38th Hawaii International Conference on Systems Science* (pp. 855-864). Gig Island (HI): IEEE Computing Society Press.

Römer, K., Schoch, T., Mattern, F., & Dubendorfer, C. (2003). T-smart identification frameworked for ubiquitous computing applications. *PerCom 2003.* Fort Worth: IEEE Press.

Ross, D. A. (2004). Cyber crumbs for successful aging with vision loss. *IEEE Pervasive Computing, 3*(2), 30-35.

Siegemund, F., & Floerkemeier, C. (2003). Interaction in pervasive computing settings using Bluetooth-enabled active tags and passive RFID technology together with mobile phones. *PerCom 2003.* Fort Worth: IEEE Press

Want, R. (2004). Enabling ubiquitous sensing with RFID. *IEEE Computer*, 84-86.

Want, R., Fishkin, K., Gujar, A., & Harrison, B. (May 1999). Bridging physical and virtual worlds with electronic tags. *ACM Conference on Human Factors in Computing Systems (CHI 99).* Pittsburgh, PA.

Want, R., Hopper, A., Falcão, V. & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems, 10*(1), 91-102.

WHO, W. H. (1997). *Who sounds the alarm: Visual disability to double by 2020.* Retrieved from http://www.who.int/archives/inf-pr-1997/en/pr97-15.html

Xu, Z., & Jacobsen, H. A., (2005). *A framework for location information processing.* 6th International Conference on Mobile Data Management (MDM'05). Ayia Napa, Cyprus.

KEY TERMS

Chi-Square Test: The Chi-square is a test of statistical significance for bivariate tabular analy-

sis (crossbreaks). This test provides the degree of confidence we can have in accepting or rejecting a hypothesis.

Ecological Context: The ecological context is a set of conditions for a user test experiment that gives it a degree of validity. An experiment with real users to possess ecological validity must use methods, materials, and settings that approximate the real-life situation that is under study.

Human-Computer Interaction: Human–computer interaction (HCI), also called man-machine interaction (MMI) or computer–human interaction (CHI), is the research field that is focused on the interaction modalities between users and computers (interface). It is a multidisciplinary subject, relating to computer science and psychology.

Location-Aware Computing: Location-aware computing is a technology that uses the location of people and objects to derive contextual information with which to enhance the application behaviour. There are two ways to acquire information about user context: requiring the user to specify it or by monitoring users and computer activity. Sensor technology, such as RFID, could enable mobile devices to extract information from user position automatically.

Mobile Tourist Guide: A mobile tourist guide is a software application with an intuitive interface, that provides users with multimedia information when and where needed during their visit to museums, city centres, parks, and so forth. Such an application runs on PDA-type terminals or on cellular phones, and could be augmented with GPRS (general packet radio service), GPS (global positioning system), and Bluetooth wireless technology. The guide allows tourists to plan routes according to preferences and ambient conditions (weather, timetables, sites of special interest, etc).

Radio Frequency Identification: Radio frequency identification (RFID) is an automatic identification method based on storing and remotely retrieving data using small and cheap devices called RFID tags or transponders. An RFID tag is an object that can be attached to objects, products, or persons to identification using radio waves. Passive tags (with a few centimeter range of sensitivity) require no internal power source, whereas active tags (with more long range of sensitivity, 100 meters) require a power source.

User-Centric Design: User-centric design is a design process that aims at realizing products that meet users' expectations. The key idea of this design methodology is to start the design strategy taking into account the user's perspective.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 657-672, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.14
Matching Dynamic Demands of Mobile Users with Dynamic Service Offers

Bernhard Holtkamp
Fraunhofer Institute for Software and Systems Engineering, Germany

Norbert Weißenberg
Fraunhofer Institute for Software and Systems Engineering, Germany

Manfred Wojciechowski
Fraunhofer Institute for Software and Systems Engineering, Germany

Rüdiger Gartmann
University of Münster, Germany

ABSTRACT

This chapter describes the use of ontologies for personalized situation-aware information and service supply of mobile users in different application domains. A modular application ontology, composed of upper-level ontologies such as location and time ontologies and of domain-specific ontologies, acts as a semantic reference model for a compatible description of user demands and service offers in a service-oriented information-logistical platform. The authors point out that the practical deployment of the platform proved the viability of the conceptual approach and exhibited the need for a more performant implementation of inference engines in mobile multi-user scenarios. Furthermore, the authors hope that understanding the underlying concepts and domain-specific application constraints will help researchers and practitioners building more sophisticated applications not only in the domains tackled in this chapter but also transferring the concepts to other domains.

INTRODUCTION

Regarding the trend towards ubiquitous computing and ambient intelligence, modern information systems basically have to support mobile users. As a first step towards fulfilling dynamic demands of mobile users, the concept of context-awareness has been introduced to enable filtering of information based on user-specific context information.

To cope with user acceptance, we abstract from context information and use a situation model. Situations are easy to understand for a user and can be derived from a set of context information, including location and time and even user profile information and other sources. They are named cognitive abstractions of context. When such situations are linked with user goals (e.g., get food when hungry), it is evident that different situations imply the need for different information and services to help a user in achieving his goals. User profiles are used for describing personal data, preferences, and interests of individual users, from which user goals can be derived.

Furthermore, we observe a growing demand to cope with dynamic service offers. Service-oriented architectures mainly integrate Web-based services from different providers. One consequence is the need to cope with unavailability of services, for example, due to broken connections or limited scopes of service validity. To enable an automatic replacement of services, that is, service roaming, service profiles are used that provide for a matching with user profiles and context information.

To enable matching of dynamic user demand and service offers on a semantic level, we use semantic technologies. This includes the development of a description model for service semantics and a semantic registry able to cope with such descriptions. The service ontology is modular, based on other ontology modules covering general concepts, situations, and the application domains. As demands from a large number of users are to be matched dynamically with service descriptions provided by a large number of service providers, the application ontology acts as a semantic reference system.

In the following, we start with the discussion of the conceptual background of our approach, followed by an outline of sample application scenarios. In the main part, we discuss the construction and use of the application ontology as a basis for a semantic matching of demand and offers and give an overview of the system architecture supporting this process. A brief summary of practical experiences gained from the deployment of the system as a mobile tourist guide follows. The chapter closes with a look at future trends.

BACKGROUND

Following Dey (2001, p. 5), "context is any information useful to characterize the situation of an entity. An entity is a person, place, or object considered relevant to the interaction between a user and an application, including the user and application themselves." Context-aware applications are able to adapt their functionality based on existing context information towards the user's environment. This includes filtering and provision of information and services being of interest to the user in his specific context, thus making applications more proactive and reducing the need for explicit user interactions. This property is of value especially for mobile applications due to the restricted interaction capabilities of mobile devices. Mobility always has a location aspect that is an important part of almost any context-aware application. In this way, mobile computing and context-awareness are good supplementations in order to provide users with the right information anywhere and anytime.

Research on context-aware applications started in the beginning of the 1990s. One of the first applications was the Active Badge System (Want, Hopper, Falcao, & Gibbons, 1992) from Olivetti Research Lab. It allowed users to locate people

in the office and to redirect incoming calls to the closest phone. This system was later in operation at Olivetti STL, Xerox EuroParc, MIT Media Lab, and Xerox PARC.

The Conference Assistant (Dey, Salber, Abowd, & Futakawa, 1999) was developed at the Georgia Institute of Technology. Its aim was to assist conference attendants. Based on user profiles including a list of research interests, the Conference Assistant displays the timetable with events highlighted that are of interest for the user. When entering a room, the Conference Assistant gives information on the presenter and shows the presentation material. The user can then make notes during the presentation, which are recorded together with additional context information, for example, the time, author, and content information useful for later retrieval.

Another type of popular context-aware applications is location-based tourist guides. There are quite a number of examples available from the mid 1990s. The Cyberguide project (Long et al., 1996) from Georgia Tech was aimed at providing information to a tourist based on his position and orientation. The user could see his position on a map. Selecting points on the map the user could get more information about his environment. A similar guide has been developed by the University of Lancaster and tested between 1996 and 1999 for visitors of the City of Lancaster (Davies, Mitchell, Cheverest, & Blair, 1998). Based on location and user preference, the visitor could get information about points of interest in the region.

Both these tourist guides were restricted to location information. The COMPASS2008 project, which has been realized based on the information-logistical service platform described in this chapter, also aims at assisting tourists in providing suitable information and services dynamically in each situation. In contrast to the previous projects, the COMPASS2008 application is not restricted to the location of the user as a context dimension. Another context dimension is time. Dependent on the current time, different time aspects are inferred, for example, activity of the user, eating time, opening hours of shops, and so forth. In addition, parts of the user profile and even external sources like event calendars are used as context dimensions. Based on the complex context model, a set of situations is derived to provide the basis for a personalized situation-aware filtering of information and services for a user.

Most of the above described applications are prototypes developed in research labs and the academic world. There are not many complex context-aware commercial solutions. However, there are several location-based services offered by mobile communication providers and service providers. Examples are route planning services, city guides, hotel and shopping guides, and location services for nearby gas stations. Most of the commercial and academic applications only use a few context dimensions, mostly location, time, and identity.

APPLICATION SCENARIOS

In this section we describe and analyze scenarios from different application domains, namely tourism and emergency management, regarding their requirements on dynamic information and service supply, ontology support, and relevant context information and user profile data.

Tourism

One application domain for ontology-based service provision is guidance for tourists. Since the behavior of tourists is not predictable and depends on various influencing factors such as personal moods, personal interests, and so forth, an intelligent system should be able to conclude the user's current needs. Furthermore, information relevant for tourists is provided by many different sources.

To be able to decide what information could be relevant for a tourist, it is crucial to detect in

which kind of situation the user currently is in. Context information and user profile information are the basis for that decision. For instance, a user standing in front of a sports stadium in which a sports event is about to start could be there accidentally, could intend to see the event, or could look for tickets. If the user profile indicates no interest in this kind of event, the first option is probably correct, otherwise the user profile could indicate whether the user has a ticket for that event. This information would either lead to provision of, for example, a navigation service to the right entrance and further information about the event (e.g., starting lists), or the system would offer an online booking service.

The COMPASS system (Weißenberg, Voisard, & Gartmann, 2006) being based in the technical infrastructure described in this chapter has recently shown in a field test in Beijing that situation detection is applicable to tourist guide systems and that this is a basis for an intelligent selection of appropriate services, which unburdens the user from searching for desired services among huge offers.

Emergency Response Support

Support for emergency response is a very demanding task. Emergency cases are always different and unpredictable, information needs depend on various parameters, and response times are always critical. A precise efficient demand-specific filtering of information from a huge information offer is needed.

The information needed highly depends on the kind of emergency and a precise recognition of the emergency case is crucial for information selection. Typically, not all relevant parameters are known initially. For instance, fighting a fire in a chemical plant is influenced by the chemical substances stored or processed here. Thus, information about the emergency case is completed gradually, and the system has to refine the provided information accordingly.

An ontology-based service selection, based on situation information such as the type of emergency and context data, is very effective to meet the mentioned requirements. An actual example is the MONA system, the Mobile Emergency Assistant (Holtkamp, Weißenberg, & Speckmann, 2005), developed for the Duisburg fire brigades. This system is fed information about the emergency, such as the location and situation (fire in different levels of escalations, car accident, rescue of jammed persons, and so forth). It has access to all internal information sources of the fire brigades and can additionally access external Web services such as geospatial mapping services to get further information if necessary. That leads to an improved information supply for the officers-in-charge and leads to a more efficient mission processing.

APPLICATION ONTOLOGY

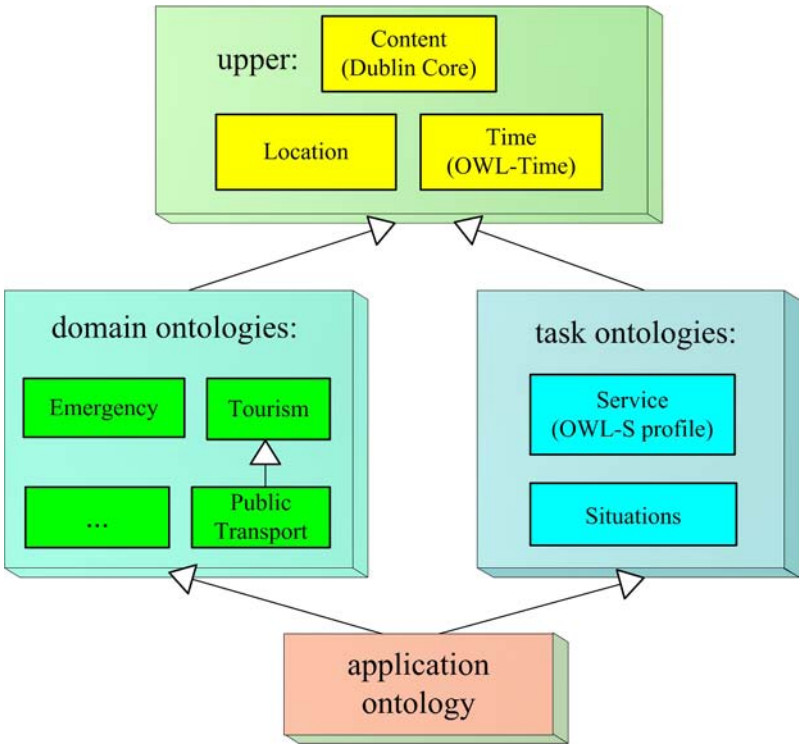
As a basis for semantic matching of dynamic offers with dynamic requests in all scenarios and as a base data model, we use a modular extensible application ontology for the description of both demands and offers. In a first step, the ontology structure is explained. Then dynamic aspects of the ontology are discussed in more detail, such as changes in user profiles, contexts or service sets, and service roaming. Finally, we have a closer look at the implementation side, including performance issues.

Ontology Structure

Following Guarino (1998), an ontology can be structured into different kinds of subontologies as depicted by Figure 1:

- The *upper ontology* is limited to generic and abstract concepts, independent of, and thus addressing a broad range of, application domains. It covers reusable dimensions like

Figure 1. Modular ontology architecture



location, time, and content, which may be refined in other ontologies.

- *Domain ontologies* specify concepts of different application domains and scenarios (e.g., tourism, emergency) and may refine concepts from the upper ontologies. For new application scenarios, mainly new domain ontologies are needed.
- *Task ontologies* code knowledge about the usage of domain ontologies, that is, they characterize computational aspects. They make generic use of domain ontologies, that is, they are independent of special domain ontologies.
- The *application ontology* at the lowest level integrates all other ontologies for the application.

Upper Ontologies

Ontology design has to keep in mind for what the ontology is used. In our case, it is intended for dynamic personalized service provision to a huge amount of concurrent users in scenarios as defined above, that is, for information logistics. Existing upper or top-level ontologies like OpenCyc (Cycorp, 2002) and SUMO (Niles & Pease, 2001) are too large for our scenarios. Hence, we took a more pragmatic approach, concentrating on the main aspects of time, location, and content, which are seen as the main dimensions of information logistics (Deiters, Löffeler, & Pfennigschmidt, 2003).

Location Ontology

For mobile applications, the location aspect is of utmost importance and the use of ontologies for this purpose has been long-praised. Fonseca, Egenhofer, Davis, & Borges (2000) summarize several such approaches. Our location ontology, however, is not a complete geo ontology, but pragmatically provides basic concepts which may be refined in (geo) domain ontologies. There are two layers: the logical or *cognitive* location concepts and their lower-level *geographic extent*, both inheriting from the root concept *Location*. The root concept of the logical layer is *Location-Name,* having subconcepts like *Country, Region*, and *AdministrativeArea,* the latter having subconcepts like *State* and *City*. Multiple inheritance is used to model entities such as *Municipalities*, being a city and a state. A tourism ontology might add concepts like *POI* (point of interest), *Hotel, Shop,* and *Restaurant,* and an emergency ontology might add concepts like *Plant* and its various parts. Instances of the higher-level concepts (i.e., the known locations) are mapped to lower-level concept *GeographicExtent,* having subconcepts like *Point, Box,* and *Polygon*. For example, a *Restaurant* instance is mapped to a *Polygon* instance having some set of points with coordinates (e.g., by using spatial extensions of a database). Using geographic relationships like containment and overlapping at the lower level, corresponding relationship for the higher-level instances can be inferred.

Time Ontology

To define temporal aspects of services and situations, a time ontology is needed. Our time ontology is structured similarly to the location ontology: both have an abstract and a physical layer. The lower physical layer is a subset of OWL-Time (Hobbs & Pustejovsky, 2003), consisting of the *TemporalEntity* subconcepts *Instant* and *Interval*, together with basic relationships (*after, before).* The additional abstract layer with root concept *PeriodicInterval* is mapped to the lower layer by timestamp patterns, which play the role of coordinates. It has subconcepts like *Yearly* and *Daily.* For example, *Yearly* is instantiated by *January*, representing the month occurring every year periodically, not a concrete month as in the lower layer. Instantiations of *Daily* concepts may even be personalized, depending on a user's context (e.g., *Sunrise*) and preferences (e.g., *Lunchtime, Dinnertime,* and *Morning*), or may be object-dependent (e.g., *TradingHours* of a shop). The personalized time *abstraction* method accesses the user profile and yields a set of known logical time concepts for a user for the current time. While the lower level of our time ontology is based on OWL-Time, the higher level is a simplification of concept *CalendarDescription*, found in some versions of OWL-Time. In the OpenCyc upper ontology it is called *RegularlyRepeatedEvent.*

Content Ontology

For the content dimension, Qualified Dublin Core (Kokkelink & Schwänzl, 2002) is often used, which is mainly a refinement of Dublin Core (DCMI, 2004), providing access to information and services at document metadata level.

Domain and Task Ontologies

The modular application ontology is open for different domain ontologies to be added, to support different scenarios. For example, a tourism ontology has different kinds of POIs, restaurants, hotels, and the like, and may use an ontology of public transport. The domain ontologies are mainly used as value pool for different properties in all other ontologies.

There is no separate *user profile ontology*, but it consists of different domain ontologies covering the interests and preferences of users in the application domain. The profile values stemming from a separate system (e.g., from LDAP)

are interpreted using the knowledge of domain ontologies. Additional rules may be meaningful when some profile attributes are to be inferred by others.

Also task ontologies can be added when needed. The main task ontologies in our case are the service ontology and the situation ontology.

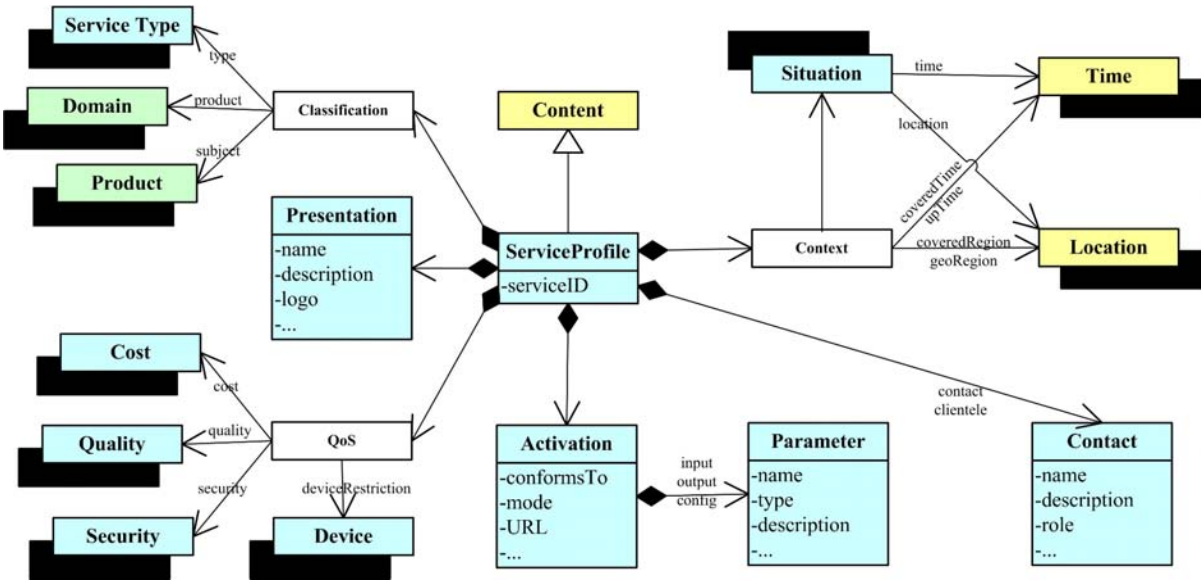
Service Ontology

Service (and information) advertisements and demands are described independently by different user groups (i.e., service provider and user), not knowing exactly the needs of each other. The most flexible way to match demands against offers is to use semantic technologies, that is, ontologies and inference. The *ServiceProfile* is a subconcept of concept *Content* having Qualified Dublin Core properties. Thus, services are special content, having content properties and additionally service-specific properties. Registered services are facts connected by their properties to instances or concepts of other ontologies, which serve as dimensions. The top level ontologies *Location* and *Time* as well as all domain and task ontologies are used as value space for different service properties. This results in a star schema, in which service properties are characterized by subprofiles, which again are characterized by their properties. For example, the *geoRegion* of a specific service may be characterized by a location instance, which itself is characterized by a set of properties and the cost aspect may be characterized by a cost subprofile.

A simplified sample service ontology of this construction is sketched by Figure 2, which is both a relational schema and an ontology, since shadowed boxes indicate the root concepts of subontologies used for a dimension. Only some sample properties are shown here.

Service retrieval is supported by different multivalued classification properties. The values

Figure 2. Modular service ontology schema



are concepts from different domain or task ontologies. Such rather orthogonal classifications together already cover much of the semantics of a service:

- **General classifications** using a service type taxonomy (property *type*), a product taxonomy (for *products* related to the service), and concepts from other domain ontologies used as service subjects (property *subject*).
- **QoS classifications** summarize nonfunctional quality of service aspects, for example, characterizing of main factors of service cost, quality, security, and possible device restrictions in the case of locally installed services (running on PDAs or home gateways).
- **Context** aspects support restricting a service to a multidimensional context and even to detected situations. This includes location and time properties for service validity (service accessibility, that is, *geoRegion*, *upTime*) and service coverage (for its result, that is, *coveredRegion*, *coveredEpoch*). An example is to find services callable *now* (validity is actual context) but delivering information (coverage) for a restaurant or event to visit this evening.
- **Contacts** and **clientele** summarize relational contact data of different parties involved in the service, like service provider and call center, as well as the clientele (for multiclientele ability).

For *presentation and activation* of retrieved services, the following aspects are specified (not necessarily in an ontology but relational, since these aspects are seldom needed for service retrieval):

- **Presentation** covers all information needed to display retrieved service at the user's frontend, and includes multilingual information such as a service name or description and possibly icons. For use of services by programs, this aspect is not relevant.
- **Activation** provides all information needed to call the service, such as an *URI*, the communication standard used (property *conformsTo*), the *mode* (active, passive, etc.) and the parameters. Since we currently do not use a process model for the service, our grounding is simple.
- **Parameter** describes each input, output, and configuration parameter of the services, for example, its name, type, and a brief description, to be used for GUI generation and parameter marshaling.

The service ontology is influenced by standards like OWL-S (Martin et al., 2004), a Web service ontology submitted to the W3C, but we focus on the service profile. However, OWL-S is not the only approach towards Web service ontologies. The Web Service Modeling Ontology (WSMO) has also been submitted to W3C and a third approach is Semantic Web Services Ontology (SWSO). There is a migration process for these approaches. However, they focus on the process model, while their service profile is only basic. In contrast, the service ontology presented here focuses on the service profile, since it is used for demand-based service retrieval primarily.

Situation Ontology

Situation detection requires a modular ontology with subontologies for all context dimensions, all combined in the situation ontology. Thus, the situation ontology extensively depends on its dimension ontologies. It consists of a hierarchy of situation profile concepts, instantiated to characterize different situations. Situations are described by semantic situation profiles, being named sets of characteristic features of situations. Situation descriptions instantiate these concepts by defining a semantic classification of an aggregate of abstracted user context, user profile, and related

information and can be inferred from these. At a given time, a user may be in zero, one, or many situation known by the system.

Ontology Usage

The application ontology is used as follows: Whenever an event occurs that might influence the service set of a user, the set is recalculated by situation detection and service matching (and possibly service roaming). These significant events include:

- **Context changes**: Our application scenarios are mobile, thus location changes occur often. We have developed an algorithm to detect significant location changes based on geographic extents of logical location concepts used in service registrations. This algorithm runs on mobile devices. Any user's GPS locations are only transmitted by his phone or PDA to the context server when a *significant* change has occurred and a new significance specification is returned. The same mechanism can be applied to other context dimensions as well. In this way, frequent context changes of mobile users lead only occasionally to service set recalculations. In a field trial in Beijing discussed in this chapter, only 86 significant context changes were produced by 15 users in approximately 4 hours in total.
- **User profile changes**: Whenever a user entered or modified his user profile, an event is fired, causing the system to react by calculating new situations and an appropriate new service set.
- **Service set changes**: Whenever a new service is provided to the public or whenever a service is changed or removed, an event is fired, causing service set recalculation.
- **Ontology changes:** Our inference engine is RDBMS-based (i.e., inference is directly executed in the relational database management system), thus ontology changes are controlled by the database and valid for the next semantic query, which is triggered.

Situation Detection

Situation detection occurs when context sensors report *significant* context changes. The abstraction and aggregation mechanisms of all dimensions are used to obtain a set of instances of the higher-level situation concepts. For example, not only the location and time may characterize a situation but also whether an action takes place at that location and time, by consulting for example a social event directory service or weather service. The resulting situation request profile is then semantically matched to all situation profiles known to the system, leading to a (possibly empty) set of situations fulfilling the request profile. A user may be in any of these situations or may be interested in being in this situation. For example, if a user is in a filled stadium, situation *Watching Competition* is detected. If it is personal lunchtime, the additional situation *Eating in Stadium* provides corresponding service offers. Only well-defined situations can be detected. If a user is in a situation not known by the system, he will only get situation-independent support.

Service Matching

The service selection process is a semantic matching of an implicitly dynamically constructed service request profile against the profiles of all known services found in the semantic registry. The request profile uses the matching situations determined previously. User profile and context are also used in the request profile, for example, to select only services matching the user's interests at the current location. Some user preferences are mapped to service types or service subjects, which requires fine-grained taxonomies of these kinds, being related to the preferences hierarchy. Other preferences are used for a personalized instantia-

tion of time ontology concepts like *Morning* and *Lunchtime*, which are used to infer situations. The matching evaluates different types of semantic relationships for all profile properties, like subclass, instance, and containment relationships. Different matching strategies can be realized by defining different semantics when using our *ModelAccess* component described below. For example, for a class-valued property (e.g., property *type* or *subject* in our service ontology), it can be defined whether it matches with subclasses, with superclasses, or with both, and to what semantic distance.

Service Roaming

Mostly, services are defined for a certain scope, which could be a geographical area they cover or certain timeframes or situations for which they are useful. Obviously, a service scope can be described based on restrictions defined on context attributes. Service matching regards the actual user context and adds it to the service request profile in order to select only services which scope covers this context. Based on context-specific service selection, service roaming aims at providing certain service functionality to a user constantly during changing contexts. Whenever the actual user context leaves the scope of the used service, a new service instance with similar functionality but with a scope fitting to the new context has to be found and invoked transparently for the user. An example is a service offering parking information for a certain city. Such a service could, for example, be used in a navigation application. If a user leaves one city and enters another, the navigation application is automatically disconnected from the currently used service and connected to the one covering the area of the new city.

Implementation Aspects

Currently, our extensible ontology comprises about 300 concepts with 1900 properties and 900 instances (among them the registered services and locations), divided into several top-level, domain, and task ontologies. It is completely stored in relational database tables, combined with relational data and accessed from all subsystems by our *ModelAccess* component. Multi-user access is controlled by a sophisticated RDBMS. We have done extensive load testing of the new architecture, which proved to scale well with the number of users (concurrent threads), the size of the ontology, and the size of the answer set. The numbers can be summarized as follows: With 100, 000 registered service profiles and about the same number of related entities as above retrieval times of about a second are achieved on a 3.2 GHz, 2 GB RAM PC, of course depending on the complexity of the query.

SYSTEM ARCHITECTURE

To support scenarios like those described, we have developed an information-logistical semantic service platform. We give a gross outline of the system architecture and the interoperation of its subsystems for semantic matching of user demands and registered offers.

Use of Profiles

All subsystems use profiles for describing entities in their interface methods. A profile is a structured set of properties covering different dimensions and characterizing an entity. Each property has a type (range) and may have several values. Samples are user profiles, service profiles, situation profiles, location profiles, and device profiles. They may be used for characterizing offers and also for describing an actual demand of such entities, and are thus a basis for semantic matching of both.

Profiles may be interpreted either directly or semantically. In a semantic profile, the range and values of the attributes are semantic categories

stemming from an ontology, which forms background knowledge for interpreting the syntactic (e.g., relational) data. The ontology can also be used to guide the process of creating or modifying profile properties (e.g., to offer allowed value for properties), and to assure consistency of stored profile properties with the ontology.

Gross Architecture

The information-logistical semantic service platform provides basic functionalities needed for intelligent demand-specific selection and provision of information and services. The key technologies used and combined by the platform are the following:

- **Personalization** is used for the selection of services and information according to a user's profile, preferences, interest, and other user-related information, and for the adaptation of filtered information to the user's needs.
- **Context and situation awareness:** Context information includes any relevant information about the user's state and his environment including the derived situations. It can be used to retrieve the information needed at the location and time or in the situation the user currently is in.
- **Open infrastructure:** The service platform uses and builds on top of existing open and distributed service infrastructures. This enables the dynamic use and selection of already existing information and services.
- **Mobile computing:** A mobile device provides users with information and services everywhere. This allows further integration of the platform in the user's daily life and work processes. The restricted communication and interaction capability of mobile devices, as well as communication costs, reliability, and security aspects have to be considered when selecting suitable presentation forms and delivery strategies.
- **Information-logistical evaluation:** All described key technologies are combined by application-specific evaluation knowledge. The evaluation component of the platform controls the appropriate selection and presentation of services and information. This may even include business processes, being out of scope for this chapter.

Main Components

The logical architecture defines overall system functionality as a cooperation of different subsystems providing specialized tasks related to the key technologies. The subsystems are application-independent and can also be used stand-alone. They all build on services and models of the core layer. As sketched, the platform developed consists of the following main components:

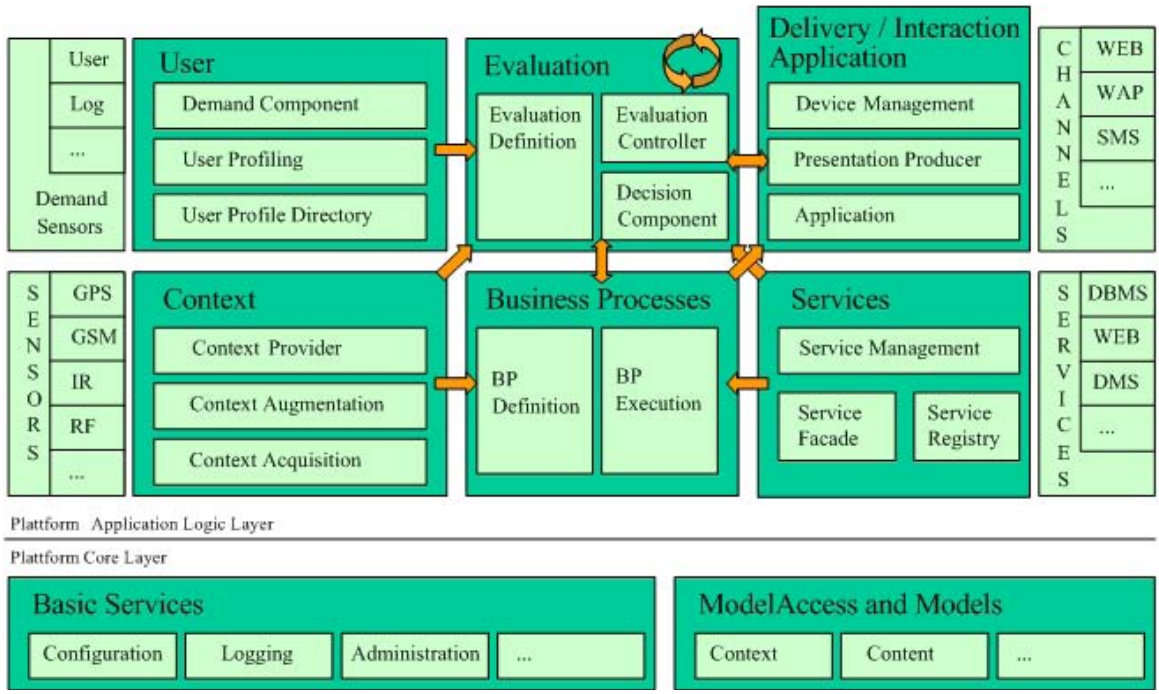
ModelAccess

Inference engines are not yet as mature as relational database systems. Especially, they are not as fast and only some of them support multithreading, which is a prerequisite for having a large number of concurrent users. Therefore, we developed a *ModelAccess* component with integrated basic and pragmatic inference support. The principle idea of *ModelAccess* is the generation and execution of *closed* structured query language (SQL) queries from semantic profiles based on ontologies that map the original semantics. The generation is based on an extensible set of registered parameterized SQL parts. Predefined parts for standard relationships like *subtype, instanceOf, partOf, geoContains* exist, and rules or characteristics for new kinds of relationships can be stored as user-defined SQL parts. The parts are selected and composed based on stored semantic metadata of table properties. The *closed* queries generated only need simple transactions handled

by the RDBMS automatically. The *ModelAccess* component was built with the main design goals to support dynamic online multi-user access to semantic data, to support efficient retrieval on voluminous persistent semantic data, and to combine relational concepts with semantic features by an abstract model access layer and runtime-engine used within all subsystems of the service platform.

Most inference engines and ontology design tools today enable to store ontologies in an RDBMS. For example, Protégé provides to store ontologies as Protégé database and others even offers to use database tables during inference (which often makes inference slower, and is intended for large ontologies that cannot completely be kept in main memory). However, *ModelAccess* is not the first inference engine using a relational database directly for inference. In Das, Chong, Eadon, and Srinivasan (2004), an approach is described to implement inference on top of an Oracle RDBMS. Semantic queries are formulated using SQL with additional operators, based on a general schema for storing concepts, properties and relationships. Due to performance reasons they perform an initial materialization of all OWL axioms (e.g., subproperties and transitivity of properties) after loading the ontology, followed by individual inferences for each semantic query. The differences to our approach are: they need initial materialization of, for example, transitivity (computing the transitive closure) to get meaningful performance and directly work on SQL level, while we use transitive discriminates (no transitive closure) and add semantic profiles and SQL generation as an optional abstraction level. In Chong, Das, Eadon, and Srinivasan (2005), the same authors describe a similar approach to realize inference based on resource description framework (RDF). Performance is optimized by

Figure 3. Gross architecture of the semantic service platform



providing indexed materialized views on the two tables of their normalized schema, which *results* in a table design similar to ours. Chong et al. (2005) conclude, "a promising storage representation is *partial* normalization" (2005, p.12).

Service Subsystem

This subsystem is responsible for selecting content and services. It is an open service infrastructure and provides functionality for management and provision of services. This includes the semantic description of services by definable ontology schemas, as described above. The semantic *ServiceRegistry* (implemented by *ModelAccess*) provides dynamic ad hoc integration of services of different kinds from third party providers. It allows for retrieval of registered services and customized service offers. The *ServiceFacade* supports use of services of any kind, provides basic mechanisms for controlling access rights and billing, and enables service roaming. Not only passive services are supported, but also subscribing active services which may fire events.

Context Subsystem

This subsystem provides functionality for definition and provision of application-specific context models, which are a machine readable representation of the part of the world relevant for an application, based on a context meta model. Normally, this model includes the user and other items like locations and relations between these objects. The subsystem provides functionality for detection and provision of context information for any entity. This includes ad hoc integration, management, selection, and use of distributed context sensors, the derivation of related context information and the detection of situations (to enable situation-awareness). The subsystem follows a layered component architecture that separates the different aspects of context detection, integration, refinement, model management and operation, and context information access.

User Subsystem

This subsystem defines the infrastructure for the personalization of any application, used for centralized management and provision of user information. This includes the provision of user data (relatively static user information like the user's name and gender), user preferences (like preferred language or modality), and user interests (in specific topics). Complex application-specific user model can be defined. Securing privacy of such information is important here. In the current version of the subsystem a lightweight directory access protocol (LDAP) server is used as the basis for management and provision of user information. It also includes a framework for integration of user profile learning components.

Evaluation Subsystem

This subsystem provides information-logistical evaluation logic. The jobs residing in the evaluation subsystem are value-added services defining the main application-specific interoperation patterns between the subsystems. They enable the evaluation of any resources, for example, information and services, its relevance for a specific user, and a delivery strategy. The subsystem executes in cooperation with the other subsystems, thus delegating special evaluation jobs to them, for example, notification of relevant context events or changes of service sets. Reusable scenarios abstract from the implementation of evaluation jobs within the other subsystems. The current implementation of the job model is based on a Boolean ECA-paradigm: events can be subscribed from the other subsystems. The condition is a Boolean expression on events that lead to the action part defining a coordinated execution of functionality using the other subsystems.

Instantiation for Application Scenarios

All application scenarios described above have common requirements on a supporting platform, namely information-logistical support by dynamic information and service supply, based on registered offers, relevant context information, and user profile data. These are the objective of the semantic service platform.

Instantiation of the platform for a new application possibly added to the set of already existing applications (ability to clientele processing) begins with extending the data models (namely ontologies). Each subsystem can be instantiated by configuration and by describing its application-dependent behavior using a domain-specific language. Then the platform component's application programming interfaces (APIs) are used by the new application, and services as well as application-specific jobs have to be registered.

PRACTICAL EXPERIENCES

As the information-logistical semantic service platform is implemented and operational, we have gained first practical experiences regarding ontology integration and performance. Having used the platform as a basis in the COMPASS2008 system that aims at providing visitors of the Olympic Games 2008 in Beijing with personalized situation-aware services, we are also able to include experiences from a recently performed field test in Beijing. During that field test the users were equipped with GPRS- und GPS-enhanced MDA Pro PDAs with the COMPASS front-end.

The field test was conducted July 8-10, 2006, in Beijing. To enable testing of the COMPASS system under real-life conditions, we defined a test scenario for Beijing. This scenario aims at covering the most common situations visitors (especially Western foreigners) experience when visiting Beijing. Regarding the resources available in the project we had to restrict the test to a part of Beijing, for example, to limit the content needed.

The test users represented Beijing Olympic visitors, that is, they acted in the role "tourist". For the field test we had foreign users from the U.S. Korea, Japan, and Italy as well as Chinese, thus representing native and non-native English speakers. Of the 15 users, 12 were male and 3 were female. Their average age was about 26 years. Approximately 40% of the users had experience in the use of PDAs and/or smartphones. The foreign users mostly had only little knowledge of the Chinese language and were to some extent familiar with Beijing. Six foreigners were in Beijing for one month or less.

In the next step, the test users were sent out to perform the test. Each user had to fulfil 16 tasks, some of them repeatedly, for example, find restaurant/coffee shop, order a drink, plan the day, pay, communicate destination to taxi driver, buy an item, and find restrooms. A COMPASS2008 team member accompanied each test user to assure a proper conduction of the different tasks.

Situation-aware service provision was considered useful by most subjects. No test person criticized it, and most people even desired to improve the feature. From the COMPASS system perspective we obtained the following results regarding the quality of the situation awareness feature: recall 95.8% (correct situation identified, only relevant services offered), precision 51.0% (only correct situation identified, redundant services offered). The recall value looks quite satisfying for the first deployment in the field. An analysis of the reason for wrong situation identifications turned out that deviation of the GPS signal led to mismatches of a user's position and the geographical locations of points-of-interest.

During the field test, the semantic service platform was integrated into the Internet infrastructure of a Chinese Internet service provider. It was fully available with no problems in providing its functionality. Even though there was a problem

in the context subsystem leading to an unnecessarily high number of context change events, there was no visible delay in providing the user with situation-specific services.

A problem faced was the unreliable and imprecise location detection of a user through GPS. Even when a user did not move, the GPS detector reported a position change of more than 30 meters. Another problem was the unreliability and limited performance of data communication using general packet radio service (GPRS) in China. Beside these problems, the technical performance of our platform during the field test was satisfying.

FUTURE TRENDS

Currently we observe a convergence of communication and information infrastructures towards Internet-based systems. This convergence eases the integration of information flows into applications. This trend is backed by service-oriented architectures where system components communicate via the Internet, residing at arbitrary locations. Consequently, information overflow of users will intensify. Customer satisfaction can only be achieved when an intelligent information supply is provided, taking care of individual user needs.

A broader adoption of semantic registries and situation-aware demand description as standards enables a guided use of offered services and service-specific contents in individual user contexts. This leads to higher acceptance on the user side and at the same time enables the forming of value chains on the business side, as service providers can establish networks where each provider covers a specific service offer that seamlessly integrates with others.

In summary, user profiles, context-awareness, and semantic service descriptions provide the basis for a demand-driven personalized information and service logistics using multistage value chains.

CONCLUSION

In this chapter, we tried to point out that ontologies and their evaluation are well suited to define a semantic reference system accessed by large heterogeneous user groups. In the COMPASS2008 project, an application ontology for Beijing Olympics tourists has been developed. On this basis a dynamic semantic matching of user demands derived from user profiles and context-driven situation detection with semantically described service offers is performed. The COMPASS pilot system proved the applicability of these concepts for situation-aware semantic Web applications in a field test in Beijing.

Although the results from the field test are satisfying, the entire development was not as smooth as it might seem. We had to solve problems on all levels, including nonsynchronized funding on the Chinese and German sides, content procurement, and system integration or technical problems when setting up the field test. Some of these problems are intrinsic to international projects with multiple partners; others are more specific, like the restricted access to digital maps in China or the impreciseness of GPS in a mega city like Beijing. The cooperative atmosphere within the consortium, however, was a major success factor. A more detailed discussion of the problems sketched above is beyond the scope of this contribution. Here we focused on ontology related issues.

The COMPASS system provides for user-individual demand description and situation-aware service filtering in the context of the Olympic Games 2008 in Beijing. The MONA prototype deploys the same technologies for situation-specific information and service provision in emergency cases, supporting emergency response teams on the spot. Here, situation awareness is used for a more precise content selection.

The insights gained from the deployment of semantic Web technologies show that they are very powerful and helpful for the development

of adequate models. On the implementation side, however, mass applications are still a critical issue as the performance of inference engines falls short compared with relational database technology. Hence, for larger applications we recommend use of ontology development tools for the conceptual phase and transferring the result in the implementation phase to a database solution.

ACKNOWLEDGMENTS

COMPASS/FLAME2008 was developed in the context of the project "Personalized Web Services on Internet III for the Olympic Games 2008 in Beijing", October, 2002 – September, 2006, supported by the German Ministry of Education and Research (BMBF Grant No. 01AK055) and the Chinese Ministry of Science and Technology (MOST).

REFERENCES

Chong, E.I., Das, S., Eadon, G., & Srinivasan, J. (2005). An efficient SQL-based RDF querying scheme. In *Proc. 31st VLDB Conf* (pp. 1216-1227). Trondheim, Norway.

Cycorp, Inc. (2002). OpenCyc selected vocabulary and upper ontology. Retrieved June 19, 2007, from *http://www.cyc.com/cycdoc/vocab/vocab-toc.html*

Das S., Chong, E.I., Eadon, G., & Srinivasan, J. (2004). Supporting ontology-based semantic matching in RDBMS. In *Proc. 30th VLDB Conference* (pp 1054-1065). San Francisco, CA: Morgan Kaufmann.

Davies, N., Mitchell, K., Cheverest, K., & Blair, G. (1998). Developing a context sensitive tourist guide. In *First Workshop on Human Computer Interaction with Mobile Devices* (GIST Tech. Rep. G98-1) (pp 17-24).

Deiters, W., Löffeler, T., & Pfennigschmidt, S. (2003). *The information logistics approach toward a user demand-driven information supply.* In D. Spinellis (Ed.), *Cross-media service delivery* (pp. 37-48). Boston, MA.

Dey, A. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal, 5*(1), 4-7.

Dey, A., Salber, D., Abowd, G.D., Futakawa, M. (1999). The conference assistant: Combining context-awareness with wearable computing. In *3rd International Symposium on Wearable Computer,* San Francisco, CA, (pp. 21-28).

Dublin Core Metadata Initiative (DCMI). (2004). Dublin core metadata element set, reference description. Retrieved June 19, 2007, from *http://dublincore.org/documents/dces/*

Fonseca, F., Egenhofer, M., Davis, C., & Borges, K. (2000). Ontologies and knowledge sharing in urban GIS. *Computer, Environment and Urban Systems, 24*(3), 251-272.

Guarino, N. (1998). Formal ontology and information systems. In *Proc. FOIS'98* (pp. 3-15). Trento, Italy: IOS Press.

Hobbs, J., & Pustejovsky, J. (2003). Annotating and reasoning about time and events. In *Proc. AAAI Spring Symposium on Logical Formalization of Commonsense Reasoning* (pp. 74-82). Menlo Park, CA: AAAI Press.

Holtkamp, B., Weißenberg, N., & Speckmann, H. (2005). MONA – A situation-aware decision support system for emergency situations. In *Proc 19th Int. Conf. EnvironInfo - Informatics for Environmental Protection,* Brno, Czech Republic (pp. 186-190).

Kokkelink, S., & Schwänzl R. (2002). Expressing qualified Dublin core in RDF/XML. Retrieved June 19, 2007, from *http://dublincore.org/documents/dcq-rdf-xml/*

Long, S., Kooper, R., Abowd, G. D., & Atkeson, C. G. (1996). Rapid prototyping of mobile context-aware applications: The cyberguide case study. In *2nd ACM International Conference on Mobile Computing and Networking* (pp 97-107).

Martin, D. (Ed.) (2004). OWL-S: Semantic markup for Web services (W3C Member Submission). Retrieved June 19, 2007, from *http://www.w3.org/Submission/OWL-S*

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *Proc FOIS'01* (pp 2-9). Ogunquit, ME.

Weißenberg, N., Voisard, A., & Gartmann, R. (2006). An ontology-based approach to personalized situation-aware mobile service supply. *Springer GeoInformatica, 10*(1), 55-90.

Want, R., Hopper, A., Falcao, V., & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems, 10*(1), 91-102.

This work was previously published in Handbook of Ontologies for Business Interaction, edited by P. Rittgen, pp. 278-293, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.15
A Multi-Agent System Approach to Mobile Negotiation Support Mechanism by Integrating Case-Based Reasoning and Fuzzy Cognitive Map

Kun Chang Lee
Sungkyunkwan University, Korea

Namho Lee
Sungkyunkwan University, Korea

ABSTRACT

This chapter proposes a new type of multi-agent mobile negotiation support system named MAM-NSS in which both buyers and sellers are seeking the best deal given limited resources. Mobile commerce, or m-commerce, is now on the verge of explosion in many countries, triggering the need to develop more effective decision support systems capable of suggesting timely and relevant action strategies for both buyers and sellers. To fulfill a research purpose like this, two artificial intelligence (AI) methods such as CBR (case-based reasoning) and FCM (fuzzy cognitive map) are integrated and named MAM-NSS. The primary advantage of the proposed approach is that those decision makers involved in m-commerce regardless of buyers and sellers can benefit from the negotiation support functions that are derived from referring to past instances via CBR and investigating inter-related factors simultaneously through FCM. To prove the validity of the proposed approach, a hypothetical m-commerce problem is developed in which theaters (sellers) seek to maximize profit by selling their vacant seats to potential customers (buyers) walking

around within reasonable distance. For experimental design and implementation, a multi-agent environment Netlogo is adopted. A simulation reveals that the proposed MAM-NSS could produce more robust and promising results that fit the characteristics of m-commerce.

INTRODUCTION

The modern mobile computing world is characterized by one of both ubiquitous connectivity and ubiquitous computational resources (Edwards, Newman, Sedivy, & Smith, 2004). Recent popular forms of mobile computing encompass omnipresent short-range communications (including both infrastructure-based technologies such as WiFi and peer-to-peer technologies such as Bluetooth), and also omnipresent long-range communications (such as cellular telephony networks). This maturing mobile environment justifies conservative estimates based on the 2000 Census report suggesting that by 2006 10% of U.S. workers will be completely mobile, with no permanent office location (Lucas, 2001). This trend will be fueling development of new mobile applications as advances in mobile technology increase coverage, data speeds, and usability (Barbash, 2001; Crowley, Coutaz, & Bérard, 2000; Parusha & Yuviler-Gavishb, 2004; Pham, Schneider, & Goose, 2000; Turisco, 2000).

In this sense, it is no wonder that **mobile commerce** (or **m-commerce**) replaces traditional forms of electronic commerce rapidly. Various types of m-commerce services include mobile shopping, location sensitive information service, traffic updates, and logistic tracking services, all of which utilize the concepts of customization, personalization, location sensitive, context awareness (Lee & Yang, 2003; Schilit, 1995; Schilit, Adams, & Want 1994; Wang & Shao, 2004; Want, Hopper, Falcao, & Gibbons, 1992; Want, Schilit, Adams, Gold, Petersen, Ellis, et al., 1995). M-commerce has been successfully activated in some industries, leading to competitive advantage (Rodgera & Pendharkarb, 2004; Varshney, 1999) and improved workflow as well as reduced costs and risk management (Miah & Bashir, 1997; Porn & Patrick, 2002; Turisco, 2000). However, such a success story is confined to specific applications where the decision support framework is not considered seriously. To reap better results from the users' view, decision makers engaged in a specific type of m-commerce should be supported more intelligently and robustly.

It cannot be overstated that decision makers under a specific m-commerce situation need more timely and robust decision support because they are in several types of contexts. For example, they cannot afford to receive detailed information from a decision support system because of the limited display capability of mobile devices they carry. Besides, they do not have enough time to consider all the related factors before making decisions because they are usually on the move. This kind of environmental limitations require that a **decision support framework** should be developed for enhancing decision making effectiveness for m-commerce users.

For this purpose, this chapter proposes a new kind of decision support framework named **MAM-NSS_**(multi-agent mobile negotiation support system) which can benefit both m-commerce buyers and sellers. MAM-NSS is based on a multi-agent mechanism in which buyers and sellers are respectively represented by agents. Each agent tries to coordinate with each other until reaching a compromised decision. Especially, the proposed MAM-NSS focuses on the fact that decision makers engaged in a specific m-commerce situation are often facing two kinds of needs: (1) to refer to past instances carefully and (2) mull over inter-related factors simultaneously. A literature survey shows that there exist few studies dealing with those research needs. To fill such a research void, this chapter proposes two important mechanisms like **case based reasoning** (CBR) and **fuzzy cognitive map** (FCM).

The proposed MAM-NSS combining CBR and FCM is therefore expected to provide more robust decision support to m-commerce decision makers irrespective of buyers and sellers. To prove the validity of the proposed approach, a hypothetical m-commerce problem is developed in which theaters (sellers) seek to maximize profit by selling their vacant seats to potential customers (buyers) walking around within reasonable distance. For experimental design and implementation, a **multi-agent** environment ***Netlogo***[1] is adopted.

BACKGROUND

Recent Trends in M-Commerce

Electronic commerce applications recently provided by mobile communication services include mobile information agents (Cabri, Leonardi, & Zambonelli, 2001, 2002; Mandry, Pernul, & Rohm, 2000-2001), online kiosks (Slack & Rowley, 2002), government applications (e.g., online selling by the postal service, Web-based electronic data interchange, or EDI, in trade applications), and direct online selling systems such as Internet-based (or Web-based) shopping mall systems and Internet-based stock trading systems. In the m-commerce and mobile communication services, their ease of use and multimedia approach to the presentation of information attract potential customers. Some countries and regions have put in tremendous efforts in pushing the development and deployment of m-commerce. These countries include the U.S., Japan, South Korea, Hong Kong, and the Scandinavian countries. There have been many different kinds of m-commerce applications deployed to businesses in these areas. In the U.S., the current rush to wireless communication methods was triggered by the U.S. Federal Communication Commission's auctioning of personal communication-service spectrum space (Senn, 2000). The collaboration of public and private sectors has facilitated the development of m-commerce businesses.

Recently, a large number of organizations have adopted m-commerce for business purposes in order to gain competitive advantages in the electronic market. To cite a few examples, NTT DoCoMo, Vodafone, Verizon, Sprint PCS, and AT&T Wireless have provided "cybermediation" for greater efficiency in supply and marketing channels through m-commerce. M-commerce can benefit business transactions by providing more efficient payment systems, shortening time to markets for new products and services, realizing improved market reach, and customization of products and services (Barnes, 2002; Senn, 2000). Besides, innovative m-commerce applications have been constantly reshaping business practices in terms of enhancing customer service, improving product quality, and lowering cycle time in business processes (Seager, 2003).

As m-commerce supports online purchasing through an electronic channel, that is, the Internet, via electronic catalogs or other innovative formats, customers procure products, services, and information through m-commerce (Bailey & Lawrence, 2001). In m-commerce, potential customers can visit various "virtual malls" and "virtual shops," and browse through their catalogues to examine products in vast detail. New areas of business opportunities for retailers, producers, and consumers can be developed from these virtual markets on m-commerce. Mobile information agents provide an effective method to support the electronic marketplace by reducing the effort involved in conducting transactions (Wang, Tan, & Ren, 2002). Mobile agents can also help search other agents for contracting, service negotiation, auctioning, and bartering (Mandry et al., 2000-2001). Agents roam through Internet sites to locally access and elaborate information and resources (Omicini & Zambonelli, 1998). The introduction of mobile agents into the electronic market scenario reduces the load and the number of necessary connections to suppliers. In

this way, the multi-agent approach is a feasible way to model and analyze complex m-commerce applications.

Among the three distinct identifiable classes of electronic commerce applications (i.e., **business-to-customer (B2C)**, **business-to-business (B2B)**, and intra-organization (Applegate, Holsapple, Kalakota, Radermacher, & Whinston, 1996), m-commerce generally falls into the B2C class. M-commerce provides Web presence with information about company products and services and facilities for both online and off-line purchasing. M-commerce also facilitates other business related activities, such as entertainment, real estate, financial investment, and coupon distribution. Usually, m-commerce sellers are required to make competitive offers in order to sell their products or services to the target customers within reasonable distance. In location-based m-commerce applications, sellers should compete with each other to appeal to potential buyers because there could be only a few buyers in a limited area. For sellers, making timely and attractive offers to buyers on the move is very challenging because the buyers continue to receive information and offers from competing sellers.

Intelligent Agents

Fundamentals

The proposed MAM-NSS is basically based on the multi-agents. Both sellers and buyers engaged in a certain m-commerce are represented by specific agents, and each agent is entitled to receiving proper decision support from the MAM-NSS. An intelligent agent (or agent) has various definitions because of the multiple roles it can perform (Applegate et al., 1996; Hogg & Jennings, 2001; Persson, Laaksolahti, & Lonnqvist, 2001; Wooldridge, 1997; Wooldridge & Jennings, 1995). An intelligent agent is simply a software program that simulates the way decision makers think and make decisions. It performs a given task based on the information gleaned from the environment to act in a suitable manner so as to complete the task successfully. It is able to adjust itself to the changes in the environment and circumstances, so that it can achieve the expected result (Paiva, Machado, & Prada, 2001).

The term "**intelligent agent**" can be disintegrated into two words: intelligence and agency. The degree of autonomy and authority vested in the agent is called its agency. It can be measured at least qualitatively by the nature of the interaction between the agent and other entities in the system in which it operates. An **agent** is an individual and it runs independently. The degree of agency will be enhanced if an agent represents a user in some way. Therefore, collaborative agents represent a higher level of agency because they cooperate with other agents or programs or entities, and so on. The agent intelligence can be interpreted as the degree of reasoning and learned behavior. It is the ability to understand the user's statement of goals and carry out the task delegated to it. Such intelligence can be easily found in the reasoning process of many decision or AI models. Intelligence enables agents to discover new relationships, connections, or concepts independently from the human user, and exploit these in anticipating and satisfying a user's needs (Bonarini & Trianni, 2001; Hu & Weliman, 2001; Schaeffer, Plaat, & Junghanns, 2001).

To retain the characteristics of "intelligence" and "agency." an intelligent agent should possess the abilities of mobility, benevolence, rationality, adaptability, and collaboration (Wooldridge, 1997). Mobility is the ability to move around an electronic network (Bohoris, Pavlou, & Cruickshank, 2000; Lai & Yang, 1998; Lai & Yang, 2000). Benevolence is the assumption that an intelligent agent does not have conflicting goals, and therefore it will always try to complete the assigned tasks (Hogg & Jennings, 2001; Jung & Jo, 2000). Rationality is the assumption that an agent will act in order to achieve its goals and will not act in such a way as to prevent its goals

from being achieved, at least insofar as its beliefs permit (Hogg & Jennings, 2001; Persson et al., 2001). Adaptability indicates that an agent should be able to adjust itself to the habits, working methods, and preferences of its user (Jung & Jo, 2000). Collaboration is an ability to cooperate with other agents so that an agent can achieve what a goal decision maker wants to attain (Jung & Jo, 2000; Lee & Lee, 1997; Wu, Yuan, Tseng, & Fuyan, 1999). This chapter places a strong emphasis on the collaboration ability. Although no single agent possesses all these abilities in a real situation, it is certain that these kinds of characteristics are those that distinguish agents from ordinary programs.

Multi-Agent System

To investigate a computational model that actually encodes and uses conflict resolution expertise, a focus can be placed on the **multi-agent framework**, which is adopted from the distributed AI problem (Bird, 1993; Chaib-Draa & Mandiau, 1992; Cooper & Taleb-Bendiab, 1998; Luo, Zhang, & Leung, 2001; Sillince, 1998; Sillince & Saeedi, 1999; Tung & Lee, 1999). Multi-agent systems have offered a new dimension for coordination in an enterprise (Bonarini & Trianni, 2001; Hu & Weliman, 2001; Kwon & Lee, 2002; Sikora & Shaw, 1998; Strader, Lim, & Shaw, 1998; Ulieru, Norrie, Kremer, & Shen, 2000; Wu, 2001). Incorporating autonomous agents into **problem-solving** processes allows improved coordination of different functional units to define tasks independently of both the user and the functional units under control (Cabri et al., 2002). Under a multi-agent system, the problem-solving tasks of each functional unit becomes populated by a number of heterogeneous intelligent agents with diverse goals and capabilities (Lottaz, Smith, Robert-Nicoud, & Faltings, 2000; Luo et al., 2001; McMullen, 2001; Ulieru et al., 2000; Wu, 2001).

The multi-agent system, in which multiple agents work collaboratively to solve specific problems, provides an effective platform for coordination and cooperation among disputing multiple entities in real world cases. For example, when a conflict occurs between buyers and sellers over a limited resource, it is difficult for a single authority or committee to reconcile it to the full satisfaction of all the entities concerned. Therefore, it is likely that the use of a multi-agent system for coordination will result in a more systematic and organized method in reality without causing unnecessary emotional and behavioral side effects.

Decision Support Mechanisms

MAM-NSS is equipped by two decision support mechanisms such as CBR and FCM. First, **CBR** is a renowned artificial intelligence methodology that provides the technological foundations for intelligent systems (Kolodner, 1993). Given a case base where a number of **past instances** are stored, CBR consists of several phases: indexing cases, retrieving the appropriate candidate cases from the case base, approximating potential solutions from them, testing whether the proposed solutions are successful, and learning to upgrade the decision quality by updating the case base and retrieval mechanism. CBR is most applicable when there is (1) no decision model available; (2) a specific decision model is too hard to acquire; or (3) when past cases are available or easy to generate. With these CBR benefits, the CBR approach is extensively used for negotiation (Kowalezyk & Bui, 1999). Considering the advantages of CBR above, MAM-NSS adopts CBR to allow agents to refer to the past relevant instances that seem to explain a part of current decision making problem before making final decisions.

Second, **FCM** is utilized to provide agents with the capability of analyzing complicated **interrelationships of all the relevant factors** by viewing them simultaneously. FCM was introduced by Kosko (1986, 1987) in which fuzzy causality concept is introduced to represent un-

certainty embedded in problem domain. In this way, FCM provides a more flexible and realistic representation of the domain knowledge. For example, Ray and Kim (2002) used it as a tool for understanding and controlling intelligent agents. Liu and Satur (1999) have used FCM as a decision support mechanism for interpreting geographic information as well as designing automatic context awareness function that is one of the important characteristics in m-commerce.

By integrating CBR and FCM, MAM-NSS is designed to provide decision makers on the move with more improved decision support functions. CBR is especially useful for m-commerce users who do not have sufficient time to consider all the constraints before making decisions. By retrieving appropriate past examples and suggesting them as a benchmarking point, CBR can help m-commerce users make fast decisions. There are many factors that are influencing m-commerce decisions either indirectly or directly. However, users cannot afford to consider all the causal relationships among those factors thoroughly in a situation when they need to move and there is not enough time. In that situation, FCM can provide an analytical and systematic way of investigating causal relationships between all the factors related to the m-commerce situation.

MAM-NSS

Basics

Figure 1 depicts a hypothetical m-commerce situation where MAM-NSS is used to provide timely decision support to m-commerce users. Since the term "m-commerce" may cause different interpretations depending on situations, we need to define m-commerce conditions more clearly to make further discussions unambiguous. First, buyers are assumed to carry mobile devices such as PDAs or mobile phones. Second, buyers cannot have access to the telecommunication network line because they are moving. Third, buyers make reservations with sellers through mobile devices. Fourth, sellers provide information about their services and goods through mobile devices.

The **negotiation process** of MAM-NSS is composed of four steps. Since MAM-NSS runs on a multi-agent framework, a buyer is represented by **B-agen**t and a seller **S-agent** from now on. In other words, each agent possesses its own generic knowledge including either buyer's or seller's basic preference that has been predefined.

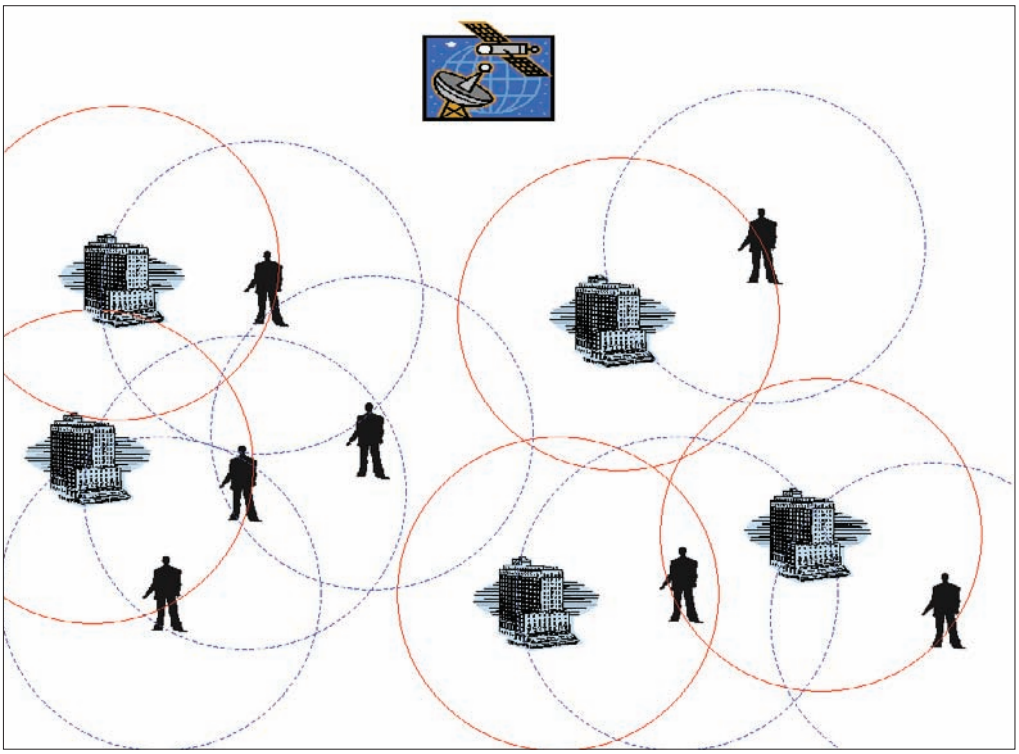
The first step is to identify buyers' and sellers' location. Such location identification is based on opt-in agreement with mobile telecommunication company.

The second step is for sellers to provide an offer to potential buyers within a fixed distance from where buyers can arrive after a reasonable time. For example, the seller's offer may include an appropriate product/service and its price quote, all of which are assumed to be obtained from CBR inference. For this CBR inference, sellers need various kinds of contextual information such as potential buyers' current location, weather, other events, and so forth.

The third step is for buyers to check the offer from sellers on their mobile devices in which buyers' personal preference is stored. Then the seller's offer is compared with the buyer's predefined preference. If the offer is not appealing, then the buyers modify the offer and send it to the seller. All this process is performed on a multi-agent basis.

The fourth step is for sellers to review the modified offer from buyers. At this time, FCM is used to induce an appropriate price level considering complicated causal relationships among qualitative and strategic factors simultaneously. Depending on marketing strategy, the sellers may change their price offer to lower or higher. Therefore, the sellers' decision can become more strategic and flexible in accordance with changing situations.

Figure 1. M-commerce situation



Architecture

MAM-NSS is composed of three entities like **B-agent**, **S-agent**, and **mobile telecommunication company**. B-agent is assumed to be downloaded into buyer's mobile device when she subscribes to the MAM-NSS service. The B-agent becomes personalized according to **buyer's personal preference** about specific products/services, and related prices, quality, brand, and other properties. Especially, B-agent specifies its own utility function following buyer's preference which is compiled from the online questionnaire when the buyer subscribes for the MAM-NSS service. Then the B-agent is stored in the memory of the buyer's carried **mobile device**. The B-agent uses this information to negotiate prices with sellers. S-agent is basically linked to the seller's back office system and negotiates with B-agent on the price of seller's products/services. S-agent's first price offer is composed referring to CBR inference, and relayed to potential buyers who are moving within a reasonable distance from the seller. When the buyer's modified price offer is entered, S-agent revises its price offer by using FCM and then feeds it back to the buyer. In this process, negotiations are going on until the final deal is struck between buyer and seller.

In the process of negotiations, the mobile telecommunication company acts as an intermediary between S-agents and B-agents. If the **sellers** and **buyers** subscribe to the MAM-NSS service

provided by the mobile telecommunication company, then they can share the information needed in negotiation such as the location of sellers and buyers, price offer, and related product/service information.

S-Agent

The ultimate goal of S-agent is to maximize profit. For this purpose, S-agent seeks a potential buyer in the range acceptable to the buyer within a specific time limit. Then it calculates the bid price of the selling product/service based on CBR, and sends the offer including price and product/service to the potential B-agents through the mobile telecommunications company. A wide variety of past selling instances are stored in the case base, and CBR uses the **similarity index** (SI) below to select the candidate case that seems to fit most with the current selling situation. Once such case is chosen successfully, then the price offer can be made appropriately referring to the price information attached to the selected case.

$$SI_i = \sqrt{\sum_{j=1}^{n} (N_j - S_{ij})^2}$$

where N_j indicates jth attribute value of a new case (j=1,2,…,n), and S_{ij} denotes jth attribute value of ith case in case base of CBR (i=1,2,…, m). Netlogo source code for implementing the CBR mechanism using SI is listed in Table 1.

If the buyer accepts the price offered by S-agent, then the deal between seller and buyer will be completed and the buyer stops negotiating. But if the buyer is not satisfied with the price offered by S-agent, the price level is adjusted and then relayed to S-agent. Finally, S-agent decides whether or not to accept the newly-adjusted price, using the FCM inference. If another price offer

Figure 2. MAM-NSS architecture

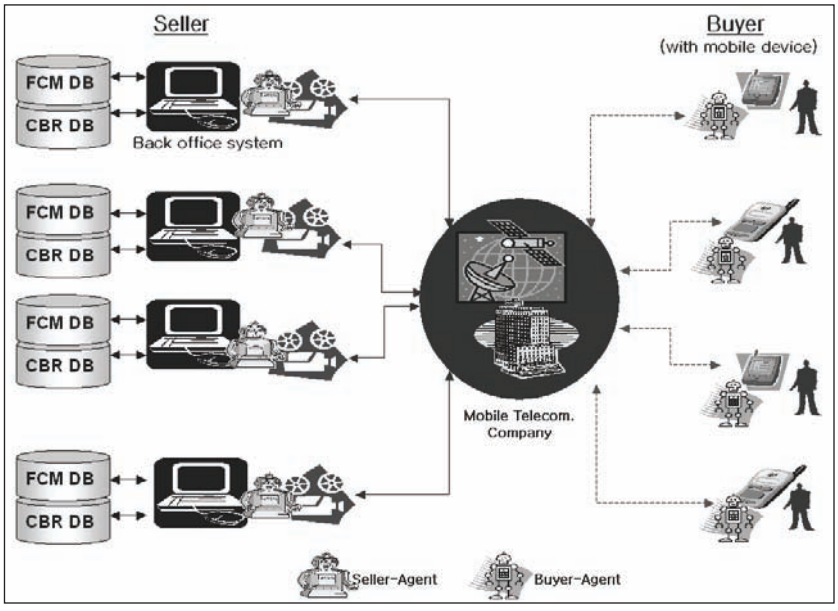


Table 1. S-agent's CBR inference mechanism

```
to change-CBR-price
locals [temp_t temp_i temp_si temp_optimal_si temp_item]
set temp_t(1)
repeat seller_number [
ask seller with [reg_number = temp_t and mobile_service = 1 and
mdss_service = 1 ]
[
   set vacant_percent int (vacant_seat_number / seat_number * 100)
set temp_i (0)
   repeat length CBR_price [
set temp_si sqrt((F1_Current_Value - item (temp_i) CBR_F1_List) ^ 2
                        + ((F2_Current_Value - item (temp_i)
CBR_F2_List) ^ 2
                        + (F3_Current_Value - item (temp_i)
CBR_F3_List) ^ 2
                        + (F4_Current_Value - item (temp_i)
CBR_F4_List) ^ 2
                        )

   if (temp_i = 0)[set temp_optimal_si (temp_si) set temp_item
(temp_i)]
```

Table 2. S-agent's FCM mechanism to adjust price offer

```
to Buyer-decision-for-new-customer-offer
set temp (0)
set temp_i (1)
repeat customer_number [
set inference_list [ ]
set inference_list lput FCM_factor1 inference_list
set inference_list lput FCM_factor2 inference_list
set inference_list lput FCM_factor3 inference_list
set inference_list lput FCM_factor4 inference_list
set inference_list lput FCM_factor5 inference_list
set inference_list lput FCM_factor6 inference_list
set inference_list lput FCM_factor7 inference_list
set inference_list lput FCM_factor8 inference_list
set inference_list lput FCM_factor9 inference_list
set inference_list lput FCM_factor10 inference_list

show inference_list

set FCM_ACCEPT_REULT (FCM_inference inference_list )

ask buyer with [id_number = temp_i ] [
 if (reserve != 1 and mdss_service = 1)   [
 set temp_utility_adjustment (utility_adjustment)
 ask seller with [reg_number = temp2 ]  [
   if (available_number_of_product > 0)  [
       if (FCM_ACCEPT_REULT = 1 )  [
          show temp_new_price + "<---- accept"
          set vacant_seat_number (vacant_seat_number - 1)
          set temp1 (1) ask customer with [id_number = temp_i ]
                                                  [set reserve (1)  set color black]
```

is made, then negotiation proceeds to the next round. This process is repeated until all sellers and buyers find their appropriate partners that meet their respective goals. In Table 2, the FCM mechanism to be used for adjusting the **price offer** is represented in the Netlogo source code.

B-Agent

The buyer represented by B-agent seeks to maximize its own utility in the process of negotiating with the suppliers. Buyers can download B-agents from the subscribed telecommunication company's site and store them into their mobile devices. B-agents incorporate the following **utility functions** where i=1,2,…,m (number of sellers) and j=1,2,…,n (number of utility factors):

$$U_i = \sum_{j=1}^{n} W_{ij} \bullet F_{ij}$$

U_i denotes ith buyer's utility, W_{ij} buyer's preference for jth utility factor, and F_{ij} ith buyer's jth utility factor. It is certain that $\sum_{j=1}^{n} W_{ij} = 1$. Examples of utility factors include not only price, product, and quality, but also **contextual information** such as the buyer's current location and environmental constraints. Table 3 shows the Netlogo source code for calculating the B-agent's utility function.

If B-agent gets the price offer from S-agent and this offer does not meet the buyer's goal utility, then the B-agent suggests a new price using the mechanism shown in Table 4. If the seller accepts the new price offered by the buyer, then the deal is complete. However, if no sellers accept this price, then the B-agent increases the price decreasing its goal utility. In this case, a new round of negotiation resumes.

EXPERIMENTS

Problem Description

Three Groups

The target problem here is that there are a number of movie theaters in an area, and customers want to go to the theater depending on their personal situations. Based on whether customers (i.e., buyers) and theaters (i.e., sellers) are using mobile devices or not, we categorize them into three groups for the sake of the experiment. Group 1, called "**non-mobile group**," is not using mobile devices. Therefore, customers either reserve tickets through non-mobile channels, such as telephone or cable Internet, or buy onsite from the box office. Theaters are assumed to contact customers through the non-mobile channels too. Group 2, called "**passive mobile group**," is using mobile devices but no negotiation functions. Therefore, customers in this group can get information about movies through mobile channels, but they cannot negotiate with theaters through agents. Theaters are also using mobile channels to send movie information to customers. Group 3, called "**active mobile group**," is assumed to use the proposed MAM-NSS for negotiation through mobile channels and agents. Customers and theaters are offering their own preferences such as price and vacant seats utilizing the negotiation mechanism based on MAM-NSS. The three groups will be compared with each other through Netlogo simulation experiments. In fact, group 1 is inappropriate for m-commerce situations because they are not reached by mobile devices. However, group 1 is included so that the other two groups can be compared.

Sellers

- Assumptions about theaters are as follows: they have 200 seats, cost $700 per show, and

Table 3. B-agent's utility calculation

```
to search-buyer
set temp (1)
set temp_id (1)
repeat customer_number [
ask customer with [reserve != 1 and id_number = temp_id] [
set temp_distance (p_distance )
set temp_price (p_price )
set temp_time (p_time )
set temp_boxoffice_ranking (p_boxoffice)
set temp_genre (p_genre )
set temp_customer_x (current_x) set temp_customer_y (current_y)
set utility (0)
set temp_selected_theoter (0)

repeat seller_number [
 ask seller in-radius-nowrap (remaining_time / time_per_patch)
   with [available_product_number > 0 and reg_number = temp1][
set actual_distance (abs (sqrt((temp_customer_x - location_x) ^ 2
+ (temp_customer_y - location_y) ^ 2 ) ))

Convert_factor_point

set temp_util (  temp_P1 * temp_point_F1
               + temp_P2 * temp_point_F2
               + temp_P3 * temp_point_F3
               + temp_P4 * temp_point_F4
               + temp_P5 * temp_point_F5
                )
if (temp_util > utility) [ set utility (temp_util)
                      set temp_selected_seller (reg_number)  ]
                           ]
                set temp1 (temp1 + 1)] set temp1 (1)
```

Table 4. B-agent's price update process

```
ask buyer with [deal !=1 ] [
 set goal_utility (Current_utility + (utility_adjustment / 100) * Current_utility )
  set temp (selected_buyer)
      ask seller with [reg_number = temp ][
        if (available_number > 0) [
        if (p_temp > 0 ) [
 set temp_price_down_request int((goal_utility -
```

start to sell vacant seats an hour before the movie begins. List price for a ticket is $7, and movie genres a theater is showing have four types. All the theaters show the movie with box office ranking from 1 to 10. Every 20 minutes, theaters in group 2 are offering discriminated pricing strategies to buyers through mobile channels, depending on the

vacancy rate: \$6.5 if vacancy rate < 40%, \$6.0 if 40% ≤ vacancy rate ≤ 50%, \$5.0 if 50% ≤ vacancy rate ≤ 60%, \$4.0 if 60% ≤ vacancy rate ≤ 70%, \$3.0 if 70% ≤ vacancy rate ≤ 80%, and \$2.0 if 80% ≤ vacancy rate. However, as noted previously, customers cannot negotiate with the theaters in group 2, indicating that they have no choice but to accept the price or not.

- Meanwhile, theaters in group 3 are offering different ranges of price using CBR inference where a case is composed of four input attributes (*current vacancy rate (%), remaining time before the show (minutes), box office ranking of the current show, approximate number of reachable customers*) and one output attribute (*ticket price*). Therefore, the price changes in accordance with the input attribute values theaters are currently facing. Theaters decide whether or not to accept the newly adjusted price offered by the buyers, using FCM as shown in Figure 3. If the FCM result is less than 1, then the theater rejects the buyer's price. Otherwise, the theater accepts the buyer's price, and the buyer's seat number and show time are specified accordingly. When conducting FCM analysis, input constructs should be transformed into an appropriate value considering the input conditions. Table 5 shows input constructs, its conditions, and transformed values.

Customers

The customer's utility function includes the following five factors: (1) D (distance from customer's current location to theater; for this experiment, it is adjusted between -18 and 18 on the Netlogo platform); (2) R (box office ranking of the movie); (3) G (movie genre); (4) P (newly

Figure 3. Group 3 theaters'

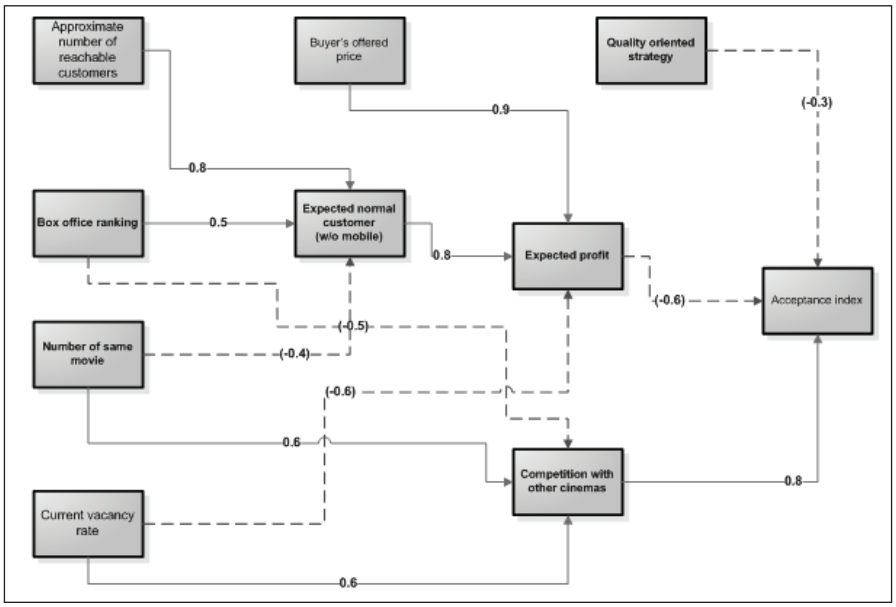


Table 5. Input constructs and conditions for theater's FCM

Input constructs	Condition	Transformed values
Approximate number of reachable customers	Many* Normal** Few***	1 0 -1
Box office ranking	Rank 1, 2 Rank 3,4 Rank 5,6 Rank 7 ~ 10	1 5 –0.5 -1
Number of same movie	4 ~ 2 ~ 3 1 0	1 0.5 -0.5 -1
Current vacancy rate	40% ~ 30% ~ 40% 20% ~ 30% 10% ~ 20% ~ 10%	1 0.7 0.5 -0.5 -1
Quality oriented strategy	Yes No	1 -1
Buyer's offered price	$7 $5 ~ $7 $4 ~ $5	1 0.7 0.5
	$3 ~ $4 $2 ~ $3 $1 ~ $2 ~ $1	0 -0.5 -0.7 -1

adjusted ticket price that customers want); and (5) T (timeliness showing whether it is the exact time that customer wants). Using the five factors like this, *i*th customer's utility function is denoted as follows:

$$U_i = W_{D_i} \cdot D_i + W_{R_i} \cdot R_i + W_{G_i} \cdot G_i + W_{P_i} \cdot P_i + W_{T_i} \cdot T_i$$

- Table 6 addresses the various conditions and their converted values for the five utility factors.

MAM-NSS Simulation

Basics

The MAM-NSS simulation prototype was performed on the Netlogo platform which is a programmable multi-agent modeling environment for simulating natural and social phenomena, and particularly well-suited for modeling complex systems that develop over time. The target m-commerce problem described previously can be well represented by multi-agents composed of B-agents, S-agents, and interactions between them for the effective negotiation. Therefore, MAM-NSS is capable of handling the problem very well.

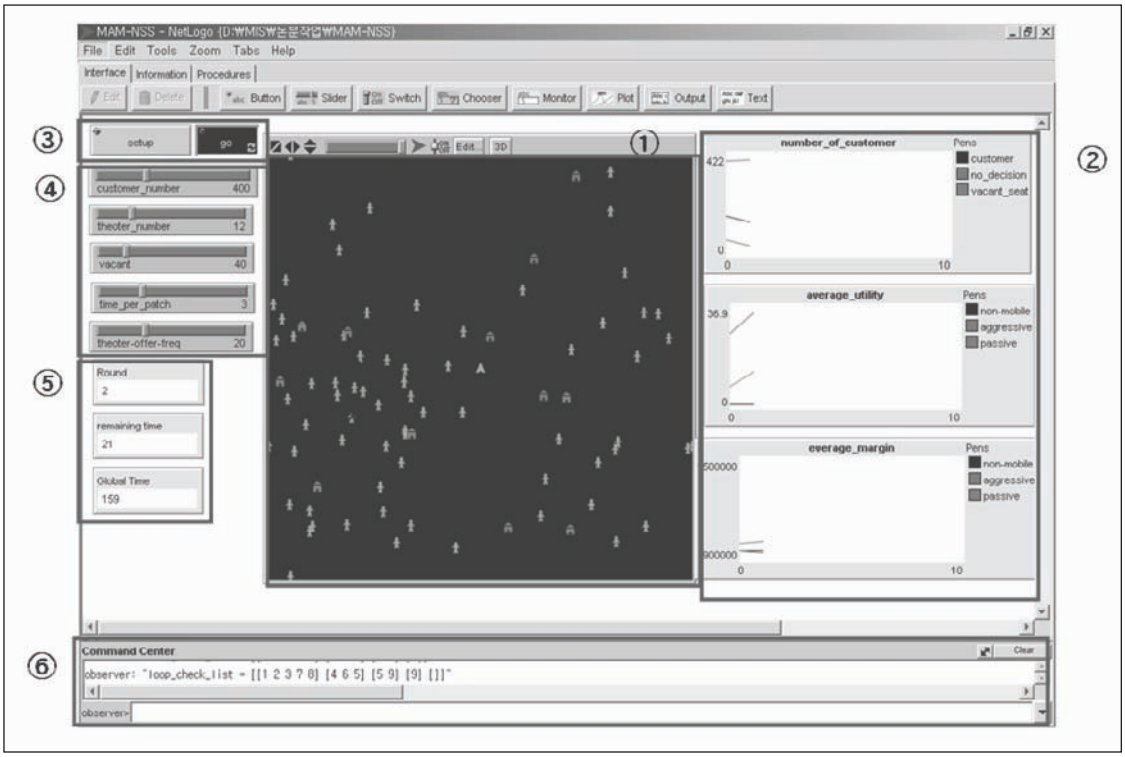
As shown in Figure 4, MAM-NSS has six types of user interface components as follows:

1. *Behavior space* shows the customer's movement and the location of theaters. The human shape indicates a customer and the house shape represents a theater. Gray color means group 1, green color group 2, and pink color group 3. All customers are designed to move one unit of position over to the random direction at a time. The customers go out of simulation after they buy tickets.
2. *Graph* monitors the change of values such as number of customers, vacant seats, number

Table 6. Customer's utility factors

Utility factor	Condition	Converted value
Distance from the theater (D)	In 20 minutes In 30 minutes In 40 minutes In 50 minutes More than 60 minutes	50 40 30 20 10
Box office ranking (R)	1,2 3,4 5,6 7,8 9,10	50 40 30 20 10
Movie genre (G)	Customer wanted Customer did not want	50 0
Ticket price (P)	For any new ticket price adjusted by customers	50 – (new ticket price / list price) * 50
Timeliness (T)	Exact time that customer wants Otherwise	50 0

Figure 4. MAM-NSS simulator



of customers who have not bought a ticket, customers' average utilities, and theaters' average margins.

3. *Control button* prepares and prompts simulation.
4. *Slider* controls the initial conditions of simulation such as number of customers, number of theaters, and maximum number of vacant seats for each theater. Each theater has a number of vacant seats falling between

0 and the "vacant" number that is set by this slide.

5. *Monitor* shows the number such as rounds of simulation, remaining/elapsed before the next show starts, and simulation time.
6. *Command center* shows temporary data generated from the agent activities.

Results and Implications

Twenty six rounds of simulation were done on MAM-NSS with the initial conditions as follows. Total rounds of simulation is 35, number of theaters 12, number of customers ranging between 100 and 600 (evenly assigned to three groups), group 2 theaters sending discriminated price every 20 minutes, and theaters starting to offer discounted price 60 minutes before the show. The simulation results with MAM-NSS are summarized in Table 7 numerically, and in Figure 5 graphically. Under a 95% confidence level, statistical results in Table 8 reveal that in terms of average utilizes and average profits, group 3 can yield the highest value compared with the other two groups. Its implication is as follows:

First, those users belonging to the passive mobile group can benefit from using the mobile devices. However, such mutual benefits increase much more when they use the negotiation support function provided by the proposed MAM-NSS.

Second, multi-agents are very convenient as well as effective for the m-commerce entities to handle them in their decision making process through the use of MAM-NSS. The reason is that agents are basically capable of autonomous operation once the entity's preference is predefined and stored into its memory. In the MAM-NSS environment, users do not have to bother themselves to interact with negotiation partners.

Third, both preference and conditions that users want their own agents to consider in the process of negotiations can be easily incorporated into agents. Since MAM-NSS is installed in the central server of the telecommunication company, it is very easy for users to use it.

Fourth, since m-commerce users are limited by narrow screen and specified functions of their mobile devices, and agents are capable of replacing users in the real negotiation process in an almost automatic manner, the use of negotiation support mechanism like MAM-NSS would greatly contribute to enhancing users' utilities and profits as well.

CONCLUDING REMARKS

To resolve the negotiation process between buyers and sellers in the context of m-commerce, we proposed a multi-agent mobile negotiation support system called MAM-NSS, in which all the buyers and sellers engaged in a m-commerce situation are represented by multi-agents embedded with each entity's preference and corresponding conditions. Its potentials were proved by the simulation experiments using the theaters' vacant seats negotiation problems. Main contributions of the MAM-NSS are as follows.

First, CBR inference is incorporated to help S-agents decide an appropriate price for a vacant seat. Without using CBR, S-agents will have difficulty finding such appropriate price which is consistent with previous decision making results. Especially, such consistency in setting price for various situations is very important to customers who want be opportunistically exploited by sellers.

Second, FCM finds its great potential in the process of negotiation, due to its generalized inference capability in a presence of a number of inter-related factors. Without FCM, decision makers would feel very stressed to consider all the complicated causal relationships among the relevant factors and expect future inference results. In this chapter, FCM was used to help S-agents accept B-agent's price offer or not.

Third, multi-agent schemes were found very meaningful in being used in the process of m-

Table 7. Simulation result

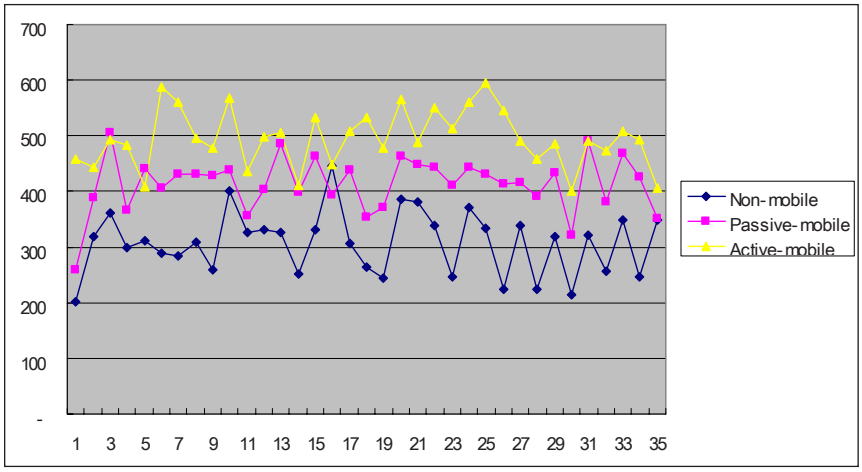
Round	Average Utility			Average Profit		
	Non-mobile	Passive-mobile	Active-mobile	Non-mobile	Passive-mobile	Active-mobile
1	202	259	458	229	349	472
2	319	388	445	238	291	347
3	361	506	492	161	455	
4	299	365	482	134	295	506
5	311	441	410	66	442	468
6	289	406	587	189	208	496
7	285	431	559	423	253	353
8	309	432	497	178	439	377
9	259	428	477	166	390	355
10	402	439	568	197	220	477
11	326	355	436	332	204	480
12	332	405	498	215	266	410
13	327	485	506	190	356	502
14	252	398	412	152	451	403
15	331	463	533	220	386	473
16	445	393	448	243	400	371
17	305	438	508	101	321	470
18	265	354	534	274	318	362
19	245	370	480	99	395	484
20	387	464	565	52	475	498
21	381	449	488	204	429	401
22	339	444	551	176	347	493
23	245	411	512	143	328	501
24	371	444	562	178	330	471
25	334	431	597	159	344	481
26	225	413	547	103	236	502
27	338	416	492	192	283	498
28	225	392	457	124	288	451
29	319	434	487	169	397	449
30	213	322	401	255	305	457
31	321	491	492	190	409	401
32	257	381	474	220	348	287
33	350	469	509	157	411	483
34	247	426	492	147	431	388
35	349	350	406	264	337	441
Average	**215**	**290**	**347**	**131**	**243**	**308**

commerce negotiation. Such multi-agent approach has been proved useful and effective in a wide variety of problems in literature, but its potentials were not proved yet in m-commerce contexts. Therefore, this study adds meaning to literature in that sense.

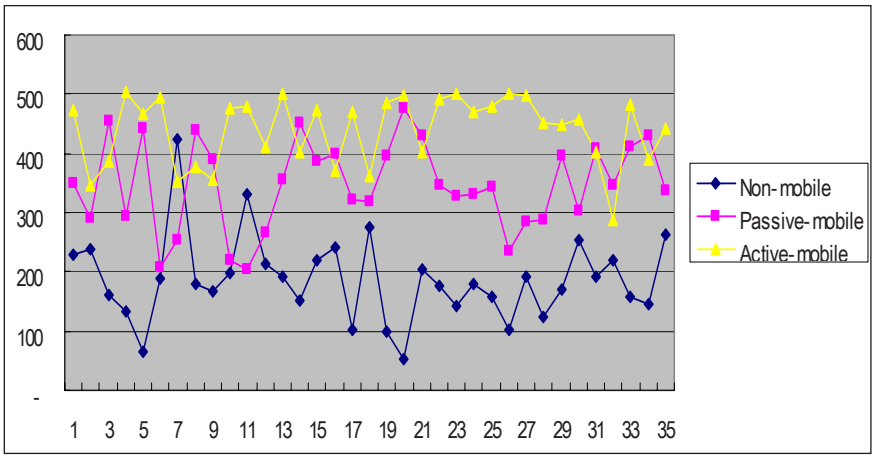
This study has several positive implications for future m-commerce research. First, m-commerce is blooming as mobile devices are providing increased convenience in users' daily activities. However, there has been no important negotiation support system to leverage the potentials of m-commerce. In that meaning, this study will shed a positive light on using the generalized multi-agent framework for designing a mobile negotiation support system. Second, we proposed practical algorithms to be used in upgrading agents' capability in problem solving. Such algorithms would be used in the other business settings with minor adjustments.

But, this study has limitations in the point that (1) all data used in experiment are not real world data; (2) we do not suggest detailed mechanisms for extracting the buyer's preference; and (3) there

Figure 5. Utilities and margins by MAM-NSS simulation



(a) Customers' average utilities



(b) Theaters' average profits

is no comparing our negotiation algorithm with other algorithms. To compare the performance of negotiation algorithms is not simple. Measures need to be developed for comparing negotiation performance before comparing existing algorithms. Additionally, performance measures need to include not only quantitative factors, but also qualitative factors. These three limitations are left as future research topics.

ACKNOWLEDGMENT

This work was supported by grant No. B1210-0502-0037 from the University Fundamental

Table 8. Results of statistical test

Utility

(I) type	(J) type	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
non-mobile	passive	-106.366	12.659	0.000	-137.812	-74.921
	active	-188.336	12.659	0.000	-219.782	-156.891
Passive-mobile	non-mobile	106.3665	12.659	0.000	74.921	137.812
	active	-81.9698	12.659	0.000	-113.416	-50.524
Active-mobile	non-mobile	188.3363	12.659	0.000	156.891	219.782
	passive	81.96982	12.659	0.000	50.524	113.416

Profit

(I) type	(J) type	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
non-mobile	passive	-159.989	16.403	0.000	-200.735	-119.242
	active	-253.081	16.403	0.000	-293.827	-212.334
Passive-mobile	non-mobile	159.989	16.403	0.000	119.242	200.735
	active	-93.092	16.403	0.000	-133.838	-52.346
Active-mobile	non-mobile	253.081	16.403	0.000	212.334	293.827
	passive	93.092	16.403	0.000	52.346	133.838

Research Program of the Ministry of Information & Communication in the Republic of Korea, 2005.

REFERENCES

Applegate, L. M., Holsapple, C. W., Kalakota, R., Radermacher, F. J., & Whinston, A. B. (1996). Electronic commerce: Building blocks of new business opportunity. *Journal of Organizational Computing & Electronic Commerce, 6*(1), 1-10.

Bailey, M. N., & Lawrence, R. L. (2001). Do we have a new economy? *American Economic Review, 91*(2), 308-312.

Barbash, A. (2001). Mobile computing for ambulatory health care: Points of convergence. *Journal of Ambulatory Care Management, 24*(4), 54-60.

Barnes, S. J. (2002). The mobile commerce value chain: Analysis and future developments. *International Journal of Information Management, 22*(2), 91-108.

Bird, S. D. (1993). Towards a taxonomy of multi-agent systems. *International Journal of Man-Machine Studies, 39*, 689-704.

Bird, S. D., & Kasper, G. M. (1995). Problem formalization techniques for collaborative systems. *IEEE Transactions on Systems, Man, and Cybernetics, 25*(2), 231-242.

Bohoris, C., Pavlou, G., & Cruickshank, H. (2000). Using mobile agents for network performance management. In *Proceedings of Network Operations and Management Symposium "The Networked Planet: Management Beyond 2000"* (pp. 637-652).

Bonarini, A., & Trianni, V. (2001). Learning fuzzy classifier systems for multi-agent coordination. *Information Sciences, 136*(1-4), 215-239.

Cabri, G., Leonardi, L., & Zambonelli, F. (2001). Mobile agent coordination for distributed network management. *Journal of Network & Systems Management, 9*(4), 435-456.

Cabri, G., Leonardi, L., & Zambonelli, F. (2002). Engineering mobile agent applications via context-dependent coordination. *IEEE Transactions on Software Engineering, 28*(11), 1039-1055.

Chaib-Draa, B. (1995). Industrial applications of distributed artificial intelligence. *Communications of the ACM, 38*(11), 49-53.

Chaib-Draa, B., & Mandiau, R. (1992). Distributed artificial intelligence: An annotated bibliography. *SIGART Bulletin, 3*(3), 20-37.

Cooper, S., & Taleb-Bendiab, A. (1998). CONCENSUS: Multi-party negotiation support for conflict resolution in concurrent engineering design. *Journal of Intelligent Manufacturing, 9*(2), 155-159.

Coursaris, C., & Hassanein, K. (2002). Understanding m-commerce. *Quarterly Journal of Electronic Commerce, 3*(3), 247-271.

Crowley, J. L., Coutaz, J., & Bérard, F. (2000). Perceptual user interfaces: Things that see. *Communications of the ACM, 43*(3), 54-64.

Edwards, W. K., Newman, M. W., Sedivy, J. Z., & Smith, T. F. (2004). Supporting serendipitous integration in mobile computing environments. *International Journal of Human-Computer Studies, 60*, 666-700.

Hogg, L. M. I., & Jennings, N. R. (2001). Socially intelligent reasoning for autonomous agents. *IEEE Transactions on Systems, Man, & Cybernetics Part A: Systems & Humans, 31*(5), 381-393.

Hu, J., & Weliman, M. P. (2001). Learning about other agents in a dynamic multiagent system. *Cognitive Systems Research, 2*(1), 67-79.

Jung, J. J., & Jo, G. S. (2000). Brokerage between buyer and seller agents using constraint satisfaction problem models. *Decision Support Systems, 28*(4), 293-304.

Kwon, O. B., & Lee, K. C. (2002). MACE: Multi-agents coordination engine to resolve conflicts among functional units in an enterprise. *Expert Systems with Applications, 23*(1), 9-21.

Lai, H., & Yang, T. C. (1988). A system architecture of intelligent-guided browsing on the Web. In *Proceedings of Thirty-First Hawaii International Conference on System Sciences* (pp. 423-432).

Lai, H., & Yang, T. C. (2000). A system architecture for intelligent browsing on the Web. *Decision Support Systems, 28*(3), 219-239.

Lee, W. J., & Lee, K. C. (1999). PROMISE: A distributed DSS approach to coordinating production and marketing decisions. *Computers and Operations Research, 26*(9), 901-920.

Lee, W. P., & Yang, T. H. (2003). Personalizing information appliances: A multi-agent framework for TV programme recommendations. *Expert Systems with Applications, 25*(3), 331-341.

Lottaz, C., Smith, I. F. C., Robert-Nicoud, Y., & Faltings, B. V. (2000). Constraint-based support for negotiation in collaborative design. *Artificial Intelligence in Engineering, 14*(3), 261-280.

Lucas, J. H. C. (2001). Information technology and physical space. *Communications of the ACM, 44*(11), 89-96.

Luo, X., Zhang, C., & Leung, H. F. (2001). Information sharing between heterogeneous uncertain reasoning models in a multi-agent environment: A case study. *International Journal of Approximate Reasoning, 27*(1), 27-59.

Mandry, T., Pernul, G., & Rohm, A. W. (2000-2001). Mobile agents in electronic markets: Opportunities, risks, agent protection. *International Journal of Electronic Commerce, 5*(2), 47-60.

McMullen, P. R. (2001). An ant colony optimization approach to addressing a JIT sequencing problem with multiple objectives. *Artificial Intelligence in Engineering, 15*(3), 309-317.

Miah, T., & Bashir, O. (1997). Mobile workers: Access to information on the move. *Computing and Control Engineering, 8*, 215-223.

Ngai, E. W. T., & Gunasekaran, A. (2005). A review for mobile commerce research and applications. *Decision Support Systems.* Retrieved August 20, 2007, from http://www.sciencedirect.com

Omicini, A., & Zambonelli, F. (1988). Co-ordination of mobile information agents in TuCSoN. *Internet Research: Electronic Networking Applications and Policy, 8*(5), 400-413.

Paiva, A., Machado, I., & Prada, R. (2001). The child behind the character. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 31*(5), 361-368.

Parusha, A., & Yuviler-Gavishb, N. (2004). Web navigation structures in cellular phones: The depth/beadth trade-off issue. *International Journal of Human-Computer Studies, 60*, 753-770.

Persson, P., Laaksolahti, J., & Lonnqvist, P. (2001). Understanding socially intelligent agents: A multilayered phenomenon. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 31*(5), 349-360.

Pham, T., Schneider, G., & Goose, S. (2000). A situated computing framework for mobile and ubiquitous multimedia access using small screen and composite devices. In *Proceedings of the Eighth ACM International Conference on Multimedia* (pp. 323-331).

Porn, L. M., & Patrick, K. (2002). Mobile computing acceptance grows as applications evolve. *Healthcare Financial Management, 56*(1), 66-70.

Rodgera, J. A., & Pendharkarb, P. C. (2004). A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies, 60*, 529-544.

Schaeffer, J., Plaat, A., & Junghanns, A. (2001). Unifying single-agent and two-player search. *Information Sciences, 135*(3-4), 151-175.

Schilit, W. N. (1995). *System architecture for context aware mobile computing.* Unpublished doctoral thesis, Columbia University.

Schilit, B. N., Adams, N. I., & Want, R. (1994). Context-aware computing applications. In *Proceedings of the First International Workshop on Mobile Computing Systems and Applications* (pp. 85-90).

Seager, A. (2003). M-commerce: An integrated approach. *Telecommunications International, 37*(2), 36.

Senn, J. A. (2000). The emergence of m-commerce. *Computer, 33*(12), 148-150.

Sikora, R., & Shaw, M. J. (1998). A multi-agent framework for the coordination and integration of information systems. *Management Science, 44*(11), 65-78.

Sillince, J. A. A. (1998). Extending electronic coordination mechanisms using argumentation:

The case of task allocation. *Knowledge-Based Systems, 10*(6), 325-336.

Sillince, J. A. A., & Saeddi, M. H. (1999). Computer-mediated communication: Problems and potentials of argumentation support systems. *Decision Support Systems, 26*(4), 287-306.

Slack, F., & Rowley, J. (2002). Online kiosks: The alternative to mobile technologies for mobile users. *Internet Research: Electronic Networking Applications and Policy, 12*(3), 248-257.

Strader, T. J., Lim, F. R., & Shaw, M. J. (1998). Information infrastructure for electronic virtual organization management. *Decision Support Systems, 23*(1), 75-94.

Tung, B., & Lee, J. (1999). An agent-based framework for building decision support systems. *Decision Support Systems, 25*(3), 225-237.

Turisco, F. (2000). Mobile computing is next technology frontier for healthcare providers. *Health Care Financial Management, 54*(11), 78-80.

Ulieru, M., Norrie, D., Kremer, R., & Shen, W. (2000). A multi-resolution collaborative architecture for Web-centric global manufacturing. *Information Sciences, 127*(1-2), 3-21.

Varshney, U. (1999). Networking support for mobile computing. *Communications of AIS, 1*(1), 1-30.

Wang, F. H., & Shao, H. M. (2004). Effective personalized recommendation based on time-framed navigation clustering and association mining. *Expert Systems with Applications, 27*(3), 365-377.

Wang, Y., Tan, K. L., & Ren, J. (2002). A study of building Internet marketplaces on the basis of mobile agents for parallel processing. *World Wide Web, 5*(1), 41-66.

Want, R., Hopper, A., Falcao, V., & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems, 10*(1), 91-102.

Want, R., Schilit, B., Adams, N., Gold, R., Petersen, K., Ellis, J., et al. (1995). *The PARCTAB ubiquitous computing experiment* (Tech. Rep. No. CSL-95-1). Xerox Palo Alto Research Center.

Wooldridge, M. (1997). Agent based software engineering. *IEEE Proceedings of Software Engineering, 144*(1), 26-37.

Wooldridge, M., & Jennings, N. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115-152.

Wu, D. J. (2001). Software agents for knowledge management: Coordination in multi-agent supply chains and auctions. *Expert Systems with Applications, 20*(1), 51-64.

Wu, G., Yuan, H., Tseng, S. S., & Fuyan, Z. (1999). A knowledge sharing and collaboration system model based on Internet. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics* (pp.148-152).

ENDNOTE

[1] http://ccl.northwestern.edu/netlogo/

This work was previously published in Agent Systems in Electronic Business, edited by E. Li & S. Yuan, pp. 218-238, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.16
Intelligent User Interfaces for Mobile Computing

Michael J. O'Grady
University College Dublin, Ireland

Gregory M. P. O'Hare
University College Dublin, Ireland

ABSTRACT

In this chapter, the practical issue of realizing a necessary intelligence quotient for conceiving intelligent user interfaces (IUIs) on mobile devices is considered. Mobile computing scenarios differ radically from the normal fixed workstation environment that most people are familiar with. It is in this dynamicity and complexity that the key motivations for realizing IUIs on mobile devices may be found. Thus, the chapter initially motivates the need for the deployment of IUIs in mobile contexts by reflecting on the archetypical elements that comprise the average mobile user's situation or context. A number of broad issues pertaining to the deployment of AI techniques on mobile devices are considered before a practical realisation of this objective through the intelligent agent paradigm is presented. It is the authors hope that a mature understanding of the mobile computing usage scenario, augmented with key insights into the practical deployment of AI in mobile scenarios, will aid software engineers and HCI professionals alike in the successful utilisation of intelligent techniques for a new generation of mobile services.

INTRODUCTION

Mobile computing is one of the dominant computing usage paradigms at present and encapsulates a number of contrasting visions of how best the paradigm should be realized. Ubiquitous computing (Weiser, 1991) envisages a world populated with artefacts augmented with embedded computational technologies, all linked by transparent high-speed networks, and accessible in a seamless anytime, anywhere basis. Wearable computing (Rhodes, Minar, & Weaver, 1999) advocates a

world where people carry the necessary computational artefacts about their actual person. Somewhere in between these two extremes lies the average mobile user, equipped with a PDA or mobile phone, and seeking to access both popular and highly specialized services as they go about their daily routine.

Though the growth of mobile computing usage has been phenomenal, and significant markets exists for providers of innovative services, there still exist a formidable number of obstacles that must be surpassed before software development processes for mobile services becomes as mature as current software development practices. It is often forgotten in the rush to exploit the potential of mobile computing that it is radically different from the classic desktop situation; and that this has serious implications for the design and engineering process. The dynamic nature of the mobile user, together with the variety and complexity of the environments in which they operate, provides unprecedented challenges for software engineers as the principles and methodologies that have been refined over years do not necessarily apply, at least in their totality, in mobile computing scenarios.

How to improve the mobile user's experience remains an open question. One approach concerns the notion of an application autonomously adapting to the prevailing situation or context in which end-users find themselves. A second approach concerns the incorporation of intelligent techniques into the application. In principle, such techniques could be used for diverse purposes, however, intelligent user interfaces (IUIs) represent one practical example where such techniques could be usefully deployed. Thus the objective of this chapter is to consider how the necessary intelligence can be effectively realized such that software designers can realistically consider the deployment of IUIs in mobile applications and services.

BACKGROUND

Research in IUIs has been ongoing for quite some time, and was originally motivated by problems that were arising in standard software application usage. Examples of these problems include information overflow, real-time cognitive overload, and difficulties in aiding end-users to interact with complex systems (Höök, 2000). These problems were perceived as being a by-product of direct-manipulation style interfaces. Thus, the concept of the application or user interface adapting to circumstances as they arose was conceived and the terms "adaptive" or "intelligent" user interfaces are frequently encountered in the literature. How to effectively realize interfaces endowed with such attributes is a crucial question and a number of proposals have been put forward. For example, the use of machine learning techniques has been proposed (Langley, 1997) as has the deployment of mobile agents (Mitrovic, Royo, & Mena, 2005).

In general, incorporating adaptability and intelligence enables applications to make considerable changes for personalization and customization preferences as defined by the user and the content being adapted (O'Connor & Wade, 2006). Though significant benefits can accrue from such an approach, there is a subtle issue that needs to be considered. If an application is functioning according to explicit user defined preferences it is functioning in a manner that is as the user expects and understands. However, should the system autonomously or intelligently adapt its services based on some pertinent aspect of the observed behavior of the user, or indeed, based on some other cue, responsibility for the system behavior moves, albeit partially, from the user to the system. Thus, the potential for a confused user or unsatisfactory user experience increases.

A natural question that must now be addressed concerns the identification of criteria that an

application might use as a basis for adapting its behavior. Context-aware computing (Schmidt, Beigl & Gellersen, 1999) provides one intuitive answer to this question. The notion of context first arose in the early 1990s as a result of pioneering experiments in mobile computing systems. Though an agreed definition of context has still not materialized, it concerns the idea that an application should factor in various aspects of the prevailing situation when offering a service. What these aspects might be is highly dependent on the application domain in question. However, commonly held aspects of context include knowledge of the end-user, for example through a user model; knowledge of the surrounding environment, for example through a geographic information system (GIS) model; and knowledge of the mobile device, for example through a suitably populated database. Other useful aspects of an end-user's context include an understanding of the nature of the task or activity currently being engaged in, knowledge of their spatial context, that is, location and orientation, and knowledge of the prevailing social situation. Such models can provide a sound basis for intelligently adapting system behavior. However, capturing the necessary aspects of the end-user's context and interpreting it is frequently a computationally intensive process, and one that may prove intractable in a mobile computing context. Indeed, articulating the various aspects of context and the interrelationships between them may prove impossible, even during system design (Greenberg, 2001). Thus, a design decision may need to be made as to whether it is worth working with partial or incomplete models of a user's context. And the benefit of using intelligent techniques to remedy deficiencies in context models needs to be considered in terms of computational resources required, necessary response time and the ultimate benefit to the end-user and service provider.

SOME REFLECTIONS ON CONTEXT

Mobile computing spans many application domains and within these, it is characterized by a heterogeneous landscape of application domains, individual users, mobile devices, environments and tasks (Figure 1). Thus, developing applications and services that incorporate a contextual component is frequently an inherently complex and potentially time-consuming endeavor, and the benefits that accrue from such an approach should be capable of being measured in some tangible way. Mobile computing applications tend to be quite domain specific and are hence targeted at specific end-users with specialized tasks or objectives in mind. This is in contrast to the one-size-fits-all attitude to general purpose software development that one would encounter in the broad consumer PC arena. For the purposes of this discussion, it is useful to reflect further on the following aspects of the average mobile user's context: end-user profile, devices characteristics, prevailing environment and social situation.

Figure 1. An individual's current activity is a notoriously difficult aspect of an individual's context to ascertain with certainty

User Profile

Personalization and customization techniques assume the availability of sophisticated user models, and currently form an indispensable component of a number of well-known e-commerce related Web sites. Personalizing services for mobile computing users is an attractive proposition in many domains as it offers a promising mechanism for increasing the possibility that the end-users receive content that is of interest to them. Though this objective is likewise shared with owners of e-commerce sites, there are two issues that are of particular importance when considering the mobile user. Firstly, mobile interactions are almost invariably short and to the point. This obligates service providers to strive to filter, prioritize, and deliver content that is pertinent to the user's immediate requirements. The second issue concerns the question of costs. Mobile users have to pay for services, which may be charged on a KB basis, thus giving mobile users a strong incentive to curtail their use of the service in question if dissatisfied.

A wide number of features and characteristics can be incorporated into user models. As a basic requirement, some information concerning the user's personal profile, for example, age, sex, nationality and so on, is required. This basic model may then be augmented with additional submodels that become increasingly domain-specific. In the case of standard e-commerce services, a record of the previous purchasing history may be maintained and used as a basis for recommending further products. Electronic tourist guides would require the availability of a cultural interest model, which as well as indicating cultural topics of interest to the user, would also provide some metric that facilitated the prioritization of their cultural interests.

Device Characteristics

Announcements of new devices are occurring with increasing frequency. Each generation successively increases the number of features offered, some of which would not be associated with traditional mobile computing devices, embedded cameras and MP3 players being cases in point. Though offering similar features and services, there are subtle differences between different generations, and indeed interim releases within the same generation, that make the life of a service provider and software professional exceedingly difficult and frequently irritating. From an interface perspective, screen size and support for various interaction modalities are two notable ways in which devices differ, and these have particular implications for the end-user experience. This problem is well documented in the literature and a number of proposals have been put forward to address this, the plasticity concept being a notable example (Thevenin & Coutaz, 1999). Other aspects in which mobile devices differ include processor, memory and operating system; all of which place practical limitations on what is computationally feasible on the device.

Prevailing Environment

The notion of environment is fundamental to mobile computing and it is the dynamic nature of prevailing environment in which the mobile user operates that most distinguishes mobile computing from the classic desktop usage paradigm. As an illustration, the case of the physical environment is now considered, though this in no way diminishes the importance of the prevailing electronic infrastructure. Scenarios in which mobile computing usage can occur are multiple and diverse. The same goes for physical environments. Such environments may be hostile in the sense that they do not lend themselves to easily accessing electronic infrastructure such as telecommunications networks. Other environments may experience extreme climatic conditions thus causing equipment to fail.

Developing a service that takes account of or adapts to the local physical environment is an attractive one. Two prerequisites are unavoidable, however. A model of the environment particular to the service domain in question must be available, and the location of the end-user must be attainable. In the former case, the service provider must construct this environmental model, possibly an expensive endeavor in terms of time and finance. In the latter case, an additional technological solution must be engaged—either one based on satellites, for example GPS, or one that harnesses the topology of the local wireless telecommunications networks. Each solution has its respective advantages and disadvantages, and a practical understanding of each is essential. However, by fulfilling these prerequisites, the service provider is in a position to offer services that take the end-users' physical position into account. Indeed, this vision, often termed location-aware computing (Patterson, Muntz & Pancake, 2003), has grasped the imagination of service providers and end-users alike. In essence, it is a practical example of just one single element of an end-user's context being interpreted and used as a basis for customizing services.

Social Situation

Developing a service that adapts to the end-user's prevailing social context is fraught with difficulty, yet is one that many people would find useful. What exactly defines social context is somewhat open to interpretation but in this case, it is considered to refer to the situation in which end-users find themselves relevant to other people. This is an inherently dynamic construct and capturing the prevailing social situation introduces an additional level of complexity not encountered in the contextual elements described previously.

In limited situations, it is possible to infer the prevailing social situation. Assuming that the end-user maintains an electronic calendar, the detection of certain keywords may hint at the prevailing social situation. Examples of such keywords might include lecture, meeting, theatre and so on. Thus, an application might reasonably deduce that the end-user would not welcome interruptions, and, for example, proceed to route incoming calls to voicemail and not alert the end-user to the availability of new email. Outside of this, one has to envisage the deployment of a suite of technologies to infer social context. For example, it may be that a device, equipped with a voice recognition system, may be trained to recognize the end-user's voice, and on recognizing it, infer that a social situation is prevailing. Even then, there may be a significant margin of error; and given the power limitations of the average mobile device, running a computationally intensive voice recognition system continuously may rapidly deplete battery resources.

ARTIFICAL INTELLIGENCE IN MOBILE COMPUTING

Artificial intelligence (AI) has been the subject of much research, and even more speculation, for almost half a century by now. Though failing to radically alter the world in the way that was envisaged, nevertheless, AI techniques have been successfully harnessed in a quite a number of select domains and their incorporation into everyday applications and services continues unobtrusively yet unrelentingly. Not surprising, there is significant interest amongst the academic community in the potential of AI for addressing the myriad of complexity that is encountered in the mobile computing area. From the previous discussion, some sources of this complexity can be easily identified. Resource management, ambiguity resolution, for example, in determining contextual state and resolving user intention in multimodal interfaces, and adaptation, are just some examples. Historically, research in AI has focuses on various issues related to these very topics. Thus, a significant body of research

already exists in some of the very areas that can be harnessed to maximum benefit in mobile computing scenarios. A detailed description of these issues may be found elsewhere (Krüger & Malaka, 2004).

One pioneering effort at harnessing the use of intelligent techniques on devices of limited computational capacity is the Ambient intelligence (AmI) (Vasilakos & Pedrycz, 2006) initiative. AmI builds on the broad mobile computing vision as propounded by the ubiquitous computing vision. It is of particular relevance to this discussion as it is essentially concerned with usability and HCI issues. It was conceived in response to the realization that as mobile and embedded artefacts proliferate, demands for user attention would likewise increase, resulting in environments becoming inhabitable, or more likely, people just disabling the technologies in question. In the AmI concept, IUIs are envisaged as playing a key role in mediating between the embedded artefacts and surrounding users. However, AmI does not formally ratify the use of any particular AI technique. Choice of technique is at the discretion of the software designer whose selection will be influenced by a number of factors including the broad nature of the domain in question, the requirements of the user, the capability of the available technology and the implications for system performance and usability.

Having motivated the need for AI technologies in mobile contexts, practical issues pertaining to their deployment can now be examined.

STRATEGIES FOR HARNESSING AI TECHNIQUES IN MOBILE APPLICATIONS

It must be reiterated that AI techniques are computationally intensive. Thus, the practical issue of actually incorporating such techniques into mobile applications needs to be considered carefully. In particular, the implications for performance must be determined as this could easily have an adverse effect on usability. There are three broad approaches that can be adopted when incorporating AI into a mobile application and each is now considered.

Network-Based Approach

Practically all mobile devices are equipped with wireless modems allowing access to data services. In such circumstances, designers can adopt a kind of client/server architecture where the interface logic is hosted on the mobile devices and the core application logic deployed on a fixed server node. The advantage of such an approach is that the designer can adopt the most appropriate AI technologies for the application in question. However, the effect of network latency must be considered. If network latency is significant, the usability of the application will be adversely affected. Likewise, data rates supported by the network in question must be considered. Indeed, this situation is aggravated when it is considered that a number of networks implement a channel sharing system where the effective data rate at a given time is directly proportional to the number of subscribers currently sharing the channel. It is therefore impossible to guarantee an adequate quality of service (QoS) making the prediction of system performance difficult. Often, the worst case scenario must be assumed. This has particular implications where the AI application on the fixed server node needs either a significant amount of raw data or a stream of data to process.

One key disadvantage of placing the AI component on a fixed server node concerns the issue of cost. There is a surcharge for each KB of data transferred across the wireless network, and though additional revenue is always welcome, the very fact that the subscriber is paying will affect their perception of application in question and make them more demanding in their expectations.

A network–based AI approach is by far the most common and has been used in quite a number of applications. For example, neural networks have been used for profiling mobile users in conversational interfaces (Toney, Feinberg & Richmond, 2004). InCa (Kadous & Sammut, 2004) is a conversational agent that runs on a PDA but uses a fixed network infrastructure for speech recognition.

Distributed Approach

In this approach, the AI component of the service may be split between the mobile device and the fixed network node. The more computationally expensive elements of the service are hosted on the fixed network node while the less expensive elements may be deployed on the device. Performance is a key limitation of this approach as the computational capacity of the devices in question as well as the data-rates supported by the wireless network can all contribute to unsatisfactory performance. From a software engineering perspective, this approach is quite attractive as distributed AI (DAI) is a mature research discipline in its own right; and a practical implementation of DAI is the multi-agent system (MAS) paradigm.

One example of an application that uses a distributed approach is Gulliver's Genie (O'Grady & O'Hare, 2004). This is a tourist information guide for mobile tourists, realized as a suite of intelligent agents encompassing PDAs, wireless networks and fixed network servers. Agents on the mobile device are responsible for manipulating the user interface while a suite of agents on the fixed server collaborate to identify and recommend multimedia content that is appropriate to the tourist's context.

Embedded Approach

As devices grow in processing power, the possibility of embedding an AI based application on the actual physical device becomes ever more feasible. The key limitation is performance, which is a direct result of the available hardware. This effectively compromises the type of AI approach that can be usefully adopted. Overtime, it can be assumed that the capability and variety of AI techniques that can be deployed will increase as developments in mobile hardware continue and the demand for ever-more sophisticated applications increases. From an end-user viewpoint, a key advantage of the embedded approach concerns cost as the number of connections required is minimized.

One example of an application that uses the embedded approach is iDorm (Hagras et al., 2004), a prototype AmI environment. This environment actually demonstrates a variety of embedded agents including fixed motes, mobile robots and PDAs. These agents collaborate to learn and predict user behavior using fuzzy logic principles and, based on these models, the environment is adapted to the inhabitant's needs.

Deployment Considerations

Technically, all three approaches are viable, but the circumstances in which they may be adopted vary. For specialized applications, the networked AI approach is preferable as it offers greater flexibility and maximum performance, albeit at a cost. For general applications, the embedded approach is preferable, primarily due to cost limitations, but the techniques that can be adopted are limited. The distributed approach is essentially a compromise, incorporating the respective advantages and disadvantages of both the networked and embedded approach to various degrees. Ultimately, the nature of the application domain and the target user base will be the major determinants in what approach is adopted. However, in the longer term, it is the embedded approach that has the most potential as it eliminates the negative cumulative effect of network vagrancies, as well as hidden costs.

Thus, for the remainder of this chapter, we focus on the embedded approach and consider how this might be achieved.

So what AI techniques can be adopted, given the inherent limitations of mobile devices? Various techniques have been demonstrated in laboratory conditions but one paradigm has been demonstrated to be computationally tractable on mobile devices: intelligent agents. As well as forming the basis of mobile intelligent information's systems, a number of toolkits have been made available under open source licensing conditions thus allowing software engineers access to mature platforms at minimum cost. Before briefly considering some of these options, it is useful to reflect on the intelligent agent paradigm.

THE INTELLIGENT AGENT PARADIGM

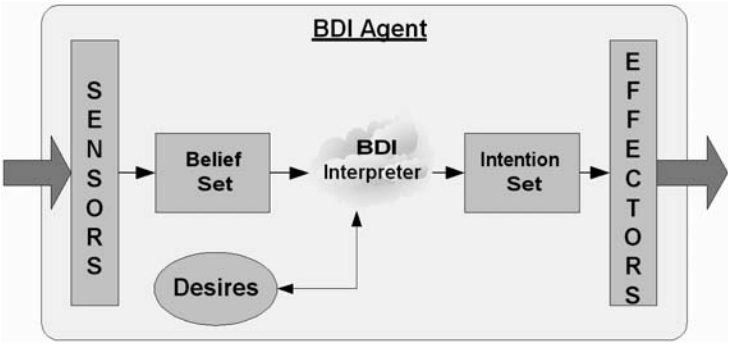
Research in intelligent agents has been ongoing since the 1970s. Unfortunately, the term agent has been interpreted in a number of ways thereby leading to some confusion over what the term actually means. More precisely, the characteristics that an arbitrary piece of software should possess before applying the term agent to it are debatable. In essence, an agent may be regarded as a computational entity that can act on behalf of an end-user, another agent or some other software artefact. Agents possess a number of attributes that distinguish them from other software entities. These include amongst others:

- **Autonomy:** The ability to act independently and without direct intervention from another entity, either human or software-related
- **Proactivity:** The ability to opportunistically initiate activities that further the objectives of the agent
- **Reactivity:** The ability to respond to events perceived in the agent's environment;
- **Mobility:** The ability to migrate to different nodes of a network as the need to fulfill its objectives dictates; and
- **Social ability:** The ability to communicate with other agents using a shared language and ontology leading to shared or collaborative efforts to achieve individual and shared objectives.

To what extent an agent possesses or utilizes each of those attributes is at the discretion of the designer. For clarity purposes, it is useful to consider agents as existing on a scale. At the lower end are so-called reactive agents. Such agents act in a stimulus-response manner, and a typical usage scenario might involve the agent monitoring for user interaction and reacting to it. Such agents are generally classified as weak agents (Wooldridge & Jennings, 1995). At the other end of the scale are so-called strong agents. Such agents maintain a sophisticated model of their environment, a list of goals or objectives, and plans detailing how to achieve these objectives. Such agents support rational reasoning in a collaborative context and are usually realized as multi-agent systems (MAS). This strong notion of agenthood is synonymous with the view maintained by the AI community.

One popular interpretation of the strong notion of agency is that of the belief-desire-intention (BDI) paradigm (Rao & Geogeff, 1995). This is an intuitive and computationally tractable interpretation of the strong agency stance. To summarize: beliefs represent what the agent knows about its environment. Note that the term environment can have diverse meanings here and may not just relate to the physical environment. Desires represent the objectives of the agent, and implicitly the raison d'être for the application. However, at any moment in time, an agent may be only capable of fulfilling some of its desires, if even that. These desires are then formulated as intentions and the agent proceeds to fulfill these intentions. The cycle of

Figure 2. Architecture of a BDI agent



updating its model of the environment, identifying desires that can be fulfilled, and realizing these intentions is then repeated for the duration of the agent's lifecycle (Figure 2).

When should agents be considered for realizing a software solution? Opinion on this is varied. If the solution can be modeled as a series of dynamic interacting components, then agents may well offer a viable solution. However, many see agents as being particular useful in situations that are inherently complex and dynamic as their native capabilities equip them for handling the myriad of situations that may arise. Naturally, there are many situations that fulfill the criteria but, for the purposes of this discussion, it can be easily seen that the mobile computing domain offers significant opportunities for harvesting the characteristics of intelligent agents.

Intelligent Agents for Mobile Computing

As the capability of mobile devices grew, researchers in the intelligent agent community became aware of the feasibility of deploying agents on such devices, and perceived mobile computing as a potentially fertile area for the intelligent agent paradigm. A common approach was to extend the functionality of existing and well-documented MAS environments such that they could operate on mobile devices. It was not necessary to port the entire environment on to the device; it was just necessary to develop an optimized runtime engine for interpreting the agent logic. In this way, the MAS ethos is persevered and such an approach subscribes to the distributed AI approach alluded to previously. A further benefit was that existing agent-oriented software engineering (AOSE) methodologies could be used. In the case of testing, various toolkits have been released by the telecommunications manufacturers that facilitate the testing of mobile applications. A prudent approach is of course to test the application at various stages during its development on actual physical devices, as this will give a more accurate indication of performance, the look and feel (L&F) of the application and so on. For a perspective on deploying agents on mobile devices, the interested reader should consult Carabelea and Boissier (2003).

While a number of environments may be found in the literature for running agents on mo-

bile devices, the following toolkits form a useful basis for initial consideration:

1. **LEAP (Lightweight Extensible Agent Platform)** (Bergenti, Poggi, Burg, et al., 2001) is an extension of the well-documented JADE platform (Bellifemine, Caire, Poggi et al., 2003). It is FIPA (http://www.fipa.org/) compliant and capable of operating on both mobile and fixed devices.
2. **MicroFIPA-OS** (Laukkanen, Tarkoma & Leinonen, 2001) is a minimized footprint of the FIPA-OS agent toolkit (Tarkoma & Laukkanen, 2002). The original FIPA-OS was designed for PCs and incorporated a number of features that did not scale down to mobile devices. Hence, MicroFIPA-OS minimizes object creation, reduces computational overhead and optimizes the use of threads and other resource pools.
3. **AFME (Agent Factory Micro Edition)** (Muldoon, O'Hare, Collier & O'Grady, 2006) is derived from Agent Factory (Collier, O'Hare, Lowen, & Rooney, 2003), a framework for the fabrication and deployment of agents that broadly conform to the BDI agent model. It has been specifically designed for operation on cellular phones and such categories of devices.
4. **JACK** is, in contrast to the three previous frameworks, a commercial product from the Agent Oriented Software Group (http://www.agent-software.com). It comes with a sophisticated development environment, and like AFME, conforms to the BDI agent model.

A detailed description of the each of these systems is beyond the scope of this discussion. However, the interested reader is referred to (O'Hare, O'Grady, Muldoon & Bradley, 2006) for a more advanced treatment of the toolkits and other associated issues.

FUTURE TRENDS

As mobile devices proliferate, and each generation surpasses its predecessor in terms of raw computational capacity and supported features, the potential for incorporating additional AI techniques will increase. In a similar vein, new niche and specialized markets for mobile services will appear. If a more holistic approach is taken towards mobile computing, it can be seen that developments in sensor technologies, fundamental to the ubiquitous and pervasive vision, will follow a similar trajectory. Indeed, the possibility of deploying intelligent agents on sensors is being actively investigated in widespread expectation that the next generation of sensors will incorporate processors of a similar capability to the current range of PDAs. Such a development is essential if the AmI vision to reach fruition.

As the possibility of incorporation of ever more sophisticated AI techniques increases, the potential for extending and refining the adaptability and IUI constructs for the support of mobile users increase. Indeed, adaptability may reach its fulfillment through the incorporation of autonomic computing precepts (Kephart & Chess, 2003). Self-configuring, self-healing, self-optimizing and self-protecting are the key attributes of an autonomic system, and it can be seen that incorporation of AI techniques may make the realization of these characteristics more attainable.

Finally, the practical issues of engineering mobile AI solutions must be considered. Mobile computing poses significant challenges to the traditional software engineering process, and the broad issue of how best to design for mobile services still needs to be resolved. The situation is exacerbated when AI technologies are included. However, it may be envisaged that as experience and knowledge of the mobile computing domain deepens and matures, new methodologies and best practice principles will emerge.

CONCLUSION

Mobile computing scenarios are diverse and numerous, and give rise to numerous challenges that must be overcome if the end-user experience is to be a satisfactory one. IUIs offers one viable approach that software designers can adopt in their efforts to make their systems more usable in what is frequently a hostile environment. However, the pragmatic issue of realizing mobile applications that incorporate intelligent techniques is of critical importance and gives rise to significant technical and design obstacles.

In this chapter, the broad issue of realizing an intelligent solution was examined in some detail. At present, the intelligent agent paradigm offers an increasingly viable proposition for those designers who wish to include intelligent techniques in their designs. To illustrate the issues involved, the intelligent agent paradigm was discussed in some detail.

As mobile developments continue unabated, the demand for increasingly sophisticated applications and services will likewise increase. Meeting this demand will pose new challenges for software and HCI professionals. A prudent and selective adoption of intelligent techniques may well offer a practical approach to the effective realization of a new generation of mobile services.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Science Foundation Ireland (SFI) under Grant No. 03/IN.3/1361.

REFERENCES

Bellifemine, F., Caire, G., Poggi, A., & Rimassa, G. (2003). *JADE — A white paper.* Retrieved January 2007 from http://jade.tilab.com/papers/2003/WhitePaperJADEEXP.pdf

Bergenti, F., Poggi, A., Burg, B., & Caire, G. (2001). Deploying FIPA-compliant systems on handheld devices. *IEEE Internet Computing, 5*(4), 20-25.

Collier, R.W., O'Hare, G.M.P., Lowen, T., & Rooney, C.F.B., (2003). Beyond prototyping in the factory of agents. In J. G. Carbonell & J. Siekmann (Eds.), *Lecture Notes In Computer Science*, 2691, 383-393, Berlin: Springer.

Carabelea, C., & Boissier O. (2003). Multi-agent platforms on smart devices: Dream or reality. Retrieved January, 2007, from http://www.emse.fr/~carabele/papers/carabelea.soc03.pdf.

Greenberg, S. (2001). Context as a dynamic construct. *Human-Computer Interaction*, 16, 257-268.

Hagras, H., Callaghan, V., Colley, M., Clarke, G., Pounds-Cornish, A., & Duman, H. (2004). Creating an ambient-intelligence environment using embedded agents. *IEEE Intelligent Systems, 19*(6) 12-20.

Höök, K. (2000). Steps to take before intelligent user interfaces become real. *Interacting with Computers, 12*, 409-426.

Kadous, M.W., & Sammut, C. (2004). InCA: A mobile conversational agent. *Lecture Notes in Computer Science*, 3153, 644-653. Berlin: Springer.

Kephart, J.O., & Chess, D.M. (2003). The vision of autonomic computing. *IEEE Computer, 36*(1), 41-50.

Kruger, A., & Malaka, R. (2004). Artificial intelligence goes mobile. *Applied Artificial Intelligence, 18*, 469-476.

Langley, P. (1997). Machine learning for adaptive user interfaces. *Lecture Notes in Computer Science*, 1303, 53-62. Berlin:Springer.

Laukkanen, M., Tarkoma, S., & Leinonen, J. (2001). FIPA-OS agent platform for small-footprint devices. *Lecture Notes in Computer Science*, 2333, 447-460. Berlin:Springer.

Mitrovic, N., Royo, J. A., & Mena, E. (2005). Adaptive user interfaces based on mobile agents: Monitoring the behavior of users in a wireless environment. *Proceedings of the Symposium on Ubiquitous Computation and Ambient Intelligence* (pp. 371-378), Madrid: Thomson-Paraninfo.

Muldoon, C., O'Hare, G.M.P., Collier, R.W., & O'Grady, M.J. (2006). Agent factory micro edition: A framework for ambient applications, *Lecture Notes in Computer Science*, 3993, 727-734. Berlin:Springer.

O'Connor, A., & Wade, V., (2006). Informing context to support adaptive services, *Lecture Notes in Computer Science*, 4018, 366-369. Berlin: Springer.

O'Grady, M.J., & O'Hare, G.M.P. (2004). Just-in-time multimedia distribution in a mobile computing environment. *IEEE Multimedia*, *11*(4), 62-74.

O'Hare, G.M.P., O'Grady, M.J., Muldoon, C., & Bradley, J.F. (2006). Embedded agents: A paradigm for mobile services. *International Journal of Web and Grid Services, 2*(4), 355-378.

Patterson, C.A., Muntz, R.R., & Pancake, C.M. (2003). Challenges in location-aware computing. *IEEE Pervasive Computing, 2*(2), 80-89.

Rao, A.S., & Georgeff, M.P. (1995). BDI agents: from theory to practice. In V. Lesser and L. Gasser (Eds.), *Proceedings of the First International Conference on Multiagent Systems* (pp. 312-319). California: MIT Press.

Rhodes, B.J., Minar, N. & Weaver, J. (1999). Wearable computing meets ubiquitous computing: Reaping the best of both worlds. *Proceedings of the Third International Symposium on Wearable Computers* (pp. 141-149). California: IEEE Computer Society

Schmidt, A., Beigl, M., & Gellersen, H-W, (1999). There is more to context than location. *Computers and Graphics, 23*(6), 893-901.

Tarkoma, S., & Laukkanen, M. (2002). Supporting software agents on small devices. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)* (pp. 565-566). NewYork: ACM Press.

Thevenin, D., & Coutaz, J. (1999). Plasticity of user interfaces: Framework and research agenda. In (M. A. Sasse & C. Johnson (Eds.), *Proceedings of IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT'99)*, (pp. 110-117). Amsterdam: IOS Press.

Toney, D., Feinberg D., & Richmond, K. (2004). Acoustic features for profiling mobile users of conversational interfaces, *Lecture Notes in Computer Science*, 3160 (pp. 394-398). Berlin: Springer.

Vasilakos, A., & Pedrycz, W. (2006). *Ambient intelligence, wireless networking, ubiquitous Computing.* Norwood:Artec House.

Weiser, M. (1991). The computer for the twenty-first century. *Scientific American*, *265*(3), 94-100.

Wooldridge, M., & Jennings, N.R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review, 10*(2), 115-152.

KEY TERMS

Ambient Intelligence: (AmI) was conceived by the Information Society Technologies Advisory Group (ISTAG) as a means of facilitating intuitive interaction between people and ubiquitous

computing environments. A key enabler of the AmI concept is the intelligent user interface.

BDI Architecture: The Belief-Desire-Intention (BDI) architecture is an example of a sophisticated reasoning model based on mental constructs that can be used by intelligent agents. It is allows the modeling of agents behaviors in an intuitive manner that complements the human intellect.

Context: Context-aware computing considers various pertinent aspects of the end-user's situation when delivering a service. These aspects, or contextual elements, are determined during invocation of the service and may include user profile, for example language, age, and so on. Spatial contextual elements, namely location and orientation, may also be considered.

Intelligent Agent: Agents are software entities that encapsulate a number of attributes including autonomy, mobility, sociability, reactivity and proactivity amongst others. Agents may be reactive, deliberative or hybrid. Implicit in the agent construct is the requirement for a sophisticated reasoning ability, a classic example being agents modeled on the BDI architecture.

Intelligent User Interface: Harnesses various techniques from artificial intelligence to adapt and configure the interface to an application such that the end-user's experience is more satisfactory.

Mobile Computing: A computer usage paradigm where end-users access applications and services in diverse scenarios, while mobile. Mobile telephony is a popular realization of this paradigm, but wearable computing and telematic applications could also be considered as realistic interpretations of mobile computing.

Multi-Agent System: A suite of intelligent agents, seeking to solve some problem beyond their individual capabilities, come together to form a multi-agent system (MAS). These agents collaborate to fulfill individual and shared objectives.

Ubiquitous Computing: Conceived in the early 1990s, ubiquitous computing envisages a world of embedded devices, where computing artefacts are embedded in the physical environment and accessed in a transparent manner.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 318-329, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.17
mCity:
User Focused Development of Mobile Services Within the City of Stockholm

Anette Hallin
Royal Institute of Technology (KTH), Sweden

Kristina Lundevall
The City of Stockholm, Sweden

ABSTRACT

This chapter presents the mCity Project, a project owned by the City of Stockholm, aiming at creating user-friendly mobile services in collaboration with businesses. Starting from the end-users' perspective, mCity focuses on how to satisfy existing needs in the community, initiating test pilots within a wide range of areas, from health care and education, to tourism and business. The lesson learned is that user focus creates involvement among end users and leads to the development of sustainable systems that are actually used after they have been implemented. This is naturally vital input not only to municipalities and governments but also for the IT/telecom industry at large. Using the knowledge from mCity, the authors suggest a new, broader definition of "m-government" which focuses on mobile people rather than mobile technology.

INTRODUCTION

All over the world, ICT technologies are used to an increasing extent within the public sector. For cities, ICTs not only provide the possibilities of improving the efficiency among its employees and its service towards tourists, citizens, and companies; it is also an important factor in the development of the city and its region, as ICTs today generally are considered to constitute the driving force of economy and social change (Castells, 1997). It is also argued that ICTs can improve efficiency, enhance transparency, control, networking and innovation (Winden, 2003). Thus, several cities are involved in projects concerning the development, testing, and implementation of ICTs. A few examples include Crossroads Copenhagen in Denmark, Testbed Botnia, and TelecomCity from the cities of Luleå and Karlskrona in

Sweden. Within all these projects, triple-helix like organizations are used involving the local municipality or national government, the local university, and the locally-based companies (Jazic & Lundevall, 2003)

Also within the City of Stockholm, there is such a project—the mCity Project. This was launched by the City of Stockholm in January of 2002, with the aim of organizing "the mobile city" through the implementation of relevant ICTs. The mCity Project consists of several small pilot projects, focusing on identifying needs in the community and creating solutions to these. In this chapter, we intend to describe this project, its organization, work processes, and the results. We also discuss the experiences made and how the project can serve as an inspiration towards a broader understanding of "m-government".

BRIEFLY ABOUT THE CITY OF STOCKHOLM

The City of Stockholm is Sweden's largest municipality with about 760,000 inhabitants,[1] but is, compared to other capitals in the world, a small city. Due to the Swedish form of government, Stockholm—as well as all other Swedish cities—has large responsibilities, including child care, primary and secondary education, care of the elderly, fire-fighting, city planning, and maintenance, and so forth. All these responsibilities are financed through income taxes, at levels set by the cities themselves, with no national interference. The operational responsibility lies, in the case of Stockholm, on 18 district councils and on 16 special administrations, depending on the issue. Through 15 different fully-owned or majority-interest, joint-stock and associated companies (hereafter called "municipal companies"), the City of Stockholm also provides water, optical fibre-infrastructure, housing (the City of Stockholm has the largest housing corporation in the country), shipping-facilities (the ports in the Stockholm area), parking, tourist information, the city theatre, the Globe Arena (for sports, concerts and other events) etc. In total, the city has an organization comprising 50,000 employees, and a yearly turn over of 31.5 billion SEK,[2] which is equivalent to about 5 million USD. For the City of Stockholm, it is only natural to engage in ICT projects of different kinds, as this could be expected to have both financial and pedagogical benefits within this large organization—just as it had for other public organizations in Sweden (Grenblad, 2003).

In fact, ICT projects are encouraged by the City of Stockholm through the Stockholm "E-Strategy". This is a visionary and strategic document, issued by the City Council[3] in the beginning of 2001 which—among other things—firmly states the role of the citizen as the central figure for all activities in the city organization; the development of mobile technologies to enhance flexibility, as well as the importance of the city acting to aid Swedish ICT industry (*The City of Stockholm's E-Strategy, 19th of February 2001*). It is the City Executive Board[4] which is responsible for implementing the resolution of the City Council, but the "E-Strategy" document also points to the responsibility of the management of the different district councils, special administrations, and municipal companies for the strategic development of ICTs within each organization. The document also describes the function of "the IT Council", which is to ensure that the e-strategy is implemented in a good way within the municipal organization, that is, not as a separate strategy, but in close contact with the activities for which the organizations are responsible.

BACKGROUND, ORGANIZATION, AND GOALS

The idea of mCity was born in 2000 when the former EU Commissioner Martin Bangeman suggested a cooperation between European cities

in order to stimulate the use of the upcoming 3G network and its services. In January 2001, a workshop was held with representatives from a number of major cities, telecom operators, vendors, and investors. A project proposal was submitted by Bangeman, suggesting that a few other European cities—Stockholm, Bremen, and Berlin among others—should start a holding corporation in order to develop and sell 3G services. However, this collaboration project did not become a reality. Instead, the City of Stockholm decided to proceed with a smaller scale project—mCity.

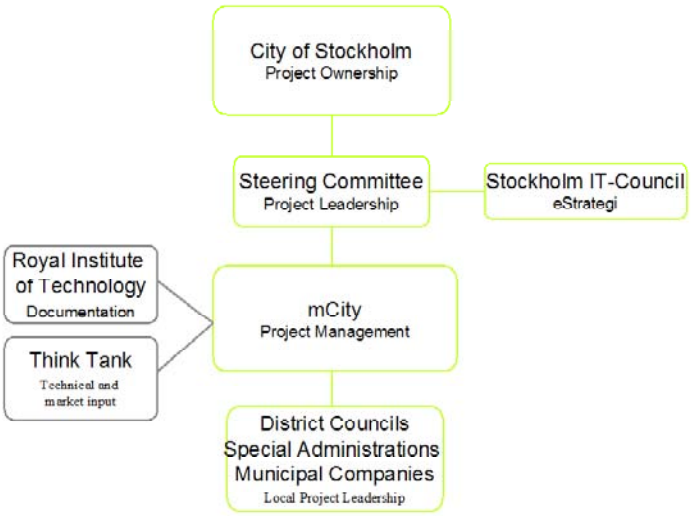
The following goals have been specified for the mCity Project in Stockholm:

- **To improve the working environment for the employees of the City of Stockholm.** By putting people in the center and letting them lead the development of mobile services, they will help develop services that will ease their own work tasks and their everyday lives.
- **To increase the quality of services for citizens.** The mCity Project strives to improve the service of the city to its citizens and visitors by improving the work environment for employees and by introducing citizen-specific solutions.
- **To stimulate the regional business (IT/Telecom).** By developing new solutions in collaboration with industry, new opportunities for the ICT industry within Stockholm, and throughout Sweden are developed, thereby creating a strong home market for companies in Stockholm.
- **To reinforce Stockholm's profile as an IT capital.** By developing new and useful mobile services, Stockholm's reputation as a leading IT capital will be further reinforced.
- **To spread the good example.** By working with small-scale test environments and small-scale tests, the results can be duplicated if successful. By involving the end users closely in the project, sustainability is ensured. An effect of more deeply involved users is that the users themselves become spokespersons for the services and actually help spread the word.

During its first year, the project was located in one of Stockholm's district councils, which meant close contact with the end users. The project manager felt, however, that in order to keep up with the ICT development in other parts of the city, the project would be better off if it could be located more centrally in the organization. Since then, the project has been moved closer to the central administrative organization in the city.

The project organization of mCity is described in Figure 1. The Steering Committee, organized with representatives from different parts of the city, for example the IT Department and the City of Stockholm Executive Office,[5] make strategic decisions about budget issues, what projects to initiate, and so forth. Different heads have chaired the Steering Committee during the course of the project. There are also members from the Stockholm IT Council in the Steering Committee, to ensure that the mCity Project follows Stockholm's E-Strategy. The different pilot projects are initiated together with district councils, special administrations, or municipal companies which undertake the responsibility of local project management in each case. The mCity Project Manager is in charge of initiating and setting up the local projects in collaboration with the local project management and then keeps track of the day-to-day development of the projects. He/She is also responsible for collecting and spreading information about the projects, and for preparing the meetings with the Steering Committee as well as implementing the decisions of the Steering Committee. In their work, he/she can also use the Think Tank, to which a number of companies within the mobile technology industry belong, to ask for advice concerning technology or market requirements/development. Finally, a researcher

Figure 1. Organization of the mCity Project



from KTH, the Royal Institute of Technology, has been responsible for documenting the project.

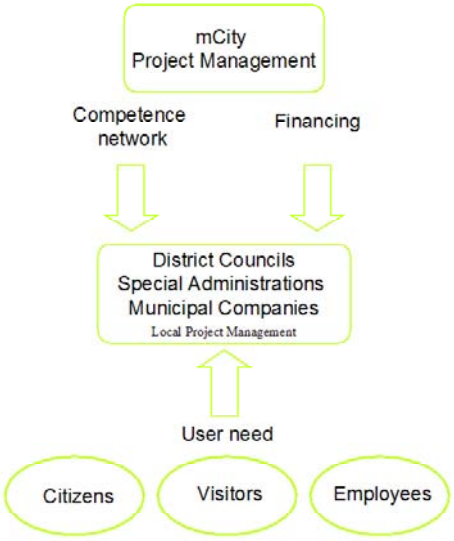
WORKING PROCESS

Within the mCity Project, services for both private and public sectors are tested and thereafter developed in a larger scale if proven relevant. The services are operated and tested in "small islands" because it makes it easier to get close to the users and to change the tested services if something needs to be improved. Using this model, mCity has been able to connect groups with specific needs with companies developing mobile services that can satisfy these needs.

End-user needs, that is, the needs of citizens, visitors and employees within the City of Stockholm, form the starting point of every initiative within mCity—see Figure 2. One way of creating situations where users can make their voices heard is by initiating hearings, focus groups, interviews, and so forth. In some cases, the mCity Project Manager has been involved in this first part of the process; in other cases, the local management of the different district councils, special administrations, or municipal companies take the initiative of formulating an application, specifying the need. The exact details of the working process have shifted, depending on the organizational setting of the project.

In the next step, the mCity Project management uses their Competence Network to form a group with technical expertise to which the user's need is presented. The group ponders about the possible technical solutions suitable to solve the problems and in this process, the end-users' knowledge of ICTs, their workload, and the financial/technical situation of the user environment are also taken into account. Depending on the situation, mCity can also contribute financially to the pilot project.

Figure 2. Working process within mCity



Through mCity, the hope is to accomplish a better every day life for end users. Therefore, the benefits of the services developed in relation to the concrete needs of users, are of high interest and hopefully, it is also possible to measure the added value. End solutions should be easy to use—it should be almost intuitive to understand how to use the provided service. This is one reason for why simple technology is mostly used in mCity Projects—technology is seldom the problem, the focus is rather on what to introduce and how to introduce it. To summarize, the working process can be described in three keywords:

- user-oriented
- benefit-driven
- simple

It should be pointed out that mCity primarily does small-scale pilot projects; when these have been launched, it is the responsibility of the district councils, the special administrations, or the municipal companies involved to decide whether to keep running the project, to enlarge it, and also to take the full operational and financial responsibility for the future project.

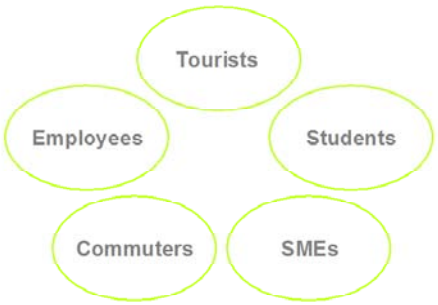
PILOT PROJECTS

mCity has started and financed several pilot projects since its launch in 2002. Different user groups are in need of different services and the largest segments identified are people who work, live in, and visit the City of Stockholm, as shown earlier in Figure 2. Through the pilot projects, these segments have been further specified, as described in Figure 3: tourists, students, SMEs, commuters, and city employees.

Tourists

The very first project within mCity was carried out in 2002 for tourists, when the official event database owned by Stockholm Visitors' Board[6] was made available via mobile Internet. The city wanted to do this in relation to its 750th anniversary which was to be celebrated that year, and it was decided that something new should be tested, which is why WAP was chosen.

Figure 3. Focus groups of the mCity Project



A few years later in 2004, another service targeting tourists was developed by mCity. This time the development process was conducted by a group of talented students taking a project course at the Royal Institute of Technology. This project, *.tourism,* was initiated by the Art Council at the Cultural Administration in order to find new ways of making information about Stockholm available through new technology. The result, a Web site with information on statues, art objects, and buildings of interest is available via mobile or fixed Internet on the address, www.explore.stockholm.se. The server recognizes if the user is accessing the Web site from a PDA, a laptop, or a mobile phone. By using XML functionality, separate interfaces for the different devices are shown, giving the user the best experience possible depending on the device used.

On the Web site, it is possible to search by the name of an object, a location, or a street. It is also possible to list all attractions within a city district. One can also make a guided tour through the Web site, and making this accessible for others to benefit from. Naturally, the personal tours have to be authorized by an administrator in order to filter non-ethic information.

Students

mStudent is a joint project venture between the Federation of Student Unions in Stockholm (SSCO), the Stockholm Academic Forum, and the City of Stockholm within the framework of mCity. The objective is to develop mobile services which are useful to 80,000 students in the Stockholm region. For example, if students can receive an SMS telling them that a lecture has been cancelled, they might not have to come to the university campus at all that day, saving time to be better used for studies or other activities.

During the spring of 2003, 28 students from eight different universities and university-colleges in the Stockholm region participated in a feasibility study to identify a number of services interesting to students. This first phase of the project was carried out together with Telia,[7] Ericsson, and Föreningssparbanken.[8] The objective was to identify mobile services that would be useful to students in their everyday life. In order to really use the most of the students' innovative minds, they were all given one of Ericsson's most modern mobile phones and were allowed to use them without limitations. This made them experts on the available services and also good judges on new services.

Today, mStudent initiates and administrates different forms of tests and evaluations of mobile services in cooperation with businesses in Stockholm. The purpose is to encourage companies and universities to develop and use improved mobile services and thereby increase the quality of service to students as a group. The activities carried out are based on the list of mobile services that the students identified as interesting in the first phase; but apart from this, mStudent has also become a testbed which tests and evaluates all types of mobile services that can be useful for students. The "test pilots" are all students from Stockholm's universities and university colleges, and mStudent gathers the students in focus groups for workshops, evaluations, and other activities. Some companies are already working together with student reference groups in order to gain feedback on their planned services.

SMEs

mCity has been involved in one project aiming toward higher use of mobile services among SMEs.[9] In one of the shopping malls in central Stockholm, Söderhallarna, the stores can use the Internet and mobile technology to communicate, both with customers and the mall administration. The choice of Söderhallarna was not a coincidence. The property is actually owned by the City of Stockholm, and it is of importance

to the mall administration to keep up with the technological development to be able to attract stores to the premises.

By working closely with the storeowners and the mall administration, mCity managed not only to improve the internal communication, but also to provide new ways of treating customer relations with the aid of mobile services. For instance, stores can now inform their customers of last-minute offers or arrivals of new products with SMS or e-mail. Also, customers can easier interact with some of the companies. One of the lunch providers receives the orders from their customers via SMS. This increases the probability of preparing the food on time when not having to take orders on the phone. The technology is also used by the Head of Marketing for the mall, in order to create VIP offers to customers, and to communicate with SME owners and other mall staff, such as janitors.

Commuters

Up-to-date traffic information, provided by the City of Stockholm and the Swedish Road Administration among others, is today available on the Internet site, www.trafiken.nu. The information can be reached via WAP and Internet, but more ways of accessing the information have been developed. To make traffic information available regardless of place or time is important since it brings the choice to commute at a given time to the commuter. The commuters can improve their itinerary and choice of transportation based on the information about the current traffic situation.

mCity is involved in several pilot projects within the traffic area all initiated with a pre-study to find out what kind of information commuters are interested in and would benefit from. In one project, mCity has financed the development of the use of dynamic voice to present information available on the Internet site. The synthetic voice starts reading the new information when a commuter calls a special telephone number available from both fixed and mobile telephones. In another project, commuters are able to subscribe to information on specific routes. The commuter submits information about specific time spans during which he/she is interested in knowing about traffic disturbances on a Web page. As soon as something happens on the route of interest on the specific time span, an SMS is sent out with this information.

Employees

mCity has initiated several SMS management systems within the municipal organizations of Stockholm. Even though the technology used often is the most basic one, the impact has been extensive. Three examples of SMS solutions developed within mCity are described in the following sub-sections.

Schools: Absence Management

A few compulsory and upper-secondary schools have been provided with an absence management system. By keying in their social security number and a four-digit code, pupils can report themselves absent into an automated solution provided by the school. The information is then automatically sent as an e-mail or an SMS to the teachers, thereby reducing administrative work. The flow of information between the school and the parents is also improved since parents may receive an SMS when the child skips class or when parents should remember to pack extra clothes for special extracurricular activities.

The Care Sector: Scheduling Services

Within the care sector, scheduling is a time-consuming effort. Now, staff can plan and book time slots through the Internet, and changes can be made by management through SMS. Positive

effects with the solution is that staff motivation has increased and the Head of Staff can now work with core activities as the administrative workload is reduced. This solution was tested together with the SMS solution described next.

The Care Sector: Substitute Management

Within the care sector, a group SMS service has been implemented to facilitate substitute management. Instead of trying to reach substitutes through regular phone calls, managers can send SMSs to groups of staff, saving several hours every time. This creates better opportunities for planning, resulting in less stress for care staff and great financial benefits for the City of Stockholm. Also, managers have discovered the possibilities of encouraging staff through group SMS; an occasional "Have a nice weekend!", or the like, is very much appreciated by the staff working in mobile care units, not seeing much of their colleagues and managers when spending much time out in the field.

This SMS system has been so successful that it has now been made available to all employees within the City of Stockholm to use and benefit from. An interesting fact is that as more people are getting the opportunity to use the system, new areas of use are discovered every day by the users themselves.

MCITY EXPERIENCES

Looking at the mCity Project, it is clear that by focusing on and involving users who traditionally are considered underdeveloped within the field of ICTs, mCity reduces the digital divide. Areas like education and the care sector present great potential for municipalities and ICT companies as large savings of time and money can be made when administrative tasks are simplified. Also, by focusing on the areas with largest potential, one can increase average levels of use and knowledge of ICTs in the organization, even if simple technology is used. Thus, even the use of SMS might be an important step toward the use of more advanced mobile services (Williamson & Öst, 2004).

By involving the end user early on, the development process becomes more time consuming. On the other hand, there seems to be a higher chance of successful development and implementation. The involvement of end users in the development of mobile services leads to the appreciation of the users who feel that their experiences are valuable and have real impact. It is important to note that the "end user" is the very person who will use the system in the end, not his or her supervisor or manager. Thus, in small-scale projects, it is often necessary to involve several levels of management, involving the ones who will use the system, the ones who can oversee work processes, as well as the ones who will pay for the system.

It is not always easy to involve people with limited skills and knowledge in technology in projects involving technology. Some people are also more skeptical of changes than others; they may have gone through several organizational changes within a short time span, or might not be interested in revising their working processes at all. This is especially obvious when implementing new technology. Thus, it is important to recognize that technological artefacts are as much social as technological objects, affecting people's way of life as time and space are changing (Brown, 2002; Glimell & Juhlin, 2001; Urry, 2000).

In order to involve the end users, the project must be presented in a way which makes it come across as a project which will lead to obvious changes for the better and not primarily as a technological project. "We're not necessarily positive to technology per se, but we are positive to all new projects and ideas that will improve our work", a manager involved in the SMS project for substitute management stated in an interview (Hallin, 2003).

The information generated through the process also provides the companies involved with valuable input on user behavior and preferences. To engage companies in an m-government project like mCity has been very rewarding for all parties, but even though the pilot projects have been too small to make it necessary to issue invitations to tender, a discussion about the delimitations of working together with the private sector in development projects has taken place within the Steering Committee. This discussion has been similar to the general discussion going on in Sweden, as several public institutions find that the Public Procurement Act makes innovation in the area of public e-services difficult (Grenblad, 2003). In Sweden, there are not many precedents concerning these kinds of simple and quick forms of cooperations between the private and the public sectors. Clear directives as to how and when companies should be involved are needed.

A final lesson from the mCity Project is that simple technology offers great possibilities. mCity has not per se been interested in testing new technology just for the sake of testing new technology; the effects should be real and readily measurable, as described previously. This said, new technological inventions may also be tested and used, as has been the case within the mStudent Project and within the early tourist project. The clue is to always have in mind who is going to use the service. Students are in the forefront when it comes to usage of technology, and tourists also tend to be open minded to use new technology when travelling. Administrators in elderly care or in the school sector might not be as mature in their use of ICTs.

The choice of technology is also often subjected to other types of limitations. When developing new systems based on new technology, you have to be able to answer a lot of questions. One is whether the service should be available for all or just for a small group of people. In the case of mCity, this has been a difficult aspect since all services are tested on a small scale, enlarged when proving relevant. In small-scale environments, technological integration is not really necessary, but when making a service available on a larger scale, it is. In the projects in elderly care and in school administration, this was clearly evident. When making the group SMS project a large-scale implementation, integration to several internal programs was necessary, such as the mail system and the identification portal. This was not impossible, but of course involved more work and thorough consideration.

In a municipality, it is also necessary to consider the cost of implementing new technology. The new services have to deliver lowered cost or some other kind of gain for the city; developing services just for fun or because they are high-tech at the moment is not good enough.

TOWARD A NEW DEFINITION OF M-GOVERNMENT

Is mCity an m-government project? Generally, "m-government" is defined as "a subset of e-government", involving the use of mobile/wireless applications in the public sector, making the public information and services available anytime, anywhere (Lallana, 2004). According to this definition, it could be questioned whether mCity is an m-government project, as there are pilot projects with other goals than the one stated earlier. The mStudent Project, for example, aims at improving the life for students in the Stockholm region by introducing new mobile services from different providers, and in the SME project, small- and middle-sized companies and their customers benefit from the mobile service introduced.

It is clear that the City of Stockholm through the mCity Project takes a broader grip on the task of providing people with the possibility of accessing public information and services, by also taking on a pedagogical role of encouraging people to use ICTs in different areas of city life, and by stimulating the ICT industry to develop

new applications as well as rethink old applications. In order to understand this approach we must establish the relationship between the mCity Project and the municipal and national ICT strategies as well as the project's relationship to the vision of Stockholm as an IT capital.

mCity in Relation to Municipal and National Strategy

As described previously, the Stockholm E-Strategy is the policy document according to which the ICT work in the city is done. On its very first page, the document points out that the globalization process inevitably will lead to a new Europe where Stockholm will face tougher competition from other European cities, and that in order to face these challenges, the use of ICTs is an important factor. "IT must help to make Stockholm more attractive by securing the city's long-term goals that Stockholm should continue to be a fine place to live and work in"(*The City of Stockholm's E-Strategy, 19th of February 2001*).

The "E-Strategy" of Stockholm is, on a municipal level, what the "24/7 Agency" is on the national level. The "24/7 Agency" was issued in 2000 by the Swedish government, aiming at extending the public sector's use of ICTs, making services available 24 hours a day, 7 days a week (*The 24/7 Delegation*). The vision entails all parts of the public sector—municipalities, county councils as well as central government—and is the Swedish government's way of trying to cope with expected demographic changes leading to a larger aging population which will demand more of a public administration with fewer employees. At the same time, citizens in general are expected to demand more value for money and a growing internationalization is thought to increase the competitive pressure on public bodies. The development of e-government in Sweden is a way of meeting these challenges (Lund) and the belief of the 24/7 Agency is that the Swedish administrative model, with independently managed central government agencies, is a factor for the success of rapid development of digital applications and e-services (Lundbergh, 2004).

Swedish authorities primarily call for the most appropriate services, not specific technologies. Thus, the name of "the 24/7 Agency" places focus on the time aspects of service-provision—public services should be provided around the clock—not on specific technologies. The question "how" is subordinated, as, "Accessibility, irrespective of time of day and geographical location, may be achieved through a range of established service channels" (Östberg, 2000). Also, the Stockholm E-Strategy is on purpose called the "E-Strategy", and not the "IT-policy", in order to shift "…focus from IT to activities and show [..] how enhanced integration of *electronic services* ('e-services') can develop the municipality's work" (*The City of Stockholm's E-Strategy, 19th of February 2001*). According to this, the E-Strategy does not prescribe certain technologies, but only points at different areas that the city should work with: Internet, information management, mobile technologies (in general), and so forth.

mCity and the Vision of Stockholm as an IT Capital

The mCity Project not only aims at developing technology which make the city available around the clock. It is also a project used to enhance the image of Stockholm as an IT capital; an image based for example on the fact that Ericsson and other major players within the ICT sector have their development offices within the area. According to the Stockholm E-Strategy, IT can play an important role in making Stockholm an attractive city for people to live and work in, and therefore, the city must take an active part in creating business opportunities for ICT companies. One of the goals stated in the E-Strategy is to, "Be one of the most attractive municipalities for relocation, start-up and running of businesses, in competition with the foremost European cit-

ies" (*The City of Stockholm's E-Strategy, 19th of February 2001,* p. 14).

Through the mCity Project, the city has given several ICT companies in the Stockholm area the opportunity to test ideas, develop new applications and market themselves in and outside the country—naturally, in compliance with the Public Procurement Act. This strife to encourage local development conveys an entrepreneurial stance which might be perceived as contrasting with the managerial practices of earlier decades which primarily focused on the local provision of services, facilities, and benefits to the population (Harvey, 1989). However, when cities find themselves competing on a global—not only on a national—arena, a new kind of city management develops, involving proactive management of the images of the city as a management tool. Today, city managers are not only active administrators of the traditional areas of responsibility (Czarniawska, 2000) and the "branding" of the city involves much more than producing colorful brochures (Ward, 1998).

mCity and M-Government

As described earlier, the mCity Project aims at creating "the mobile city", as this is thought to be a good place for people to live, work, and spend their holidays in, since mobility means flexibility. But this does not necessarily mean that the project only deals with the development of mobile technologies, which makes information and services of the City of Stockholm available. "Mobile" here does not refer to the technology, but to the people using it, and "the mobile city" is the city where people have the flexibility to do what they want, where and when they choose. The mobile city can be achieved by the city becoming a role model, using mobile technology for its own activities, for example in schools, in homes for the elderly, or through mobile services which give commuters information about traffic, but also by stimulating the use of mobile technology in general, for example by encouraging students in the Stockholm area to ask for and use mobile services.

It is also obvious, that for mCity, the traditional m-government definition is not sufficient, as the city itself is not limited to its municipal organization. As we have showed earlier, the projects within mCity involve cooperation with both national institutions (for example, within the traffic projects), regional institutions (for example, within the mStudent Project) as well as private companies. Thus, rather than focusing on technology or the municipal organization, mCity focuses on people, and to see this project as an m-government project is to broaden the definition of m-goverment from only encompassing the use of mobile/wireless applications in the public sector, making the public information and services available anytime, anywhere. And rather than having the municipal organization as the starting point for its activities, the city, as it is perceived by its citizens, visitors, and employees, is the unit from where the project takes off. Thus, we suggest a new definition of m-government:

A public body which supports the mobility of its people, by providing its services when and where the people need them, and by supporting the development of whatever wireless technologies are needed, and the education of people in these.

THE FUTURE OF M-CITIES

It has been argued that the organizing capacity of a city determines whether the city will be able to develop in a sustainable way, and that the ability to include ICTs is becoming a more important aspect of the organizing capacity of cities (Winden, 2003). This, we believe is true.

Once a small, local initiative, mCity has grown into a project which covers many application areas. Through the project, it has become clear that mobile services can help Stockholm simplify routines, minimize administration, save both time

and money, and make life a bit easier for people thus contributing to a better working and living environment by improving the service quality offered by the city. These results further strengthen the notion that the building of m-government is important, probably not only for cities, but for all public bodies. But in order to be successful, a people's perspective has to be adopted and the traditional borders of the public body might have to be challenged. To start with people rather than with technology or with the organization, is an important prerequisite for success.

REFERENCES

The 24/7 Delegation. (No. Dir 2003:81). Retrieved November 19, 2006, from http://www.sou.gov.se/24timmarsdel/PDF/Eng%20version.pdf

Brown, B. (2002). Studying the use of mobile technology. In B. Brown, N. Green, & R. Harper (Eds.), *Wireless world. Social and interactional aspects of the mobile age* (pp. 3-15). London: Springer.

Castells, M. (1997). *The power of identity* (vol. 2). Oxford: Blackwell.

The City of Stockholm's E-Strategy. (2001, February 19). Available through the City of Stockholm +46 (0)8-508 00 000. The Swedish version can be retrieved (last retrieval, November 20, 2006) from http://www.stockholm.se/files/16100-16199/file_16185.pdf

Czarniawska, B. (2000). The European capital of the 2000s: On image construction and modeling. *Corporate Reputation Review, 3,* 202-217.

Glimell, H., & Juhlin, O. (Eds.). (2001). *The social production of technology. On the everyday life with things.* Göteborg: BAS.

Grenblad, D. (2003). *Growth area – E-services in the public sector, analyses of the innovation system in 2003.* Vinnova (The Swedish Agency for Innovation Systems).

Hallin, A. (2003). *Mobile technology and social development – Dialogic spaces in msociety.* EGOS Annual Conference, Copenhagen, July 2-5.

Harvey, D. (1989). From managerialism to entrepreneurialism: The transformation in urban governance in late capitalism. *Geografiska Annaler, 71B*(1), 3-17.

Jazic, A., & Lundevall, K. (2003). *mWatch – A survey on mobile readiness in the Baltic Sea Region*. Presented at the 5th Annual Baltic Development Forum Summit, Riga, Latvia. Retrieved November 20, 2006, from http://www.bdforum.org/download.asp?id=49

Lallana, E. C. (2004). *eGovernment for development, mgovernment definition on and models page*. Retrieved January 13, 2005, from http://www.e-devexchange.org/eGov/mgovdefn.htm

Lund, G. *The Swedish vision of 24-hour public administration and e-government – Speech by Gunnar Lund, Minister for International Economic Affairs and Financial Markets, held December 9th.* Unpublished manuscript.

Lundbergh, A. (2004). *Infra services – A Swedish way to facilitate public e-services development: MEMO.* The Swedish Agency for Public Management.

Urry, J. (2000). *Sociology beyond societies, mobilities for the twenty-first century.* London & New York: Routledge.

Ward, S. V. (1998). *Selling places. The marketing and promotion of towns and cities 1850-2000.* New York: E & Fn Spon.

Williamson, S., & Öst, F. (2004). *The Swedish telecommunications market 2003.* (No. PTS-ER-2004-24), The Swedish National Post and Telecom Agency.

Winden, W. v. (2003). *Essays on Urban ICT Policies.* Rotterdam: Erasmus University Rotterdam.

Östberg, O. (2000). *The 24/7 agency. Criteria for 24/7 agencies in the networked public administration.* Stockholm: The Swedish Agency for Administrative Development.

RELEVANT WEB SITES

www.stockholm.se/mCity
www.stockholm.se/english/
www.mstudent.se
www.telecomcity.org
www.testplats.com
www.24-timmarsmyndigheten.se
www.pts.se
www.trafiken.nu
www.explore.stockholm.se

ENDNOTES

1. All of Sweden has about nine million inhabitants.
2. 2004.
3. The City Council is the supreme decision-making body in the City of Stockholm, consisting of 101 members from the six parties represented in the council, and are elected by the Stockholmers every 4th year.
4. The City Executive Board consists of 13 members, who proportionally represent the parties in the City Council.
5. The Office of the City Executive Board.
6. The municipal company in Stockholm providing service to visitors.
7. The largest telecom operator in Sweden today known as TeliaSonera after a merge with the Finish company Sonera.
8. One of the major bank corporations in Sweden.
9. Small- and middle-sized enterprises.

This work was previously published in Mobile Government: An Emerging Direction in E-Government, edited by I. Kushchu, pp. 12-29, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 8.18
Mobile Speech Recognition

Dirk Schnelle
Technische Universität Darmstadt, Germany

ABSTRACT

This chapter gives an overview of the main architectures for enabling speech recognition on embedded devices. Starting with a short overview of speech recognition, an overview of the main challenges for the use on embedded devices is given. Each of the architectures has its own characteristic problems and features. This chapter gives a solid basis for the selection of an architecture that is most appropriate for the current business case in enterprise applications.

OVERVIEW

Voice-based interaction is a common requirement for ubiquitous computing (UC). However, the idea of having speech recognition on wearable devices is not simply copying the recognizer to such a device and running it. The limitations of the device, especially computational power and memory, pose strong limitations that cannot be handled by desktop size speech recognizers. This chapter gives a brief overview of the different architectures employed to support speech recognition on wearable devices. A background in speech recognition technology is helpful in order to understand them better, but is not required. At some points you will be provided with pointers to the literature to achieve a better understanding. A detailed understanding of the available architectures is needed to select the appropriate architecture for the enterprise, if it wants to support audio-based applications for mobile workers. The selection process has to consider the available resources, such as servers, wireless network, the software that has already been bought in order to save the investment, and to be able to justify the decision to invest more money in required infrastructure.

Most of the figures use UML 2.0 as a means of communicating architectural descriptions. The diagrams are easy to read, even if the reader is not familiar with this modeling language. The UML specification can be obtained from the Object Management Group (OMG, 2006).

A speech recognizer has the task of transcribing spoken language into a text (see Figure 1). The input is the *speech signal,* the human voice

that is recorded, for example, with a microphone. The textual output, in this case "*one two three*," is called an *utterance*.

The architecture of a speech recognizer has not changed over the past decades. It is illustrated in Figure 2 based on Jelinek (2001).

It comprises the main components of recognizers as they are used today, regardless of the technology used. They are available as pure software solutions or implemented in hardware to gain speed. In the following sections we focus only on the main components involved. Some recognizers may use additional components or components that are slightly different. However, the architectures presented show the main functionalities of each of them and discuss the main challenges that have to be faced when applied to mobile devices.

The *signal processor* generates real valued vectors σ_i from a speech signal, obtained from a microphone. They are also called feature vectors. Currently, most speech recognizers use at least 13

Figure 1. Speech recognition

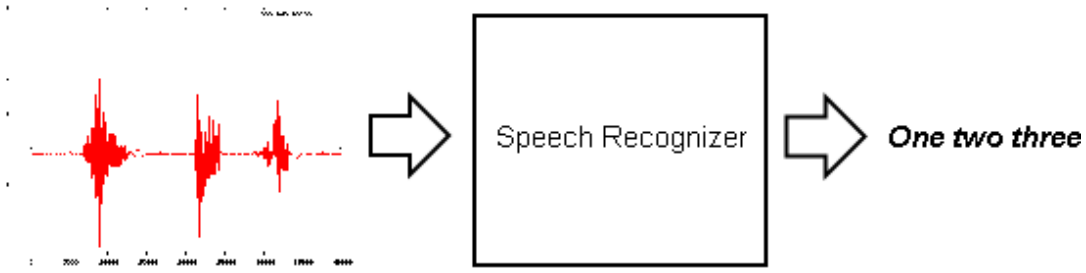
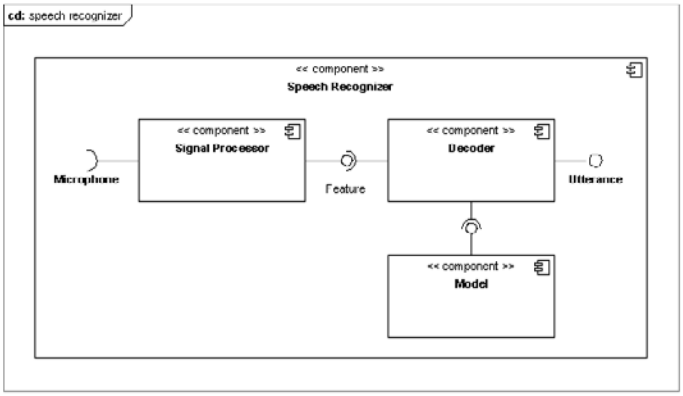


Figure 2. General architecture of a speech recognizer



features in each vector. We will have a closer look at them in the section "Evaluation of Sphinx 3.5." Normally computation of the features happens at regular intervals, that is, every 10msec., where the feature vectors are passed to the *decoder* to convert it into the utterance. The *decoder* uses the *model* for decoding. In the simplest case, the *model* contains a set of prototypes ρ_i, which are of the same kind as σ_i. Then, the *decoder* finds the ρ_i closest to σ_i for a given distance function *d*

$$a_i = \min_{j=1}^{k} d(\sigma_i, \rho_j) \quad (1)$$

a_i is the acoustic symbol for σ_i, which is emitted to the rest of the recognizer for further processing.

For **word-based speech recognizers** these acoustic symbols are the single words. For the example shown in Figure 1, this would be the concatenation of $\{a_1$=*one*, a_2=*two*, a_3=*three*$\}$.

A **phoneme-based speech recognizer** would output a concatenation of phonemes for each word. Phonemes are small sound units. The word *this* comprises the following phonemes $\{a_2$=*TH*, a_2=*I*, a_2=*is*, a_2=*S*$\}$. Obviously this output requires some post processing to obtain an output comparable to word-based recognizers that can be used by an application.

The benefit of phoneme-based speech recognizers is that they are generally more accurate, since they reduce the decoding problem to small sound units, and that they are more flexible and can handle a larger vocabulary more easily. Remember the first attempts in writing, starting from symbols for each word over symbols for each syllable to the letters that we find today.

This chapter is organized as follows: the section "Speech Recognition on Embedded Devices" gives an overview about the limitations of embedded devices to address speech recognition functionality. Then the two main architectures to work around these limitations are presented, which will be discussed in the two following sections in more detail. The section "Parameters of Speech Recognizers in UC" names some aspects that are needed to rate the solutions presented in the section "Service Dependent Speech Recognition" and the section "Device Inherent Speech Recognition." The section "Future research directions" concludes this chapter with a summary and an overview of the required computational resources on the device and the server for the architectures discussed.

SPEECH RECOGNITION ON EMBEDDED DEVICES

Speech recognition is computationally expensive. Advances in computing power made speech recognition possible on off-the-shelf desktop PCs beginning in the early 1990s. Mobile devices do not have that computing power and speech recognizers do not run in real time. There are even more limitations, which will be discussed later in this section. Moore's law states that memory size and computational performance increase by a factor of two every 18 months.

Frostad (2003) writes:

"Most of what is written on speech is focused on server based speech processing. But there is another speech technology out there that's powerful enough to sit on a stamp-sized microchip. It's called "embedded speech." Advances in computing power gave server side speech the power boost it needed in the early 90s. Now that same rocket fuel is launching embedded speech into the limelight."

Although computing power is increasing on these smaller computers, making it possible to run a small recognizer, performance is still not efficient enough to enable speech recognizers off-the-shelf on such devices. The attempt to use speech recognition on a mobile device, such as a computer of PDA size or a mobile phone, encounters the same problems that were faced

on desktop PCs years ago and which have been solved by the growth of computing power. The following section gives an overview of these limitations.

Limitations of Embedded Devices

The development of all applications, especially speech recognition applications for embedded devices has to tackle several problems, which arise as a result of the computational limitations and hardware resources on the device. These limitations are:

- **Memory:** Storage Capacity on embedded devices, such as a PDA or a cell phone, is very limited. This makes it impossible to have large *models.*
- **Computational Power:** Although the computational power of embedded devices has grown continuously over the last few years, it is still far from that what is available on desktop size PCs. The *signal processor* and the *decoder* perform computationally intense tasks.
- **Power Consumption:** Battery lifetime is a scarce resource on embedded devices. The device will stop working if the battery is empty. Since speech recognition is computationally intensive, processing consumes a lot of energy.
- **Floating Point:** Most processors for PDAs, like the Strong ARM or XScale processor, do not support floating-point arithmetic. It has to be emulated by fixed-point arithmetic, which is much slower than direct support. The value vectors σ_i are real valued and most state-of-the-art recognizers work with statistical methods. Thus, support of floating point arithmetic is essential and emulation results in loss of speed. Moreover, this may lead to a loss of precision. Especially signal processing is a critical task, since the quality of the output has a direct impact on the preserved information. Jelinek (2001) states that "Bad processing means loss of information: There is less of it to extract."

In the following the approaches to work around these limitations will be discussed. A short overview of the technology used is given to understand how they cope with the challenges of embedded devices.

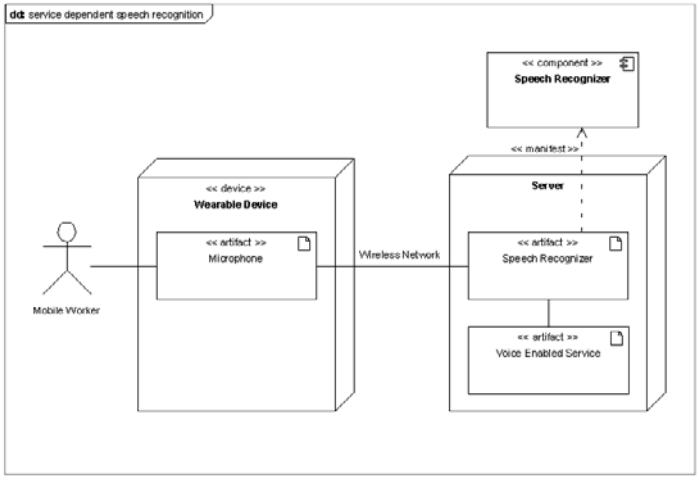
Main Architectures for Speech Recognition on Embedded Devices

Progress in speech recognition has made it possible to have it on embedded devices. Cohen (2004) states that, "Although we did not know in the 1990s all of the tricks we know today, we can use 1990s-like computing resources … to good advantage to compute a task which would have been difficult in 1990, but is simpler today because of our technical advancements." However, the limitations of mobile devices that were mentioned in the previous section still pose a lot of challenges. There have been several attempts to deal with them and enable speech recognition on embedded devices. An overview of these approaches is given in the following sections. We concentrate on the most common approaches that can be divided into two main categories:

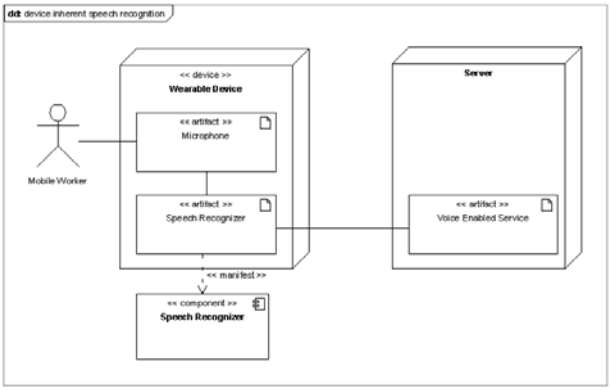
- Service dependent speech recognition, Figure 3(a)
- Device inherent speech recognition, Figure 3(b)

The main difference between these two architectures is the node where the *speech recognizer* component is deployed. Architectures for service dependent speech recognition will be introduced in the section "Service Dependent Speech Recognition" and those for device inherent speech recognition in the section "Device Inherent Speech Recognition."

Figure 3. Deployment of voice enabled service usage with mobile devices



(a) Service dependent speech recognition



(b) Device inherent speech recognition

Zaykovskiy (2006) proposed another categorization. He distinguishes:

- Client,
- Client-server, and
- Server-based architectures.

The main reason for his differentiation is the location of the *signal Processor* and the *decoder.* In the service-oriented view of ubiquitous computing it makes more sense to emphasize the ability to have speech recognition as a network service or as an independent functionality of the

device itself. This is a fundamental fact in smart environments, where services can be inaccessible while the user is on the move. Bailey (2004) requires that "there need to be clear boundaries between the functionality of the device, and the functionality of the network." The technological orientation of these approaches confirms this differentiation. Whereas service dependent speech recognition deal with APIs for remote access to a speech recognizer, device inherent speech recognition uses the techniques of desktop size speech recognition technology to enable speech recognition on the device itself.

Parameters of Speech Recognizers in UC

In order to rate the different architectures, we need an understanding of the core parameters. This section will give a short overview of these parameters.

- **Speaking Mode:** Word boundaries are not easy to detect. The presence of pauses is not enough, since they may not be present. Early speech recognizers forced the user to pause after each word. This is called *isolated word* recognition. If there are no such constraints, the speech recognizer is able to process *continuous speech.*
- **Speaking Style:** This parameter states if a speech recognizer for continuous speech is able to process *read speech*, meaning a very precise and clear pronunciation, or if it is capable of processing *spontaneous speech*, as used when we talk to each other.
- **Enrollment:** Some speech recognizers require an initial training before they can be used. This training is used to adapt to the speaker in order to achieve higher accuracy. These recognizers are called *speaker dependent.* This concept is often used on desktop PCs, but is also possible in UC, where the device is personalized. The opposite case is *speaker independent* speech recognizers that are trained to work with multiple speakers. Thus they have a lower accuracy. This concept is used, for example, in telephony applications. There are only a few scenarios that really require speaker independence with embedded devices. For these applications, speaker-independent systems do not have an advantage over speaker-dependent systems, but can benefit from a better accuracy.
- **Vocabulary:** The size of the vocabulary is one of the most important factors, since this strongly influences the way in which users can interact with the application. A vocabulary is said to be *small* if it contains up to 20 words. A *large* vocabulary may contain over 20,000 words.
- **Perplexity:** Perplexity defines the number of words that can follow a word. This is an important factor if the recognizer has to decode an utterance consisting of multiple words and tries to find the path with the lowest error rate.
- **SNR:** SNR is the acronym of **s**ignal-to-**n**oise-**r**atio. It is defined as the ratio of a given transmitted signal to the background noise of the transmission medium. This typically happens where the microphone also captures some noise from the background, which does not belong to the signal to decode.
- **Transducer:** A transducer is the device that converts the speech into a digital representation. For speech recognition this may be, for instance, a noise-cancelling headset or telephone. Each of them features different characteristics, such as the available bandwidth of the voice data or the ability to cut background noise as with a noise cancelling headset. In UC environments noise-cancelling headsets are typically used.

In a UC world there are some additional parameters depending on the location, where

recognition takes place. These parameters, as presented in Bailey (2004) are:

- **Network Dependency:** One major aspect is the dependency on a network resource. A recognizer located on the device will not need any network to be operated, while a recognizer streaming raw audio data to a server (see the section "Audio Streaming") will not work without a network. Apart from the technical aspect, the user expects the device to work and may not be able to distinguish between a non-functional recognizer and missing network connectivity if the device is "broken."
- **Network Bandwidth:** Available network bandwidth is a scarce resource. Architectures performing recognition on the device have a more compact representation of the data that has to be transmitted than those architectures streaming pure audio data.
- **Transmission Degradation:** With the need to transmit data from the mobile device to a server, the problem of transmission degradation arises. Failures, loss of packets or corrupted packets while transmitting the data means a loss of information. If the raw audio is transmitted to a server, recognition accuracy goes down.
- **Server Load:** In a multi-user scenario it is important that the application scales with an increasing number of users.
- **Integration and Maintenance:** Embedded devices are hard to maintain, especially if parts of the functionality are implemented in hardware. A server, on the other hand, is easy to access and bug fixes are available for all clients at once. This issue goes in a similar direction to the discussion of centralized server architectures versus rich clients.
- **Responsiveness:** A must for speech recognition is that the result is available in real time. This means that the result of the recognition process must be available as fast as possible.

In the following sections the different architectures will be characterized according to these parameters.

SERVICE DEPENDENT SPEECH RECOGNITION

The architectures presented in this section have in common that they require network connectivity to work.

Audio Streaming

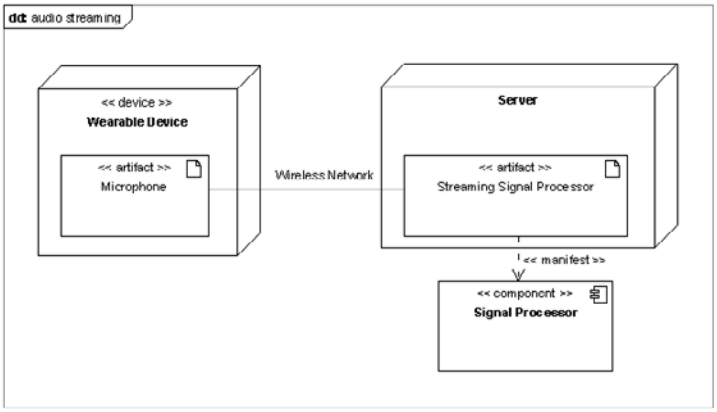
An immediate idea to solve the performance bottleneck on embedded devices is not to perform the recognition process on the device itself. A general model of the recognizer, shown in Figure 4, uses the audio recording capabilities of the embedded device as a microphone replacement to record the raw audio as the input for the *signal processor.*

The audio is streamed over the wireless network, for example, Wi-Fi or Bluetooth, to the *signal processor* on a server. This allows the use of a full-featured recognizer with a large vocabulary running on the server. A disadvantage of this solution is that a stable wireless network connection is required. Another disadvantage is a possibly very large amount of data streamed over the network. Since recognition is not performed on the embedded device, we have all the benefits of a desktop-size speech recognizer at the cost of high network traffic.

MRCP

Since server side speech recognition is mainly an immediate idea, developers tend to implement this architecture on their own using a proprietary

Figure 4. Architecture of an audio streaming speech recognizer



protocol, which makes it unusable with other applications that do not know anything about that proprietary protocol. In addition, real time issues are generally not considered which can result in misrecognition. A standard for server side speech recognition that has been adopted by industry is MRCP. MRCP is an acronym for **m**edia **r**esource **c**ontrol **p**rotocol. It was jointly developed by Cisco Systems, Nuance Communications and Speechworks and was published by the IETF as an RFC (Shanmugham, 2006).

MRCP is designed as an API to enable clients control media processing resources over a network to provide a standard for audio streaming. Media processing resources can be speech recognizers, text-to-speech engines, fax, signal detectors and more. This allows for a use in distributed environments, for example, a small device that accesses a recognizer on the network.

The specification is based on RTSP in Schulzrinne (1998), the **r**eal **t**ime **s**treaming **p**rotocol, as a MIME-type the **M**ultipurpose **I**nternet **M**ail **E**xtension. MIME is used to support, for example, no-text attachments in e-mail messages. RTSP defines requests, responses, and events needed to control the media processing resources. The protocol itself is text based. Mechanisms for the reliable exchange of binary data are left to protocols like SIP, the **S**ession **I**nitiation **P**rotocol, or RTSP. SIP enables control of sessions such as Internet telephone calls.

A media server that can be accessed by RTSP mechanisms controls all resources, in this case, recognizer and synthesizer.

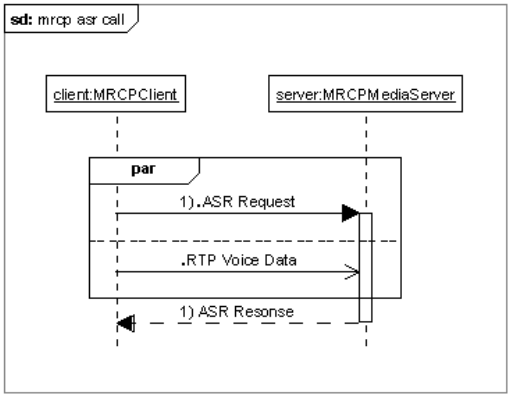
Figure 5 shows a simplified view on the messages that are exchanged in an **a**utomatic **s**peech **r**ecognition (ASR) request and a **t**ext-**t**o-**s**peech (TTS) request.

In an *ASR request*, the MRCP client initiates the request and delivers the voice data via RTP in parallel. The recognition process is executed on the *MRCP Media Server* and the result of the recognition is delivered to the client as the *ASR Response*.

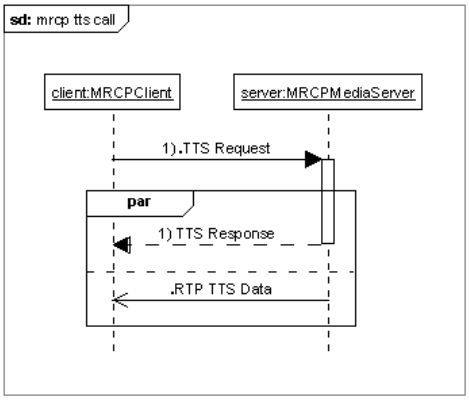
In a *TTS request*, the MRCP client initiates the request. The *MRCP Media Server* answers with a *TTS response* and delivers the synthesized voice data via RTP in parallel.

Figure 5. Simplified view on MRCP requests for ASR and TTS

(a) MRCP ASR request



(b) MRCP TTS request



Distributed Speech Recognition

Another possibility to enable speech recognition on mobile devices uses an architectural compromise. Since full-featured recognizers are hard to implement on embedded devices and streaming of raw audio data produces too much network traffic, ETSI, the **E**uropean **T**elecommunication **S**tandard **I**nstitute, introduced a solution to perform parts of the recognition process on the device and the rest is handled on a server. This architecture is called *Distributed Speech Recognition* (DSR). Pearce (2000) named the component, which is deployed on the device the *DSR Front-end* and the component deployed on the server the *DSR Backend*. This concept is shown in Figure 6.

An obvious point for such splitting is the separation of *signal processor* and *decoder*. Instead of sending all the audio over the network, the feature vectors ρ_i are computed on the embedded device and sent to the *decoder* on a server. In order to reduce the amount of data and to ensure a secure transmission, the data is compressed and a CRC value is added. The architecture of the DSR Frontend is shown in Figure 7.

The DSR Backend, shown in Figure 8, checks the CRC value and decompresses the data before it is passed to the *decoder*.

In this way, the computational capabilities of the device are used for the tasks of the *signal processor* in the *DSR Front-end*, whereas the *decoder* and the *model* reside on the server in the *DSR Backend*. The architecture is a result of discussion between multiple companies, that is, Nokia and Motorola in the Aurora project. The data exchange of DSR Frontend and DSR Backend is standardized by ETSI. This specification includes the features used, CRC check and their compression. Compared to pure audio streaming, the transmitted data is reduced without much loss of information. This also means that the error rates in transmission are reduced. As a positive consequence, DSR also works with lower signal strength, as shown in Figure 9.

Figure 6. Architecture of a distributed speech recognizer

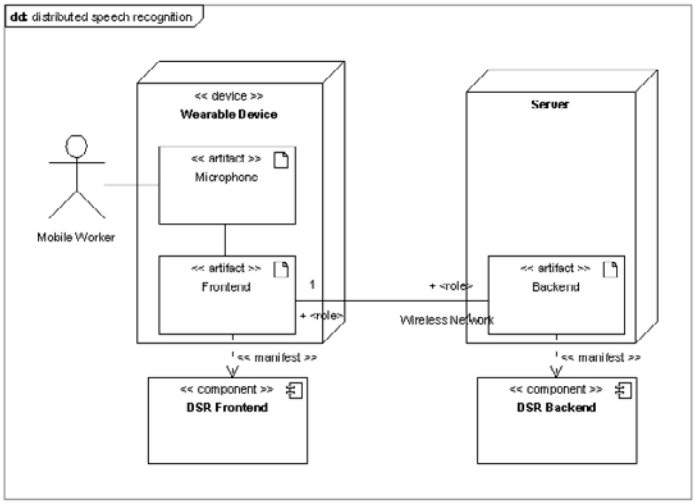
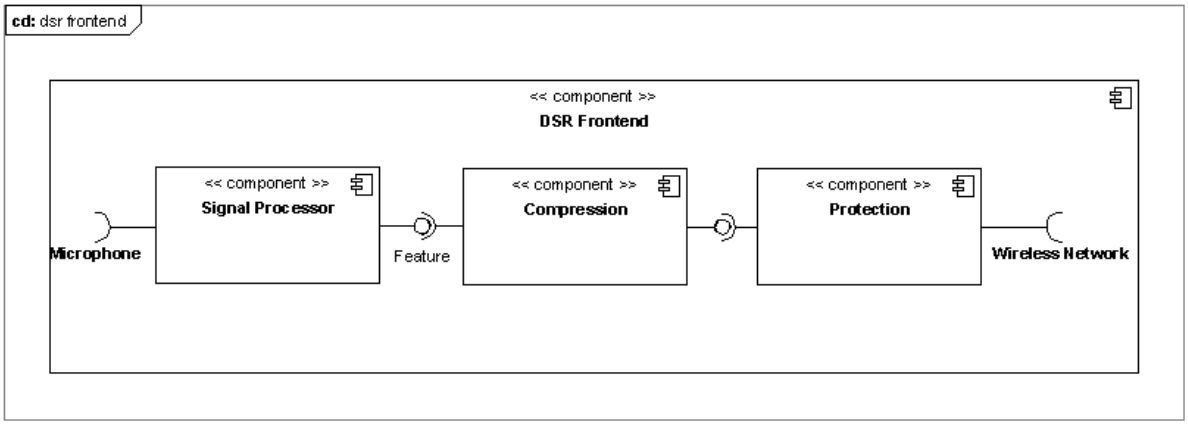


Figure 7. DSR front-end



The experiment was conducted in the Aurora project and is described in more detail in Pearce (2000). The figure shows the recognition performance of DSR compared to a mobile speech channel. The measurement proves that recognition still works with lower signal quality. A great advantage of this technology over streaming solutions like audio streaming or MRCP is the reduced network

Figure 8. DSR backend

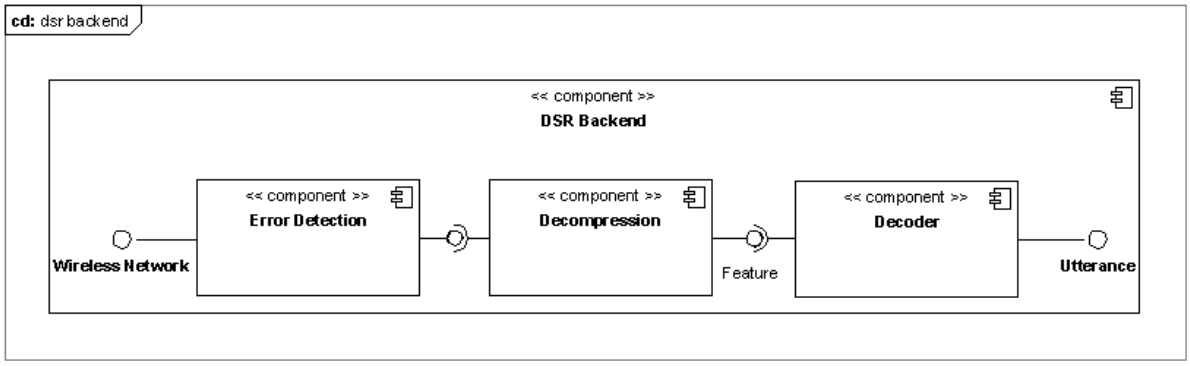
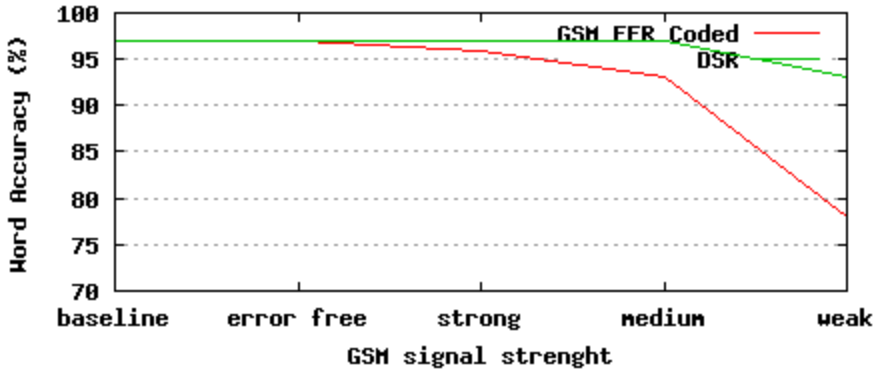


Figure 9. Performance of DSR with channel errors according to Pearce (2000)



traffic. Like MRCP it defines a standard, but with less acceptance. Speech recognition can be used in various environments, as long as they are compliant to the DSR standard. Unlike MRCP it relies on computation on the device, decreasing its chances of being established in a company's network in contrast to a pure protocol. Again, the recognition has all the features of a desktop size recognizer.

As a negative point, the set of feature vectors is a compromise. This also means that other or additional features used in specific recognizers cannot be transmitted using this technology.

ETSI promises a better use of available resources and better transmission. The following section gives some insight into the computational requirements.

Evaluation of Sphinx 3.5

Sphinx is an open source speech recognizer from Carnegie Mellon University. It was DARPA funded and was used in many research projects in speech recognition.

The anatomy of Sphinx can be divided into three phases:

1. Front-end processing,
2. Gaussian probability estimation, and
3. Hidden Markov evaluation.

The Gaussian phase and the HMM phase are part of the *decoder*. A closer look at it is given in the section "Hidden Markov Models." Mathew (2002) gives an overview of the time that Sphinx 3.5 spends on each phase (see Figure 10).

Obviously, the front-end processing constitutes the smallest part of the computation to be performed. This shows that this is an ideal candidate to be performed by smaller devices, as it is done with DSR. Consequently, Mathew et al. (2000) consider it to be not worthy of further investigation, stopping their analysis at this point. They focus more on the optimization of the latter two phases.

Front-end processing usually comprises the computational steps shown in Figure 11.

The following paragraphs show how these steps are handled in Sphinx. A more general view can be found in the literature, for example, in Schukat-Talamazzini (1995).

Processing starts with a speech signal, as captured, for example, from a microphone. An example of such a speech signal is shown as the input to the speech recognizer in . The transformation into a digital representation, also called *quantization*, means also a loss of information, but this can not be avoided.

Figure 10. Profile information of Sphinx 3.5 phases according to Mathew (2002)

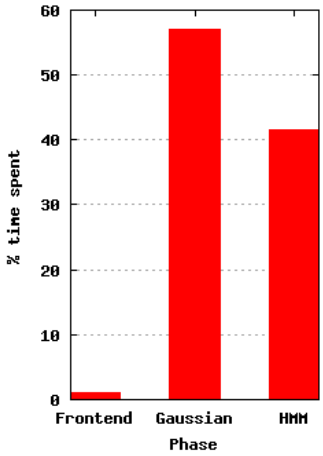
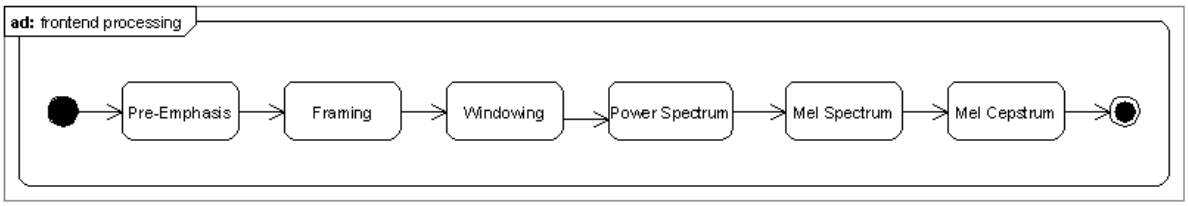


Figure 11. Front-end processing

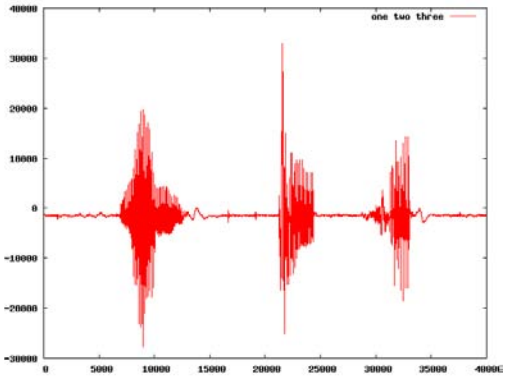


- **Pre-Emphasis:** In this step the quantized signal is filtered. This step becomes necessary from the observation that the signal is weaker in higher frequencies, which can be solved using a digital high-pass filter. Figure 12 shows an example of the speech signal and the effect of this filtering step.
- **Framing:** The input signal is divided into overlapping frames of *N* samples. The frame shift interval, that is, the difference between the starting points of consecutive frames, is *M* samples.
- **Windowing:** The **F**ast **F**ourier **T**ransformation (FFT) is known from the domain of signal processing to compute the spectrum of a signal. FFT requires a periodical signal, it is assumed that the time segment continues to be periodical. Since speech changes over time, we try to get segments of the signal, where it can be considered to be constant. These time segments last from 5-30 ms. An example for such a windowing function is the *Hamming Window*. The following figure shows four such time segments of the utterance. It is noticeable that the signal is smoothed to the borders of the time segments.
- **Power Spectrum:** For speech recognition, the discrete case of the FFT, the **d**iscrete **f**ourier **t**ransformation (DFT) is used. The output of the DFT usually consists of a power of 2 of complex numbers. The power spectrum is computed by the squared magnitude of these complex numbers. The following figure shows the power spectrum for the word *one* of the utterances.
- **Mel Spectrum:** The next step is a filtering step to filter the input spectrum through individual filters. One of these filters is the Mel filter. An impression of this filter is given in the following figure.

The output is an array of filtered values, typically called Mel-spectrum, each corresponding to the result of filtering the input spectrum through an individual filter. Therefore, the length of the output array is equal to the number of filters created.

Figure 12. Pre-emphasis of the speech signal

(a) Quantized speech signal



(b) Pre-emphasized speech signal

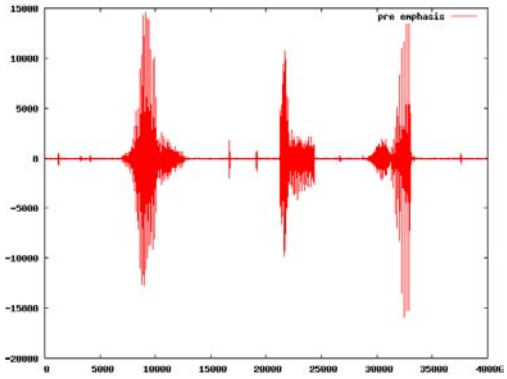


Figure 13. Framing

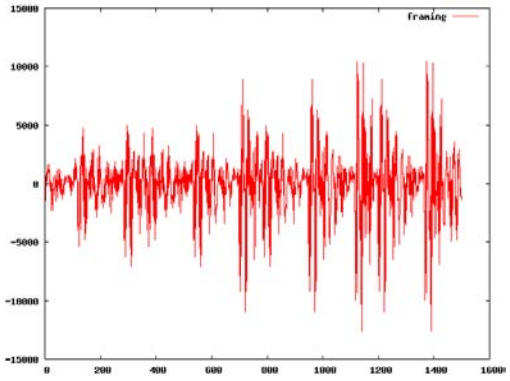


Figure 14. Windowing

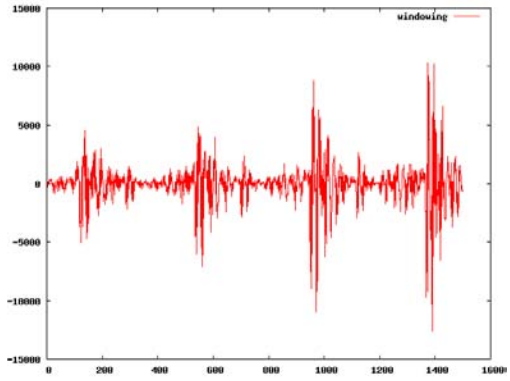
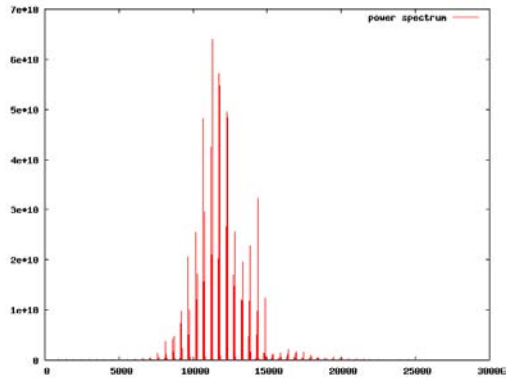


Figure 15. Power spectrum



Mel Cepstrum

Davis (1980) showed that Mel-frequency cepstral coefficients present robust characteristics that are good for speech recognition. The artificial word cepstrum is obtained by reversing the letter order in the spectrum to emphasize that this is an inverse transformation. These cepstral coefficients are computed via a discrete cosine transform.

Sphinx uses 16-bit raw audio data as input and produces 13 cepstral parameters as output for each time segment. In order to determine

Figure 16. Mel filter

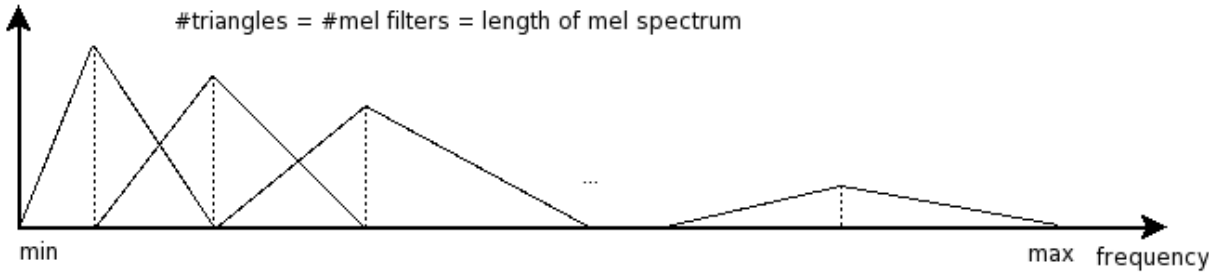


Figure 17. Mel-Spectrum

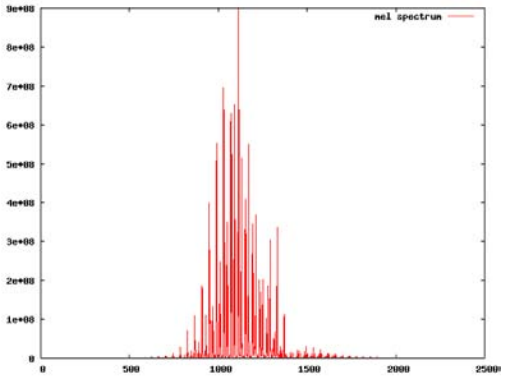


Figure 18. Mel cepstrum

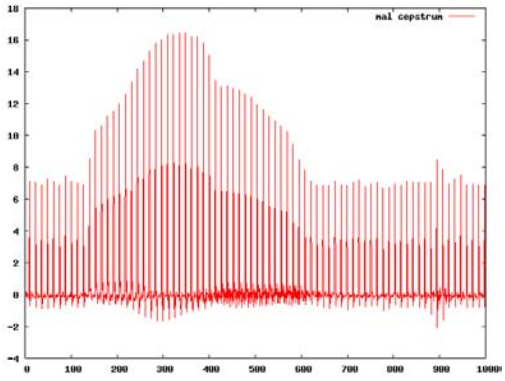
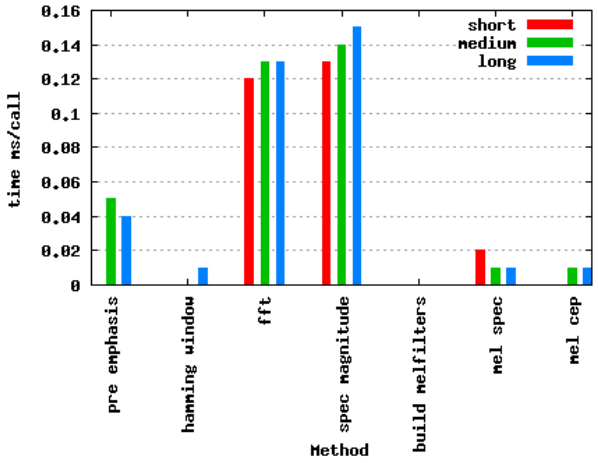


Figure 19. Profile information of sphinx 3.5 front-end



the execution time consumption by individual parts of Sphinx, we used a profiling tool to get detailed information on functions and routines on a Sparc processor-based platform. The profiling was done with three audio files of different lengths as input:

- Short (2.05 sec),
- Medium (6.02 sec), and
- Long (30.04 sec).

The profiling result is shown in Figure 19.

Obviously, the computation of the power spectrum, which comprises the methods fft and spec magnitude, consumes most of the time. Both are part of the power spectrum computation. Tuning this method can speed up the computation a great deal. Alternatively, it can be replaced by a hardware solution, such as a **d**igital **s**ignal **p**rocessor (DSP). This issue will also be addressed in the section "Hardware-Based Speech Recognition."

The process becomes more complicated if the device does not support floating-point operations. Junqua (2001) mentions, "While most automatic speech recognition systems for PC use are based on floating-point algorithms, most of the processors used in embedded systems are fixed-point processors. With fixed-point processors there is only a finite amount of arithmetic precision available, causing an inevitable deviation from the original design." A study by Delaney (2002) showed that for Sphinx 2 a Strong ARM simulator spent over 90% of the time on the floating-point emulation. These results can be transferred to Sphinx 3.5, since they use the same code base for front-end processing.

A way in which to solve these issues is to substitute floating-point arithmetic by fixed-point arithmetic. This is done using scaled integers to perform basic math functions. The scaling factor, that is, the location of the decimal point, must be known in advance and requires careful decision. For adding two numbers, the number n of bits after the decimal point must line up. A multiplication of two numbers yields a number with $2n$ bits after the decimal point.

Unfortunately, this also means a loss of information and the risk of overflowing the register size of 32 bits. This is especially important for the computation of the power spectrum and the Mel-spectrum. Delaney (2002) suggests changing the computation for the Mel-spectrum using a square root to compute the Mel coefficients. It is guaranteed that the square root results in small values, which means that the result of multiplication is small. They also suggest storing the Mel coefficients in a lookup table to avoid the computationally complex calculations of the square root. An experiment conducted in Huggins (2005) showed that feature extraction on a Sharp Zaurus had a 2.7-fold gain in speed using this method. The loss in precision for the result in computing the Mel Cepstrum increased from 9.73% to 10.06%.

DEVICE INHERENT SPEECH RECOGNITION

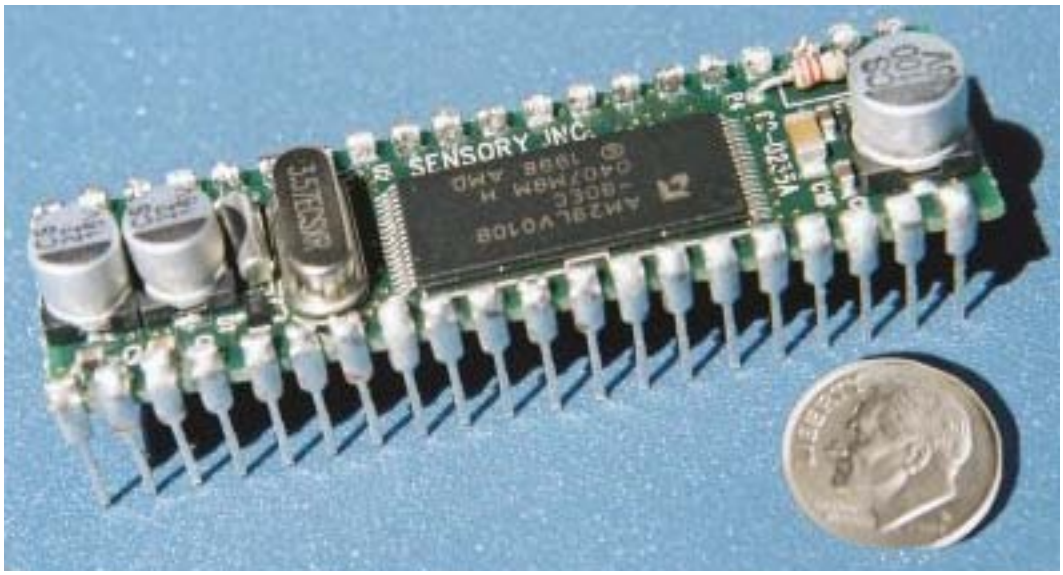
In contrast to the architectures described above, those described in this section are handled on the device only, without the need for a server or service from the network. These architectures are also often referred to as *software-only* and *embedded* architectures (Eagle, 1999; Frostad, 2003). Embedded architectures require the existence of a dedicated DSP. They reside as hardware-based speech recognition, since the term *embedded* is totally overloaded with the meanings of a DSP, an embedded device or embedded into an application. So this architecture does not deal only with software-based architectures, but also include partial or full support with hardware.

Hardware-Based Speech Recognition

Some manufacturers offer designated chips for mobile devices. An example of such a chip is shown in Figure 20.

The technology that is used in these chips differs. All software-based speech technologies for device inherent speech recognition, as described in the following sections, can be found implemented as a port to a DSP. It is even possible to replace just certain parts of the recognizer, that is, the FFT computation for the feature extraction in DSR, with a hardware solution. The main advantage is that a hardware-based solution does not have the runtime problems of software-based approaches,

Figure 20. Sensory voice recognition module



since the hardware is designed to address this specific problem. This is gained at the cost of less flexibility. The range of hardware implementations is as broad as the underlying technology. It starts from a fixed vocabulary used in toys through dynamic time warping, the technology used in most mobile phones, up to programmable DSPs like the sensory chip shown in.

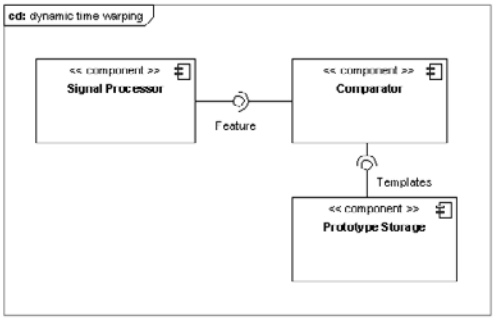
Advantages and drawbacks of these solutions are not discussed in this section, since they are inherited from the technology used. Benefits and drawbacks of the architectures are discussed in the corresponding sections.

Dynamic Type Warping

One of the earliest approaches of enabling speech recognition is the dynamic time warping (DTW). The architecture is shown in Figure 21.

The *signal processor* is responsible for the feature analysis of the raw speech signal. The computational steps are the same as the front-end

Figure 21. Dynamic time warping



processing of DSR (see). An output of the feature analysis component is a feature vector of a test utterance $\sigma = (\sigma_1,..., \sigma_n)^T$ which is compared in the *comparator*, which replaces the *decoder*, with all reference feature vectors $\rho_i = (\rho_{i,1},..., \rho_{i,m})^T$ stored in the *prototype* storage, replacing the *model* of the utterances ρ_i in the set of trained utterances with the help of a distance function $d(\sigma_i, \rho_j)$ that was already mentioned in the section "Overview." Usually the prototypes are gained in a single recording. The features of this recording are computed, stored in the *prototype storage* and associated with the output. If the distance of the currently spoken word to the template is too big $d(\sigma_i, \rho_j) > \mu$, it is likely that no prototype matches the utterance. In this case, the comparator rejects the input.

The problem of calculating the distance from σ_i to ρ_j with the help of a distance function $d(\sigma_i, \rho_j)$ consists of two parts:

1. Definition of a distance function to calculate the distance of two related feature vectors
2. Definition of a time warping function to define a relationship between the elements of σ_i and ρ_j

Multiple distance functions exist and are used. For a Gaussian distribution, Mahalanobis distance is used. Since this is complex and we do not have many computational resources on the device, Euclidean distance is more common. This requires that the features be normalized to unity variance.

The problem with a pairwise distance calculation is that it is unlikely that the lengths of the template and of the input are the same, that is, the length of the *o* in *word* may vary. DTW uses dynamic programming to find an optimal match between two sequences of feature vectors allowing for stretching and compression of sections [see Sakoe (1990)]. The template word having the least distance is taken as a correct match, if its value is smaller than a predetermined threshold value μ.

The technique of comparison with a template word makes this an ideal candidate for isolated word recognition with a small vocabulary, but unsuitable for continuous speech. Since the templates are generally taken in a single recording, DTW is also speaker dependent with little computational effort. The computational requirements are slightly higher than those for DSR, see the section "Distributed Speech Recognition," but lower than those for hidden Markov models (see next section), or artificial neural networks, see the section "Artificial Neural Networks."

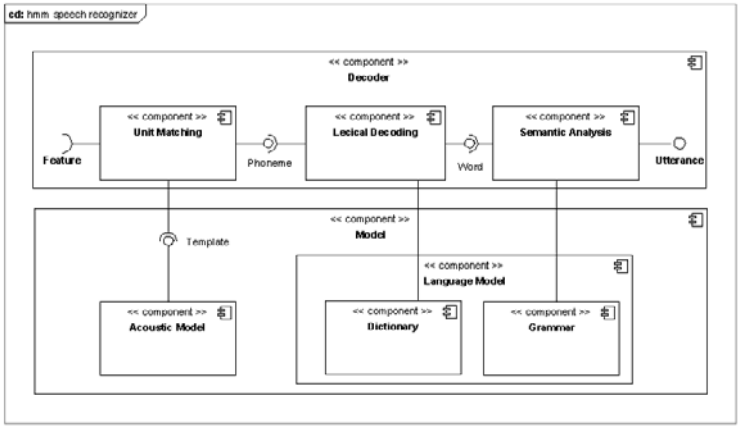
Hidden Markov Models

Most modern speech recognizers are based on **h**idden **M**arkov **m**odels (HMM). An overview of the architecture of a HMM based recognizer is shown in Figure 22, which is in fact a phoneme-based recognizer.

It is also possible to use HMM-based recognition for word-based models. In this case, the architecture is slightly different, as Schukat-Talamazzini (1995) points out. More about the basics of Markov chains and their use can be obtained from the literature, for example, Rabiner (1989). Although this approach is very old, it is still the most successful approach for speech recognition.

Instead of using the computed features as a seed for the states, most recognizers use **v**ector **q**uantization (VQ) to reduce the data rate. Since speech recognition deals with a continuous signal, a certain amount of data arrives periodically. This is called the *data rate*. Since HMM decoding is time consuming, a lower data rate promises real time performance. Furthermore, the storage size is reduced, since only the codebook is stored instead of the cepstral parameters. A *codebook* stores the mapping of the feature vectors as they are computed from the speech signal to a discrete label. Thus the codebook is a discrete representation of the continuous speech data.

Figure 22. Architecture of a HMM-based recognizer



Unit Matching

HMMs are the core of the *unit matching* component. They are described as a tuple $\lambda = (S, A, B, \pi, V)$ with

- $S = \{s_1,..., s_n\}$ representing a set of states,
- $A = \{a_{i,j}\}$ representing a matrix of transition probabilities, where $a_{i,j}$ denotes the probability $p(s_j, s_i)$ for the transition from state s_i to s_j,
- $B = \{b_1,..., b_n\}$ representing a set of output probabilities, where $b_i(x)$ denotes the probability $q(x|s_i)$ to observe x in state s_i and
- O as a set of observations, which means the domain of b_i.

A schematic view on a HMM is given in the following figure.

The probability of observing an output sequence $O = O_1O_2...O_r$ is given by

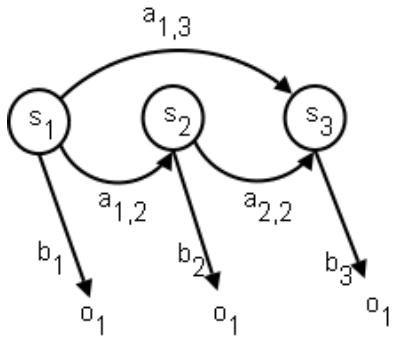
$$P(O = O_1O_2...O_r) = \sum_{\{s_1,s_2,...s_T\}} \prod_{i=1}^{T} p(s_i \mid s_{i-1})q(x_i \mid s_{i-1}) \quad (2)$$

Rabiner (1989) raises three basic questions that are to be solved for speech recognition with HMMs.

1. Given the observation sequence $O = O_1O_2...O_r$ and a model λ, how do we efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model?
2. For decoding, the question to solve is, given the observation sequence $O = O_1O_2...O_r$ and the model λ how we choose a corresponding state sequence $Q = q_1, q_2,...q_T$, which is optimal in some meaningful sense (i.e., best "explains" the observations)?
3. How do we adjust the model parameters λ to maximize $P(O|\lambda)$?

The first problem is also known as the *evaluation problem*, but it can also be treated as a *scoring*

Figure 23. Schematic view on a HMM



problem. In that case, the solution to this problem allows us to choose the model that best explains the observations.

The third problem tries to optimize the model parameters to describe a given observation sequence as well as possible. This optimization can be used to *train* the model. Training means to adapt the model parameters to observed training data.

The most important one for speech recognition is the second problem, since it tries to find the *correct* state sequence.

A well-known approach to this is the Viterbi algorithm, based on dynamic programming (DP) to find the most likely sequence of hidden states. A more detailed description, related to speech recognition, can be found in the literature, for example, Rabiner (1989) and Jelinek (2001). The Viterbi algorithm tries to find the best score, which means the highest probability, along a single path at time *t*, also known as *trellis.*

Computational Optimization

The Viterbi algorithm is computationally intensive, especially for larger vocabularies. It requires roughly $|A_u|$ multiplications and additions, where $|A_u|$ is the number of transitions in the model (Bahl, 1993). In order not to search the entire Viterbi trellis, the number of branch-out search candidates can be limited using beam-search, as Lowerre (1976) points out.

The idea is to eliminate all states from the trellis that have a probability above a certain threshold, which depends on the maximum probability of the states at this stage. This reduces the number of states without affecting the values, if the threshold is appropriately chosen.

Novak et al. (2003) suggest an even more aggressive pruning with their two-pass strategy. Instead of using the probabilities directly, they convert them to probabilities based on their rank. Thus, the probability space is bounded and the values of the best and worst state output probabilities remain the same for each time frame. Instead of computing the output probabilities, they simply take a single value from the tail of the ranked probability distribution. This is based on the approach described in Bahl et al. (1993) where the authors claim a speedup by a factor of 100.

There are many more attempts to simplify the computational effort of Viterbi search. Most of them try to replace multiplications by additions, which are faster to compute (Ming, 2003). Usually these attempts increase speed at the cost of accuracy and/or memory demands.

Lexical Decoding and Semantic Analysis

The result of the unit matching is a scoring for the different recognition hypotheses. The next two steps help to determine the word chain with the highest probability with respect to the constraints imposed by the language model. For word-based recognition with HMMs, the recognition process is finished at this point.

In the *lexical decoding* phase those paths are eliminated that do not have an existing word in the dictionary. In an alternative approach, using a so-called *statistical grammar*, the sequences

are reduced a couple of phonemes in a row, for example, trigrams. The output of the latter case is a list of trigrams, ordered according to their probability. This is not suitable for isolated word recognition. The next step is *syntactic analysis*, where those paths are eliminated that do not match an allowed sequence of words from the dictionary.

These steps do not require intensive computation except for fast memory access to the dictionary and the grammar. Again, smaller vocabularies offer a faster result and require less memory.

The word or utterance with the highest probability in the remaining list of possible utterances is taken as the recognition output.

HMM-based recognition is computationally intensive, but shows good results in isolated word recognition as well as continuous speech. If the HMM is trained well, it is also a suitable technology for speaker independent recognition.

Artificial Neural Networks

Artificial **n**eural **n**etworks (ANN) is a method in computer science that is derived from the way the human brain works. The goal is to create a system that is able to learn and that can be used for pattern classification. More detailed information about ANN is given in the chapter "Socionics & Bionics: Learning from 'The Born.'" The use of ANN for classification and their use in speech recognition can be found in the literature, for example, Cholet (1999).

Expectations were very high when ANNs were discovered as a means for speech recognition. Modelling of speech recognition by artificial neural networks doesn't require *a priori* knowledge of the speech process and this technique quickly became an attractive alternative to HMM (Amrouche, 2006).

Neural nets tend to be better than HMMs for picking out discrete words, but they require extensive training up front [see Kumagai (2002)].

An output of an artificial neuron (see the chapter "*Socionics & Bionics: Learning from 'The Born'*") in multi-layer perceptron (MLP) networks is computed via

$$f_z = \sum_{i=1}^{n} w_i x_i \qquad (3)$$

There is nearly no optimization to reduce the large amount of calculations that have to be done to compute the output of a complex multilayer perceptron. The good point is that there are only additions and multiplications. The bad point is that there are too many of them, which makes it unusable on devices with a lower CPU frequency. A way out of this dilemma is the use of proprietary hardware, as used in hardware-based speech recognition, consult the section "Hardware-Based Speech Recognition".

Nowadays ANNs play a minor role in continuous speech recognition, but are still used in hybrid architectures with HMM-based recognition. In contrast to HMMs, which try to achieve their goal based on statistical and probability models, ANNs deal with classification. They are able to classify a given pattern into phonemes, but are not ideal for the processing of continuous speech. This means that neural networks are used as a replacement for various pieces of a HMM-based recognizer. This is more a philosophical difference with little relevance to use in embedded environments.

As an example of such a network, we look at the multilayer perceptron developed by Bourlard (1992). This network has nine 26-dimensional feature vectors to compute 61 phonemes with 500-4000 neurons in a hidden layer. It allows computing the a posteriori probability $p(q_k|x_i)$. The Viterbi algorithm requires $p(x_i|q_k)$, which can be guessed using the Bayes theorems via

$$p(x_i \mid q_k) = \frac{p(q_k \mid o_T)\, p(o_T)}{p(q_k)} \qquad (4)$$

where $p(o_T)$ can be treated as a constant for all classes and $p(q_k)$ is the frequency distribution of all phonemes, which is known in advance.

FUTURE RESEARCH DIRECTIONS

This chapter gave an overview of the challenges of implementing speech recognition on mobile devices. None of the architectures is ideal in all aspects. Most researchers in speech recognition hope that embedded devices will become powerful enough to have enough performance running off-the-shelf speech recognizers on embedded devices. However, this attitude does not solve the problems that users have if they want to use speech recognition today.

Currently, there are two main approaches to enabling speech recognition on future mobile devices. The first one is followed by hardware engineers who are trying to improve the performance of embedded devices. The development of better hardware will not be able to solve all the challenges in a short time, but will at least address some of them. One aspect is the lack of support for floating point arithmetic, which is also present for rendering of graphical interfaces. Others, like limited memory capacity, will persist. Evolution in recognition performance currently entails a shorter battery life. This is where research in the domain of electrical engineering is required. An alternative is presented by streaming technologies like MRCP that are increasingly used, for example, on mobile phones. Here, standards are needed that allow the distribution of mass data over a wireless connection, or better reduce the traffic.

The second approach is research in speech recognition, looking for tricks to enable speech recognition on these devices with limited capabilities. Advancements in recognition technology are needed too to overcome the challenges. The current approaches have the drawback that they are accompanied by a loss of precision. Here we need better strategies.

The first steps have been taken, but more work is needed to make the vision of voice input on embedded devices come true.

SUMMARY

There are multiple architectures and technologies for implementing speech recognition on mobile devices. They can be divided into *service dependent speech recognition* and *device inherent speech recognition*. Service dependent architectures require a service running on the network to move the computational burden from the client. These architectures offer the same potential as desktop speech recognition at the cost of environmental dependencies. Speech recognition on the device is independent of services running on the network, but pose high computational effort to the mobile device.

This is also the main reason why the speech recognition parameters of service dependent speech recognition cannot be determined exactly. They depend on the technology used on the server side, resulting in full network dependency and high server load. The values of the additional parameters for UC are generally worse than those for device inherent speech recognition.

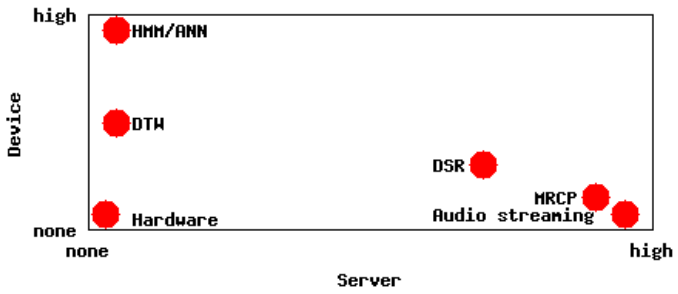
The required network bandwidth is better for DSR than for the streaming architectures, since it aims at reducing the transmitted data. As a consequence, transmission degradation and server load is slightly better.

HMM and ANN based recognizers offer the greatest flexibility and have the best scores for the parameters of speech recognition systems. This is the main reason why service dependent speech recognition performs better in that area.

The transducer is in all cases a noise-canceling microphone. Thus SNR is not a crucial factor.

Implemented on the device, these technologies require too many resources to achieve the same performance as their counterparts on the server. This results in smaller vocabularies, smaller

Figure 24. Distribution of computational resources



models and lower perplexity. They have generally a lower recognition rate than server implementations. In addition, implementations may not have real time capabilities, resulting in a low scoring for responsiveness. The decisive factor is the use of computational resources.

Figure 24 gives a graphical representation of how the type of architecture influences the distributed use of computational resources on the device and on the server.

Hardware-based speech recognition seems to be an appropriate candidate to enable speech recognition on mobile devices, but its rigidity makes it impossible to address multiple application scenarios. Thus it has the worst value for integration and maintenance. DTW requires fewer resources on the device than HMM or ANN, but is highly speaker dependent. It requires enrollment, and supports only isolated word-based recognition, which makes it unusable for certain scenarios.

This analysis can serve as a decision criterion for the architecture to implement or to use. None of the architectures is ideal in all contexts. Especially server dependent architectures require a higher invest, hampering their use in enterprise applications.

REFERENCES

Bahl, L.R., Genneraro, S.V., Gopalakrishnan, P.S., & Mercer, R.L. (1993). A fast approximate acoustic match for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing, 1*(1), 59-67.

Bailey, A. (2004). *Challenges and opportunities for interaction on mobile devices* (Tech. Rep.). Canon Research Centre Europe Ltd.

Bourlard, H., Morgan, N., Wooters, C., & Renals, S. (1992). CDNN: A Context Dependent Neural Network for Continuous Speech Recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 2, pp. 349-352).

Chollet, G. (Ed.). (1999). *Speech processing, recognition and artificial neural networks.* Berlin: Springer.

Cohen, J. (2004). Is embedded speech recognition disruptive technology? *Information Quarterly, 3*(5), 14–16.

Delaney, B., Jayant, N., Hans, M., Simunic, T., & Acquaviva, A. (2002). A low-power, fixed-point, front-end feature extraction for a distributed speech recognition system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 1, pp. 793-796).

Eagle, G. (1999, June/July). Software-only vs. embedded: Which architecture is best for you? *Speech Technology Magazine*.

Frostad, K. (2003, April). The state of embedded speech. *Speech Technology Magazine*.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A., Ravishankar, M., & Rudnicky, A. (2005). *Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices* (Tech. Rep.). Carnegie Mellon University.

Jelinek, F. (2001). *Statistical methods for speech recognition* (3rd ed.). Cambridge, MA: MIT Press.

Junqua, J.-C. (2000). *Robust speech recognition in embedded system and PC applications*. Norwell, MA: Kluwer Academic Publishers.

Kumagai, J. (2002, September 9). Talk to the machine. *IEEE Spectrum Online*.

Lowerre, B. (1976). *The HARPY speech recognition system*. Unpublished doctoral dissertation, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, USA.

Mathew, B.K., Davis, A., & Fang, Z. (2002, November 11). *A Gaussian probability accelerator for SHINX 3* (Tech. Rep. No. UUCS-03-02). Salt Lake City, UT: University of Utah.

Ming, L.Y. (2003). *An optimization framework for fixed-point digital signal processing*. Unpublished master's thesis, The Chinese University of Hong Kong.

Novak, M., Hampl, R., Krbec, P., Bergl, V., & Sedivy, J. (2003). Two-pass search strategy for large list recognition on embedded speech recognition platforms. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 200–203).

Pearce, D. (2000a, May 5). *Enabling new speech driven series for mobile devices: An overview of the ETSI standard activities for distributed speech recognition front-ends* (Tech. Rep.). Motorola Labs.

Pearce, D. (2000b, May 5). Enabling new speech driven services for mobile devices: An overview of the ETSI standard activities for distributed speech recognition front-ends. *AVIOS 2000: The Speech Applications Conference*, San Jose, CA, USA.

Rabiner, L.R. (1989, February 2). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Rabiner, L.R. (1997). Applications of speech recognition to the area of telecommunications. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 501-510).

Rabiner, L.R., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice Hall PTR.

Sakoe, H., & Chiba, S. (1990). Dynamic programming algorithm optimization for spoken word recognition. A.Waibel & K.-F. Lee (Eds.), *Readings in speech recognition* (pp. 159–165). Morgan Kaufmann Publishers, Inc.

Schukat-Talamazzini, E.G.(1995). *Automatische spracherkennung*.

Schulzrinne, H., Rao, A., & Lanphier, R. (1998, April 4). *Real time streaming protocol*. Retrieved May 19, 2006, from http://www.rfc-archive.org/getrfc.php?rfc=2326. Shanmugham, S., Monaco, P., & Eberman, B. (2006, April 4). *A media resource control protocol (MRCP)*. Re-

trieved from http://www.rfc-archive.org/getrfc.php?rfc=4463.

Zaykobskiy, D. (2006). Survey of the speech recognition techniques for mobile devices. In *Proceedings of the 11th International Conference on Speech and Computer*, St. Petersburg, Russia.

ADDITIONAL READING

Amrouche, A., & Rouvaen, J. M. (2006). Efficient system for speech recognition using general regression neural network. *International Journal of Intelligent Technology, 1*(2), 183–189.

Burke, D. (2007). *Speech processing for IP networks: Media resource control protocol (MRCP).* Wiley & Sons.

Chollet, G., DiBenedetto, G., Esposito, A., & Benedetto, G.D. (1999). *Speech processing, recognition and artificial neural networks.* Springer.

Chugh, J., & Jagannathan, V. (2002). Voice-Enabling Enterprise Applications. In *Proceedings of the 11th IEEE International Workshops on Enabling Technologies* (pp. 188–189). Washington, DC: IEEE Computer Society.

Digital speech: Coding for low bit rate communication systems. (1994). John Wiley & Sons, Ltd.

Dynkin, E.B. (2006). *Theory of Markov processes.* Dover Publications.

IEEE (Ed.). (1999). *Speech coding for telecommunications 1999 IEEE workshop.* IEEE Press.

Held, G. (2002). *Voice and data internetworking. Voice over IP gateways.* McGraw-Hill Professional.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition.* New Jersey: Prentice Hall.

Jurafsky, D., & Martin, J.H. (2003). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition.* Prentice Hall.

Kahrs, M., & Brandenburg, K. (Eds.). (1998). *Applications of digital signal processing to audio and acoustics.* Springer-Verlag.

Loizou, P.C. (2007). *Speech enhancement: Theory and practice.* Taylor & Francis Ltd.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Minker, W., & Bennacef, S. (2004). *Speech and human-machine dialog.* Springer US.

Nakagawa, S., Okada, M., & Kawahara, T. (Eds.). (2005). *Spoken language systems.* IOS Press.

Niemann, H. (1990). *Pattern analysis and understanding.* Springer-Verlag.

Novak, M., Hampl, R., Krbec, P., Bergl, V., & Sedivy, J. (2003). Two-pass search strategy for large list recognition on embedded speech recognition platforms. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 1, pp. 200–203).

Oppenheim, A.V., Schafer, R.W., & Buck, J.R. (1999). *Discrete-time signal processing.* Prentice Hall.

Sieworik, D.P. (2001, September 9). *Mobile access to information: Wearable and context aware computers* (Tech. Rep.). Carnegie Mellon University.

Waibel, A., & Lee, K.-F. (Eds.). (1990). *Readings in speech recognition.* Morgan Kaufmann Publishers, Inc.

Wang, Y., Li, J., & Stoica, P. (2005). *Spectral analysis of signals: The missing data case*. Morgan & Claypool Publishers.

William R.G., & Mammen, E.W. (1975). *The art of speaking made simple*. London: Doubleday.

This work was previously published in Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises, edited by M. Mühlhäuser and I. Gurevych, pp. 397-420, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.19
Voice-Enabled User Interfaces for Mobile Devices

Louise E. Moser
University of California, Santa Barbara, USA

P. M. Melliar-Smith
University of California, Santa Barbara, USA

ABSTRACT

The use of a voice interface, along with textual, graphical, video, tactile, and audio interfaces, can improve the experience of the user of a mobile device. Many applications can benefit from voice input and output on a mobile device, including applications that provide travel directions, weather information, restaurant and hotel reservations, appointments and reminders, voice mail, and e-mail. We have developed a prototype system for a mobile device that supports client-side, voice-enabled applications. In fact, the prototype supports multimodal interactions but, here, we focus on voice interaction. The prototype includes six voice-enabled applications and a program manager that manages the applications. In this chapter we describe the prototype, including design issues that we faced, and evaluation methods that we employed in developing a voice-enabled user interface for a mobile device.

INTRODUCTION

Mobile devices, such as cell phones and personal digital assistants (PDAs), are inherently small, and lack an intuitive and natural user interface. The small keyboards and displays of mobile devices make it difficult for the user to use even the simplest of applications. Pen input is available on PDAs, but is difficult to use on handheld devices.

Voice input and output for mobile devices with small screens and keyboards, and for hands- and eyes-free operation, can make the user's interaction with a mobile device more user friendly. Voice input and output can also facilitate the use of Web Services (Booth, Hass, McCabe, Newcomer, Champion, Ferris, & Orchard, 2004) from a mobile device, making it possible to access the Web anytime and anywhere, whether at work, at home, or on the move. Global positioning system (GPS) technology (U.S. Census Bureau, 2006)

can provide location information automatically for location-aware services.

Many everyday applications can benefit from voice-enabled user interfaces for a mobile device. Voice input and voice output for a mobile device are particularly useful for:

- Booking theater and sports tickets, making restaurant and hotel reservations, and carrying out banking and other financial transactions
- Accessing airline arrival and departure information, weather and traffic conditions, maps and directions for theaters, restaurants, gas stations, banks, and hotels, and the latest news and sports scores
- Maintaining personal calendars; contact lists with names, addresses, and telephone numbers; to-do lists; and shopping lists
- Communicating with other people via voice mail, e-mail, short message service (SMS), and multimedia message service (MMS).

It is important to provide several modes of interaction, so that the user can use the most appropriate mode, depending on the application and the situation. The prototype system that we have developed supports client-side, voice-enabled applications on a mobile device. Even though the applications support multimodal input, allowing keyboard and pen input, we focus, in this chapter, on voice input and on multimodal output in the form of voice, text, and graphics. The prototype includes a program manager that manages the application programs, and six voice-enabled applications, namely, contacts, location, weather, shopping, stocks, and appointments and reminders.

BACKGROUND

A multimodal interface for a mobile device integrates textual, graphical, video, tactile, speech, and/or other audio interfaces in the mobile device (Hjelm, 2000; Oviatt & Cohen, 2000). With multiple ways for a user to interact with the applications, interactions with the device become more natural and the user experience is improved. Voice is becoming an increasingly important mode of interaction, because it allows eyes- and hands-free operation. It is essential for simplifying and expanding the use of handheld mobile devices. Voice has the ability to enable mobile communication, mobile collaboration, and mobile commerce (Sarker & Wells, 2003), and is becoming an important means of managing mobile devices (Grasso, Ebert, & Finin, 1998; Kondratova, 2005).

The increasing popularity of, and technological advancements in, mobile phones and PDAs, primarily mobile phones, is leading to the development of applications to fulfill expanding user needs. The short message service (SMS) is available on most mobile phones today, and some mobile phones provide support for the multimedia messaging service (MMS) to exchange photos and videos (Le Bodic, 2002). The mobile phone manufacturers are no longer focused on making a mobile phone but, rather, on producing a mobile device that combines phone capabilities with the power of a handheld PC. They recognize that the numeric keypad and the small screen, common to mobile phones of the past, do not carry over well to handheld PCs (Holtzblatt, 2005).

With the emergence of Web Services technology (Booth et al., 2004), the Web now provides services, rather than only data as it did in the past. Of the various Web Services available to mobile users today, the map application seems to be the most popular, with online map services available from Google (2006) and Yahoo! (2006b). Much progress has been made in creating the multimodal Web, which allows not only keyboard and mouse navigation but also voice input and output (Frost, 2005).

GPS technology (U.S. Census Bureau, 2006) already exists on many mobile devices, and can be used to provide location-aware services (Rao

& Minakakis, 2003), without requiring the user to input geographical coordinates, again contributing to user friendliness.

Speech recognition technology (Rabiner & Juang, 1993) has been developed over many years, and is now very good. Other researchers (Kondratova, 2004; Srinivasan & Brown, 2002) have discussed the usability and effectiveness of a combination of speech and mobility. Currently, handheld voice-enabled applications use short commands that are translated into functional or navigational operations. As observed in Deng and Huang (2004), speech recognition technology must be robust and accurate, and close to human ability, to make its widespread use a reality. Noisy environments present a particular challenge for the use of speech recognition technology on mobile devices and, therefore, multimodal interactions are essential. For example, the MiPad system (Deng, Wang, Acero, Hon, Droppo, Boulis, et al., 2002; Huang, Acero, Chelba, Deng, Droppo, Duchene, Goodman, et al., 2001) uses a strategy where the user first taps a "tap & talk" button on the device and then talks to the device.

Distributed speech recognition (Deng, et al., 2002), in which the speech recognition happens at a remote server exploits the power of the server to achieve fast and accurate speech recognition. However, studies (Zhang, He, Chow, Yang, & Su, 2000) have shown that low-bandwidth connections to the server result in significant degradation of speech recognition quality. In contrast, *local speech recognition* (Deligne, Dharanipragada, Gopinath, Maison, Olsen, & Printz, 2002; Varga, Aalburg, Andrassy, Astrov, Bauer, Beaugeant, Geissler, & Hoge, 2002) utilizes speech recognition technology on the mobile device, and eliminates the need for high-speed communication. Local speech recognition limits the kinds of client handsets that are powerful enough to perform complicated speech processing and, thus, that can be used; however, the computing power of mobile handsets is increasing.

THE PROTOTYPE

The prototype that we have developed allows mobile applications to interact with the user without the need for manual interaction on the part of the human. Speech recognition and speech synthesis software are located on the mobile device, and make the interaction with the human more user friendly. The prototype that we have developed processes natural language sentences and provides useful services while interacting with the user in an intuitive and natural manner. A user need not form a request in a particular rigid format in order for the applications to understand what the user means.

For our prototype, we have developed six application programs and a Program Manager. These applications are Contacts, Location, Weather, Shopping, Stocks, and Appointments and Reminders applications. The Program Manager evaluates sentence fragments from the user's request, determines which application should process the request, and forwards the request to the appropriate application.

The prototype is designed to interact with a human, using voice as the primary means of input (keyboard, stylus, and mouse are also available but are less convenient to use) and with voice, text, and graphics as the means of output. The speech recognizer handles the user's voice input, and both the speech synthesizer and the display are used for output. Characteristics of certain applications render a pure voice solution infeasible. For example, it is impossible to convey the detailed contents of a map through voice output. However, voice output is ideal when it is inconvenient or impossible for the user to maintain visual contact with the display of the mobile device, and it is possible to convey information to the user in that mode. Voice output is also appropriate when the device requests confirmation from the user.

Thus, an appropriate choice of speech recognition and speech synthesis technology is vital to

the success of our prototype. Our choices were constrained by:

- The processing and memory capabilities of typical mobile devices
- The need for adaptability to different users and to noisy environments

The use of speech recognition and speech synthesis technology on a mobile device is different from its use in call centers, because a mobile device is associated with a single user and can learn to understand that particular user.

The Underlying Speech Technology

The prototype uses SRI's DynaSpeak speech recognition software (SRI, 2006) and AT&T's Natural Voices speech synthesis software (AT&T, 2006). It currently runs on a handheld computer, the OQO device (OQO, 2006). We chose this device, rather than a cell phone, because it provides a better software development environment than a cell phone.

Speech Recognition

The DynaSpeak speech recognition engine (SRI, 2006) is a small-footprint, high-accuracy, speaker-independent speech recognition engine. It is based on a statistical language model that is suitable for natural language dialog applications. It includes speaker adaptation to increase recognition accuracy for individuals with different accents or tone pitches. It can be configured so that it performs speech recognition specific to a particular individual. DynaSpeak is ideal for handheld mobile devices, because of its small footprint (less than 2 MB of memory) and its low computing requirements (66 MHz Intel x86 or 200 MHz Strong Arm processor).

DynaSpeak supports multiple languages, adapts to different accents, and does not require training prior to use. It incorporates a Hidden Markov Model (HMM) (Rabiner & Juang, 1993). In an HMM, a spoken expression is detected as a sequence of phonemes with a probability associated with each phoneme. A probability is also associated with each pair of phonemes, that is, the probability that the first phoneme of the pair is followed by the second phoneme in natural speech. As a sequence of phonemes is processed, the probability of each successive phoneme is combined with the transition probabilities provided by the HMM. If the probability of a path through the HMM is substantially greater than that of any other path, the speech recognizer recognizes the spoken expression with a high level of confidence. When the response is below an acceptable confidence threshold, the software seeks confirmation from the user or asks the user questions.

The HMM is augmented with grammars for the particular applications that are required for understanding natural language sentences (Knight, Gorrell, Rayner, Milward, Koeling, & Lewin, 2001). When the user says a new word, the word can be added to the vocabulary dynamically. The HMM is also extended by adapting the vocabulary of the speech recognizer to the current and recent past context of interactions of the user with the applications.

Accuracy of the speech recognition system can be increased by training it for the voice of the particular user. There are two kinds of training, explicit and implicit. *Explicit training* requires the user to read a lengthy script to the device, a process that is likely to be unpopular with users. *Implicit training* allows the device to learn to understand better its particular user during normal use. Implicit training can be provided in two modes, confirmation mode and standard mode.

In *confirmation mode*, the system responds to a user's sentence, and the user confirms or corrects the response. If the user corrects the sentence, the learning algorithm tries to match a rejected, lower probability, interpretation of the original sentence with the user's corrected intent. If a match is found, the learning algorithm adjusts

the HMM transition probabilities to increase the probability of selecting the user's intent. Initially, a new user of the system will probably prefer confirmation mode.

In *standard mode*, the system does not confirm sentences for which there is one interpretation that has a much higher probability than any other interpretation. If no interpretation has a high probability, or if several interpretations have similar probabilities, the speech recognition system responds as in confirmation mode. More experienced users of the system are likely to use standard mode.

The success of implicit training strategies depends quite heavily on starting with a speech recognizer that is well matched to the individual speaker. It is possible, from relatively few sentences, to classify a speaker and then to download, to the mobile device, an appropriate initial recognizer for subsequent implicit training.

DynaSpeak can be used with either a *finite-state grammar* or a *free-form grammar*. We used the finite-state grammar because it offers greater control over parsed sentences. The tendency for DynaSpeak to accept or reject spoken sentences is heavily influenced by the complexity of the grammar. The *complexity of the grammar* is quantified by the number of paths by which an accepting state can be reached. The greater the complexity of the grammar, the higher is its tendency to accept an invalid spoken request. Conversely, the lower the complexity of the grammar, the higher is its tendency to reject a valid spoken request. To minimize the complexity of the grammar and to improve speech recognition accuracy, each application has its own relatively simple grammar. The program manager determines which applications are involved in a sentence and then reparses the sentence using the appropriate grammars.

Speech Synthesis

Natural Voices (AT&T. 2006) is a speech synthesis engine that provides a simple and efficient way of producing natural (rather than electronic) sounding device-to-human voice interactions. It can accurately and naturally pronounce words and speak in sentences that are clear and easy to understand, without the feeling that it is a computer that is speaking.

Natural Voices supports many languages, male and female voices, and the VoiceXML, SAPI, and JSAPI interface standards. Using Natural Voices, we created text-to-speech software for our prototype that runs in the background and accepts messages in VoiceXML format. Each message contains the name of the voice engine (i.e., "Mike" for a male voice and "Crystal" for a female voice) and the corresponding text to speak.

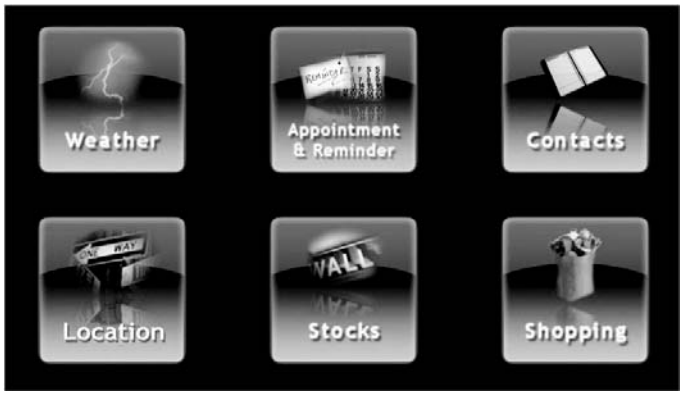
Managed Applications

For the prototype we developed six multimodal applications (contacts, location, weather, shopping, stocks, appointments, and reminders) that use speech as the main form of input. The stocks, maps, and weather applications exploit existing Web Services on the Internet. Communication with those Web Services uses a local WiFi 802.11 wireless network. The program manager controls the operation of the applications. The graphical user interface for the program manager with the six applications is shown in Figure 1. We now present an explanation of the functionality of each application and its role in the overall system.

Contacts

The contacts application stores personal information regarding friends and acquaintances in a database, including their addresses and phone numbers. The contacts application is a mobile extension of a physical contact list or address book that is controlled by voice input. It retrieves data from Microsoft Office Outlook® to populate the database when in docking mode. After using the mobile device and possibly entering new contact information, the user can synchronize informa-

Figure 1. The GUI of the program manager, showing six applications



tion on the mobile device with that on a desktop or server computer. The contacts application is configured to interact with other applications that require information about names, addresses, phone numbers, and so forth. The contacts grammar is the least complex of the application grammars that we developed. The contacts vocabulary grows linearly as contacts are added to the user's contact list.

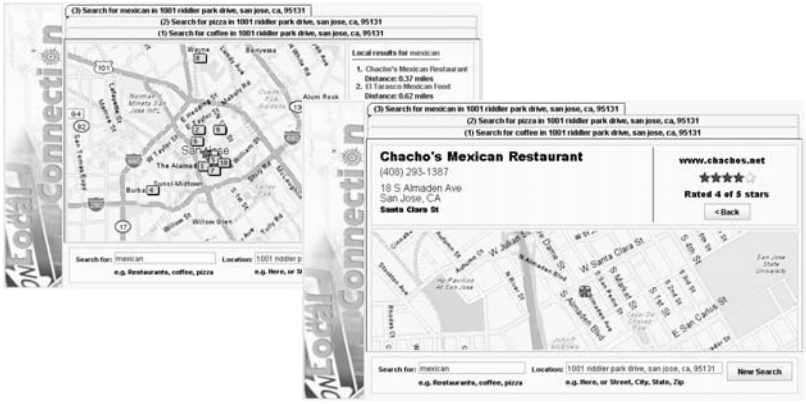
Location

The Location application allows the user to search for restaurants, movie theaters, banks, and so forth, in a given area, using the Yahoo! LocalSearch Web Service (2006b). For example, if the user says to the mobile device "Search for a Mexican restaurant in 95131," the location application on the mobile device sends a Web Service request to Yahoo! LocalSearch, gets back the results, and presents up to 10 results to the user in list form. The user can then view additional information about a single location by indicating the location's number in the presented list. For example, the user can choose to view additional information about "Chacho's Mexican Restaurant" by speaking, "Get more information about number one." On processing this request, the location application presents the user with detailed information about the restaurant including its phone number, address, and a detailed street map showing its location. Figure 2 shows a screen shot of the graphical user interface for the location application.

The location application is loosely coupled with the contacts application to provide responses related to individuals listed in the user's contact list. For example, the request, "Search for a movie theater around Susan's house" uses the contacts grammar to determine the location of Susan's house and replaces the phrase "Susan's house" with the specific address so that the actual search request looks something like this: "Search for a movie theater around 232 Kings Way, Goleta, CA, 93117." The location application then searches for a movie theater in the vicinity of that address.

The location application is also loosely coupled with a GPS module that is contacted when the user has a question related to the user's current location. For example, if the user says "Look for a pizza place around here.", the word "here" is

Figure 2. An example graphical user interface for the location application



recognized by the application and replaced with the GPS coordinates of the user's current location. The location application then sends a Web Service request to Yahoo! LocalSearch, which returns a map of the user's current location, indicating where the user is, along with the 10 nearest pizza places. The Yahoo! LocalSearch Web Service is ideal to use with GPS because of its ability to locate positions on the map on the basis of longitude and latitude. With GPS, the user is no longer limited to requests involving a particular city or zip code. The user now has the ability to create requests that are truly location-aware.

Compared to the grammars of the other applications, the location grammar is one of the most complex. For information like maps and lists, it is desirable to use a graphical or textual display, as well as speech output, in a multimodal user interface. Thus, the most appropriate kind of output can be chosen, depending on the kind of information, the capabilities of the mobile device, and the context in which the user finds himself or herself.

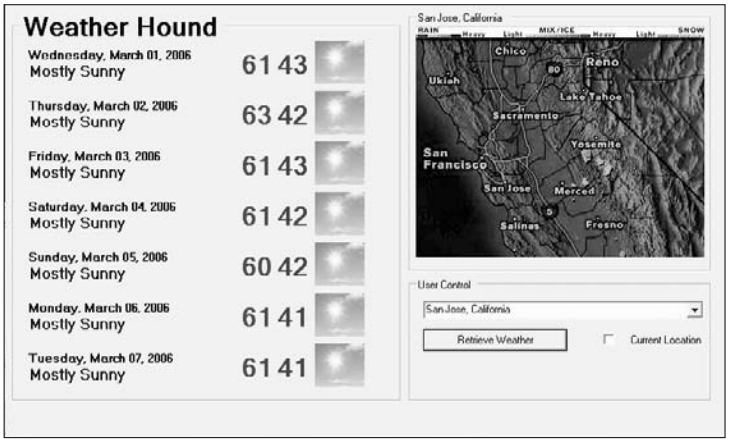
Weather

The weather application supplies weather forecasts obtained from the Web Service provided by the National Weather Service (NOAA, 2006). It allows the user to query for weekly, daily, and 3-day weather information in major U.S. cities using voice input. It allows the user either to select a city or to use the user's current location, as the location for which the weather forecast is to be retrieved from the National Weather Service. The weather application knows the geographical coordinates of dozens of cities in the continental United States. It references those coordinates when the user requests a weather forecast from the National Weather Service for one of those cities.

A user can say "Tell me the weather forecast in San Jose," which then uses "today" as the starting time of the forecast, and produces the graphical user interface for the weather application shown in Figure 3.

Because the weather application operates on a mobile device, it is necessary to be able to deter-

Figure 3. An example graphical user interface for the weather application



mine the user's location dynamically. If the user asks "What's the weather like here two days from now?", the weather application consults the GPS module to obtain the geographical coordinates of the user, contacts the Web Service, and responds with the high and low predicted temperatures and an indication that there is a change to cloudy in Santa Barbara. Thus, the user does not need to provide his/her current location or to obtain the weather forecast for that location.

Our prototype takes into account the many ways in which a person can convey, semantically, equivalent requests in English. For example, a user can ask for the weather in many ways including "What is the weather in Boston like?" or "Tell me what the forecast is like in Boston." These two requests are semantically equivalent because they both contain the same essential parameter, namely the Boston location.

Shopping

The shopping application provides the user with a service capable of reducing the time that the user spends on grocery shopping and the associated stress. The shopping application maintains shopping lists, recipes, and floor plans of supermarkets. The multimodal interface includes speech, text, and graphics, which makes the shopping application easy to use. Figure 4 shows a screen shot of the graphical user interface for the shopping application.

The shopping application allows a user to update his/her shopping list and to forward it to another user. When a user issues a command, like "Remind John to go grocery shopping," the contacts application is used to find John's phone number or e-mail address in the user's contact list. A dialog box then appears asking the user if he/she wants to send, to John, not only a reminder to go shopping but also the shopping list. If so, the shopping list, consisting of the product ids and the quantities of the items needed, is formatted in XML, and appended to the message containing the reminder. The message is then sent to John's shopping application.

The shopping application also displays graphically the floor plan of the supermarket and the

Figure 4. An example graphical user interface for the shopping application

location of items in the store, as shown in Figure 4. This feature provides assistance to the user without the need for the user to contact an employee of the supermarket. The shopping application also allows the user to retrieve recipes while shopping, possibly on impulse, for an item that is on sale. A newly chosen recipe is cross-referenced with the current shopping list, so that needed items can be added automatically. The shopping application has the largest grammar of the applications that we developed, with a vocabulary that depends on the items that the user has purchased recently.

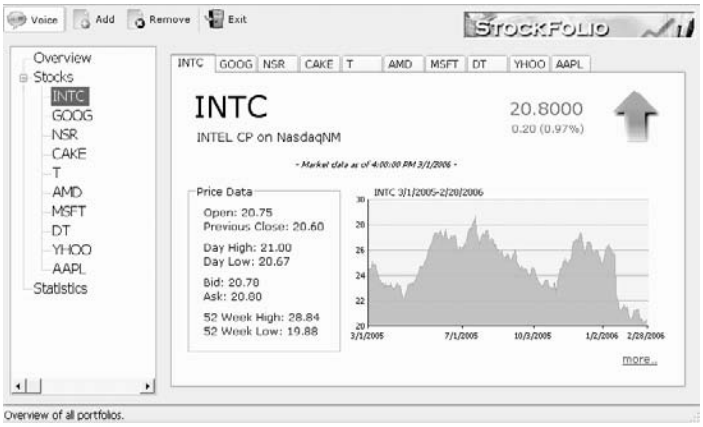
Stocks

The stocks application allows the user to manage his/her stock portfolio using voice input and output. The objective of the stocks application is to monitor stock fluctuations, rather than to trade stocks. The stocks application exploits the Yahoo! Finance Web service (2006a) to store and update stock information in a database. It stores the most recent stock information in the database so that it can reply to the user's requests when connectivity to the Yahoo! Finance Web Service is limited. Although such stored data can be somewhat stale, it allows the user to obtain information whenever the user requests it. The vocabulary of the stocks application grows to match the user's portfolio each time the user adds a new stock.

Appointments and Reminders

The appointments and reminders application manages the user's calendar and allows the user to send reminders to other people. It supports time-based requests of various forms, for example, "Remind me to go to the dentist on Monday," "Remind me to see the dentist on August 15th," and "Remind me to see the dentist a week from today." It displays an easily readable schedule, so that the user can recall what is planned for the day. The appointments and reminders application interacts with other applications, such as the shopping application. For example, the request "Remind John to go shopping on Monday" sends a reminder to John, along with the current shopping list, if the user wishes to forward that information. It also supports

Figure 5. An example graphical user interface for the stocks application



reminders to the user that are location-aware using GPS, for example, if the user is in the vicinity of a supermarket. The appointments and reminders application is an extension of a calendar service. It links to Microsoft Office Outlook®, and updates scheduled appointments and reminders when in the vicinity of the user's desktop.

Program Manager

The program manager evaluates sentence fragments from a user's request, identifies keywords that determine which application or applications should process the request, reparses the sentence using the grammars for those applications, and forwards the parsed request to the appropriate application. If more than one user is involved, the program manager on one user's mobile device sends messages to the program manager on another user's mobile device, which then handles the request.

The program manager leverages DynaSpeak and a weighted keyword recognition algorithm to break down recognized sentences into application-specific fragments. Those fragments are then processed by the appropriate applications, and are subsequently merged to form the final sentence meaning. This process allows the program manager to handle requests that involve more than one application, for example, "Search for a gas station around Paul Green's house." The parsing of this sentence, using the location grammar, requests a search centered on a location that the location grammar cannot itself provide. The program manager must recognize a keyword from the contacts grammar, parse the sentence using that grammar, and query the contacts application for the address of Paul Green's house. The response to the query is then sent to the location application to obtain the location of the gas station nearest his house.

Graphical User Interface

The graphical user interface (GUI) of thepProgram manager, shown in Figure 1, displays the current running application programs and allows the user to select an application by using voice or keyboard

input. The GUI provides buttons that appear gray when an application has not been started and blue after startup. If the user makes a spoken request that requires an application to display a result, the display for that application is topmost and remains topmost until the user issues another request or a timeout occurs. Whenever the GUI is displayed, the user must provide a keyword in a spoken request to wake up the program manager, or click on one of the application-specific buttons on the display.

EVALUATION

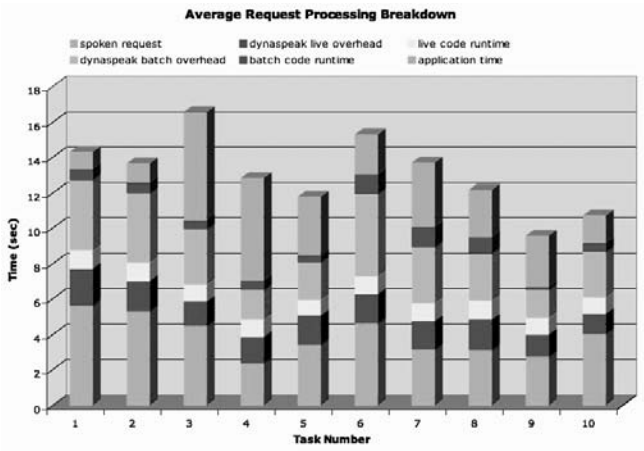
Several experiments were performed to collect qualitative and quantitative data to evaluate the prototype system. Although it is difficult to determine a clear boundary between the user interface and the speech recognizer, it is important to evaluate the user interface and the speech recognizer separately, so that the qualitative and quantitative data gathered from the experiments are not mixed, leading to inconclusive results.

Thus, the experiments were designed as a classical "Don't mind the man behind the curtain" study. In this type of study, the user interacts with a system that is identical to the actual system except that the experiment is being controlled by someone other than the user. The man behind the curtain controls what is spoken as responses to the user's requests and changes the current screen to an appropriate graphical response. This method was used, so that the responses to the qualitative questions would not be biased by the accuracy of the speech recognizer.

To evaluate the system quantitatively, the program manager was instrumented with time segment metrics and data were collected for several performance metrics, including:

- Total time a participant took to complete all tasks
- Overhead of the DynaSpeak speech recognizer during live and batch recognition
- Runtime overhead of the program manager without DynaSpeak
- Spoken length of a request vs. processing time

Figure 6. Processing overhead per task



The results are shown in Figure 6. The time segment metrics represent the runtime complexity of the code associated with the speech recognition and processing. The amount of time taken by each segment adds to the delay associated with the user's request. If any of the time segments has a large duration, the user might become irritated. By measuring each segment separately, the bottleneck in the system can be determined.

The speech processing time increases with the size of the grammar. However, by means of a multiphase procedure that uses keywords organized and weighted by application relevance, the grammar size and the speech processing time can be improved. After live recognition, the system provides a keyword-associated request, which it processes for application weights and then reprocesses using an application-specific grammar, possibly more than once with different grammars. This procedure increases both the speed and the accuracy of the speech recognition, by decreasing the size of the grammar size in the initial phase.

An alternative approach (Kondratova, 2004) is to force the user to make repeated requests, possibly from a menu, with responses by which the device asks for the next step or for more information, so that the device arrives at a better understanding of the user's request. Such an approach introduces navigational complexity for the user. Reducing the speech processing time by creating a complex navigational structure is not the best way to improve usability of the system.

The speech recognizer works better for some speakers than for other speakers. The accuracy of the results can be improved by tuning the speech recognition parameters and enabling learning capabilities. However, the developers of DynaSpeak advise against modification of the speech recognition parameters and use of learning until a relatively high success rate is achieved. For appropriately selected users, quite good speech recognition and understanding can be achieved without using learning capabilities. However, speech recognition accuracy can only improve if voice profiling is combined with learning.

Ambient noise and microphone quality also affect speech recognition accuracy. The internal microphone in the OQO device is of rather poor quality. To ameliorate this problem, a Jabra© Bluetooth headset, was used to provide noise cancellation and reduce the distance between the microphone and the user's mouth. In addition, when the confidence score from DynaSpeak falls below an acceptable threshold, the program manager seeks confirmation from the user or asks for clarification. These mechanisms greatly improve the accuracy of the speech recognizer.

The accuracy of speech recognition is degraded when the grammar contains words that are phonetically similar. During preliminary experiments for the shopping application, we had problems recognizing differences between similar sounding requests like "Add lamb to my shopping list" and "Add ham to my shopping list." These problems arise particularly when users are non-native English speakers or when they have accents. Creating more specific requests can reduce the phonetic similarity, for example, by saying "Add a lamb shank to my shopping list" and "Add a ham hock to my shopping list." However, modifying requests in such a way is undesirable because the requests are then less intuitive and natural.

The location, weather, and stocks applications all use Web Services and require communication over the Internet and, thus, have longer application runtimes than the other Web Services. The location application is written in Java, which runs more slowly than C#. Both the weather application and the stocks application cache data associated with previous requests to take advantage of timing locality. Location requests are different because the caching of maps can involve a large usage of the memory, and users are not inclined to perform the same search twice. Memory is a precious commodity on a handheld device and needs to

be conserved; thus, the location application is coded so that it does not cache maps resulting from previous queries.

To evaluate the qualitative aspects of the system, we performed a user study with participants from diverse backgrounds of education, ethnicity, and sex. The user study was completed with 10 individuals performing 10 tasks resulting in 100 request results. The participants were given a questionnaire that assessed their general impressions about the prototype, with the results shown in Table 1.

After analyzing the averaged responses of the participants, we found several trends. The participants' scores are not strongly correlated with speech recognition accuracy. Participant G gave the system a high score, but was one of the two participants who encountered the most speech recognition problems. Participant B gave the system a low score despite good speech recognition.

The participants agreed that speaking to a mobile handheld device as if it were a human is not comfortable. It is difficult to get used to interacting with a computer that can understand tasks that would be commonplace for humans. The participants were relatively pleased with the GUI interface design and felt the system is relatively easy to use. However, the ease-of-use metric needs to be taken lightly. Ease of use can be assessed more concretely by measuring the number of times a user must repeat a command.

The scores for response appropriateness and relevance are high, indicating that the spoken responses of the applications were well crafted. The scores related to recommending the service to friends and daily life helpfulness are relatively high, from which one might infer that the participants would purchase a device providing the speech-enabled applications. However, this conclusion is not necessarily justified. The participants were not enthusiastic about having to pay for such a device or for such services. However, most participants in the study were quite pleased with the prototype system and found the user interface helpful and easy to use.

Table 1. Responses to the questionnaire

Questions	A	B	C	D	E	F	G	H	I	J	Mean
Was it comfortable talking to the device as if it were a human?	3	3	4	3	4	3	4	5	3	5	3.7
Was the GUI aesthetically pleasing?	5	4	4	5	5	5	5	5	5	4	4.7
Were the request responses appropriate and easy to understand?	3	3	5	5	5	4	4	5	4	4	4.2
Were the spoken responses relevant to your requests?	5	4	5	5	5	3	4	5	4	5	4.5
Was the system easy to use?	4	5	3	4	4	5	4	5	4	5	4.3
Do you think the services would be helpful in your daily life?	4	4	4	5	4	4	4	4	4	5	4.2
Would you recommend a system like this to your friends?	3	3	4	5	4	4	5	5	4	5	4.2
Would you buy the software if it were available for your phone?	3	2	4	5	3	3	5	5	3	5	3.8

FUTURE TRENDS

Integration of multiple applications, and multiple grammars, is not too difficult for a small number of applications that have been designed and programmed to work together, as in our prototype. However, future systems will need to support tens or hundreds of applications, many of which will be designed and programmed independently. Integration of those applications and their grammars will be a challenge.

Currently, speech-enabled applications typically use short commands from the human that are translated into navigational or functional operations. More appropriate is speech recognition technology that supports a more natural, conversational style similar to what humans use to communicate with each other (McTear, 2002).

A mobile device that listens to its owner continuously can provide additional services, such as populating the user's calendar. For example, when a user agrees to an appointment during a conversation with another person, the mobile device might recognize and automatically record the appointment, possibly confirming the appointment later with its user. Similarly, the mobile device might note that the user habitually goes to lunch with the gang at noon on Mondays, or that the user leaves work promptly at 5pm on Fridays. With existing calendar systems, the user often does not record appointments and other commitments, because it is too much bother using the human interfaces of those systems, greatly reducing the value of the calendar.

A useful capability of speech recognition systems for mobile devices is being able to recognize intonation and emotional overtones. "The bus leaves at 6" is, overtly, a simple declaration, but appropriate intonation might convert that declaration into a question or an expression of disapproval. Existing speech recognition systems do not yet recognize and exploit intonation. Similarly, the ability to recognize emotional overtones of impatience, uncertainty, surprise, pleasure, anger, and so forth, is a valuable capability that existing speech recognition systems do not yet provide.

Speech recognition requires a relatively powerful processor. Typical cell phones contain a powerful digital signal processor (DSP) chip and a much less powerful control processor. The control processor operates continuously to maintain communication with the cellular base stations. The DSP processor uses a lot of power and imposes a significant drain on the battery and, thus, analyzes and encodes speech only during calls. The DSP processor is capable of the processing required for speech recognition, although it might need more memory.

For mobile devices, battery life is a problem, particularly when speech recognition or application software requires a powerful processor. The limit of 2 hours of talk time for a cell phone is caused at least as much by the power drain of the DSP processor as by the power needed for wireless transmission. The DSP processor might be needed for speech processing for more than 2 hours per day. There are several possible solutions to this problem, namely, larger batteries, alcohol fuel cells, and DSP processors with higher speeds, reduced power consumption, and better power management.

Background noise remains a problem for speech recognition systems for mobile devices, particularly in noisy environments. The quality of the microphone, and the use of a headset to decrease the distance between the microphone and the speaker's mouth, can improve speech recognition accuracy.

CONCLUSION

The use of voice input and output, in addition to text and graphics and other kinds of audio, video, and tactile interfaces, provides substantial benefits for the users of mobile devices. Such multimodal

interfaces allow individuals to access information, applications, and services from their mobile devices more easily. A user no longer has to put up with the annoyances of a 3-inch keyboard, nested menus, or handwriting recognition, nor does the user need to have a tethered desktop or server computer in order to access information, applications, and services. Providing multiple ways in which the users can interact with the applications on mobile devices brings a new level of convenience to the users of those devices.

REFERENCES

AT&T. (2006). *Natural voices.* Retrieved from http://www.natural voices.att.com/products/

Booth, D., Hass, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., & Orchard, D. (2004). *Web services architecture*. Retrieved from http://www.w.3.org/Tr/WS-arch

Deligne, S., Dharanipragada, S., Gopinath, R., Maison, B., Olsen, P., & Printz, H. (2002). A robust high accuracy speech recognition system for mobile applications. *IEEE Transactions on Speech and Audio Processing, 10*(8), 551-561.

Deng, L., & Huang, X. (2004). Challenges in adopting speech recognition. *Communications of the ACM, 47*(1), 69-75.

Deng, L., Wang, K., Acero, A., Hon, H., Droppo, J., Boulis, C., Wang, Y., Jacoby, D., Mahajan, M., Chelba, C., & Huang, X. D. (2002). Distributed speech processing in MiPad's multimodal user interface. *IEEE Transactions on Speech and Audio Processing, 10*(8), 605-619.

Frost, R. A. (2005). Call for a public-domain SpeechWeb. *Communications of the ACM, 48*(11), 45-49.

Google. (2006). *Google Maps API.* Retrieved from http://www.google.com/apis/maps

Grasso, M. A., Ebert, D. S., & Finin, T. W. (1998). The integrality of speech in multi-modal interfaces. *ACM Transactions on Computer-Human Interaction, 5*(4), 303-325.

Hjelm, J. (2000). *Research applications in the mobile environment. Wireless information service*. New York, NY: John Wiley & Sons.

Holtzblatt, K. (2005). Designing for the mobile device: Experiences, challenges, and methods. *Communications of the ACM, 48*(7), 33-35.

Huang, X., Acero, A., Chelba, C., Deng, L., Droppo, J., Duchene, D., Goodman, J., Hon, H., Jacoby, D., Jiang, L., Loynd, R., Mahajan, J., Mau, P., Meredith, S., Mughal, S., Neto, S., Plumpe, M., Stery, K., Venolia, G., Wang, K., & Wang, Y. (2001). MiPad: A multimodal interaction prototype. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1,* 9-12.

Knight, S., Gorrell, G., Rayner, M., Milward, D., Koeling, R., & Lewin, I. (2001). Comparing grammar-based and robust approaches to speech understanding: A case study. In *Proceedings of Eurospeech 2001, Seventh European Conference of Speech Communication and Technology* (pp. 1779-1782). Aalborg, Denmark.

Kondratova, I. (2004, August). Speech-enabled mobile field applications. In *Proceedings of the IASTED International Conference on Internet and Multimedia Systems,* Hawaii.

Kondratova, I. (2005, July). Speech-enabled handheld computing for fieldwork. In *Proceedings of the International Conference on Computing in Civil Engineering*, Cancun, Mexico.

Le Bodic, G. (2002). *Mobile messaging, SMS, EMS and MMS.* John Wiley & Sons.

McTear, M. (2002). Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys, 34*(1), 90-169.

National Oceanic and Atmospheric Administration (NOAA). (2006). *National Weather Service*. Retrieved from http://www.weather.gov/xml/

OQO. (2006). *The OQO personal computer*. Retrieved from http://www.oqo.com

Oviatt, S., & Cohen, P. (2000). Multi-modal interfaces that process what comes naturally. *Communications of the ACM, 43*(3), 45-53.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Sarker, S., & Wells, J. D. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM, 46*(12), 35-40.

SRI (2006). *DynaSpeak*. Retrieved from http://www.speechatsri.com/products/sdk.shtml

Srinivasan, S., & Brown, E. (2002). Is speech recognition becoming mainstream?. *Computer Magazine*, (April), 38-41.

U.S. Census Bureau. (2006). *Precision of GPS*. Retrieved from http://www.census.gov/procur/www /fdca/library/mcd/7-29%20 MCD_WG_ hardware_subteam_report.pdf

Varga, I., Aalburg, S., Andrassy, B., Astrov, S., Bauer, J. G., Beaugeant, C., Geissler, C., & Hoge, H. (2002). ASR in mobile phones—An industrial approach. *IEEE Transactions on Speech and Audio Processing, 10*(8), 562-569.

Yahoo! LocalSearch. (2006a). Retrieved from http://www.local.yahooapis.com/LocalSearchService/V3/localSearch

Yahoo! Finance. (2006b). Retrieved from http://finance.yahoo.com/rssindex

Zhang, W., He, Y., Chow, R., Yang, R., & Su, Y. (2000, June). The study on distributed speech recognition system. In *Proceedings of the IEEE International Conference on Acoustical Speech and Signal Processing* (pp. 1431–1434), Istanbul, Turkey.

KEY TERMS

Global Positioning System (GPS): A system that is used to obtain geographical coordinates, which includes a GPS satellite and a GPS receiver.

Hidden Markov Model (HMM): A technique, based on a finite state machine that associates probabilities with phonemes, and pairs of phonemes, that is used in speech recognition systems, to determine the likelihood of an expression spoken by a user of that system.

Location Aware: An application that is based on a particular physical location, as given by geographical coordinates, physical address, zip code, and so forth, that determines the output of the application.

Mobile Device: For the purposes of this chapter, a handheld device, such as a cell phone or personal digital assistant (PDA), that has an embedded computer and that the user can carry around.

Multimodal Interface: The integration of textual, graphical, video, tactile, speech, and other audio interfaces through the use of mouse, stylus, fingers, keyboard, display, camera, microphone, and/or GPS.

Speech Recognition: The process of interpreting human speech for transcription or as a method of interacting with a computer or a mobile device, using a source of speech input, such as a microphone.

Speech Synthesis: The artificial production of human speech. Speech synthesis technology is also called text-to-speech technology in reference to its ability to convert text into speech.

Web Service: A software application identified by a Uniform Resource Indicator (URI) that is defined, described, and discovered using the eXtensible Markup Language (XML) and that supports direct interactions with other software applications using XML-based messages via an Internet protocol.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 446-460, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.20
Voice Driven Emotion Recognizer Mobile Phone:
Proposal and Evaluations

Aishah Abdul Razak
Multimedia University, Malaysia

Mohamad Izani Zainal Abidin
Multimedia University, Malaysia

Ryoichi Komiya
Multimedia University, Malaysia

ABSTRACT

This article proposes an application of emotion recognizer system in telecommunications entitled voice driven emotion recognizer mobile phone (VDERM). The design implements a voice-to-image conversion scheme through a voice-to-image converter that extracts emotion features in the voice, recognizes them, and selects the corresponding facial expression images from image bank. Since it only requires audio transmission, it can support video communication at a much lower bit rate than the conventional videophone. The first prototype of VDERM system has been implemented into a personal computer. The coder, voice-to-image converter, image database, and system interface are preinstalled in the personal computer. In this article, we present and discuss some evaluations that have been conducted in supporting this proposed prototype. The results have shown that both voice and image are important for people to correctly recognize emotion in telecommunications and the proposed solution can provide an alternative to videophone systems. The future works list some modifications that can be done to the proposed prototype in order to make it more practical for mobile applications.

INTRODUCTION AND MOTIVATION

Nonverbal communication plays a very important role in human communications (Komiya, Mohd Arif, Ramliy, Gowri, & Mokhtar, 1999). However,

in telephone systems, only audio information can be exchanged. Thus, using telephony, the transmission of nonverbal information such as one's emotion would depend mostly on the user's conversation skills. Although the importance of nonverbal aspects of communication has been recognized, until now most research on nonverbal information concentrated on image transmission such as transmission of facial expression and gesture using video signal. This has contributed to the emergence of a videophone system, which is one of the most preferred ways to exchange more information in communication. Such services, however, require a wide bandwidth in order to provide real time video that is adequate for a natural conversation. This is often either very expensive to provide or difficult to implement. Besides, in a videophone system, the user has to be fixed in front of the camera at the correct position during the conversation, so that the user's image can be captured and transmitted correctly. This limitation does not happen in the normal telephone system.

Another approach is to use model-based coding (Kidani, 1999). In this approach, instead of transmitting video signals containing an image of the user, only the human action data such as the facial expressions, movement of the mouth, and so on acquired using a microphone, a keypad, and other input devices, are transmitted over the network. When these data are received by the receiver, the polygon coordinate data for each facial feature is recalculated in accordance with the displacement rules and the person's expression is synthesized.

Our approach is similar to the second approach in a sense that a synthesize image is used for the facial expression reconstruction at the receiver side. However, the difference is that only voice is transmitted and the emotion data is extracted from the received voice tone at the receiving side. This is based on the idea that, voice, besides for communication, it is also an indicator of the psychological and physiological state of a speaker. The identification of the pertinent features in the speech signal may therefore allow the evaluation of a person's emotional state. In other words, by extracting the emotion information from the voice of the speaker, it is possible to reconstruct the facial expression of that speaker. Thus, based on this voice-to-image conversion scheme, we propose a new system known as voice driven emotion recognizer mobile phone (VDERM), as seen in Figure 1. This system uses a voice-to-image converter system at the receiver side that identifies the emotional state of the received voice signal and selects the corresponding facial expression of that particular emotion from the image bank to be displayed. Using this approach, only audio transmission is required. Therefore, the existing second generation (2G) mobile phone infrastructures can be used. Another advantage is that the user does not need to be fixed in front of the camera during the conversation because there is no need for image transmission.

VOICE TO IMAGE CONVERSION

Referring to Figure 1, the voice-to-image conversion for this system is done at the receiving side. The conversion scheme can be divided into two parts: the emotion recognition and facial expression reconstructor. These two processes are done by the voice-to-image converter.

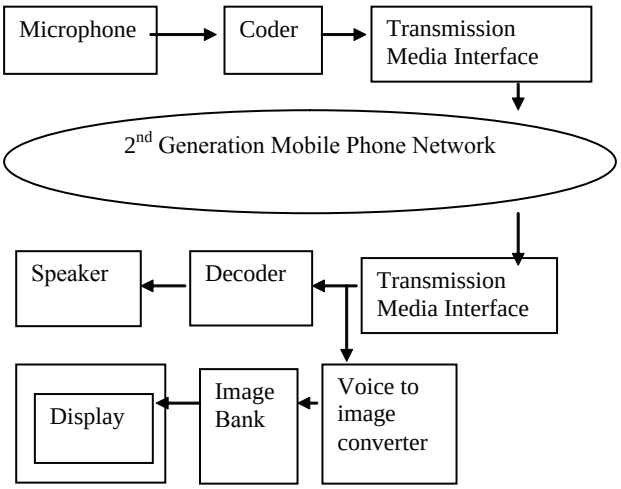
Emotion Recognition

Before we come out with the emotion recognizer design, first we have to deal with these three issues:

1. What kind of emotion to be recognized?

How many and what types of emotional states should be recognized by our system is an interesting yet difficult issue. Besides, there is no widely accepted definition and taxonomy of emotion; it

Figure 1. Basic block diagram of VDERM system



should also be kept in mind that a single emotion can be uttered in different ways. Scherer (1986) distinguishes different categories in a single emotion, for instance, the category "cold anger/ irritation" and the category "hot anger/rage." In our study, we have chosen to consider emotions as discrete categories (Ekman, 1973; Izard & Carroll, 1977; Plutchik, 1980). Six basic emotions defined by Cornelius (1996), that is, happiness, sadness, anger, fear, surprise, and disgust, have been chosen as the emotions to be recognized and reconstructed.

2. What are the features to represent emotion?

Determining emotion features is a crucial issue in the emotion recognizer design. This is because the recognition result is strongly dependant on the emotional features that have been used to represent the emotion. All the studies in this area point to the pitch (fundamental frequency) as the main emotion feature for emotion recognition. Other acoustic features are vocal energy, frequency, spectral features, formants (usually only one or the first two formants F1 and F2 are considered), and temporal features (speech rate and pausing) (Banse & Scherer, 1996). Another approach to feature extraction is to enrich the set of features by considering some derivative features, such as linear predictive coding cepstrum (LPCC) parameters of signal (Tosa & Nakatsu, 1996). Our study has adopted this approach and uses linear predictive analysis (Rabiner & Schafer, 1978) to extract the emotion features. A detailed analysis has been done on selected emotion parameters (Aishah, Izani, & Komiya, 2003a, 2003b, 2003c). Based on these analyses, a total of 18 features (as in Table 1) have been chosen to represent the emotion features. The 18 features are pitch (f_0), jitter (jt), speech energy (e), speech duration (d), and 14 LPC coefficients (a_1- a_{14}). The LPC coefficients are included because we intended to use LPC analysis for the extraction algorithm. Besides,

Table 1. Speech features and description

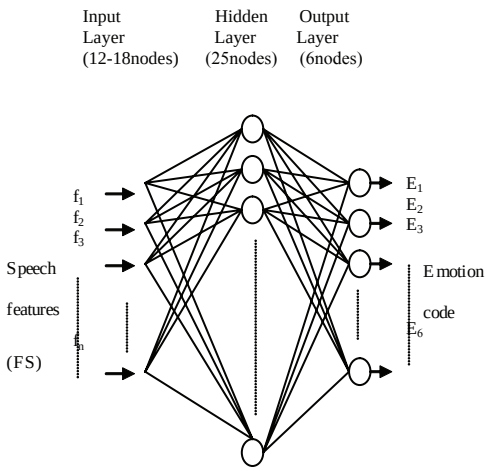
No	Feature	Symbol used	Description
1	Energy	e	Average energy of the speech signal
2	LPC Coefficient	$a_1, a_2, a_3 \ldots a_{14}$	The weighting coefficient used in the linear prediction coding analysis.
3	Duration	d	Duration of the speech signal
4	Pitch	f_0	Fundamental frequency (oscillation frequency) of the glottal oscillation (vibration of the vocal folds).
5	Jitter	jt	Perturbation in the pitch

it represents the phonetic features of speech that are often used in speech recognition

3. What technique to be used for recognition?

There are many methods that have been used for emotion recognition/classification. For instance, Mcgilloway, Cowie, Douglas-Cowie, Gielen, Westerdijk, and Stroeve (2000) have compared and tested three classification algorithms, namely linear discriminant, support vector machine (Schölkopf, Burges, & Smola, 1998), and quantization (Westerdijk & Wiegerinck, 2000). Others are using fuzzy model, K-nearest neighbors, and neural networks (Petrushin, 1999). Among all, perhaps the most common and popular method of emotion recognition is neural network. However, the configuration of the networks differs from one researcher to another, as discussed by Morishima and Harashima (1991), Nakatsu, Nicholson, and Tosa (1999), Petrushin (1999), and Tosa and Nakatsu (1996). In this article we applied neural network configuration as described by NETLAB (Nabney, 2001) for the recognition technique. It uses a 2-layer multilayer perceptron (MLP) architecture with 12-18 elements in the input vector which correspond to the speech features, 25 nodes in the hidden layer, and 6 nodes in the output layer which correspond to the six elements of output vector, the basic emotions. This configuration is illustrated in Figure 2.

Figure 2. The neural network configuration



The weights are drawn from a zero mean, unit variance isotropic Gaussian, with variance scaled by the fan-in of the hidden or output units as appropriate. This makes use of the MATLAB function RANDN and so the seed for the random weight initialization can be set using RANDN ("STATE," S) where S is the seed value. The hidden units use the TANH activation job.

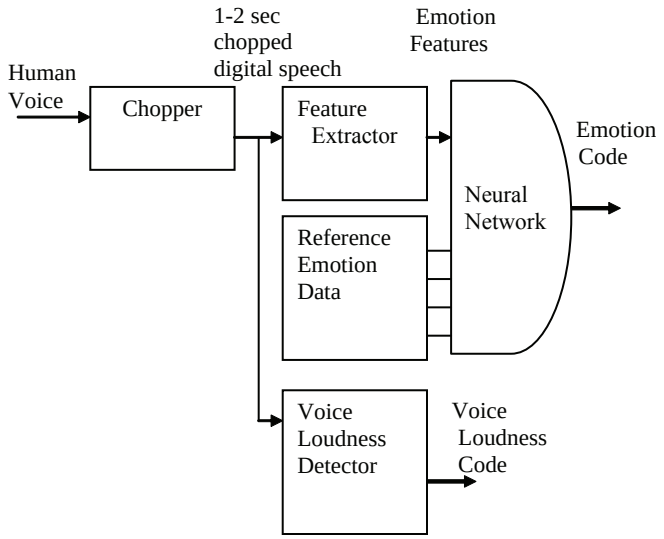
During the training, the weights are adjusted iteratively using a scaled conjugate gradient algorithm (Fodslette, 1993) to minimize the error function, which is the cross-entropy function with softmax as the output activation function. In our experiment, 1,000 iterations are found to be sufficient to achieve an acceptable error rate.

Once our network is trained, we test the network using test samples and calculate the recognition rate. The speech features from test samples are fed into the network and forward propagate through the network to generate the output. Then, the resulted classification performance for the predicted output (output which is recognized by the network) is compared to the target output and displayed in the confusion matrix table. A detail discussion on the result of our experiment using neural network approach is presented by Aishah, Azrulhasni, and Komiya (2004). It is found that an emotion recognition rate of 60% is achievable using the neural network method and this result is sufficient based on a human recognition rate done on the same experiment data.

Once we have dealt with the above issues, we come out with an emotion recognition process as shown in Figure 3. Basically this part extracts the emotion content of the speech received from the transmitting user and recognizes it. In addition to emotion, we also identify the voice loudness level of the speech so that it can be used to control the opening of the mouth shape on the facial images later on.

First, the continuous human voice would be chopped into 2 second speech and undergo preprocessing where it is normalized by its maximum amplitude and the d.c. component is removed.

Figure 3. Block diagram of emotion recognition process



Next, it will go through a feature extractor process where the sample is segmented into 25 msec frames with a 5 msec overlap. LPC analysis is then carried out on these series of frames and the outputs are the 14 LPC coefficients, first reflection coefficient, and energy of the underlying speech segment and energy of the prediction error. Using these outputs, the remaining parameters are determined in the next stage. Speech duration is determined by first classifying the speech frames as voiced or unvoiced using the prediction error signal by simply setting a threshold. The first frame that is classified as voiced will mark the beginning of the speech period. After the beginning of the speech period, if the frame is classified as unvoiced for few consecutive frames, the speech is decided to be ended. The length of the speech period is calculated to get the speech duration. The pitch period for each frame that lies within the speech period is calculated using the cepstrum of the voiced prediction error signal. If an abrupt change in the pitch period is observed, that period is compared to previous pitch periods, and then low-pass filtered (or median filtered) to smooth the abrupt change. With the perturbation in the pitch, jitter is then calculated using pitch perturbation order 1 method, which is obtained by taking the backward and forward differences of perturbation order zero. All the calculation is developed using MATLAB with the use of speech processing and synthesis toolbox (Childers, 1999).

It should be noted that the features extracted so far are based on frame-by-frame basis. As a result, for each sample, it might have many sets of features depending on the number of frames that lie within the speech period of that particular sample. Since we need to standardize the entire sample to have only one feature set, the average of each feature over the frame size is calculated for each sample. The final feature set (FS) for sample n (s_n), consisting of 18 elements is given as

$$\text{FS for } s_n = (e_n, a_{1n}, a_{2n}, a_{3n} \ldots a_{14n}, d_n, f_{0n}, jt_n) \quad (1)$$

On the other hand, a copy of the chopped digital speech will be sent to the voice loudness level detector to detect the loudness level. The outputs of this emotion recognition process are emotion code (from the neural network) and voice loudness code (from the voice loudness detector).

Facial Expression Reconstructor

Figure 4 shows the process involved in facial expression reconstructor. First, the code processor will process the emotion code and voice loudness code and convert it to the equivalent image ID used in the database. The code conversion would also depend on which image the user would like to use represented by the model ID.

For the first prototype of this system, we have used Microsoft Access for the image database and Visual Basic is used as its interface. For each images stored in the database, there is a unique ID tagged to it. The ID is generated sequentially by the system automatically whenever an image is uploaded into the system. Before starting the conversation, the user must first choose which image to be used (for example, male or female model) and once the image is selected, the code processor will make necessary conversions on the received emotion code and voice loudness code to match the ID range for that image. Accordingly, the ID number is sent to the image database, and the image with a matching ID retrieved from the database is then displayed. This image database by default consists of 24 images of female models and 24 images of male models. For each model, the images consist of six basic emotions and each emotion has four levels of voice presented by the opening of mouth shape. In addition we also include different eye shapes which are generated randomly among the four levels.

System Interface

Figure 5 shows the main interface of the facial expression reconstructor system. It consists of a

Figure 4. Block diagram of facial expression reconstructor process

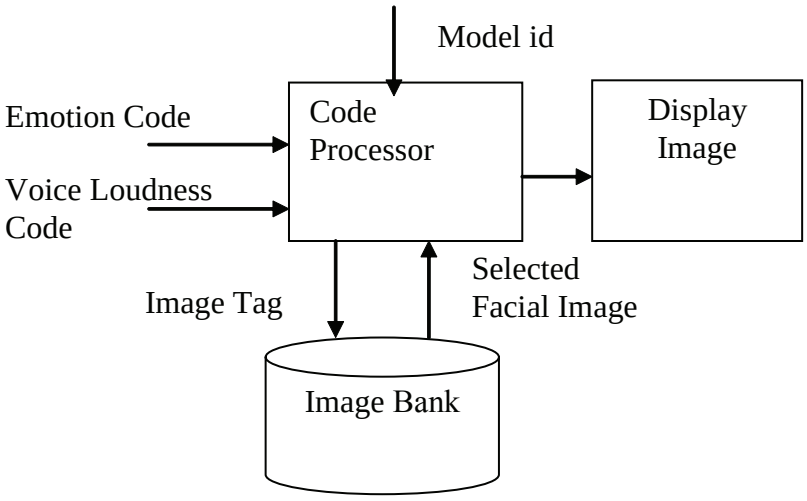
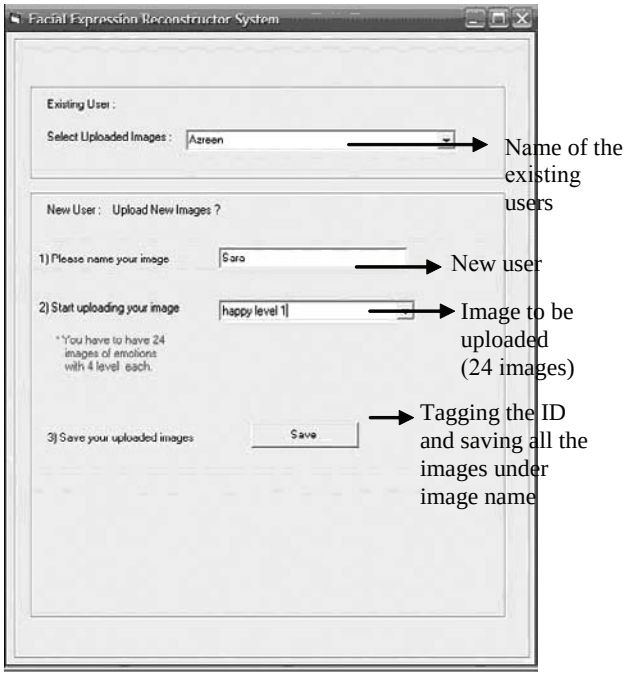


Figure 5. Main interface of facial expression re-constructor

display displaying the facial expression images and buttons for the user to select which model to be used, such as a male or female model. If they click on the "Male" button, the image of male model will be used. Similarly, when they click on the "Female" button, the female image will be displayed, as shown in Figure 5.

In addition to the prestored male and female model, users can also use their own images by using the "Upload own Image" button. When a user clicks on it, a second interface that is shown in Figure 6 will pop up. It has the buttons that show the name of the users that already have their own images uploaded and stored in the system. For new users who want to upload their own image, they can type in their name or any name that will represent them. This will act as the name of the folder which will store all the 24 images of the person. Then users can start to upload their own images which consist of all the six basic emotions which are happiness, sadness, anger, disgust, fear, and surprise with four levels each. The user must have all the 24 images in order to have a complete set of images. At this section, once the user chooses, for example "happy level 1," it will pop up a dialog box and the user will search and upload the appropriate image and save it. Each of the newly uploaded images will

Figure 6. Interface for uploading new image



be tagged accordingly with an image ID. After all the 24 images are completely uploaded, users have to click on the "Save as" button for the final step. This is the part where all the images will be put under the folder that has been created and the image ID range for this image name will be recorded and saved for future use. Once this is done, the user can use it by selecting the name from the "uploaded images" box.

EVALUATIONS

Preliminary Evaluations

There are three preliminary evaluations conducted in the process of completing our first prototype. The details of the evaluations are discussed in the next subsections.

PE I

The main objective of this evaluation are to select good quality of emotional voice samples for our voice database. The samples would later be used for training and testing our neural network model for emotion recognition. The second objective is to see how accurate a person can recognize the emotion content in a speech by just listening to the speech/voice. The result of this human recognition rate will be used as a guide for the recognition rate expected to be achieved by our recognition system later on. The evaluation involves two steps:

1. Collecting voice sample

Four short sentences frequently used in everyday communication that could be expressed in all emotional states without semantic contradictions

are chosen as the utterance. The sentences are: "Itu kereta saya" (In Malay language), "That is my car" (In English), "Sekarang pukul satu" (In Malay language), and "Now is one o'clock" (In English). Two languages are used for the utterance because we want to compare the emotional parameters for both languages and determine how much difference in language could influence the way emotions are expressed. Then, we asked acting students to utter the same utterance 10 times, each for different emotional states and also in neutral voice as a reference.

2. Listening test

The emotionally loaded voice samples are then randomized and each sample is repeated three times within short intervals followed by a 30 second pause. The series of stimuli are presented via headphones to 50 naive listeners to evaluate the emotional state within seven categories: happiness, sadness, anger, fear, disgust, surprise, and not recognizable emotional state.

PE II

The main objective of this evaluation is to detect the most appropriate and highly recognizable facial expressions images that can be used to represent the six basic human emotions for our image database. Another objective is to see how accurate humans can detect emotion based on only the facial expression images, without any audio support. The evaluation also involves two steps:

1. Image collection

For this purpose we have selected three male models and three female models. The facial expressions of the six basic human emotions, which are happy, sad, angry, surprise, disgust, and fear, portrayed by each model are captured, focusing from the neck and above. The background used is a blank white wall with natural daylight so that the model image is clear and focused. The size of images taken are 1000x1600, using a Sony Cyber shot digital camera. The images are then cropped and resized to 6x4 inches size.

2. Viewing test

All the images are randomized and presented to 20 assessors consisting of 10 males and 10 females volunteers. The images are displayed for 2 seconds each with a 3 second gap between images. The assessors are asked to identify the emotion of the given images and then the recognition rates for each image and emotion state are calculated.

PE III

The main objective of this experiment is to verify that the combination of voice and image can improve the capability of humans to correctly recognize an emotion and thus justify the importance of a VDERM system.

For the purpose of this evaluation, we have selected three emotional voice samples for each emotion and matched it with the corresponding images. Around 50 assessors have participated in this evaluation and the recognition rate is calculated and analyzed. They are also asked to answer some questions to reflect the importance of voice and facial expressions in effective communications.

System Evaluation

The main objective behind this system evaluation is to evaluate the reliability and feasibility of the idea of a VDERM system using the developed prototype. This evaluation tries to get some feedback from the user on how efficient the system can improve the message conveyed during a conversation, is the displayed image synchronous with the intended emotion of the speech, and how can the interface be further improved according

to the user's specification. Responses from the assessors on their perception of the VDERM system are important to determine better research direction for the proposed system. For this initial evaluation, the prototype of the VDERM system has been implemented into a personal computer. The coder, voice-to-image converter, image database, and system interface are preinstalled in the personal computer.

Experimental Set-Up

A subjective assessment technique is chosen due to the practicability and suitability. A total of 20 assessors consisting of experts and nonexperts take part in the evaluation test. Assessors first went through a demo on the idea behind a VDERM system and followed with a briefing about the evaluation form contents and the way evaluation must be done. The list of evaluation item is given in Table 2. A sample of a 40 second one-way conversation was played, first using a female model and then a male model with a 15 second gap. The facial expression images switched between different emotions and mouth shapes depending on the emotion content and loudness level of the conversation at every two seconds. Assessors were asked to pretend that they were having a conversation with the model and then they were required to rate the quality of each of the evaluation items based on the scale given in Table 3. Then, they were asked to give comments and suggestions on the grade given for each evaluation item.

Table 2. List of evaluation item for preliminary system evaluation

No.	Item
1	Overall
2	Image accuracy in displaying the facial expression
3	Image synchronization with the speech
4	Features and interface
5	Quality of the images
6	Emotion recognition capability

Table 3. Grading scale

Scale	Quality
0	Worse
1	Average
2	Good
3	Best

RESULTS AND ANALYSIS

PE I

From this evaluation, we have achieved an average recognition rate of 62.33%. The average recognition rate is in line with what has been achieved by other studies using different languages (i.e., around 55-65%) (Morishima & Harashima, 1991; Nakatsu et al., 1999; Petrushin, 1999). This result has proven that even a human is not a perfect emotion recognizer. This is because recognition of emotions is a difficult task due to the fact that there are no standard ways of expressing and decoding emotion. Besides, several emotional states may appear in different scale and have very similar physiological correlates, which result the same acoustic correlates. In an actual situation, people solve the ambiguities by using the context and/or other information. This finding indicates that we shall not try to have our machine to achieve a perfect recognition rate. The human recognition rate is used as a guideline towards achieving the satisfactory rate for computer recognition.

Table 4. Confusion matrix table of PE I

Intended emotion	Response from the assessors					
	Happy	Sad	Anger	Disgust		Surprise
Happy	**68**	2	5	8	12	5
Sad	7	**61**	9	3	18	2
Anger	4	21	**46**	11	7	11
Disgust	4	1	5	**77**	7	6
Fear	9	12	5	15	**54**	5
Surprise	8	1	5	8	10	**68**

Table 4 shows the confusion matrix table that is achieved in PE I. The confusion matrix table of PE I suggests how successful the actors were in expressing the intended emotion and which emotions are easy or difficult to realize. We see that the most easily recognizable emotion based on this experiment is disgust (77%) and the least easily recognizable category is anger (46%). A high percentage of confusion occurs in sad-fear (18%) and anger-sad (21%).

A total of 200 samples which have the highest recognition rate are selected for each emotion. This has resulted in 1,200 samples for the whole voice database.

PE II

From the results in Table 5, it is concluded that among all the facial expression images of emotion, the easiest expressions detected by assessors is happy. This is due to the fact that happiness is the most common emotion shown publicly by humans and it is usually expressed with a smile and bright eyes. Thus it is not difficult to identify

Table 5. Confusion matrix table of PE II

Intended emotion	Response from the assessors					
	Happy	Sad	Anger	Disgust	Fear	Surprise
Happy	**90**	0	0	0	0	10
Sad	0	**80**	0	5	15	0
Anger	0	0	**60**	35	0	5
Disgust	0	0	35	**40**	0	15
Fear	0	5	0	5	**60**	30
Surprise	10	0	0	0	30	**60**

Table 6. Summary of facial gestures according to emotions

EMO-TION	FACIAL GESTURES
Happy	Smile, laughter
Sad	Down turned mouth and eyes, tears
Angry	Eyes bulging, mouth tighten
Disgust	Wrinkled nose, lowered eyelids and eyebrow, raised upper lip
Fear	Eyes squinting
Surprise	Eyes bulging, raised eyebrows, mouth shaped "O"

a happy face even without an audio support. The least recognizable emotion is found to be disgust (40%) and it is often confused with anger (35%). This is because the expressions for both of these emotions are very similar, shown through the "hostile" face; the tight line of the mouth and the squinting eyes. Another thing to note is that based on image, the diversification of confusion in a particular emotion is less compared to recognition based on voice only.

We have also done some analysis on the images which are highly recognizable, and together with the feedback from the assessors, we have identified some main facial gestures which significantly contribute to the recognition of certain emotion. This is summarized in Table 6. Based on the results of Human Evaluation II, we have identified 1 male and 1 female model which have the highest recognizable images to be our model. Then the images of six emotions are recaptured according to the significant facial gestures and each emotion is further developed into four levels of mouth shape, resulting in 24 images for each models.

PE III

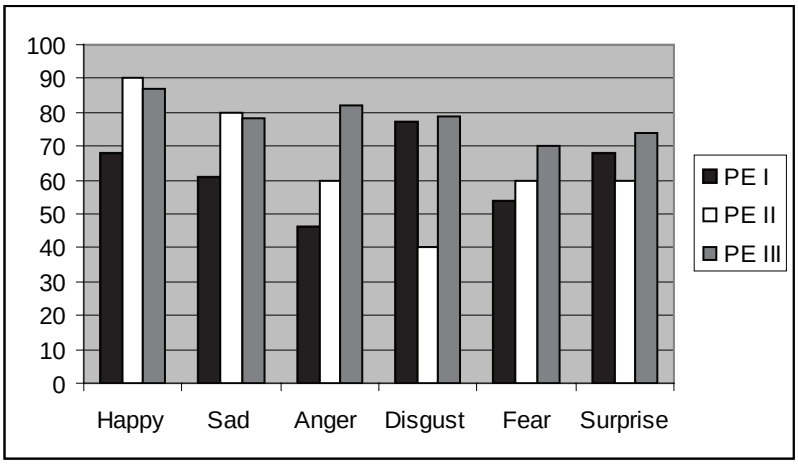
From the confusion matrix Table 7, it is illustrated that the recognition rate for all emotions are quite high (70%-87%). According to assessors, fear is difficult to recognize (70%) compared to other emotions as the expression accompanying the voice of fear can be confused or sad. The most easily recognizable emotion is happy (87%), as the cheerful voice is supported by the smiling face.

Figure 7 compares the recognition rate achievable in all the three preliminary evaluations. On average, PE I achieved an average recognition rate of 62.33% with individual recognition rates ranging between 46% and 77%. PE II shows a slightly higher average recognition rate (65%) compared to PE I with a wider range of individual recognition rate (between 40% and 90%). A significant increase in average and individual recognition is clearly seen in PE III with an average recognition of 78.3% and individual recognition concentration between 70% and 87%. The results have clearly illustrated that PE III has the most percentage of correctness in emotion identification by assessors. This shows that combination of both what we hear (voice) and what we see (image) can greatly improve human capability of identifying the emotions. Thus, this result has justified the importance of the proposed system, which is to combine the image and voice to improve the naturalness and efficiency of telecommunication.

Table 7. Confusion matrix table of PE III

Intended emotion	Response from the assessors					
	Happy		Anger	Disgust		Surprise
Happy	**87**	0	0	0	0	13
Sad	0	**78**	0	0	22	0
Anger	0	0	**82**	18	0	0
Disgust	0	6	15	**79**	0	0
Fear	0	16	7	0	**70**	7
Surprise	26	0	0	0	0	**74**

Figure 7. Comparison between recognition rate achieved in PE I, II and III



On top of the recognition rate presented above, below we have summarized the findings that were collected from the asessors feedback.

1. Comfortable way to express and detect emotion

For emotion expression (Figure 8), 45% agreed that facial expression is the most comfortable way to express emotion, followed by voice tone (34%), and body gesture (21%).

The pattern is also the same for emotion detection (Figure 9) but the percentage for emotion

detection by facial expression is higher at 64%. This is followed by voice tone at 30% and body gesture at 6%.

2. Medium of communication

The results in Figure 10 show that telephone/ mobile phones are the most popular medium of communication nowadays with 48% as the majority, followed by Internet instant messenger (40%) and short messages service (SMS) (12%). This is because telephones are the most convenient and widely available medium of communication.

3. Importance of audio (voice tone) and video (facial expression) information for effective communication

The result in Figure 11 show that 98% agreed that both audio and video information are important for effective communication.

4. Reliability of emotion extraction from voice

The result in Figure 12 show that 64% have agreed that it is reliable to extract emotion from voice.

The result of human emotion recognition based on audio, video, and both highlights two important points. The first one is that emotion recognition is a difficult task and humans can never perfectly recognize emotion. This is because the way people express and interpret emotion might differ from one another, even though some researches have found that it does follow some standard pattern (Cosmides, 1983). The second point is that the ability of humans to correctly recognize emotion is increased to more than 70% when both audio and video information of the speaker are present.

It is also found that facial expression is the most comfortable way for people to express and detect emotion during communication. Thus the presence of facial expressions is very important in communication.

Figure 8. Comfortable way to express emotion

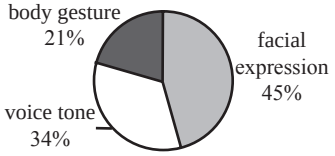


Figure 9. Preferable way to detect emotion

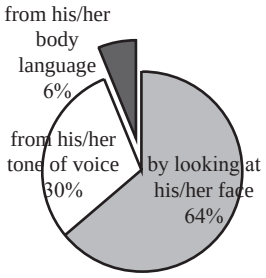


Figure 10. Popular medium of communication

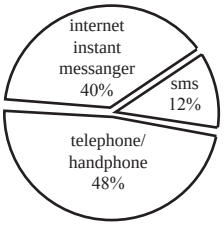


Figure 11. Importance of audio (voice tone) and video (facial expression) information for effective communication

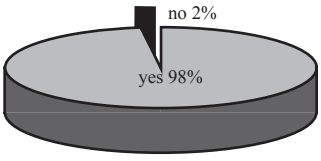
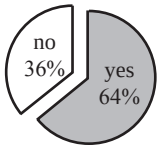


Figure 12. Reliability of emotion extraction from voice



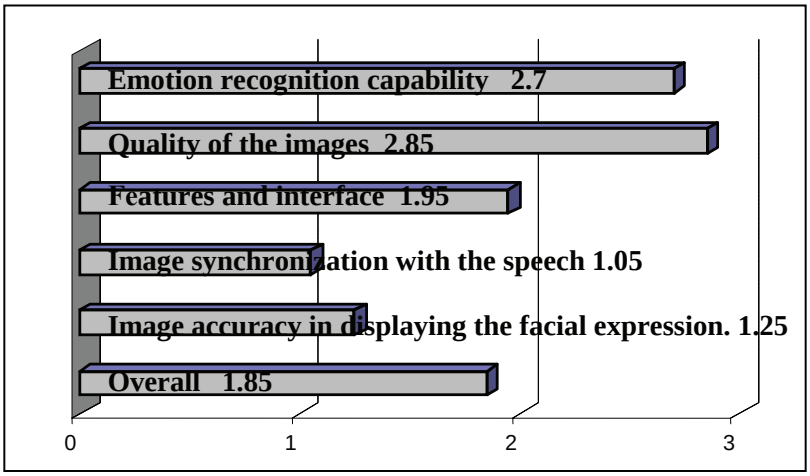
Overall, the results of the preliminery evalutions highlight the importance of facial expressions in providing effective communication, and thus, it is very important to incorporate facial expressions in today's telephone system, as proposed in the VDERM system.

System Evaluation

An average score of each evaluation item was calculated and a bar chart was plotted to represent the results. The preliminary evaluation results are illustrated in Figure 13. From the chart, it is shown that the quality of the images used have the highest scale average with value of 2.85. This shows that the assessors are happy with the quality of the images used, which are clear and focused. This level of quality is not achievable when using a videophone system because the image is transmitted real time and thus subject to transmission quality and the position of the user during the conversation. The next highest scale average is on the emotion recognition capability. This is expected because this system provides both audio and visual information of the speaker which help the assessors to recognize the emotion more easily compared to just having the audio information, as in the case of normal phone conversation.

The third highest score item goes to features and interface design. Overall the assessors think that the system is user friendly because the system has a straight forward design that does not confuse a user, even if the user is not familiar with a computer system. Another point is that the interface buttons have direct and clear instructions to be followed by the user. Moreover, most of the assessors find that the feature to upload their own image is very interesting because it gives the user customization to make the system personal to them. However, a few are concerned with the image database size as more images are being uploaded, and the difficulty of having a set of 24 images before can use their own image. Some also suggested that the features can be improved by having interactive interface using JAVA and personalized skins for the background.

Figure 13. Result for system evaluation



The average score for overall system performance is 1.85. Many of the assessors agreed that this system can improve the efficiency of telecommunication system because the presence of both audio and visual have given them more clues on the emotions being conveyed. Besides they also found that having both elements made the conversation more fun and interesting. In addition, the automatic detection of emotion from voice is also an interesting attempt because most of the currently available chat/Web cam applications require the user to manually select the emotion to be/being conveyed. However, the assessors believed that the system still has a lot of room for improvement, especially in the aspect of image accuracy and image synchronization.

As shown on the chart, image accuracy in displaying the facial expression and image synchronization with the speech has an average score of less than 1.5. The main reason for this is that the level of image for each emotion currently used is only four, which has resulted in switching between images which seems less smooth and the lip movement does not appear to be synchronized with the speech. This is an important issue to address because if the images and voice do not synchronize, the user might not be able to catch the emotion being conveyed. By having more levels for a particular emotion, switching between different emotion states can be smoother, thus the user can have more time to identify the emotion of the images.

FUTURE DESIGN ISSUES

The main improvement in the system is concerned with the accuracy of the images being displayed. Since our intention is to provide real time visual images as in video conferencing, it is very important that the image switching appears to be smooth and synchronized with the speech and emotion content of the speech. One simple modification is to have more levels to represent the voice loudness. However, increasing the level means increasing the number of images needed and consequently can increase the size of our database, which is not desirable.

One possible solution to this is to have the personal images of the user deleted after the call is terminated. However, the problem with this is that the user might need to upload their image every time before the user can start a conversation, which can be time consuming and not practical.

The other advanced alternative for facial expression reconstruction is to use a 3D artificial face model, which can generate facial expressions based on the coded emotion, instead of switching between still pictures. This method is more advanced and complicated but it can provide a more natural facial expressions. For this method, software will be used to convert a person's photo into a face model and a program will be developed to generate the facial expressions on the model based on the emotion code and voice loudness. Using this technique, there is no need for a large image database, which might be more appropriate for application on a mobile phone.

CONCLUSION

In general, based on the results, the system has achieved its main objective, that is, to improve telecommunication by providing voice together with facial expressions and provide an alternative to videophone systems. However it still has a lot of room for improvement in the aspect of interface design and the accuracy of the images being displayed. The evaluation that we have conducted so far was tested on a personal computer. In order to apply the system on a mobile phone, the issues pertaining to the image database size should be thoroughly dealt with.

REFERENCES

Aishah, A. R., Azrulhasni, M. I., & Komiya, R. (2004). A neural network approach for emotion recognition in speech. In *Proceedings of the 2nd International Conference on Artificial Intelligence in Engineering and Technology (ICAIET2004)* (pp. 910-916).

Aishah, A. R., Izani, Z. A., & Komiya, R. (2003a). A preliminary analysis for recognizing emotion in speech. In *Proceedings of IEEE Student Conference On Research and Development (SCOReD 2003).*

Aishah, A. R., Izani, Z. A., & Komiya, R. (2003b). Emotion pitch variation analysis in Malay and English voice samples. In *Proceedings of the 9th Asia Pacific Conference on Communications (APCC2003)* (Vol. 1, pp. 108-112).

Aishah, A. R., Izani, Z. A., & Komiya, R. (2003c). Towards automatic recognition of emotion in speech. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT2003).*

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*, 614-636.

Childers, D. G. (1999). *Speech processing and synthesis toolboxes.* New York: John Wiley & Sons.

Cornelius, R. R. (1996). *The science of emotion: Research and tradition in the psychology of emotion.* Upper Saddle River, NJ: Prentice-Hall.

Cosmides, L. (1983). Invariance in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 864-881.

Ekman, P. (1973). *Darwin and facial expression: A century of research in review.* New York: Academic Press.

Fodslette, M. M. (1993). A scaled conjugate gradient algorithm for fast-supervised learning. *Neural Networks, 6*, 525-533.

Izard, & Carroll, E. (1977). *Human emotions.* New York: Plenum Press.

Kidani, Y. (1999). Video communication system using portrait animation. In *Proceedings of the IEEE Southeastcon '99* (pp. 309-314).

Komiya, R., Mohd Arif, N. A., Ramliy, M. N., Gowri Hari Prasad, T., & Mokhtar, M. R. (1999). A proposal of virtual reality telecommunication system. In *Proceedings of the WEC'99* (pp. 93-98).

Mcgilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, C. C. A. M., Westerdijk, M. J. D., & Stroeve, S. H. (2000). Approaching automatic recognition of emotion from voice: A rough benchmark. In *Proceedings of the ISCA Workshop on Speech and Emotion* (pp. 207-212).

Morishima, S., & Harashima, H. (1991). A media conversion from speech to facial image for intelligent man-machine interface. *IEEE J. on Selected Areas in Comm., 9*(4), 594-600.

Nabney, I. (2001). *Netlab: Algorithms for pattern recognition, advances in pattern recognition.* London: Springer-Verlag.

Nakatsu, R., Nicholson, J., & Tosa, N. (1999). *Emotion recognition and its application to computer agents with spontaneous interactive capabilities.* Paper presented at the International Congress of Phonetic Science (pp. 343-351).

Petrushin, V. A. (1999). Emotion in speech recognition and application to call centers. In *Proceedings of the ANNIE '99.*

Plutchik, R. (1980). *Emotion: A psycho-evolutionary synthesis.* New York: Harper and Row.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals.* Eaglewood Cliffs, NJ: Prentice-Hall.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*, 43-165.

Schölkopf, C. J. C., Burges, A. J., & Smola (1998). *Advances in kernel methods: Support vector learning.* Cambridge, MA: MIT Press.

Tosa, N., & Nakatsu, R. (1996). Life-like communication agent-emotion sensing character MIC and feeling session character MUSE. In *Proceedings of the IEEE Conference on Multimedia* (pp. 12-19).

Westerdijk, M., & Wiegerinck, W. (2000). Classification with multiple latent variable models using maximum entropy discrimination. In *Proceedings of the 17th International Conference on Machine Learning* (pp. 1143-1150).

This work was previously published in the International Journal of Information Technology and Web Engineering, edited by G. Alkhatib, Volume 3, Issue 1, pp. 53-69, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).

Chapter 8.21
Mobile Multimedia for Speech and Language Therapy

Nina Reeves
University of Gloucestershire, UK

Sally Jo Cunningham
University of Waikato, New Zealand

Laura Jefferies
University of Gloucestershire, UK

Catherine Harris
Gloucestershire Hospitals NHS Foundation Trust, UK

ABSTRACT

Aphasia is a speech disorder usually caused by stroke or head injury (Armstrong, 1993). Related communication difficulties can include word finding, speaking, listening, writing, and using numbers (FAST, 2004). It is most commonly acquired by people at middle age or older, as a result of stroke or other brain injury. Speech and language therapy is "the process of enabling people to communicate to the best of their ability" (RCSLT, 2004). Treatment, advice, and support are provided based on assessment and monitoring activities that conventionally are carried out in face-to-face sessions. This chapter considers issues in providing technology to continue to support aphasic patients between therapy sessions, through multimedia applications for drill-and-practice in vocalizing speech sounds. Existing paper therapy aids are generally designed to be used under the guidance of a therapist. Multimedia applications enable people with aphasia to practise spoken language skills independently between sessions, and mobile multimedia speech and language therapy devices offer still greater promise for blending treatment and support into an aphasic person's daily life.

INTRODUCTION

Current trends in the demography of the developed world suggest that increased longevity will lead to a larger population of patients needing rehabilitation services after a stroke (Andrews & Turner-Stokes, 2005). An essential part of these services is speech and language therapy (SLT) (NHS, 2004) to enable the patient with aphasia to return to the community and live as independently as possible. At present, even in countries where SLT is a well-developed profession, resources in terms of staff and mobile communication devices for loan are limited (Harris, 2004). Therapy generally cannot offer a "cure" for aphasia; instead, the goals of therapy are to support the person in capitalizing on remaining language ability, regaining as much of their prior language skills as possible, and learning to use compensatory methods of communication.

This chapter describes the existing therapy methods based on paper materials and mobile electronic devices commonly called augmentative and alternative communication (AAC) devices and proposes the development of software solutions which could be delivered flexibly via readily available mobile devices such as personal digital assistants (PDA) used in a stand alone mode or via Internet delivered services. These could be designed to suit the needs of not only the patients and their carers, but also those of the professional speech and language therapists (SLTs) who could tailor and monitor the treatment more regularly than presently possible. The process of creating and evaluating prototype applications with SLTs is described and recommendations are made for the direction of future research and development.

CONVENTIONAL SPEECH AND LANGUAGE THERAPY AIDS

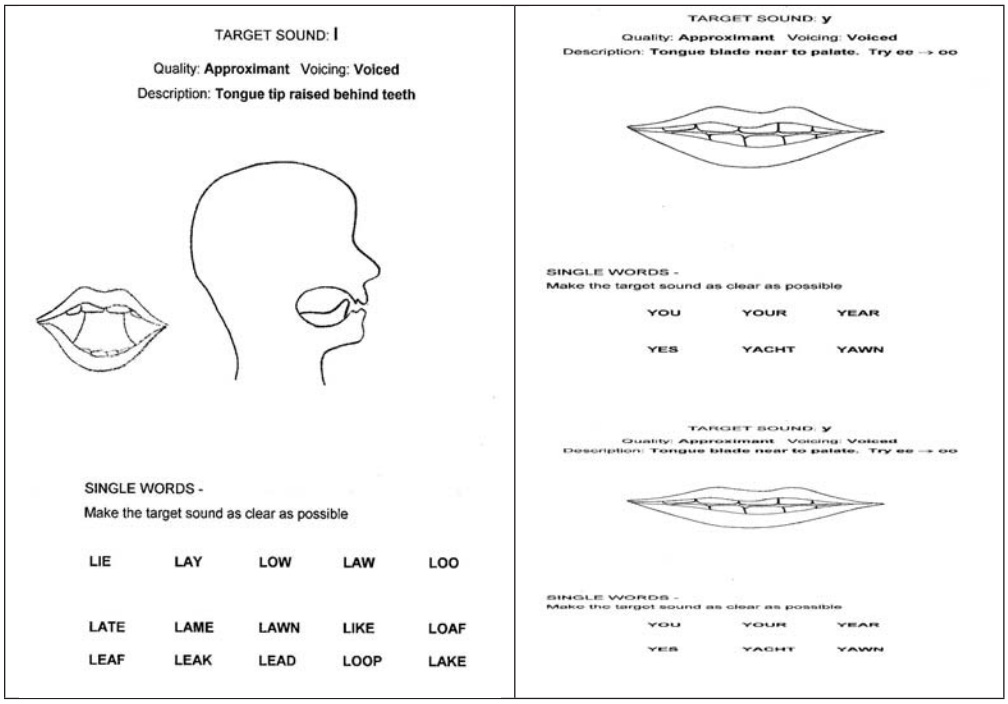
Paper-based representations of lip and tongue positions for sounds are a venerable and common speech and language therapy aid. These are generally provided as line drawings, illustrating how specific sounds are made; the aids are intended for use both during therapy and in practice sessions outside of therapy. Figure 1 is typical of this type of therapy material.

While useful, this paper-based material has obvious limitations. The line drawings are static, and fail to accurately represent the movements necessary to make speech sounds (Harris, 2004). No spoken explanation or auditory reinforcement is possible—this would be provided by the speech therapist, during a session or by a carer. The latter would be untrained and may in certain circumstances reinforce errors. Aphasia frequently includes impairments in processing written language. Notice that these sheets include a relatively large amount of text—some of it possibly redundant ("Make the target sound as clear as possible"), some echoing in a less accessible form the line drawings ("Tongue tip raised behind teeth"), some vocabulary that is highly technical and likely to be unintelligible without training in speech therapy ("Quality: Approximant"). These issues can put off users from practising on their own, or can diminish the effectiveness of their practice.

The problem of confining effective therapy to formal, therapist-facilitated sessions is significant. A meta-analysis of evaluations of aphasia therapy (covering 864 individuals) concluded that concentrated therapy over a shorter period of time has a greater positive impact on recovery than less concentrated therapy over a longer period (Bhogal, Teasell, & Speechley, 2003). Clearly, face-to-face therapy with a trained therapist is the ideal, but availability is a bottleneck for aphasia treatment. Even in an economically well-developed country such as the UK, there are only, on average, 0.6 SLTs per 10 beds in rehabilitation units (Andrews & Turner-Stokes, 2005)

Early attempts at software support for SLT have had variable and limited success (Burton, Meeks, & Wright, 1991), in part because of hardware and

Figure 1. Sample paper-based therapy material



development environment limitations, and in part because there was a limited body of knowledge about the accessibility and usability for aphasic users. As will be discussed next, these barriers are less significant today.

MOBILE DIGITAL COMMUNICATION AND THERAPY TOOLS

Two broad categories of software speech and language therapy applications exist for use by people with aphasia:

- **Drill-and-practice software** that offer instruction and the opportunity to practise language skills. A typical application contains a variety of standard instructional material, and also allows the therapist to record additional words, phrases, or other utterances for playback and practice by the patient.
- **Compensatory software** that provides alternative means for the user to communicate, for example by producing audio, image-based, or written messages for the user. The user selects an appropriate message for playback or display in situations where communication is required—for example, when dealing with commercial organizations or government departments.

Both types of software now commonly include multimedia facilities. Earlier applications were severely limited in scope by tiny (by today's standards) storage available with standard PCs. Larger hard drives, DVD storage, and high-bandwidth Internet connections now support applications that can include video and audio display of large practice sets, audio recording of user practice sessions (for both immediate feedback to the user, and for later evaluation by the therapist), icons for navigating the interface, and speech generation facilities to reduce the need for pre-recording messages used to communicate with others.

Both application categories also have potential to increase their effectiveness by moving to mobile devices. Compensatory communication devices, or AAC devices, would obviously be more useful if they were small and light enough to be easily carried along into conversational situations—so that the user does not have to struggle with a laptop while shopping. Similarly, portable drill-and-practice devices would allow users to practise frequently during the odd breaks that are inevitably sprinkled through the day, and would allow patients to continue therapy when away from home for more extended periods such as vacations (Glykas & Chytas, 2004). However, current commercially-available portable AAC devices are special-purpose pieces of equipment; as such, they are considerably more expensive than standard PDAs and mobiles, even with monochrome screens. As a consequence, portable AAC tools are not widely used at present.

Moving implementations of AAC tools from special-purpose hardware to standard, general purpose devices holds promise for supporting the development of cheaper, more readily available devices. Recent PDA models are now viable platforms for speech and language therapy applications. Screens, while small, now have a high enough resolution to provide a crisp display of images and line drawings. Memory remains relatively small but is sufficient to store a selection of drill exercises and common conversational phrases and sentences. Indeed, there is evidence that including smaller datasets makes these applications more, rather than less, usable. Experience with information display for non-literate (Deo, Nichols, Cunningham, Witten, & Trujillo, 2004) and communication impaired users (Dunlop, Cunningham, & Jones, 2002) emphasizes that these users primarily depend on browsing rather than search to navigate the application. Browsing forces the user to rely on memory to find, and return to, desired documents or exercises, and a too-rich set of options to select from can quickly lead to frustration when the user cannot remember the location of a previously retrieved item or cannot efficiently navigate to a new item. For this type of application, it appears that the limitation is the capacity of the user's memory, rather than the device's storage. Some patients with aphasia have short-term memory problems so the retrieval issue is even more important.

ACCESSIBILITY ISSUES

Funding is one significant accessibility issue for speech and language therapy tools. The cost of a device, software, and training often puts these applications out of reach for many individuals, and local health authorities have extremely limited supplies, if any, for loan (Harris, 2004).

It is clear, however, that the provision of a mobile speech and communication aid is not sufficient in and of itself to ensure that it will be used, and useful, in everyday settings. One study in the UK and the Netherlands demonstrated that, for carefully selected patients, these applications can effectively support therapy—however, 11 of the 28 patients did not choose to use the device in non-clinical settings (van de Sandt-Koenderman, Wiegers, & Hardy, 2005).

Accessibility can also be limited by interface and interaction design. People with aphasia often experience difficulties with reading and understanding text, and may experience a related visual

problem that makes reading tiring and error-prone. Investigations into accessibility and usability of software for people with aphasia have produced several recommendations for effective applications (FAST, 2004; Queensland Aphasia Groups, 2001a, 2001b):

- Interfaces should include few, and simple, words, in large print.
- White space should dominate—text should be as widely spaced as possible.
- When possible, include images and icons to explain the words, or to serve as substitutes/reminders of the words.
- Text, images, and functionality should be pertinent to the needs and interests of people with aphasia—not to therapists, caregivers, or members of the medical community. Note that a separate interface may be required for these other potential users of the application (for example, to allow therapists to create new drill exercises).
- Interaction should not be keyboard dependent, and should be designed for ease of use with alternative input/output devices such as screen readers and switches. Aphasia frequently makes construction of text via the keyboard difficult or impossible, and the condition causing the aphasia (for example, a stroke) may also create physical disabilities that hamper keyboarding.
- Applications should be interactive rather than static information displays, and if possible should support the user in expressing his/her own thoughts (rather than literally putting words into the user's mouth with a standard set of utterances).

DEVELOPMENT OF MULTIMEDIA PROTOTYPES

Two prototypes of a speech therapy tool, SoundHelper, were developed. The tool is intended to support drill and practice in phonetic sounds. In both versions, the user selects a sound for practice, and the sound is represented both through audio and through a demonstration of mouth placement for producing that sound. The design of these prototypes was informed by a speech and language therapist and by an expert in interaction design. On the advice of the therapist the prototypes focus on vowel sounds, as these are the first sounds which are usually addressed in therapy for a person newly aphasic after a stroke or other brain injury.

The SoundHelper interface is organized around the familiar and common "folder" metaphor. The top level of the system presents the folder "containing" all of the exercises (Figure 2a). At the next level, folders are also used to organize and group different classes of sound—for example, exercises pertaining to vowels (Figure 2b).

The two demonstration displays implemented are an animated line drawing (Figure 3a) and a short video clip (Figure 3b). The prototypes were developed using Adobe Premiere to capture the video and Macromedia Flash MX to compress the video clips. Macromedia Flash was then used to rotoscope or motion capture the lip and facial movements, to produce the animated line drawings. The vector graphics produced will automatically rescale to fit a browser window. On a PC monitor, this allows a life-size representation of the animation, creating a "mirror" effect for the user. Automatic scaling also allows the display to be easily ported to a smaller screen such as that of a PDA or other mobile device (Figure 4).

Simplicity and limited use of text are the guiding principles of the interface design, in conformity to interface design principles for users with aphasia. No text or keyboard input is required to use the system. The circular buttons are labelled with speech utterances—in this case, vowel sounds. When the user clicks on a button, an audio recording of the sound is played in synch with the video or animation. If the user cannot read even this limited amount of text, then the

Figure 2a. Top level of SoundHelper

Figure 2b. A second-level folder, organizing exercises based on vowel sounds

small number of options helps the user to rely on memory to locate the desired sound. The feedback from selecting a sound is immediate—the audio and visual demonstration—and so this high degree of interactivity supports learning and exploring the application. A black circle around the selected button highlights the current sound, and the appropriate phonetic text is displayed on the folder tab in large font (here, "ay") to provide emphatic feedback. The pause button allows user to develop their understanding and practise the different stages of making the sound.

Note that the video and animation only include the lower part of the face. This further reduces the cognitive load for users by eliminating distracting elements such as the expression of the eyes (Bulthoff, Graf, Scholkopt, Simoncelli, & Wichmann, 2004).

A more complex exercise or interface would require additional support for the user. One possibility is to provide audio "mouseover" help—that is, for mouse pointer contact with screen elements to trigger the playing of an audio file containing spoken help or reading out button labels, rather than the conventional display of help text (Deo et al., 2004). Textual mouseover displays are unlikely to be useful for users with aphasia—the text is by default displayed in a small font, and users with aphasia frequently find it difficult to interpret text.

The audio in the prototypes was recorded by a female speaker of a similar age to the median age of the intended evaluators of the prototypes (female speech and language therapists). The video prototype is based on a native English speaker of the same gender, race, and median age as the intended evaluators. The importance of matching audio and visual display to the intended user is discussed in the next section.

The prototypes were evaluated in an empirical study with 20 professional speech and language therapists in the southwest region of England. The evaluators rated both prototypes as very easy to use and as potentially more helpful to their patients than conventional paper-based handouts. Eighteen of the evaluators (90%) could identify current patients who might benefit from using this type of application—but only if it were to be made available on a portable device (17 evaluators; 85%). This study was part of an action research

Figure 3a. Prototype line drawing, pronouncing "ay"

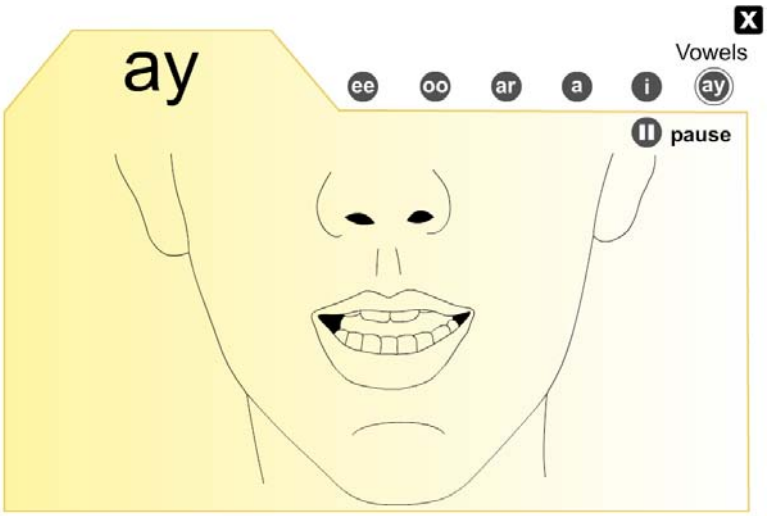
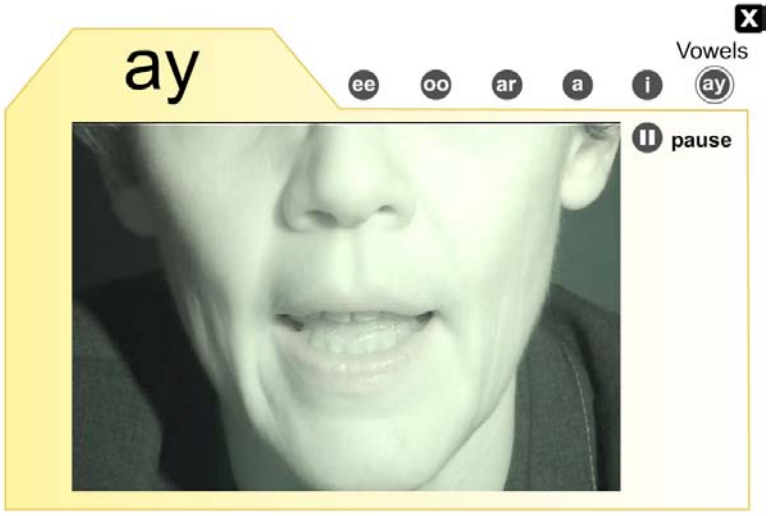
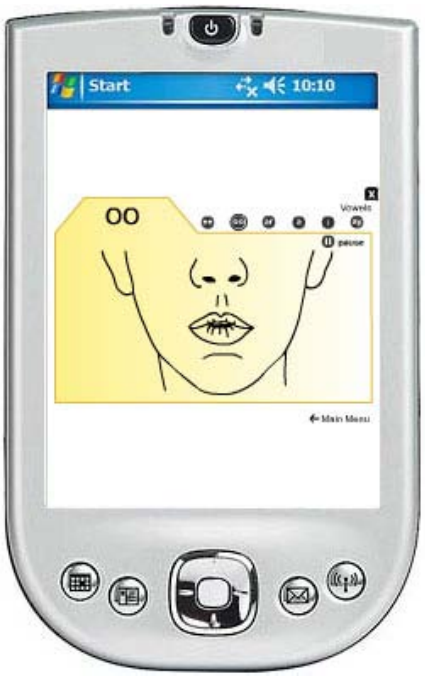


Figure 3b. Prototype video, pronouncing "ay"



project with the next stage being a refinement of the application to include 3D representations of the formation of sounds made inside the mouth and, provided ethical approval is granted, evaluation with a set of case studies is planned.

Figure 4. Sample display as a PDA application



SOCIO-CULTURAL ISSUES IN SPEECH AND LANGUAGE THERAPY APPLICATION DESIGN

Voice characteristics of the speaker featured in speech and language therapy software should ideally match those of the user in significant factors such as age, cultural group, and sex, as it is commonplace for users of communication aids to adopt the intonation and accent of the recorded voice (Harris, 2004). Recording the voice of a family member or friend is therefore usually avoided, as this can cause confusion for the patient.

The user's cultural membership can also have a significant effect on the choice of voice for a speech and language therapy tool of this type—it is important that the user be able to understand and identify with the accent of the audio. The user's nationality or culture can also affect the example sounds and words to include in the application. New Zealand English provides extreme examples of the accommodations in the exercises that must be possible; for instance, for a New Zealand user the phrase "fish and chips" would be a poor choice for practice in the short i sound as the Kiwi accent usually renders that as "fush and chups", and vowel neutralisation causes word pairs such as "full" and "fill" to be indistinguishable when spoken. Alternative pronunciations are generally seen by individuals as incorrect, with their own pronunciation viewed as the only correct way to speak (Maclagan, 2000). Users are unlikely to accept software that uses pronunciation they regard as "not speaking properly".

The user should also be able to identify with the speaker images, or at the very least should not feel excluded or offended by them. Again, this points to the need to tailor this type of speech and language therapy software to groups of users. A system based on realistic videos, such as the prototype in Figure 3b, obviously would require a greater degree of customisation. The line drawings of Figure 3a are suitable for a broader intended target audience—the animation is less identifiable by gender, race, and age (though not entirely generic). The creation of icons and representations of humans that are completely culturally neutral is not possible, but guidelines exist to reduce the level of cultural bias (del Galdo & Nielsen, 1996).

Note that these problems with providing appropriate matches to the user in voice and appearance exist for face-to-face therapy as well. Again, New Zealand provides an extreme case in point: New Zealand speech is so distinctive that it can be difficult for New Zealand born therapists to work in other English-speaking countries, and conversely New Zealand born clients report difficulties in understanding non-New Zealand therapists (Maclagan, 2000). The creation of multimedia therapy tools offers the chance to provide tailored examples and exercises to an

extent rarely possible in conventional therapy, where the number and age/sex/cultural distribution of therapists is rarely large enough to allow a patient to choose a therapist exactly matching his or her own background.

INTERNET-ENABLED THERAPY

The Internet has been suggested for use with speech and language applications in two ways: for delivering therapy and for monitoring therapy. Internet delivery of exercises is highly appealing for the use of the prototypes on mobile devices, as this could address the problem of relatively small memory availability on PDAs and other mobiles: exercises could be streamed in on demand. This solution may be too financially costly for many users, however. Alternatively, selected exercises might be periodically uploaded via a "sync" with a PC or other larger storage device. Remote monitoring possibilities include storing logs or summaries of user sessions in central database, for therapists to later examine (Glykas & Chytas, 2004). This type of monitoring does not provide the capacity for the immediate feedback that is available in a face-to-face session, but does allow the therapist to maintain awareness of the user's progress between sessions. Therapist feedback could be delivered via the Internet, or the monitoring could be used to inform the next face-to-face session.

The potential of Internet-enabled therapy devices for supporting between-session monitoring or care raises several concerns—most notably, that the increased use of technology in therapy could raise barriers between the therapist and patient, and could lead to the dehumanisation of the professional-patient relationship. Issues of legal liability for the efficacy of this more attenuated version of treatment have yet to be fully resolved. There is also a perceived risk that other areas of care could suffer if scarce health funds are diverted to this potentially costly form of telemedicine (including cost of devices, development of multimedia therapy software, tailoring of software to individual patients, Internet transmission costs, therapist time devoted to remote monitoring, and so forth) (Rosenberg, 2004).

A more general anxiety voiced by the therapists evaluating the prototypes of the earlier section, *Development of Multimedia Prototypes*, is that of the ability of therapists to customize therapies as delivered via the Internet, or for that matter through any multimedia application or device. Therapists perceive that therapy as delivered conventionally is tailored to an individual's needs, although it is difficult to see how the current generic paper-based materials are personalized to any great extent. Clearly these new technologies have raised expectations, and any system deployed would have to provide facilities for therapists to inspect, approve, and modify the therapy support being offered.

FURTHER DIRECTIONS FOR RESEARCH AND DEVELOPMENT

The preliminary evaluation with SLTs illustrated that although they were positive about the application, there are clear limitations to the prototypes already developed. In particular, SLTs identified the need for context words and illustrations for each phonetic sound and help with the production of sounds involving movements within the mouth. For example, the plosive sound "t" where the tongue is placed against the front upper palate. There is also a need for an SLT to be able to tailor the learning materials for a particular patient and monitor their progress.

It should be possible to create 3D models of the mouth rather than the vertical sectional drawings currently used. The model could then be rotated by the patient to see different viewpoints. The context words and images could be addressed relatively easily by building up a repository, or digital library, of learning objects which could be chosen by the SLT or added to by them in the

same way that current mobile AAC devices can be tailored for a particular patient. A suitable system for this needs to be developed in a cooperative design project with SLT users which will allow the SLT to create drill and practice exercises, capture video and audio together with a further system to allow a patient to upload video of their progress for the SLT to monitor. The technical problems of suitable codecs to compress video need to be addressed although some of this work has already been done in the context of British Sign Language learning (Andrews, 2005). The advent of full 3G services on mobile devices could be utilised to develop a fully mobile therapy service. However, the ergonomics of the devices used will need careful design as it is widely accepted that current mobile phones are unsuitable for use by elderly users due to the small size of their interaction buttons (Goodman, Dickinson, & Syme, 2004).

There are, therefore, a range of future directions in which research in mobile multimedia applications to enable people with aphasia to practise spoken language skills independently between sessions could progress. Blended approaches to therapy are likely to be positively received by SLTs to enable them to support patients with aphasia more flexibly than at present.

REFERENCES

Andrews, J. (2005). Using SignLab for formative and summative assessment. *The University of Bristol Learning Technology Support Service Fifth Annual National Conference, Bristol, June 20. Retrieved April 24, 2006, from* http://www.ltss.bris.ac.uk/vleconf05/Speakers/andrews.doc

Andrews, K., & Turner-Stokes, L. (2005). *Rehabilitation in the 21st century: Report of three surveys.* London: Royal Hospital for Neurodisability.

Armstrong, L. (1993). Assessing the older communication-impaired person. In J. R. Beech, & L. Harding (Eds.), *Assessment in speech and language therapy* (pp. 163-166). London: Routledge.

Bhogal, S. J., Teasell, R. W., & Speechley, M. R. (2003). Intensity of Aphasia therapy, impact on recovery. *Stroke*, (34), 987-993.

Bulthoff, H. H., Graf, A. B. A., Scholkopt, B., Simoncelli, E. P., & Wichmann, F. A. (2004). Machine learning applied to perception: Decision-images for gender classification. *Advances in Neural Information Processing Systems, 17.* Retrieved April 24, 2006, from http://www.cns.nyu.edu/pub/eero/wichmann04a.pdf

Burton, E., Meeks, N., & Wright, K. (1991). Opportunities for using computers in speech and language therapy: A study of one unit. *British Journal of Disorders in Communication, 26*(2), 207-217.

Deo, S., Nichols, D. M., Cunningham, S. J., Witten, I. H., & Trujillo, M. F. (2004). Digital library access for illiterate users. *Proceedings of the 2004 International Research Conference on Innovations in Information Technology,* Dubai (UAE), October (pp. 506-516). United Arab Emirates: UAE University.

Dunlop, H., Cunningham, S. J., & Jones, M. (2002). A digital library of conversational expressions: Helping profoundly disabled users communicate. *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL), Portland (Oregon, USA), July 14-18 (pp. 273-274). New York: ACM Press.

Foundation for Assistive Technology (FAST). (2004, April). *Reporting on assistive technology in a rapidly changing world* (pp. 11-14). Retrieved April 24, 2006, from http://www.fastuk.org/RAPID.pdf

del Galdo, E. M., & Nielsen, J. (Eds.) (1996). *International user interfaces.* London: John Wiley & Sons.

Glykas, M., & Chytas, P. (2004). Technology assisted speech and language therapy. *International Journal of Medical Informatics, 73,* 529-541.

Goodman, J., Dickinson, A., & Syme, A. (2004). Gathering requirements for mobile devices using focus groups with older people. *Designing a More Inclusive World, Proceedings of the 2nd Cambridge Workshop on Universal Access and Assistive Technology (CWUAAT),* Cambridge, UK, March. Retrieved April 24, 2006, from http://www.computing.dundee.ac.uk/projects/UTOPIA/publications/navigation_workshop.pdf

Harris, C. (2004). Progressing from paper towards technology. *Communication Matters, 18*(2), 33-37.

Maclagan, M. (2000). Where are we going in our language? New Zealand English today. *New Zealand Journal of Speech-Language Therapy, 53-54,* 14-20.

NHS. (2004). *Allied health professionals.* Retrieved April 24, 2006, from http://www.nhscareers.nhs.uk

Queensland Aphasia Groups. (2001a). *Web developer's guidelines.* Retrieved April 24, 2006, from http://www.shrs.uq.edu.au/cdaru/aphasiagroups/Web_Development_Guidelines.html

Queensland Aphasia Groups. (2001b). *What is aphasia-friendly?* Retrieved April 24, 2006, from http://www.shrs.uq.edu.au/cdaru/aphasiagroups/Aphasia_Friendly.html

Rosenberg, R. S. (2004). *The social impact of computers (3rd ed.).* USA: Elsevier Academic Press.

Royal College of Speech and Language Therapists (RCSLT). (2004). *What do speech and language therapists do?* Retrieved January 15, 2006, from http://www.rcslt.org/whatdo.shtml

van de Sandt-Koenderman, M., Wiegers, J., & Hardy, P. (2005, May). A computerised communication aid for people with aphasia. *Disability Rehabilitation, 27*(9), 529-533.

This work was previously published in Mobile Multimedia Communications: Concepts, Applications, and Challenges, edited by G. Karmakar and L. Dooley, pp. 74-84, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.22
A Proposed Tool for Mobile Collaborative Reading

Jason T. Black
Florida A&M University, USA

Lois Wright Hawkes
Florida State University, USA

ABSTRACT

This chapter presents a tool for collaborative e-learning using handheld devices that incorporates pair communication via text and speech input. It discusses the current state of e-learning for mobiles and illustrates the lack of such tools in reading comprehension domains. It then describes the tool development as a model for interface design, communication strategies, and data manipulation across mobile platforms. It is argued that such a tool can enhance e-learning among children, due to freedom of movement and variety of input (text and speech). The design is centered on a proven paper-based collaborative learning methodology which should strengthen its effectiveness. A paper prototype test that assisted in determining optimum interface layout and confirming that speech input was preferred among children is described. The system was developed and designed using creative strategies for interface layout and data manipulation. Lessons learned and plans for additional research are discussed.

INTRODUCTION

Collaboration is an important aspect of today's educational learning environment, and the infusion of technology has given rise to various studies in the area of computer-supported collaborative learning (CSCL), computer-supported collaborative work (CSCW), and computer-supported intentional learning environments (CSILE) (Jones, Dircknick-Holmfield, & Lindstrom, 2005; Scardamelia & Bereiter, 1996). The systems developed through these studies have been effectively implemented to produce major gains in comprehension of material in the math and science curriculum, but have yet to explore these benefits when applied

to domains which are not math and science. The investigation of how to efficiently apply emerging technology in such environments is resulting in innovations in a wide range of systems and platforms, including handheld computers and other mobile devices.

One of the disciplines that could benefit significantly from such advancements is reading comprehension. At the present, it is apparent that reading comprehension has emerged as a major problem area in American society (Vaughn, Klingner, & Bryant, 2001). It is important to note that there are several reading comprehension tools available for the desktop platform, but the problem becomes enormous when attempting to transfer such applications to the handheld platform. There are many obstacles that must be overcome, such as limited screen real estate, smaller memory capacity, smaller processing power, and limited and often more difficult input mechanisms (such as stylus and virtual keyboard). These obstacles have led developers to steer away from this handheld platform and instead focus on the more common personal computer environment. Yet, research is indicating that the handheld computer is becoming a more viable and attractive platform due to the smaller cost, portability, durability, and increasing advancements in wireless technology (Soloway & Norris, 1999). Additionally, many scientists are investigating more innovative ways to utilize this technology and make it much more readily available to children from diverse backgrounds (MIT Media Lab, 2006).

Question-answer relationships (QAR) is a very successful learning methodology for developing reading comprehension skills (Royer & Richards, 2005; Outz, 1998; Raphael, 1986). QAR has been beneficial to educational research in that QAR not only has demonstrated the ability to improve comprehension skills of student participants, but has also shown effective implementation of peer-assisted learning strategies. There has not been a significant effort to place QAR in a computerized reading environment, and it is worth investigating whether applying QAR to a handheld learning environment would produce a more efficient reading comprehension software platform.

Thus, this chapter makes the case for collaborative reading comprehension on a mobile platform by illustrating the absence of current research in this area, describes a paper-prototype study for an interface model for collaborative reading comprehension, and then presents a handheld tool supporting collaborative reading using text and speech communication. The tool is designed using QAR as a foundation, and presents a model for development of such systems on mobile platforms. An emphasis is placed on speech input, which can further increase the robustness of user input and collaboration as a result, particularly when implemented for children.

RELATED WORK

Mobile Collaboration in Learning Environments

The explosion of mobile learning (m-learning) in educational environments is largely due to the massive influx of these portable devices in society, and more directly, in the classroom. Mobile learning takes place when users communicate wirelessly via handheld devices (phones, Personal Digital Assistants (PDAs), tablets, etc.) in the process of learning—in other words, learning that takes place with the aid of handhelds (Attewell, 2005). And, since collaboration is a natural and significant extension of a robust learning environment, it is natural to consider ways to facilitate mobile cooperation in learning activities. The mobile environment is rich with a plethora of communication tools (chat, instant messaging, shared workspaces, e-mail, and voice input/out-

put) that make collaborative work a simple and efficient endeavor (Issacs, 2002). It is essential for researchers to explore a wide range of scenarios employing these tools in an effort to improve student learning outcomes. This research takes a look at one such endeavor.

Question-Answer Relationships

Several programs have been implemented that have shown significant development in reading comprehension skills. Among the most successful is Question-Answer Relationships (Royer & Richards, 2005; McIntosh & Draper, 1996; Raphael, 1986), which has been shown to be a particularly effective supplement to a classroom reading program. Question-Answer Relationship teaches students to read by recognizing relationships between questions and possible sources of information, either in the text or in the reader's background. In this technique, readers are asked to read a passage and answer questions about what was read. Then, readers are required to identify the category to which each question belongs: Right There Questions (answer is explicitly in the text), Think-and-Search Questions (answer is implicitly in the text), the Author and You Questions (the answer requires you to use inference to arrive at the answer), and On Your Own Questions (the answer is entirely based on your background knowledge). Several studies have shown that students were capable of generating and answering questions that enhanced their comprehension and led to independent processing and development of knowledge (Royer & Richards, 2005; Outz, 1998). Yet, these approaches have not been incorporated in a desktop or handheld reading comprehension learning environment. It is worth investigating whether doing so will create an electronic comprehension tool which can reinforce through practice, techniques introduced by a human teacher, and hence address these issues (Vaughn, Klingner & Bryant, 2001).

Speech Recognition in Mobile Environments

This increase in the use of mobile devices has created an environment where various types of users are interacting, and as a result, researchers must utilize the full suite of modalities (or modes of input) to facilitate communication (Nanavati, Rajput, Rudnicky, & Siconni, 2006). Almost all mobile devices are equipped for voice input, making speech recognition a viable means of capturing data. In many cases, to compensate for the limited memory and power on these smaller devices, a form of *distributed speech recognition* is implemented (Schmandt, Lee, Kim, & Ackerman, 2004). In such an environment, speech is captured on the mobile device and sent across a wireless network to a server, where processing is done. The translated text is then returned to the device for use. While there are many issues to consider when utilizing this strategy (such as quality of speech, network traffic, noise, etc.), the scope of this work is to present an interface mechanism for facilitating voice input in a mobile collaborative learning session.

OVERVIEW OF SYSTEM DEVELOPMENT

Discussion of Paper Prototype Testing

Paper (also called "low-fidelity" or "lo-fi") prototyping is an interface development strategy that utilizes paper-based designs of the system and interactions with potential users with such system to arrive at an optimum design plan (Snyder, 2003). In paper prototype testing, users are asked to interact with the paper-based interface on a series of popular system tasks, with a designer playing the role of "Computer." The "Computer" mimics the actions and sounds of the system while

the user progresses through these tasks, and user choices and behavior are recorded. After the session, the user is questioned in order to learn his or her cognitive processes in making decisions and the results of these answers are used to design and implement the user interface, complete with modifications indicated in the test. The attempt here is to create a "living" prototype—one that is changing to better fit the designs and recommendations of the testers involved—in order to eventually obtain the optimum design methodology for all involved. Researchers have demonstrated the benefits seen in the application of paper prototype testing—preemptive user feedback (changes are suggested before development has begun), rapid iterative development (changes can be incorporated "on the fly"), and enhanced developer/user communication (Snyder, 2003).

Participants in Paper Prototype Test

To obtain a model for interface development, a paper prototype study was conducted (Black, Hawkes, Jean-Pierre, Johnson, & Lee, 2004) involving elementary school students from a local after-school center. Five students were selected based on their background with computers (two had had experience with handhelds, one had moderate experience, and two had no experience), and age (two were in grade 2, two were in grade 3, and one in grade 4). This number of subjects is consistent with Snyder's recommendations of effective numbers of subjects in such tests, which is recommended to be between five and seven (Snyder, 2003; Nielson and Landauer, 1993). Subject #1 was a fourth grader who was a good and constant reader, and was the only subject that had familiarity with a PDA, though not much exposure. Subject #2 was a third grader who had some experience with computers, but was not a strong reader. Subject #3 was the youngest of the group, a second grader with very little computer experience, but was a strong reader. Subject #4 was a third grader who had very little computer experience and was also not a very strong reader. Subject #5 was the oldest of the group, a fourth grader who had computer experience, but was having trouble reading at grade level. All of the students had some exposure to the basic features of a computer application—buttons, passwords, pointers, and so forth—and were eager to participate in the study.

Apparatus

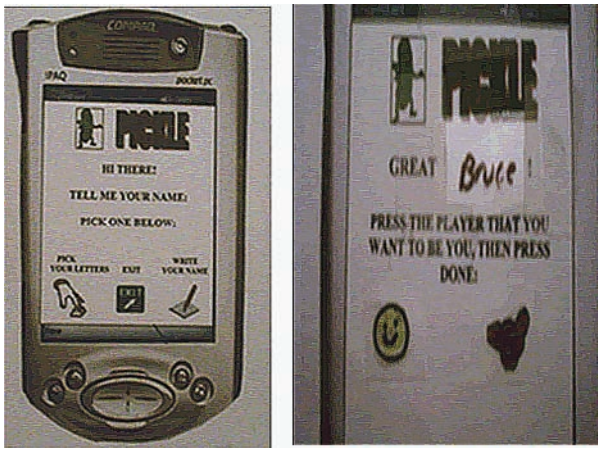
The test was designed using the iPAQ ™ PDA as a model (see Figure 1). A picture of the device was taken, and then scanned and printed, so that the actual size and shape of the PDA could be used in the testing. Then, cut-outs of screens to be presented (as well as buttons, menus, scroll bars, etc.) were designed and used as interchangeable interface components to be presented to subjects during completion of tasks.

Test Design

Screen mock-ups of five basic tasks were created: 1) logging into the system and selecting a partner (for collaboration), 2) reading a story, 3) answering questions, 4) e-mailing the teacher for help, and 5) chatting with their partner. Researchers participated in the test in the roles of the computer (one person transitioned screens as the computer would), and observers (who took notes on user actions and tendencies). Subjects were tested in 30 minute sessions, with two tests conducted on a given day, as recommended by Snyder (2003). During a test session, each subject was introduced to the concept of the test and why it was being conducted. They were then seated at a desk with the "computer" present as well, and the observers looking on.

Each subject was then asked to complete each of the tasks listed, with the question-answer exchanges following each task. After the completion of all five tasks, the subjects were thanked for

Figure 1. Image of mock-up of screen used in Paper Prototype test



their participation and were free to leave. Notes from observers were discussed between tests, and modifications were proposed for future tests.

Test Results

During testing, three of the five subjects preferred writing on the screen for input as opposed to the other presented forms (keypad or speech). This was not surprising, since most of these students had little computer experience, and thus would be more comfortable writing (at least initially) than using the innovative input techniques presented. Of the two subjects that did choose to pick their letters using the keypad, one subject (Subject #4) had trouble navigating the keypad and began pressing buttons on the bottom of the PDA instead of the buttons on the keypad. This action resulted in "beeps" from the "computer" indicating actions that were not allowed by the system, which further confused the subject. But what was gathered from this subject was that he was familiar with the GameBoy™ handheld computer games, which use the directional keypad on the bottom of the device for manipulation of all applications. Research has shown that students' experiences with such gaming devices can be very productive design focuses for scientists developing applications for handhelds. All subjects except Subject #1 had trouble finding the icon during the player selection phase. All subjects had no trouble finding and clicking the "Done" button.

All subjects reported no problem in reading text on the small, handheld-mock up screens, and indicated that the process was enjoyable. Subjects also had no problem transitioning between screens (done by clicking NEXT and BACK buttons presented on the interface). The chat/messaging task perhaps provided the most valuable feedback. All except one subject chose to speak their message instead of the other input features (Subject #4 chose to use the keypad in all writing tasks), indicating that this will likely be a popular feature of the application. The voice input would be very helpful to younger readers, who often do not have the ability to type or write

well, and would prefer an alternative to user input. However, two of the subjects were confused when faced with the submenu that appears with the speak feature, which asks them to click on the microphone to begin speaking and to click send to transmit the message. The two subjects were unsure what buttons to press and when to press send. But once this was explained, the subjects were able to complete the task. The reliance on speech for inputting validated an earlier hypothesis which suggested that due to the age of users and simplicity of the action, speech would be chosen more often by younger users (Black, et. al, 2004). The results of this test were then used as blueprints for actual screen development.

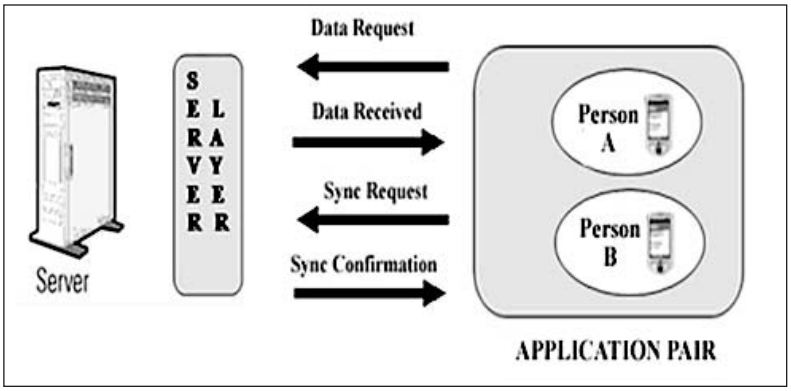
System Overview

The interface layout was developed using an application development toolkit that allows for rapid prototyping of applications for mobile devices and provides a series of emulators that can present a simulation of the application running in or on its intended device platform. Upon completion, the finished application can easily be ported to the actual target device, and run as needed. The prototype uses a client-server approach for wireless communication (SEIR-TEC, 2002), which implies that devices are served by a central access point or base station, and communicate with the central access point through the network (see Figure 2). In this configuration, users will send computation-intensive operations to the more powerful server (such as the speech recognition) for remote processing and then download application-specific tasks to the handhelds (Omojokun, 2002). And, when usage has completed, the handhelds synchronize with the server, uploading user session information back to the server for storage in the database.

System Design

The process flow of the application is constructed on the question-answer relationship collaborative reading model mentioned earlier. In implementing this model, students read a passage on the handheld display and then are asked to answer questions which are downloaded from the server about the text just read. Once questions are answered correctly, students must then identify the type of questions that are presented (Right There, Think and Search, Author and You, or On Your Own), based on Raphael's Taxonomy of Questions

Figure 2. High-level system architecture



(Raphael, 1982). Students work individually on reading and question-answering tasks, but are allowed to collaborate during periods of reaching a consensus on the correct answer (collaboration is "turned on" when questions are answered incorrectly). Students communicate by chatting with their partner as needed, as well as utilizing a shared workspace for group reflection. Students can also record personal notes in the personal journal as they progress through the lessons, as well as interact with the instructor through e-mail. All of the activities will be done on the PDA, with lessons downloaded wirelessly from the server to the PDA as requested, and student progression data being stored on the system server.

The application makes use of the standard high-level interface components—forms for user fill-in, canvases for drawing and painting of both text and images, and checkboxes and radio buttons that register user action. The screen is also touch-sensitive, allowing for stylus input at various points on the interface. Textual input is

Figure 3. Introductory screens for user input

Figure 4. Story and question-answering screens

handled via stylus, keypad, and both the letter recognizer and transcriber (written text using the stylus), and is processed by device standards. The letter recognizer allows users to, after a training period, write letters using the stylus, which are "recognized" by the system and translated to their typed form. The transcriber works similarly, with users writing letters on a text pad, these letters are converted to their typed format. Both have significant learning curves on their usage, but once mastered can serve as very convenient input techniques. Figure 3 shows the initial user login screen (which is essentially a form with text fields and images that behave as buttons), a screen for selecting an icon (which uses images painted on a canvas), and a screen for registering that the user is ready to begin (again, images painted on a canvas).

These screens also make use of the standard mobile menus, which appear at the bottom of the screen just above the device soft buttons. Users can activate these menus either by clicking on them with the stylus or pressing the corresponding soft button.

The story and question screens are similarly done, with images and radio buttons dominating the device display. The user is also presented with icons at the bottom that allow him or her to activate the various system functions (using the *Diary* for personal reflection; sending a message to the instructor, using the *Group Workbook*, etc.). These are illustrated in Figure 4.

Collaboration Components

The collaborative features implemented in the prototype system are shared workspaces, e-mail messaging, personal reflection, and chat services. The shared workspace is implemented as a *group workbook* that is visible to both participants of a team. Each user sees an up-to-the-second image of the workbook and changes made (by entering data in the workbook) are broadcast to each user's device. The system synchronizes access to this feature, locking it while it is being written to so that changes can be implemented before additional writes are allowed. The workbook enables each user to jot down notes that may assist the team in answering questions in later sessions. These notes can be entered either by keypad or writing (textual) or by speaking the text (voice) (see Figure 5).

Figure 5. Workbook and diary screens

The personal journal is used for reflection as the user moves through sessions. Each journal is seen only by that user and is updated upon request. All additional entries into the journal are added to its previous contents, similar to writing in a paper journal. The entries are recorded by being sent to the server for storage upon completion of the session period. Upon the next login, the current contents of the journal are sent to the client should any new entries be desired. As with all other methods of input, the user can provide the journal entry either through textual or voice input (see Figure 5).

The system allows users to send e-mail to the instructor in the event that assistance is needed. The sending of the message is implemented by the client sending a message to the server and the server forwarding that message to the instructor's e-mail address. A record of the transmittal is also stored by the server for reporting purposes. Again, the message can be either in the form of text or voice input (see Figure 6).

Chat Service

The chat service is implemented similar to the standard chat service hosted by any Internet service provider. The system registers that a user wishes to chat and sends a message to the user's partner that he or she wishes to chat. Once a confirmation has been received by the partner, the chat session begins, with users typing in messages (or entering them via voice input) and those messages being displayed on the screen. These messages are also recorded by the server for reporting. When chat is no longer desired, the user indicates this and the session resumes from its previous point. The chat provides the users the opportunity to reach a consensus on certain learning tasks and facilities the peer-tutoring methodology, both techniques present in successful collaborative environments (see Figure 6).

Speech Recognition Strategy

Since the system is designed to be adaptable to a variety of environments, it may be the case that the keypad, recognizer, and transcriber are too complex in a setting of younger users. Thus, the system also allows for Speech Input, where users are allowed to speak their messages into the system, and these messages are converted to text and displayed on the screen (or sent via e-mail if needed). This is done to take into account that

Figure 6. Chat and help screens

younger users may not be good typists or even know how to spell well, but may still wish to enter data.

The ability to provide speech input for users is a major component of the architecture. The current literature does not indicate any use of this feature in studies involving handhelds and collaboration. In the math and science-centered applications, input is often simple, with users asked to select items or to enter numbers as part of equations. This poses a problem in non-science domains, where input may often be sentence-structured and much more verbose. There needs to be an additional method of providing this type of input, and speech or voice input fills this need adequately.

The application environment implements a strategy for dealing with speech input and/or speech output. The system receives the data sent to the server and runs it through a speech recognizer program to produce written text. The written text is then sent back to the application to be displayed on the screen. The server is responsible for handling speech requests sent by the Diary, Help, Journal, Workbook, and Chat applications and funneling them to the appropriate mechanism.

DISCUSSION

It is important to note that while this work demonstrates that this type of interface can indeed be developed for mobile devices, the actual testing of this system in a live classroom will occur in additional studies. The researchers are currently working with school teachers to develop a curriculum model that can incorporate such a system, in an effort to determine if its application would be effective in improving reading skills of younger students. This is a daunting task, but one that holds much promise for both computer scientists and educators as well.

Prior to implementation of a complete system, there are issues to be examined related to data management, data modeling, logging of system and user operations and functions, and server-side management. There are also issues regarding data security and reliability of data to consider. And, while the screen size is likely appropriate for beginning readers and younger users, there may be an issue regarding displaying of material on the device for more advanced users. The current model only displays one page of data at a time, and pages are turned and not scrolled. More advanced users would likely want to remain on a page and simply scroll down or up to view additional material. This would call for some device other than the PDA, or at the very least, in a revision of the type of screen layout. However, the focus of this project is younger children, and thus the designed system is very appropriate.

CONCLUSION AND FUTURE WORK

The development of this collaborative mobile learning system and its implementation on the actual target handheld devices is an indication that this type of architecture is possible and proved both challenging and rewarding. Each of the desired screen designs was capable of being constructed in the chosen language and the interaction between screens was simple to maintain. The communication between devices and between device and server was easily maintained via a wireless network and access to a server machine. Using a Web server allows for testing and demonstration of the system in any environment where wireless Web access is available.

The implementation of QAR in a collaborative platform was also successful. As mentioned earlier, QAR requires individuals to work in pairs, which is accommodated by the interface in this system. QAR also expects students to not only read, but also answer questions and then identify categories of questions. The multiple screens developed in this system also accomplish this task. And, since QAR is an extremely successful

tool (in a paper-based environment) in enhancing reading comprehension skills, it is rewarding to note that the interface presented does not take away from the functionality and robustness of the methodology, but stresses it very well.

Speech input is a very significant feature of any collaborative environment and the tools included in the system provide for that capability. Students are able to speak words of communication with partners, and these words are indeed translated and presented on the screen. This is a major component since mobile devices often have challenging input techniques (using a stylus for large volumes of text can be very cumbersome). This system addresses and solves this issue as well.

The next step is to implement this system in the actual classroom, with the assistance of grade school teachers and administrators, in the effort to study its effect on reading comprehension skills. It is believed that students using this system will become better readers, and that the system's integration into the classroom learning setting will be unobtrusive and seamless. Since most reading comprehension software is developed for the desktop environment, utilization of such a system in this mobile platform could prove very exciting and rewarding, serving to fill a much needed void in the collaborative learning spectrum.

Overall, the system presented in this work provides one possible approach to developing collaborative learning environments on intermittent devices, successfully providing an architecture for modeling interfaces for smaller, more limited machines. This research is just scratching the surface of what is capable for reading comprehension software, showing that tomorrow is promising for addressing the crisis of improving children's reading skills nationally.

REFERENCES

Attewell, J. (2005). Mobile learning: Reaching hard-to-reach learners and bridging the digital device. In G. Chiazzese, M. Allegra, A. Chifari, & S. Ottaviano (Eds.), *Methods and technologies for leaning.* (pp. 361-365). Southampton: WIT Press.

Black, J., Hawkes, L., Jean-Pierre, K., Johnson, I. & Lee, M. (2004, September 13-16). A paper prototype study of the interface for a children's collaborative handheld learning application. In *Proceedings of Mobile HCI 2004,* Glasgow, Scotland.

Isaacs, E., Walendowski, A., & Ranganathan, D. (2002). Mobile instant messaging through Hubbub. *Communications of the ACM, 45* (9), 68-72.

Jones, C., Dirckinck-Holmfeld, L., & Lindström, B. (2005). CSCL The next ten years—A view from Europe. In T. Koschmann, D. Suthers, & T-W. Chan (Eds.), *Computer supported collaborative learning 2005: The next ten years!* Mahwah, NJ: Lawrence Erlbaum Associates.

McIntosh, M.E., & Draper, R. J. (1996). Using the question-answer relationship strategy to improve students reading of mathematics texts. *The Clearing House, 154*, 161.

MIT (2006). *MIT Media Laboratory.* Retrieved from http:// www.media.mit.edu/about/overview.pdf.

Nanavati, A., Rajput, N., Rudnicky, A. I., & Sicconi, R. (2006, September). Workshops: SiMPE: speech in mobile and pervasive environments. In *Proceedings of Mobile HCI 2006.* Helsinki, Finland.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of Conference on Human Factors in Computing Systems INTERCHI '93,* Amsterdam (pp. 206-213). New York: ACM Press.

Ouzts, D. (1998). Enhancing the connection between literature and the social studies using the question-answer relationship. *Social Studies & The Young Learner, 10*(4), 26-28.

Raphael, T. (1982). Question-answering strategies for children. *The Reading Teacher, 36,* 186- 19.

Raphael, T. (1986). Teaching question answer relationships, revisited. *The Reading Teacher, 39,* 516-522.

Royer, R., & Richards, P. (2005). Revisiting the treasure hunt format to improve reading comprehension. Retrieved from http://www.iste.org/Content/NavigationMenu/Research/NECC_Research_Paper_Archives/NECC_2005/Royer-Regina-NECC05.pdf

Scardamalia, M., & Bereiter, C. (1996). Student communities for the advancement of knowledge. *Communications of the ACM, 39*(4), 36-37.

Schmandt, C., Lee, K. H., Kim, J., & Ackerman, M. (2004, June). Impromptu: Managing networked audio applications for mobile users. In *Proceedings of MobiSys 2004* Boston, Massachusetts.

Snyder, C. (2003). *Paper prototyping: The fast and easy way to design and refine user interfaces.* San Francisco: Morgan Kaufmann Publishers.

Soloway, E., & Norris, C. A. (1999). *Schools don't want technology, schools want curriculum.* Retrieved in 2003 from http://www.cisp.org/imp/june_99/06_99soloway-insight.htm

SouthEast Initiatives Regional Technology in Education Consortium (SEIR-TEC) News-Wire. (2002). *Using handheld technologies in schools, 5*(2).

Trifonova, A., & Ronchetti, M. (2005). Prepare for a bilingualism exam with a PDA in your hands. In G. Chiazzese, M. Allegra, A. Chifari, & S. Ottaviano (Eds.), *Methods and technologies for learning* (pp. 343-347). Southampton: WIT Press.

Vaughn, S., Klingner, J. K., & Bryant, D. P. (2001). Collaborative strategic reading as a means to enhance peer-mediated instruction for reading comprehension and content area learning, *Remedial and Special Education, 22*(2), 66-74.

KEY TERMS

Collaborative Learning: An environment where students work alone or in groups to complete a set of tasks, usually lessons, where they assist each other in learning.

Computer-Supported Collaborative Learning (CSCL): The study of users collaborating in a computerized environment on learning tasks.

Computer-Supported Intentional Learning Environments (CSILE): Database software that provides tools for organizing and storing knowledge as a means of sharing information and thoughts with peers, supporting both individual and collaborative learning.

Computer-Supported Collaborative Work (CSCW): The study of how people work with computers and how they can work with each other using them.

Distributed Speech Recognition: The process of capturing speech on a mobile device and transporting it via wireless network to a server to be processed ("recognized") or translated, and subsequently returning the translated speech to the mobile device.

Mobile Learning (M-Learning): Users communicating wirelessly via handheld devices in the process of learning.

Question-Answer Relationships (QAR): An instructional methodology for enhancing reading comprehension skills by teaching students to answer questions and generate their own questions based on text.

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 1068-1078, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.23
Mobile Decision Support for Time-Critical Decision Making

F. Burstein
Monash University, Australia

J. Cowie
University of Stirling, UK

INTRODUCTION

The wide availability of advanced information and communication technology has made it possible for users to expect a much wider access to decision support. Since the context of decision making is not necessarily restricted to the office desktop, decision support facilities have to be provided through access to technology anywhere, anytime, and through a variety of mediums. The spread of e-services and wireless devices has increased accessibility to data, and in turn, influenced the way in which users make decisions while on the move, especially in time-critical situations. For example, on site decision support for fire weather forecasting during bushfires can include real-time evaluation of quality of local fire weather forecast in terms of accuracy and reliability. Such decision support can include simulated scenarios indicating the probability of fire spreading over nearby areas that rely on data collected locally at the scene and broader data from the regional and national offices. Decision Support Systems (DSS) available on mobile devices, which triage nurses can rely on for immediate, expert advice based on available information, can minimise delay in actions and errors in triage at emergency departments (Cowie & Godley, 2006).

Time-critical decision making problems require context-dependent metrics for representing expected cost of delaying an action (Greenwald & Dean, 1995), expected value of revealed information, expected value of displayed information (Horvitz, 1995) or expected quality of service (Krishnaswamy, Loke, & Zaslavsky, 2002). Predicting utility or value of information or services is aimed at efficient use of limited decision making time or processing time and limited resources to allow the system to respond to the time-critical situation within the required time frame. Sensitivity analysis (SA) pertains to analysis of changes in output due to changes in inputs (Churilov et al.,

1996). In the context of decision support, traditionally SA includes the analysis of changes in output when some aspect of one or more of the decision model's attributes change, and how these affect the final DSS recommendations (Triantaphyllou & Sanchez, 1997). In time-critical decision making monitoring, the relationship between the changes in the current input data and how these changes will impact on the expected decision outcome can be an important feature of the decision support (Hodgkin, San Pedro, & Burstein, 2004; San Pedro, Burstein, Zaslavsky, & Hodgkin, 2004). Thus, in a time-critical decision making environment, the decision maker requires information pertaining to both the robustness of the current model and ranking of feasible alternatives, and how sensitivity this information is to time; for example, whether in 2, 5, or 10 minutes, a different ranking of proposed solutions may be more relevant. The use of graphical displays to relay the sensitivity of a decision to changes in parameters and the model's sensitivity to time has been shown to be a useful way of inviting the decision maker to fully investigate their decision model and evaluate the risk associated with making a decision now (whilst connectivity is possible), rather than at a later point in time (when perhaps a connection has been lost) (Cowie & Burstein, 2006).

In this article, we present an overview of the available approaches to mobile decision support and specifically highlight the advantages such systems bring to the user in time-critical decision situations. We also identify the challenges that the developers of such systems have to face and resolve to ensure efficient decision support under uncertainty is provided.

MOBILE DECISION SUPPORT

Recent work on mobile decision support focuses on the implementation of knowledge-based services on hand-held computers. Work on mobile clinical support systems, for example, addresses different forms of intelligent decision support such as knowledge delivery on demand, medication consultant, therapy reminder (Spreckelsen et al., 2000), preliminary clinical assessment for classifying treatment categories (Michalowski, Rubin, Slowinski, & Wilk, 2003; San Pedro, Burstein, Cao, Churilov, Zaslavsky, & Wassertheil, 2004), and providing alerts of potential drugs interactions and active linking to relevant medical conditions (Chan, 2000). These systems also address mobility by providing intelligent assistance on demand, at the patient's bedside or on-site.

Research on location-based mobile support systems uses search, matching, and retrieval algorithms to identify resources that are in proximity to the location of the mobile users and that satisfy multi-attribute preferences of the users and the e-service providers. Examples of such location-based systems are those that recommend best dining options to mobile users (Tewari et al., 2001), locate automatic teller machines nearby (Roto, 2003), and locate nearest speed cameras and intersections using GPS-enabled mobile devices. Most of these mobile decision support systems use intelligent technologies and soft computing methodologies (e.g., rule-based reasoning, rough sets theory, fuzzy sets theory, multi-attribute utility theory) as background frameworks for intelligent decision support. However, none of these systems address the issue of quality of data or quality of decision support while connected or disconnected from the network or consider a comprehensive measure of reliability of data as part of supporting time-critical decision-making.

It should be noted that not all real-time decision situations, which could benefit from mobile DSS, are also constringed by the period of time, in which this decision support should be provided. For example, if a decision problem is more of a strategic, rather than operation nature, the time factor could be less critical, hence, more time can be devoted to improve the quality of data before the final decision has to be accepted by the user. In this article, we mainly address the needs of

operational decision making, when time before the final choice is made is limited. In such situations an aggregate measure of quality of data (QoD) is particularly important when aiming to enhance a level of user's confidence and trust.

In recent papers (Hodgkin et al., 2004; San Pedro, Burstein, & Sharp, 2003) the issue of QoD has been addressed. A framework has been proposed for assessing QoD as an indicator of the impact of mobility in decision-making (Burstein, Cowie, Zaslavsky, & San Pedro, 2007; Cowie & Burstein, 2006; Hodgkin et al., 2004; San Pedro et al., 2003). QoD is based on multiple parameters which measure user-specific factors, current technology-related factors, and some factors which can be learned based on past experiences with similar problem situations. By providing a QoD alerting service from the mobile device, the mobile decision maker can be warned against making decisions when QoD falls below a predetermined threshold or when QoD becomes critically low. The assumption made is that a decision maker should feel more confident with a decision when QoD is high, or be alerted when QoD becomes lower than acceptable.

In mobile DSS, QoD can be calculated incrementally at every stage of the decision making process, as the mechanism for alerting the user when more data and/or resources are needed before the best option can be selected. For example, following Simon's classical decision making principal phases (Simon, 1960), when describing a decision situation, QoD can be used to judge how accurate the set of data collected is at the Intelligence phase; when designing alternative actions, QoD can assist in making sure the user is satisfied with the range of possibilities the user is presented with for a choice; when a model is applied for selecting the best alternative, the final output includes a full and explicit representation of the QoD, which is derived as an aggregate of the ones used in the previous stages (Burstein et al., 2007).

Providing the user with a measure of QoD is extremely important in a mobile decision making environment as the focus of decision support moves from strategic long term decision analysis to just-in-time operational decision support (Hodgkin et al., 2004; Malah, 2000). In addition, in a mobile environment, data timeliness, completeness, reliability, and relevance have to be considered as contributing factors of QoD metrics. For example, in the area of contingency management a "good enough" feasible decision achievable "on the spot," anytime, anywhere is often preferable to a perfect solution that may require extensive additional computational resources as well as time.

Further important issues to consider in time-critical DSS are diversity and heterogeneity of data both at the input and output points of the system. For example, in a medical emergency, context DSS needs to meet the varying information and resource needs of the personnel at the site and yet be able to support their physical mobility requirements. The mobile computing approaches seem to provide an ideal environment to reconcile varying data sources while identifying the best form of delivery of time-critical information. For example, aiding ambulance staff involving in the transfer of patients to the most appropriate hospital in the minimal amount of time (Burstein, Zaslavsky, & Arora, 2005; Cowie & Godley, 2006).

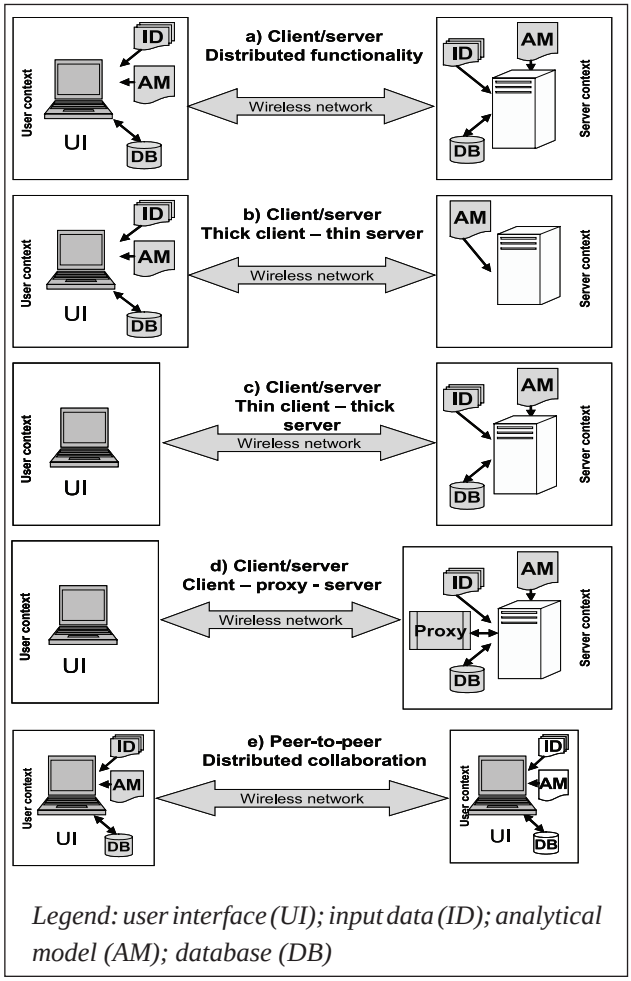
APPROACHES TO AND REQUIREMENTS OF TIME-CRITICAL DECISION SUPPORT

Mobile decision support systems can be implemented in a number of ways, depending on user requirements, available technological resources, frequency of data access, urgency of data retrieval, and so forth. Most of these mobile support systems use intelligent technologies and soft computing

methodologies, for example, multicriteria decision analysis, rule-based reasoning, rough sets theory (Michalowski et al., 2003; San Pedro et al., 2003), fuzzy sets theory (Zadeh, 1994), and multi-attribute utility theory (Keeney, 1992) as background frameworks for being able to learn about the environment in which the decision is taking place and decide upon appropriate support. Location-based decision support is one of many context-aware applications in which systems "can discover or take advantage of contextual information (such as user location, time of day, nearby people and devices, and user activity" (Chen & Kotz, 2000, p. 1).

Time-critical decision making problems require establishing system architectures for such systems that allow infrastructure for fault handling and system recovery (Saksena, da Silva, & Agrawala, 1994), for supporting wireless connectivity (Ahlund & Zaslavsky, 2002),

Figure 1. Mobile decision support architectures (Burstein et al., 2008)



Legend: user interface (UI); input data (ID); analytical model (AM); database (DB)

and provision of network security (Ghosh, 1998; Reis, Hoye, Gilbert, & Ryumae, 2003). Such infrastructures are essential for developing systems that can handle uncertainties due to unreliable communications, possible disconnections from the network, and other technical difficulties that might delay the action or response to a time-critical situation.

As research into such technology is relatively new, optimal architectures for various decision contexts, design configurations, and potential future applications are yet to be investigated. Some generic architectures are proposed and described, for example, by Burstein et al. (2007). They consider how standard DSS component, that is, database (DB), user interface (UI), and analytical model (AM) (Aronson, Liang, & Turban, 2005) can be arranged in mobile decision support architecture. Burstein et al. (2007) describe five possible types of mobile DSS implementation architectures as illustrated in Figure 1.

Portable devices can act as computational platforms for task specific applications, collecting, storing, and providing data for analytical processing. The use of device specific resources or server resources creates a distinction between possible types of DSS that can be provided (Navarro, Schulter, Koch, Assuncao, & Westphall, 2006). Mobile decision support can be client-based, server-oriented, proxy-based, or distributed across an ad hoc network of similar peer devices (Bukhres, Pitoura, & Zaslavsky, 2003). The type of architecture depends on where information is stored and where computations are performed. These varying implementations have associated advantages and disadvantages. The developers can choose between different implementations depending on the user context and technical infrastructure available at the time. A context which requires access to significant amounts of information would be more likely to use a server architecture given the limited processing and information storage capabilities of small portable devices. On the other hand, decision support in situations where some preliminary information processing could be performed based on some aggregated information can be deployed directly onto a portable device. Mobile and distributed decision support improves a system's fault tolerance and reduced support requirements.

Currently, the most popular implementation is where the functionality is distributed across a client-server environment with a user interface (UI) located on the user's portable device. In this case the data is distributed across both client and server; while user-sensitive data resides on the user device, the bulk of the data, including historical databases, are located on a server (Burstein et al., 2008). In this configuration, the Analytical Model (AM) is also distributed across client and server, with the user device performing elementary computations and delegating more complex and resource-intensive computations to the server.

Thick client-thin server and vice-versa represent more extreme cases and therefore more rare configurations. Given the high likelihood of disconnections in the wireless environment, some systems may use the concept of proxy architecture where a proxy process is located on a server-side representing a client and, if connectivity is good, allowing data and computations to be simply channelled between server and client. However, if a client becomes disconnected (e.g., driving through a tunnel), then the proxy assumes full functionality of the client and caches data and results until the client reconnects. With proliferation of peer-to-peer computing, it is now becoming possible to form ad hoc networks of similar devices, discovered at a time of need, in order to consolidate resources and perform the AM in a distributed computing environment in a cost-efficient manner.

These alternative architectural approaches and set up options enable enough flexibility and scalability for any possible DSS application or scenario. For time-critical DSS, any of these configurations should provide adequate assistance as long as the decision can be reached with a reason-

able level of confidence within the required time constraints. Moreover, it is essential to utilize the most current data available while also providing additional information on sensitivity of the selected option to minor variations in context data or more dynamic model characteristics (Cowie & Burstein, 2007)

FUTURE TRENDS

The last decade has seen significant advances in the way humans interact with technology. Users of computers are no longer constrained to the office desktop, but can have much more flexible access to technology almost anywhere, anytime. The spread of e-services and wireless devices has facilitated a new way of using computers, increased accessibility to data, and, in turn, influenced the way in which users make decisions while on the move.

Availability of such an infrastructure coupled with an increase in the demands of users provide new opportunities for developing improved, "just in time" support for decision making. Mobile decision support can be incorporated as an integral component of a mobile commerce infrastructure as a means of an enhanced communication (Carlsson et al., 2006) or as a part of an operational decision support environment for a mobile manager.

Mobile DSS can benefit from storing real-time data and then re-using it at the time of calculating the requested best option (Michalowski et al., 2003). Such systems have been proven to perform well in both stand alone and distributed environments, where they can share historical information to deal with a lack of information when supporting time-critical decisions. Shim, Warkentin, Courtney, Power, Sharda, and Carlsson (2002) suggest that the availability of Web-based and mobile tools within a wireless communication infrastructure will be a driving force in further development of decision support, facilitating decision support for decision makers "wherever they may be" (p. 112).

Financial decisions present a good example of when real-time decision support could be beneficial (Hartmann & Bretzke 1999). Decision makers, who require getting a real-time update on their financial position in order to make the best use of their money, can be supported by mobile DSS, which will utilise new as well as transaction history data in calculating the options (Burstein et al., 2008).

CONCLUSION

The realities of the changing way in which we make decisions and the advances in mobile technology create multiple challenges and opportunities for decision makers and computer system developers alike. Making decisions on the move under uncertainty requires decision support systems that can adequately provide up-to-date, context specific, complete information in a way that is understandable and useful to the decision maker.

Observing the development of DSS for the future, Shim et al. (2002) envisage that use of mobile devises for providing decision support will lead to greater collaboration and allow the achievement of true ubiquitous access to information in a timely manner. We wholeheartedly concur with this opinion and look forward to developers of DSS embracing current and future technologies that facilitate mobile decision support, building on well-founded methodologies to model decision situations with better precision. In addition, we believe that by accommodating a measure of the quality of the information provided, decisions on the move can be supported just as effectively as those made behind the desk.

REFERENCES

Ahlund, C., & Zaslavsky, A. (2002, October 7-8). Support for wireless connectivity in hot spot areas. In *Proceedings of the International Conference on Decision Support Systems, the First MOST International Conference,* Warsaw-Poland (pp. 152-164). 8-10 June, Atlanta, GA.

Aronson, J., Liang, T., & Turban, E. (2005). *Decision support systems and intelligent systems.* Upper Saddle River, NJ: Pearson Education, Inc.

Bukhres, O., Pitoura, E., & Zaslavsky, A. (2003). Mobile computing. In J. Blazewicz, W. Kubiak, T. Morzy, & M. Rusinkiewicz (Eds.), *Handbook on data management in information systems* (Vol. 3). Springer Verlag.

Burstein, F., Cowie, J., Zaslavsky, A., & San Pedro, J. (2008). Support for real-time decision-making in mobile financial applications. In Burstein & Holsapple (Eds.), *Handbook for decision support systems* (Vol. 2). Springer Verlag.

Burstein, F., Zaslavsky, A., & Arora, N. (2005, July). Context-aware mobile agents for decision-making support in healthcare emergency applications. In *Proceedings of the Workshop on Context Modeling and Decision Support, at the Fifth International Conference on Modelling and Using Context, CONTEXT'05,* Paris, France (pp. 1-16). Retrieved December 14, 2007, from http://ceur-ws.org/Vol-144

Burstein, F., San Pedro, J. C., Zaslavsky, A., Hodgkin, J.(2004, July 12-13). Pay by cash, credit or EFTPOS? Supporting the user with mobile accounts manager. In *Proceedings of the Third International Conference on Mobile Business, m>Business 2004,* (pp. 1-13). Institute of Technology and Enterprise, Polytechnic University, New York.

Carlsson, C., Carlsson, J., & Walden, P. (2006). Mobile travel and tourism services on the Finnish market. In *Proceedings of the 24th Euro CHRIE Congress.*

Chan, A. (2000) WWW+ smart card: Towards a mobile health care management system. *International Journal of Medical Informatics, 57,* 127-137.

Chen, G., & Kotz, D. (2000). *A survey of context-aware mobile computing research* (Tech. Rep. TR2000-381), Darmouth Computer Science. Retrieved December 14, 2007, from http://citeseer.nj.nec.com/chen00survey.html

Churilov L., Sniedovich M., &Byrne A. (1996) On the concept of Hyper Sensitivity Analysis in Decision Making. In the *Proceedings of the First Asian-Pacific Decision Science Institute Conference,* (pp.1157-1160). Hong Kong University of Science and Technology, Hong Kong.

Cowie, J., & Burstein, F. (2007). Quality of data model for supporting mobile decision making. *Decision Support Systems Journal, Decision Support Systems, 43,* 1675-1683.

Cowie, J., & Godley, P. (2006, May). Decision support on the move: Mobile decision making for triage management. In *Proceedings of the 8th International Conference on Enterprise Information Systems: Artificial Intelligence and Decision Support Systems,* Paphos, Cyprus (pp. 296-299). INSTICC Press.

Ghosh, A.K. (1998). E-*commerce security weak links, best defenses* (pp. 21-26). New York: Wiley Computer Publishing.

Greenwald, L., & Dean, T. (1995). Anticipating computational demands when solving time-critical decision-making problems. In K. Goldberg, D. Halperin, J.C. Latombe & R. Wilson (Eds.), *The algorithmic foundations of robotics.* Boston, MA: A. K. Peters. Retrieved December 14, 2007, from http://citeseer.nj.nec.com/greenwald95anticipating.html

Hartmann, J., & Bretzke, S. (1999). *Financial services for the future—mobile, flexible, and agent-based*. Retrieved December 14, 2007, from http://citeseer.ist.psu.edu/correct/287340

Hodgkin, J., San Pedro, J., & Burstein, F. (2004, July 1-3). Quality of data model for supporting real-time decision-making. In *Proceedings of the 2004 IFIP International Conference on Decision Support Systems (DSS2004)*. Prato, Italy.

Horvitz, E. (1995, February). *Transmission and display of information for time-critical decisions* (Tech. rep. MSR-TR-9513). Microsoft Research. Retrieved December 14, 2007, from http://citeseer.nj.nec.com/horvitz95transmission.html.

Keeney, R.L. (1992). *Value-focused thinking*. Harvard University Press.

Krishnaswamy, S., Loke, S.W., & Zaslavsky, A. (2002). Application run time estimation: A QoS metric for Web-based data mining service providers. In *Proceedings of ACM SAC*. ACM Press.

Malah, E.G. (2000). *Decision support and data-warehouse systems*. McGraw Hill.

Michalowski, W., Rubin, S., Slowinski, R., & Wilk, S. (2003). Mobile clinical support system for pediatric emergencies. *Decision Support Systems, 36*, 161-176.

Navarro, F., Schulter, A., Koch, F., Assuncao, M., & Westphall, C. (2006). Grid middleware for mobile decision support systems. In *Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies*, (pp. 125-125), *ICN/ICONS/MCL*.

Reis, L., Hoye, D., Gilbert, D., & Ryumae, M. (2000).*Online banking and electronic bill presentment payment are cost effective*. Retrieved December 14, 2007, from citeseer.nj.nec.com/402007.html

Roto, V. (2003). *Search on mobile phones*. Retrieved December 14, 2007, from http://home.earthlink.net/~searchworkshop/docs/Roto-SearchPositionPaper.pdf

Saksena, M.C., da Silva, J., & Agrawala, A.K. (1994). Design and implementation of Maruti-II. In S. Son (Ed.), *Principles of real-time systems*. Englewood Cliffs, NJ: Prentice-Hall. Retrieved December 14, 2007, from http://citeseer.nj.nec.com/saksena94design.html

San Pedro, J. C., Burstein, F., Cao, P. P., Churilov, L., Zaslavsky, A., Wassertheil, J. (2004, July 01-03). Mobile decision support for triage in emergency departments. In *The 2004 IFIP International Conference on Decision Support Systems (DSS2004) Conference Proceedings*, (pp. 714-723). Monash University, Melbourne, Victoria Australia.

San Pedro, J. Burstein, F., & Sharp, A. (2005). Toward case-based fuzzy multicriteria decision support model for tropical cyclone forecasting [Special Issue on Tools for Decision Support Systems]. *European Journal of Operational Research, 160*(2), 308-324.

Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems, 33,* 111–126

Simon, H.A. (1960). *The new science of management decision*. New York: Harper and Row.

Spreckelsen, C. Lethen, C., Heeskens, I., Pfeil, K., and Spitzer, K. (2000) *The roles of an intelligent mobile decision support system in the clinical workflow*. Retrieved January 23, 2008, from citeseer.nj.nec.com/spreckelsen00roles.html

Tewari, G., & Maes, P. (2001). A generalized platform for the specification, valuation, and brokering of heterogeneous resources in electronic

markets. In J. Liu & Y. Ye (Eds.), *E-commerce agents* (LNAI 2033, pp. 7-24). Springer-Verlag.

Triantaphyllou, E., & Sanchez, A. (1997). A sensitivity analysis for some deterministic multi-criteria decision making methods. *Decision Sciences, 28*(1), 151-194.

Zadeh, L.A. (1994). Fuzzy logic, neural networks, and soft computing. *Communication of the ACM, 37*(3), 77-78.

KEY TERMS

Mobile Decision Support: Providing support for a decision maker who has access to a mobile device for their decision support, is possibly on the move, and is possibly in a time-critical environment.

Mobile Devices: Mobile devices are portable computers that facilitate access to information in much the same way as a desktop computer. Typically such devices use a small visual display for user output and either some form of keypad, keyboard, or touch screen for user input.

Quality of Data: A measure of the quality of the data being used to assist the decision maker in making a decision. The quality of data measure is an aggregate value which encompasses information about technical factors of the mobile device (such as connectivity) as well as information pertaining to the completeness, accuracy of the data provided, reliability and relevance of the data provided.

Time-Critical Decisions: The idea that the context of a decision, its parameters, options, and best outcomes, are dependent on when the decision is made. A good outcome at one point in time is not necessary a good outcome at a later point in time if the decision is time-critical.

Uncertainty: Decision making under uncertainty occurs when the decision maker has no clear view on how different outcomes of a decision fair in comparison to each other. There is no obvious ranking of outcomes from best to worst.

This work was previously published in Encyclopedia of Decision Making and Decision Support Technologies, edited by F. Adam and P. Humphreys, pp. 638-644, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.24
OFDM Transmission Technique:
A Strong Candidate for the Next Generation Mobile Communications

Hermann Rohling
Hamburg University of Technology, Germany

ABSTRACT

The orthogonal frequency division multiplexing (OFDM) transmission technique can efficiently deal with multi-path propagation effects especially in broadband radio channels. It also has a high degree of system flexibility in multiple access schemes by combining the conventional TDMA, FDMA, and CDMA approaches with the OFDM modulation procedure, which is especially important in the uplink of a multi-user system. In OFDM-FDMA schemes carrier synchronization and the resulting sub-carrier orthogonality plays an important role to avoid any multiple access interferences (MAI) in the base station receiver. An additional technical challenge in system design is the required amplifier linearity to avoid any non-linear effects caused by a large peak-to-average ratio (PAR) of an OFDM signal. The OFDM transmission technique is used for the time being in some broadcast applications (DVB-T, DAB, DRM) and wireless local loop (WLL) standards (HIPERLAN/2, IEEE 802.11a) but OFDM has not been used so far in cellular communication networks. The general idea of the OFDM scheme is to split the total bandwidth into many narrowband sub-channels which are equidistantly distributed on the frequency axis. The sub-channel spectra overlap each other but the sub-carriers are still orthogonal in the receiver and can therefore be separated by a Fourier transformation. The system flexibility and use of sub-carrier specific adaptive modulation schemes in frequency selective radio channels are some advantages which make the OFDM transmission technique a strong and technically attractive candidate for the next generation of mobile communications. The objective of this chapter is to describe an OFDM-based system concept for the fourth generation (4G) of mobile communications and to discuss all technical details when establishing a cellular network which requires synchronization in time and frequency domain with sufficient accuracy. In this cellular environment a flexible frequency division multiple

access scheme based on OFDM-FDMA is developed and a radio resource management (RRM) employing dynamic channel allocation (DCA) techniques is used. A purely decentralized and self-organized synchronization technique using specific test signals and RRM techniques based on co-channel interference (CCI) measurements has been developed and will be described in this chapter.

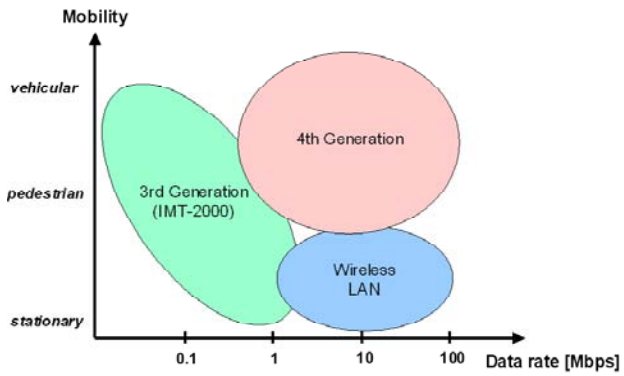
INTRODUCTION

In the evolution of mobile communication systems approximately a 10-year periodicity can be observed between consecutive system generations. Research work for the current 2nd generation of mobile communication systems (GSM) started in Europe in the early 1980s, and the complete system was ready for market in 1990. At that time the first research activities had already been started for the 3rd generation (3G) of mobile communication systems (UMTS, IMT-2000) and the transition from the current second generation (GSM) to the new 3G systems will be observed this year. Compared to today's GSM networks, these new UMTS systems will provide much higher data rates, typically in the range of 64 to 384 kbps, while the peak data rate for low mobility or indoor applications will be 2 Mbps.

The current pace, which can be observed in the mobile communications market, already shows that the 3G systems will not be the ultimate system solution. Consequently, general requirements for a 4G system have to be considered which will mainly be derived from the types of service a user will require in future applications. Generally, it is expected that data services instead of pure voice services will play a predominant role, in particular due to a demand for mobile IP applications. Variable and especially high data rates (20 Mbps and more) will be requested, which should also be available at high mobility in general or high vehicle speeds in particular (see Figure 1). Moreover, asymmetrical data services between up- and downlink are assumed and should be supported by 4G systems in such a scenario where the downlink carries most of the traffic and needs the higher data rate compared with the uplink.

To fulfill all these detailed system requirements the OFDM transmission technique applied in a wide-band radio channel is a strong candidate for an air interface in future 4G cellular systems due to its flexibility and adaptivity in the techni-

Figure 1. General requirements for 4G mobile communication systems



cal system design. From these considerations, it already becomes apparent that a radio transmission system for 4G must provide a great flexibility and adaptivity at different levels, ranging from the highest layer (requirements of the application) to the lowest layer (the transmission medium, the physical layer, that is, the radio channel) in the ISO-OSI stack. Today, the OFDM transmission technique is in a completely matured stage to be applied for wide-band communication systems integrated into a cellular mobile communications environment.

OFDM TRANSMISSION TECHNIQUE

Radio Channel Behaviour

The mobile communication system design is in general always dominated by the radio channel behaviour (Bello, 1963; Pätzold, 2002). In typical radio channel situation, multi-path propagation occurs (Figure 2) due to the reflections of the transmitted signal at several objects and obstacles inside the local environment and inside the observation area. The radio channel is analytically described unambiguously by a linear (quasi) time invariant (LTI) system model and by the related channel impulse response $h(\tau)$ or alternatively by the channel transfer function $H(f)$. An example for these channel characteristics is shown in Figure 3, where $h(\tau)$ and $H(f)$ of a so-called wide-sense stationary, uncorrelated scattering-channel (WSSUS) are given.

Due to the mobility of the mobile terminals the multi-path propagation situation will be continuously but slowly changed over time which is described analytically by a time variant channel impulse response $h(\tau, t)$ or alternatively by a frequency selective and time dependent radio channel transfer function $H(f, t)$ as it is shown in Figure 5 by an example. All signals on the various propagation paths will be received in a superimposed form and are technically characterized by different delays and individual Doppler frequencies which lead finally to a frequency selective behaviour of the radio channel, see Figure 4. The other two system functions, the Delay Doppler function, $v(\tau, f_D)$ and the Frequency Doppler function, $U(f, f_D)$ can be used as an alternative description

Figure 2. Multi-path propagation scenario

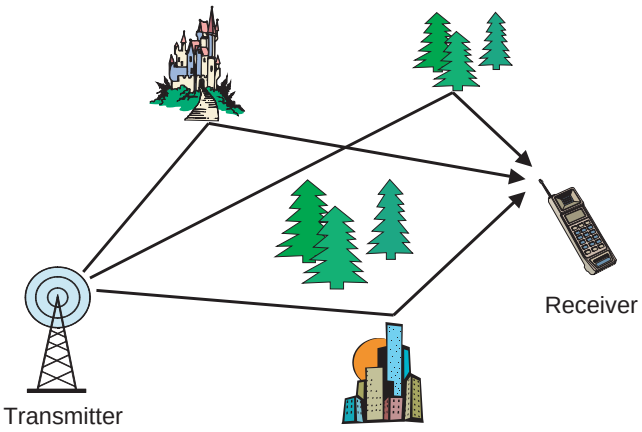


Figure 3. Impulse response and channel transfer function of a WS-SUS channel

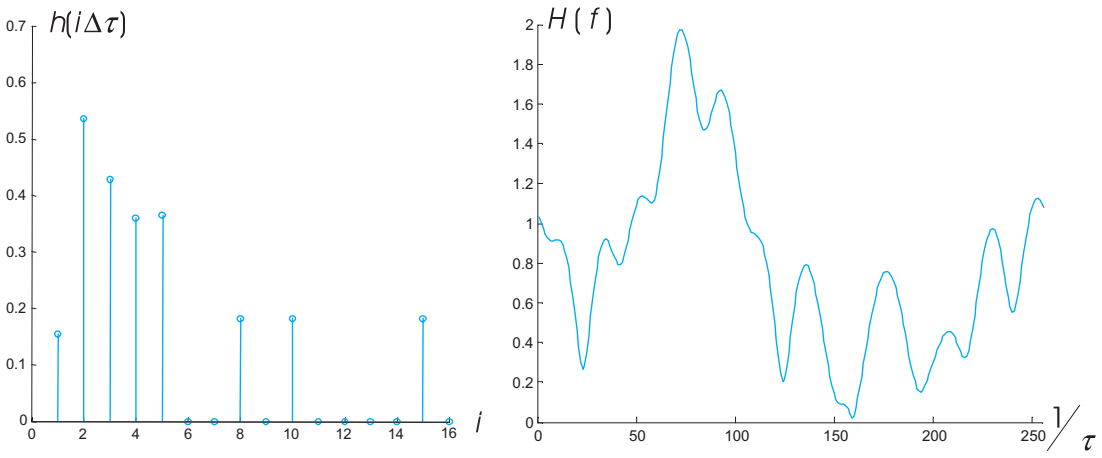
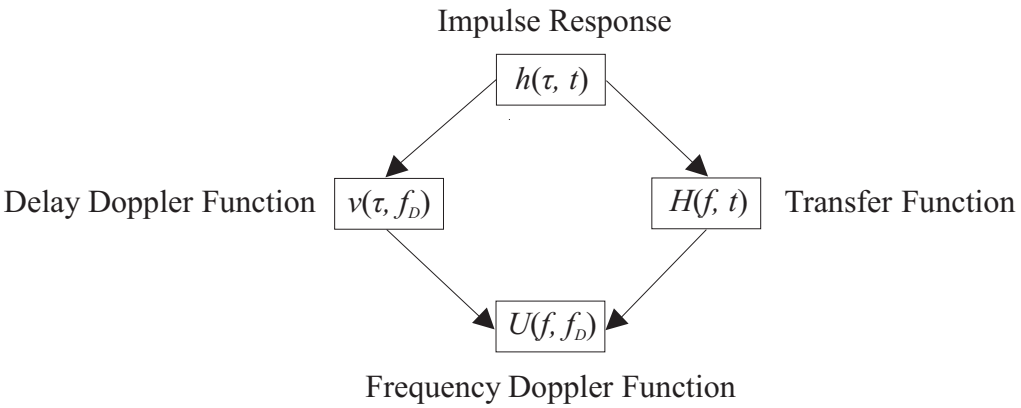


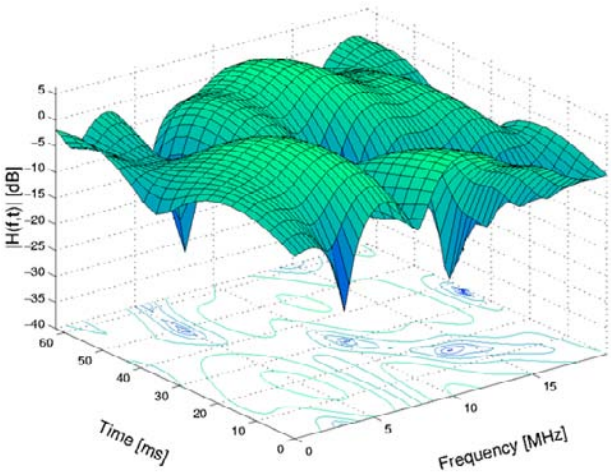
Figure 4. Relationships between different system functions



of the radio channel behaviour. The Delay Doppler function $v(\tau, f_D)$ describes the variation of the channel impulse response related to certain values of the Doppler frequency f_D. This means the channel delays change due to alteration of the relative speed between a mobile terminal and the base station. The Frequency Doppler function $U(f, f_D)$ models the same effects for the channel behaviour in the frequency domain.

The radio channel can roughly and briefly be characterised by two important system parameters: the maximum multi-path delay τ_{max} and

Figure 5. Frequency-selective and time-variant radio channel transfer function



the maximum Doppler frequency $f_{D\max}$ which are transferred into the coherence time T_C and the coherence bandwidth B_C of the radio channel:

$$T_C = \frac{1}{f_{D\max}}, \quad B_C = \frac{1}{\mathrm{t}_{\max}} \tag{1}$$

Over time intervals significantly shorter than T_C, the channel transfer function can be assumed to be nearly stationary. Similarly, for frequency intervals significantly smaller than B_C, the channel transfer function can be considered as nearly constant. Therefore it is assumed in this chapter that the coherence time T_C is much larger compared to a single OFDM symbol duration T_S and the coherence bandwidth B_C is much larger than the distance Δf between two adjacent sub-carriers:

$$B_C \gg \Delta f, \quad \Delta f = \frac{1}{T_S}, \quad T_S \ll T_C \tag{2}$$

This condition should be always fulfilled in well-dimensioned OFDM systems and in realistic time variant and frequency selective radio channels.

There are always technical alternatives possible in new system design phases. But future mobile communication systems will in any case require extremely large data rates and therefore large system bandwidth. If conventional single carrier (SC) modulation schemes with the resulting very low symbol durations are applied in this system design, it will be observed that very strong inter-symbol interference (ISI) is caused in wide-band applications due to multi-path propagation situations. This means for high data rate applications the symbol duration in a classical SC transmission system is extremely small compared to the typical values of maximum multi-path delay $\tau_{\max}$ in the considered radio channel. In these strong ISI situations a very powerful equalizer is necessary in each receiver, which needs high computation complexity in a wide-band system.

These constraints should be taken into consideration in the system development phase for a new radio transmission scheme and for a new 4G air interface. The computation complexity for the necessary equalizer techniques to overcome all these strong ISI in a SC modulation scheme increases exponentially for a given radio channel with increasing system bandwidth and can be extremely large in wide-band applications. For that reason alternative transmission techniques for broadband applications are of high interest.

Alternatively, the OFDM transmission technique can efficiently deal with all these ISI effects, which occur in multi-path propagation situations and in broadband radio channels. Simultaneously the OFDM transmission technique needs much less computation complexity in the equalization process inside each receiver. The performance figures for an OFDM based new air interface for the next generation of mobile communications are very promising even in frequency selective and time variant radio channel situations.

Advantages of the OFDM Transmission Technique

If a high data rate is transmitted over a frequency selective radio channel with a large maximum multi-path propagation delay τ_{max} compared to the symbol duration, an alternative to the classical SC approach is given by the OFDM transmission technique. The general idea of the OFDM transmission technique is to split the total available bandwidth B into many narrowband sub-channels at equidistant frequencies. The sub-channel spectra overlap each other but the sub-carrier signals are still orthogonal. The single high-rate data stream is subdivided into many low-rate data streams in the several sub-channels. Each sub-channel is modulated individually and will be transmitted simultaneously in a superimposed and parallel form.

An OFDM transmit signal therefore consists of N adjacent and orthogonal sub-carriers spaced by the frequency distance Δf on the frequency axis. All sub-carrier signals are mutually orthogonal within the symbol duration of length T_S if the sub-carrier distance and the symbol duration are chosen such that $T_S = 1 / \Delta f$. For OFDM-based systems the symbol duration T_S is much larger compared to the maximum multi-path delay τ_{max}. The k-th unmodulated sub-carrier signal is described analytically by a complex valued exponential function with carrier frequency $k\Delta f$, $\tilde{g}_k(t),\ k = 0,\ldots,N-1$.

$$\tilde{g}_k(t) = \begin{cases} e^{j2\pi k\Delta f t} & \forall t \in [0, T_S] \\ 0 & \forall t \notin [0, T_S] \end{cases} \qquad (3)$$

Since the system bandwidth B is subdivided into N narrowband sub-channels, the OFDM symbol duration T_S is N times larger as in the case of an alternative SC transmission system covering the same bandwidth B. Typically, for a given system bandwidth the number of sub-carriers is chosen in a way that the symbol duration T_S is sufficiently large compared to the maximum multi-path delay τ_{max} of the radio channel. On the other hand, in a time-variant radio channel the Doppler spread imposes restrictions on the sub-carrier spacing Δf. In order to keep the resulting inter-carrier interference (ICI) at a tolerable level, the system parameter of sub-carrier spacing Δf must be large enough compared to the maximum Doppler frequency $f_{D\max}$. In Aldinger (1994), the appropriate range for choosing the symbol duration T_S as a rule of thumb in practical systems is given as (compare with Equation (2)):

$$4\tau_{max} \le T_S \le 0.03 \frac{1}{f_{D,\max}}. \qquad (4)$$

The duration T_S as of the sub-carrier signal $\tilde{g}_k(t)$ is additionally extended by a cyclic prefix (so-called guard interval) of length T_G which is larger than the maximum multi-path delay τ_{max}

in order to avoid any ISI completely which could occur in multi-path channels in the transition interval between two adjacent OFDM symbols (Peled & Ruiz, 1980).

$$g_k(t) = \begin{cases} e^{j2\pi k\Delta ft} & \forall t \in [-T_G, T_S] \\ 0 & \forall t \notin [-T_G, T_S] \end{cases} \qquad (5)$$

$$= e^{j2\pi k\Delta ft} rect\left(\frac{2t+(T_G-T_S)}{2T}\right)$$

The guard interval is directly removed in the receiver after the time synchronization procedure. From this point of view the guard interval is a pure system overhead and the total OFDM symbol duration is therefore $T = T_S + T_G$. It is an important advantage of the OFDM transmission technique that ISI can be avoided completely or can be reduced at least considerably by a proper choice of OFDM system parameters.

The orthogonality of all sub-carrier signals is completely preserved in the receiver even in frequency selective radio channels which is an important advantage of the OFDM transmission technique. The radio channel behaves linear and in a short-time interval of a few OFDM symbols even time invariant. Therefore the radio channel behaviour can be described completely by a linear and time invariant (LTI) system model characterized by the impulse response $h(t)$. The LTI system theory gives the reason for this important system behaviour that all sub-carrier signals are orthogonal in the receiver even when transmitting the signal in frequency selective radio channels. All complex valued exponential signals (e.g., all sub-carrier signals) are Eigenfunctions of each LTI system and therefore Eigenfunctions of the considered radio channel which means that only the signal amplitude and phase will be changed if a sub-carrier signal is transmitted in the linear and time invariant radio channel.

The sub-carrier frequency is not affected at all by the radio channel transmission which means that all sub-carrier signals are even orthogonal in the receiver and at the output of a frequency selective radio channel. The radio channel interferes only amplitudes and phases individually but not the sub-carrier frequency of all received sub-channel signals. Therefore all sub-carrier signals are still mutually orthogonal in the receiver. Due to this important property the received signal which is superimposed by all sub-carrier signals can be split directly into the different sub-channel components by a Fourier transformation and each sub-carrier signal can be demodulated individually by a single tap equalizer in the receiver.

At the transmitter side each sub-carrier signal is modulated independently and individually by the complex valued modulation symbol $S_{n,k}$, where the subscript n refers to the time interval and k to the sub-carrier signal number in the considered OFDM symbol. Thus, within the symbol duration time interval T the time continuous signal of the n-th OFDM symbol is formed by a superposition of all N simultaneously modulated sub-carrier signals.

$$s_n(t) = \sum_{k=0}^{N-1} S_{n,k} g_k(t - nT) \qquad (6)$$

The total time continuous transmit signal consisting of all OFDM symbols sequentially transmitted on the time axis is described analytically by the following equation:

$$s(t) = \sum_{n=0}^{\infty} \sum_{k=0}^{N-1} S_{n,k} e^{j2\pi k\Delta f(t-nT)} rect\left(\frac{2(t-nT)+(T_G-T_S)}{2T}\right) \qquad (7)$$

The analytical transmit signal description shows that a rectangular pulse shaping is applied for each sub-carrier signal and each OFDM symbol. But due to the rectangular pulse shaping, the spectra of all the considered sub-carrier signals are sinc-functions which are equidistantly located

on the frequency axis, for example, for the *k*-th sub-carrier signal the spectrum is described in the following equation:

$$G_k(f) = T \cdot \mathrm{sinc}\left[\pi T\left(f - k\Delta f\right)\right] \text{ where}$$
$$\mathrm{sinc}(x) = \frac{\sin(x)}{x} \tag{8}$$

The typical OFDM-Spectrum shown in Figure 6 consists of N adjacent sinc-functions, which are shifted by Δf in the frequency direction.

The spectra of the considered sub-carrier signals overlap on the frequency axis, but the sub-carrier signals are still mutually orthogonal which means the transmitted modulation symbols $S_{n,k}$ can be recovered by a simple correlation technique in each receiver if the radio channel is assumed to be ideal in a first analytical step:

$$\frac{1}{T_s}\int_0^{T_S} g_k(t)\overline{g_l(t)}dt = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases} = \delta_{k,l} \tag{9}$$

$$S_{n,k} = \frac{1}{T_S}\int_0^{T_S} s_n(t)\overline{g_k(t)}dt = \frac{1}{T_S}\int_0^{T_S} s_n(t)e^{-j2\pi k\Delta ft}dt \tag{10}$$

where $\overline{g_k(t)}$ is the conjugate complex version of the sub-carrier signal $g_k(t)$. The following equations show the correlation process in detail:

$$Corr = \frac{1}{T_S}\int_0^{T_S} s_n(t)\overline{g_k(t)}dt = \frac{1}{T_S}\int_0^{T_S}\sum_{m=0}^{N-1} S_{n,m}g_m(t)\overline{g_k(t)}dt$$
$$= \sum_{m=0}^{N-1} S_{n,m}\frac{1}{T_S}\int_0^{T_S} g_m(t)\overline{g_k(t)}dt = \sum_{m=0}^{N-1} S_{n,m}\delta_{m,k} = \underline{\underline{S_{n,k}}} \tag{11}$$

In practical applications the OFDM transmit signal $s_n(t)$ is generated in a first step and in the digital baseband signal processing part of the transmitter as a time discrete signal. Using the sampling theorem while considering the OFDM transmit signal inside the bandwidth $B = N\Delta f$, the transmit signal must be sampled with the sampling interval $\Delta t = 1/B = 1/N\Delta f$. The individual samples of the transmit signal are denoted by $s_{n,i}$, $i = 0, 1, \ldots, N-1$ and can be calculated as follows (see Equation (7)):

$$s(t) = \sum_{k=0}^{N-1} S_{n,k}e^{j2\pi k\Delta ft}$$
$$s(i\Delta t) = \sum_{k=0}^{N-1} S_{n,k}e^{j2\pi k\Delta f(i\Delta t)}$$
$$s_{n,i} = \sum_{k=0}^{N-1} S_{n,k}e^{j2\pi ik/N} \tag{12}$$

This Equation (12) describes exactly the inverse discrete Fourier transform (IDFT) applied to the complex valued modulation symbols $S_{n,k}$ of all sub-carrier signals inside a single OFDM symbol.

The individually modulated and superimposed sub-carrier signals are transmitted in a parallel way over many narrowband sub-channels. Thus, in each sub-channel the symbol duration is quite large and can be chosen much larger as compared to the maximum multi-path delay of the radio channel. In this case each sub-channel has the property to be frequency non-selective.

Figure 7 shows the general OFDM system structure in a block diagram. The basic principles of the OFDM transmission technique have already been described in several publications like Bingham (1990) and Weinstein and Ebert (1971). In the very early and classical multi-carrier system considerations like Chang (1966) and Saltzberg (1967), narrowband signals have been generated independently, assigned to various frequency bands, transmitted, and separated by analogue filters at the receiver. The new and modern aspect of the OFDM transmission technique is that the various sub-carrier signals are generated digitally and jointly by an IFFT in the transmitter and that

Figure 6. OFDM spectrum which consists of N equidistant sinc-functions

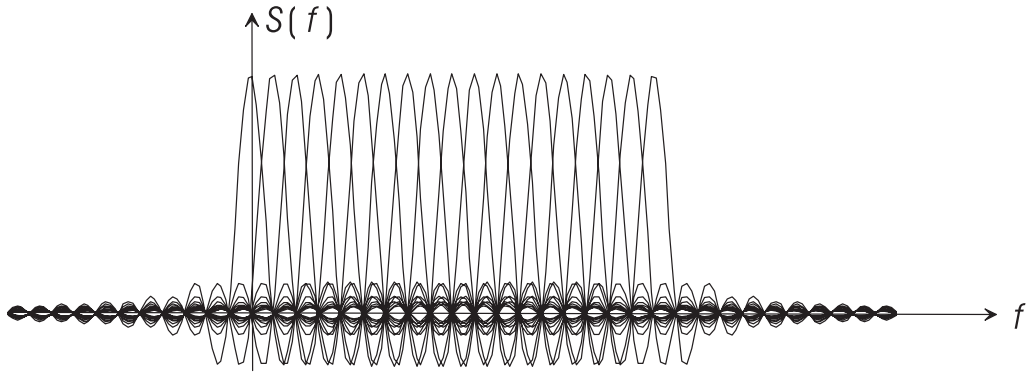
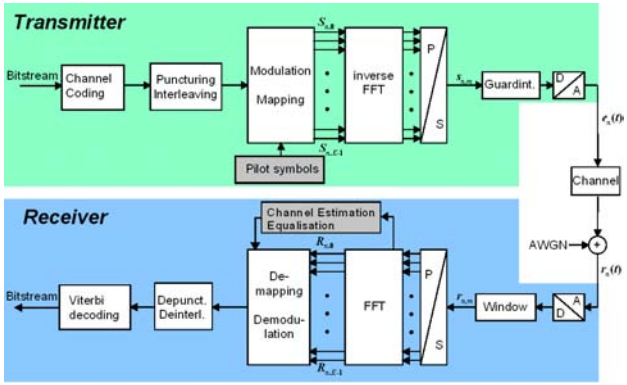


Figure 7. OFDM system structure in a block diagram



their spectra strongly overlap on the frequency axis. As a result, generating the transmit signal is simplified and the bandwidth efficiency of the system is significantly improved.

The received signal is represented by the convolution of the transmitted time signal with the channel impulse response $h(t)$ and an additive white Gaussian noise term:

$$r_n(t) = s_n(t) * h_n(t) + n_n(t) \quad (13)$$

Due to the assumption, that the coherence time T_C will be much larger than the symbol duration T_S the received time continuous signal $r_n(t)$ can be separated into the orthogonal sub-carrier signal components even in frequency selective fading

situations by applying the correlation technique mentioned in Equation (10):

$$R_{n,k} = \frac{1}{T_S}\int_0^{T_S} r_n(t)e^{-j2\pi k\Delta ft}dt \qquad (14)$$

Equivalently, the correlation process at the receiver side can be applied to the time discrete receive signal at the output of an A/D converter and can be implemented as a DFT, which leads to the following equation:

$$R_{n,k} = \frac{1}{N}\sum_{i=0}^{N-1} r_{n,i}e^{-j2\pi iik/N} \qquad (15)$$

In this case $r_{n,i} = r_n(i\cdot\Delta t)$ describes the *i*-th sample of the received time continuous base-band signal $r_n(t)$ and $R_{n,k}$ is the received complex valued symbol at the DFT output of the *k*-th sub-carrier.

If the OFDM symbol duration T is chosen much smaller than the coherence time T_C of the radio channel, then the time variant transfer function of the radio channel $H(f, t)$ can be considered constant within the time duration T of each modulation symbol $S_{n,k}$ for all sub-carrier signals. In this case, the effect of the radio channel in multi-path propagation situations can be described analytically by only a single multiplication of each sub-carrier signal $g_k(t)$ with the complex transfer factor $H_{n,k} = H(k\Delta f, nT)$. As a result, the received complex valued symbol $R_{n,k}$ at the DFT output can be described analytically as follows:

$$r_n(t) = s_n(t) * h_n(t) + n_n(t)$$
$$r_{n,i} = s_{n,i} * h_{n,i} + n_{n,i}$$

$$R_{n,k} = S_{n,k}H_{n,k} + N_{n,k} \qquad (16)$$

where $N_{n,k}$ describes an additive noise component for each specific sub-carrier generated in the radio channel. This equation shows the most important advantage of applying the OFDM transmission technique in practical applications. Equation (16) describes the complete signal transfer situation of the OFDM block diagram including IDFT, guard interval, D/A conversion, up- and down-conversion in the RF part, frequency selective radio channel, A/D conversion and DFT process in the receiver, neglecting non-ideal behaviour of any system components.

The transmitted Symbol $S_{n,k}$ can be recovered, calculating the quotient of the received complex valued symbol and the estimated channel transfer factor $\tilde{H}_{n,k}$:

$$S_{n,k} = \frac{R_{n,k} - N_{n,k}}{H_{n,k}}, \quad \tilde{S}_{n,k} = R_{n,k}\frac{1}{\tilde{H}_{n,k}} \qquad (17)$$

It is obvious that this one tap equalization step of the received signal is much easier compared to a single carrier system for high data rate applications. The necessary IDFT and DFT calculations can be implemented very efficiently using the Fast-Fourier-Transform (FFT) algorithms such as Radix 2^2, which reduces the system and computation complexity even more.

It should be pointed out that especially the frequency synchronization at the receiver must be very precise in order to avoid any inter-carrier interferences (ICI). Algorithms for time and frequency synchronization in OFDM-based systems are described in Classen and Meyr (1994) and Mizoguchi et al. (1998), for example and will be considered in the section, *Self-Organized Cell Synchronization.*

Besides the complexity aspects, another advantage of the OFDM technique lies in its high degree of flexibility and adaptivity. Division of the available bandwidth into many frequency-non-selective sub-bands gives additional advantages for the OFDM transmission technique. It allows a sub-carrier-specific adaptation of transmit parameters, such as modulation scheme (PHY mode) and transmit power (cp. Water Filling) in

accordance to the observed and measured radio channel status. In a multi-user environment the OFDM structure offers additionally an increased flexibility for resource allocation procedures as compared to SC systems (Hanzo et al., 2003).

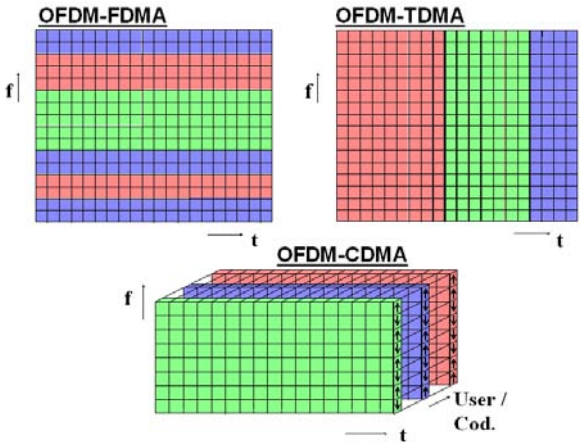
The important system behaviour that all sub-carrier signals are mutually orthogonal in the receiver makes the signal processing and the equalization process realized by a single-tap procedure very simple and leads to a low computation complexity.

OFDM COMBINED WITH MULTIPLE ACCESS SCHEMES

A very high degree of flexibility and adaptivity is required for new mobile communication systems and for the 4G air interface. The combination between multiple access schemes and OFDM transmission technique is an important factor in this respect. In principle, multiple access schemes for the OFDM transmission technique can be categorized according to OFDM-FDMA, OFDM-TDMA, and OFDM-CDMA (Kaiser, 1998; Rohling & Grünheid, 1997). Clearly, hybrid schemes can be applied which are based on a combination of these techniques. The principles of these basic multiple access schemes are summarized in Figure 8, where the time-frequency plane is depicted and the user specific resource allocation is distinguished by different colours.

These access schemes provide a great variety of possibilities for a flexible user specific resource allocation. In the following, one example for OFDM-FDMA is briefly sketched (cf. Galda, Rohling, Costa, Haas, & Schulz, 2002). In the case that the magnitude of the channel transfer function is known for each user the sub-carrier selection for an OFDM-FDMA scheme can be processed in the BS for each user individually which leads to a multi-user diversity (MUD) effect. By allocating a subset of all sub-carriers with the highest SNR to each user the system performance can be improved. This allocation technique based on the knowledge of the channel transfer function shows a large performance advantage and a gain in quality of service (QoS). Nearly the same flexibility in resource allocation is possible in OFDM-CDMA systems. But in

Figure 8. OFDM transmission technique and some multiple access schemes



this case the code orthogonality is destroyed by the frequency selective radio channel resulting in multiple access interferences (MAI), which reduces the system performance.

TECHNICAL PROPOSAL AND EXAMPLE FOR A 4G DOWNLINK INTERFACE

Taking all these important results from the previous sections into consideration, a system design example is considered in this section. OFDM system parameters for a 4G air interface are considered and three different multiple access schemes inside a single cell are compared quantitatively. A bandwidth of 20 MHz in the 5.5 GHz domain is assumed. The assumed multi-path radio channel has a maximum delay of $\tau_{max} = 5$ μs (the coherence bandwidth is therefore $B_C = 200$ kHz). Additionally, a maximum speed of $v_{max} = 200$ km/h is assumed, which yields a maximum Doppler frequency of $f_{Dmax} = 1$ kHz and a coherence time of $T_C = 1$ ms. Table 1 shows an example for the system parameters of a 4G air interface.

For the considered OFDM based system three different multiple access concepts have been analysed and compared. The first proposal is based on a pure OFDM-TDMA structure, while the second one considers an OFDM-FDMA technique with an adaptive sub-carrier selection scheme, as described in the section, *OFDM Transmission Technique Combined with Multiple Access Schemes*. The third one is based on OFDM-CDMA where the user data are spread over a subset of adjacent sub-carriers.

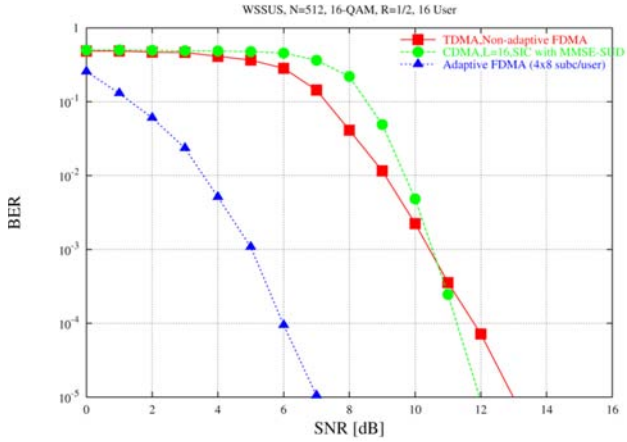
In this case MAI occur and an interference cancellation technique implemented in each MT is useful. To compare the different multiple access schemes, Figure 9 shows the bit error rate (BER) for an OFDM system with the system parameters shown in Table 1.

A single cell situation has been considered in this case with a perfect time and carrier synchronization. As can be seen from this figure, the best performance can be achieved by an OFDM-FDMA system which exploits the frequency selective fading of the mobile radio channel by allocating always the best available sub-carrier to each user. Note that a channel adaptive FDMA scheme requires a good prediction of the channel transfer function which has been considered to be perfect in this comparison. If a non-adaptive FDMA technique was used (i.e., fixed or random

Table 1. Proposal of OFDM system parameters

PARAMETER	VALUE
FFT Length	$N_C = 512$
Guard interval length	$N_G = N_C/8 = 64$
Modulation technique	16-QAM
Code rate	R=1/2, m=6
FDMA	Best available sub-carrier is selected.
TDMA	
CDMA Spreading matrix	Walsh-Hadamard (L=16)
CDMA detection technique	SUD with MMSE, MUD with soft interference cancellation plus MMSE

Figure 9. BER results for a coded OFDM system employing different multiple access techniques



allocation of sub-carriers), the performance would be comparable to the OFDM-TDMA curve.

In the case of an OFDM-TDMA system the frequency selectivity of the radio channel can be exploited by the Viterbi decoder in conjunction with bit interleaving. A pure coded OFDM-CDMA system which utilizes an orthogonal spreading matrix with minimum mean square error (MMSE) equalization and single user detection (SUD) to exploit the diversity of the channel suffers from MAI due to loss of code orthogonality in frequency selective fading. A performance improvement can be achieved for an OFDM-CDMA scheme applying multi-user detection (MUD) techniques. By successively removing inter-code/-user interference using MUD procedure, a gain of approximately 2 dB can be achieved. But still an OFDM-FDMA system outperforms an optimized OFDM-CDMA system. Additionally, OFDM-CDMA technique has a much higher computational complexity in the MUD scheme.

TECHNICAL PROPOSAL AND EXAMPLE FOR A 4G UPLINK INTERFACE

As shown in the preceding paragraph, there are several system proposals published for an OFDM-based downlink procedure for broadcast and communication systems respectively. But by designing an OFDM uplink transmission scheme some important and additional technical questions will come up. Therefore, OFDM-based uplink systems are still under consideration and research (Rohling, Galda, & Schulz, 2004). As a contribution to this topic, an OFDM-based multi-user uplink system with *M* different users inside a single cell is considered in this section.

Each user shares the entire bandwidth with all other users inside the cell by allocating exclusively a deterministic subset of all available sub-carriers inside the considered OFDM system. This user specific sub-carrier selection process allows to

share the total bandwidth in a very flexible way between all mobile terminals. Hence, as a relevant multiple access scheme an OFDM-FDMA structure is considered in which each user claims the same bandwidth or the same number of sub-carriers inside the total bandwidth respectively. Due to the assumed perfect carrier synchronization and resulting sub-carrier orthogonality in the receiver any multiple access interference (MAI) between different users can be avoided. The sub-carrier allocation process can either be designed to be non-adaptive or adaptive in accordance with the current radio channel state information (CSI).

Since the OFDM transmission signal results from the superposition of a large number of independent data symbols and sub-carrier signals the envelope of the complex valued baseband time signal is in general not constant but has a large peak-to-average ratio (PAR). The largest output power value of the amplifier will therefore limit the maximum amplitude in the transmit signal. Additionally, non-linear distortions due to clipping and amplification effects in the transmit signal will lead to both in-band interferences and out-of-band emissions (Brüninghaus & Rohling, 1997). Therefore, in the downlink case each base station will spend some effort and computation power to control the transmit signal amplitude and to reduce the PAR. The objective is in this case to minimize the resulting non-linear effects or even to avoid any interferences.

But for the uplink case it is especially important to design a transmit signal with low PAR to reduce computation complexity in the mobile terminal and to avoid any interference situation caused by non-linear effects of the amplification process.

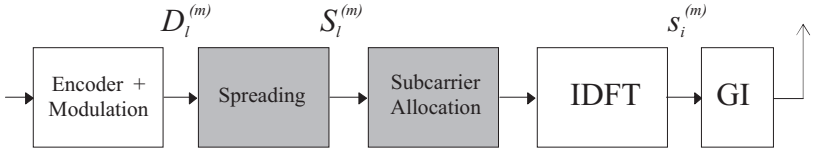
It will be shown in this section that an OFDM-FDMA system based on an equidistant sub-carrier selection procedure combined with an additional sub-carrier spreading technique will reduce the resulting PAR significantly (Brüninghaus & Rohling, 1998) for the uplink procedure. Furthermore, this proposal will lead to a modulation technique which becomes technically very simple and where the transmit signal consists of a periodic extension and multiple repetition of all modulation symbols. This is the result of the duality between multi-carrier CDMA and single carrier transmission technique as described in Brüninghaus and Rohling (1998).

Figure 10 shows the general structure of an OFDM uplink signal processing in the mobile terminal, which will be considered in this section. In this block diagram there are two main components in the OFDM-based modulation scheme which will be treated in the design process of a multi-user uplink system: The sub-carrier selection technique, and a user specific spreading scheme applied to the user's selected sub-carrier subset, respectively.

The last two blocks in the block diagram show the characteristic IDFT processing and the guard interval (GI) insertion, that are common in all OFDM-based transmitter schemes.

In the uplink system model of a multi-user, OFDM-FDMA based scheme, an arbitrary num-

Figure 10. Block diagram of a multi-user OFDM-FDMA uplink system

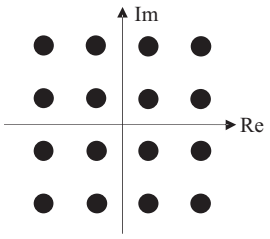


ber of M different users are considered inside a single cell and each user allocates exclusively L different sub-carriers which are considered inside the entire system bandwidth for data transmission. The total number of all considered sub-carriers inside the system bandwidth of the transmission scheme is therefore $N_C = L \cdot M$.

The input data stream for each mobile user terminal m, $m = 0, \ldots, M-1$, is convolutionally encoded in a first step. Afterwards, the bit sequence is mapped onto a modulation symbol vector $\vec{D}^{(m)} = \left(D_0^{(m)}, D_1^{(m)}, \ldots, D_{L-1}^{(m)}\right)$ of L complex valued symbols $D_l^{(m)}$ from a given modulation alphabet with 2^Q different modulation symbols inside the constellation diagram. An example for such a modulation alphabet is given in Figure 11 for a 16-QAM.

In this section, a non-differential, higher level modulation scheme is assumed for the uplink case.

Figure 11. 16-QAM as an exemplary alphabet with modulation symbols $D_l^{(m)}$



Each user transmits $L \cdot Q$ bits per OFDM symbol. It is assumed in this section without any loss of generality that each user transmits the same data rate or the same number of modulation symbols per OFDM signal respectively.

Sub-Carrier Allocation Process

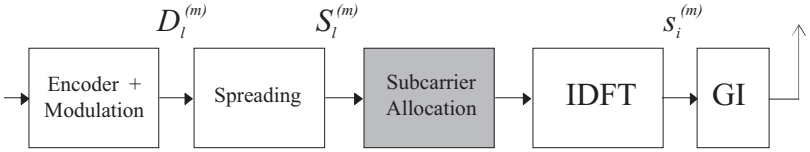
The first important question in the OFDM-FDMA multi-user uplink system design is the user specific sub-carrier selection scheme. This process is responsible for sharing the bandwidth between M different users, see Figure 12.

There is a large degree of freedom in this system design step to allocate exclusively a subset of L specific sub-carriers to each user. This can either be done by a random or a deterministic allocation scheme. Alternatively there are proposals made for adaptive sub-carrier selection schemes to increase the resulting system capacity (Gross, Karl, Fitzek, & Wolisz, 2003; Shen, Li, & Liu, 2004; Toufik & Knopp, 2004).

In this paragraph a very specific non-adaptive sub-carrier selection procedure is proposed. In this case the allocated sub-carrier subset is equidistantly located on the frequency axis over the entire system bandwidth. This approach is shown in Figure 13 and will be pursued in the following.

In this multi-user uplink system each user m allocates exclusively in total L sub-carriers which are in each case placed in an equidistant way on the frequency axis. The selected L sub-carriers are modulated with L complex valued transmit sym-

Figure 12. Block diagram for an OFDM-FDMA based system with sub-carrier selection process



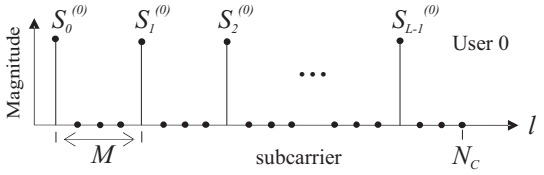
bols $S_l^{(m)}$, described and denoted by the transmit symbol vector $\vec{S}^{(m)}$. The proposed non-adaptive sub-carrier selection and modulation process does not need any radio channel state information (CSI) at the transmitter side.

Due to this specific sub-carrier selection process based on equidistantly located sub-carriers on the frequency axis the resulting OFDM uplink transmit time signal $s_i^{(m)}$ of any user has a periodic structure with period length L and consists in any case of an M-times repetition time signal, see Figure 14.

Equation (18) describes the relation between the sub-carrier transmit symbols $S_l^{(0)}$ and a single period of the resulting OFDM transmit time signal $s_i^{(0)}$ for user 0 analytically.

$$s_i^{(0)} = \frac{1}{\sqrt{L}} \sum_{l=0}^{L-1} S_l^{(0)} e^{j2\pi il/L} \quad \text{for } i = 0,1,\ldots,L-1 \tag{18}$$

Figure 13. Equidistantly allocated subset of L sub-carriers for a single user in a multi-user environment



This relation in equation is simply an IDFT applied to the transmit symbols $S_l^{(0)}$, as shown in Equation (19).

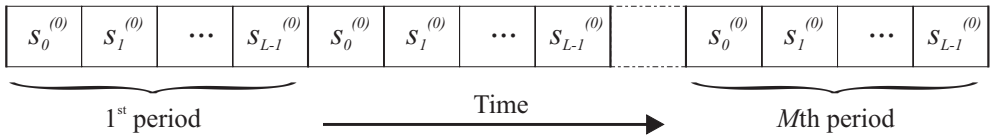
$$\begin{pmatrix} s_0^{(0)} \\ s_1^{(0)} \\ \cdots \\ s_{L-1}^{(0)} \end{pmatrix} = \text{IDFT} \begin{pmatrix} S_0^{(0)} \\ S_1^{(0)} \\ \cdots \\ S_{L-1}^{(0)} \end{pmatrix} \tag{19}$$

Because the sub-carrier subset of a single user is assumed to be allocated equidistantly over all N_C sub-carriers inside the entire bandwidth (Figure 13), it can be shown that an N_C-IDFT processing of the sub-carrier transmit symbols $S_l^{(0)}$ inside the OFDM transmitter leads to the same M-times repetition of the user time signal $s_i^{(0)}$ as shown in Figure 14. The periodicity of the transmit signal is directly related to the selection process of equidistantly located sub-carrier on the frequency axis.

Sub-Carrier Spreading Technique

This paragraph addresses the second design element of an OFDM-FDMA based system: a spreading technique applied to the user's selected sub-carriers, see Figure 15. There are several well-known spreading techniques, which can be integrated into an OFDM-based transmission technique (Kaiser, 2002; Linnartz, 2000). Analo-

Figure 14. OFDM-FDMA based periodic transmit time signal with period length L and M-times repetition



gous to other MC-CDMA systems, described in Kaiser (2002) and Linnartz (2000), the vector $\vec{D}^{(m)}$ of L modulation symbols (see Figure 11) is spread in this case over L sub-carriers which are exclusively allocated to user m applying an unitary spreading matrix [C].

This results in a transmit sub-carrier symbol vector $\vec{S}^{(m)} = \left(S_0^{(m)}, S_1^{(m)}, \ldots, S_{L-1}^{(m)}\right)$ consisting of L complex valued transmit symbols $S_l^{(m)}$, $l = 0,\ldots, L-1$. The spreading operation can be denoted mathematically by the following matrix multiplication where each complex valued transmit symbol $S_l^{(m)}$ is calculated by the sum of L user specific modulation symbols $D_l^{(m)}$ weighted by L orthogonal code vectors $\vec{C}_l = \left(C_{l,0}, C_{l,1}, \ldots, C_{l,L-1}\right)$ with $l = 0,\ldots, L-1$:

$$\begin{pmatrix} S_0^{(m)} \\ S_1^{(m)} \\ \vdots \\ S_{L-1}^{(m)} \end{pmatrix} = \begin{bmatrix} C_{0,0} & C_{0,1} & \cdots & C_{0,L-1} \\ C_{1,0} & \ddots & & C_{1,L-1} \\ \vdots & & & \vdots \\ C_{L-1,0} & \cdots & \cdots & C_{L-1,L-1} \end{bmatrix} \cdot \begin{pmatrix} D_0^{(m)} \\ D_1^{(m)} \\ \vdots \\ D_{L-1}^{(m)} \end{pmatrix} \tag{20}$$

The spreading Matrix [C] consists of L orthogonal spreading codes. It can be designed, for example, by a Walsh-Hadamard matrix like in Kaiser (2002) and Linnartz (2000) or by a DFT matrix as described in Brüninghaus and Rohling (1997, 1998). Both matrix types fulfill the requirements for unity and orthogonality. Examples for these matrices are shown in Figure 16. In the considered multi-user uplink system, only a DFT matrix based spreading technique will be used, because of the resulting benefits in combination with an equidistant sub-carrier allocation scheme.

After the spreading process, the sub-carrier specific transmit symbols $S_l^{(m)}$ are mapped onto L sub-carrier signals which are exclusively allocated to user m. In principle, the user specific sub-carrier subset can be composed of any L out of N_C sub-carriers that have not been assigned to another user.

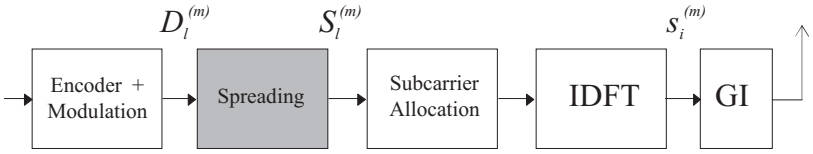
Combination of Spreading and Sub-Carrier Allocation

As explained in the previous paragraph, the spreading technique applied to the modulation symbols

Figure 16. Examples for Walsh-Hadamard (left) and DFT spreading matrix

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & e^{-j\pi/2} & e^{-j\pi} & e^{-j\frac{3}{2}\pi} \\ 1 & e^{-j\pi} & e^{-j2\pi} & e^{-j3\pi} \\ 1 & e^{-j\frac{3}{2}\pi} & e^{-j3\pi} & e^{-j\frac{9}{2}\pi} \end{bmatrix}$$

Figure 15. Block diagram of a multi-user OFDM-FDMA uplink system with additional spreading technique



$D_l^{(m)}$ is considered in a way, that a DFT-Matrix can be used as spreading matrix [C]. Therefore, the relation between modulation symbols $D_l^{(0)}$ and sub-carrier transmit symbols $S_l^{(0)}$ are described analytically by Equation (21):

$$\begin{pmatrix} S_0^{(0)} \\ S_1^{(0)} \\ \cdots \\ S_{L-1}^{(0)} \end{pmatrix} = \begin{bmatrix} \\ [C] \\ \\ \end{bmatrix} \cdot \begin{pmatrix} D_0^{(0)} \\ D_1^{(0)} \\ \cdots \\ D_{L-1}^{(0)} \end{pmatrix} = \mathrm{DFT} \begin{pmatrix} D_0^{(0)} \\ D_1^{(0)} \\ \cdots \\ D_{L-1}^{(0)} \end{pmatrix} \tag{21}$$

If this DFT-based spreading technique is combined with the earlier explained, equidistant sub-carrier selection process the transmit time signal $s_i^{(0)}$ can be calculated directly by the M-times repetition of modulation symbol vector $\vec{D}^{(0)}$ which consists of L complex valued modulation symbols, see Figure 17. Therefore, it is needless to process the DFT spreading matrix and the IFFT in the OFDM system structure explicitly, which reduces the computation complexity in the mobile terminal, see Equation (22). Hence, a single period of the resulting time signal $s_i^{(0)}$ is directly given by the calculated modulation symbols $D_l^{(0)}$.

$$\begin{pmatrix} s_0^{(0)} \\ s_1^{(0)} \\ \vdots \\ s_{L-1}^{(0)} \end{pmatrix} = \mathrm{IDFT} \begin{pmatrix} S_0^{(0)} \\ S_1^{(0)} \\ \vdots \\ S_{L-1}^{(0)} \end{pmatrix} = \mathrm{IDFT} \left(\mathrm{DFT} \begin{pmatrix} D_0^{(0)} \\ D_1^{(0)} \\ \vdots \\ D_{L-1}^{(0)} \end{pmatrix} \right) = \begin{pmatrix} D_0^{(0)} \\ D_1^{(0)} \\ \vdots \\ D_{L-1}^{(0)} \end{pmatrix} \tag{22}$$

In almost all OFDM systems, a cyclic prefix of length N_G will be added to the transmit time signal $s_i^{(0)}$ to avoid any ISI. Therefore, the so-called guard interval is also an integral part of the multi-user uplink system described in this paragraph. Thus, the structure of the OFDM-FDMA multi-user uplink system depicted in Figure 18 can be simplified. Figure 18 shows the functionality of the overall system in detail. It becomes clear that because of the cancellation of DFT spreading and IDFT calculation these components can be completely removed in the technical realization. They are replaced by a simple repetition process of the considered user specific modulation symbols $D_l^{(0)}$.

Figure 17. Periodic transmit signal for the multi-user uplink system: Symbols $s_i^{(0)}$ and modulation symbols $D_l^{(0)}$

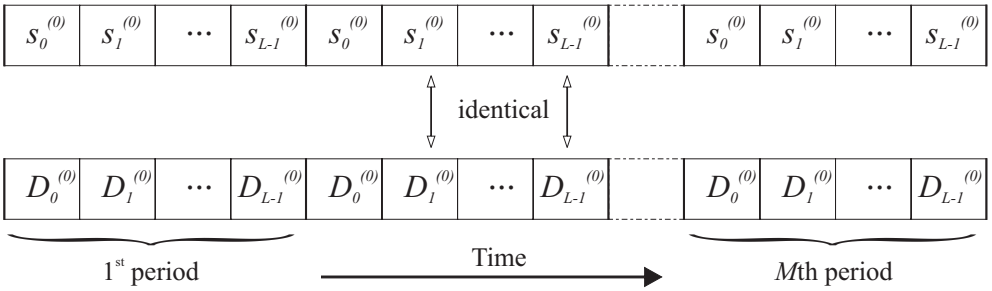
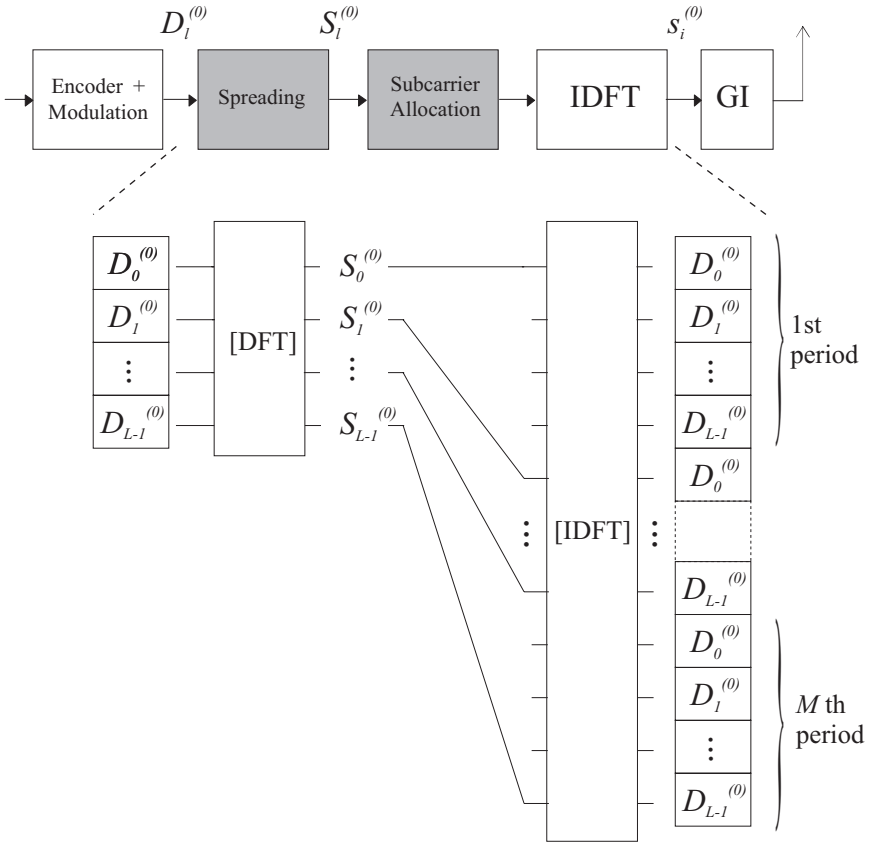


Figure 18. OFDM-FDMA based uplink system including a DFT spreading matrix applied to a set of equidistant sub-carriers



Multi-User Case

The extension from a single to an arbitrary user m is straight forward and will be described in the following. Another user m also allocates an equidistantly spaced subset of all sub-carriers which is shifted in the frequency space by m sub-carriers, see Figure 19.

Any frequency shift results in a multiplication of the transmit time signal $s_i^{(m)}$ with a complex valued signal $e^{j2\pi im/N_C}$, see Equation (23).

Figure 19. Shifting the total sub-carrier subset in a multi-user environment

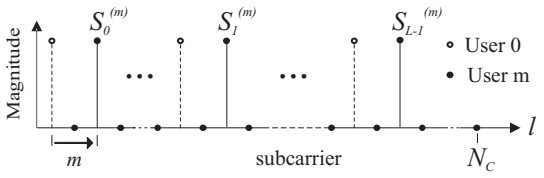
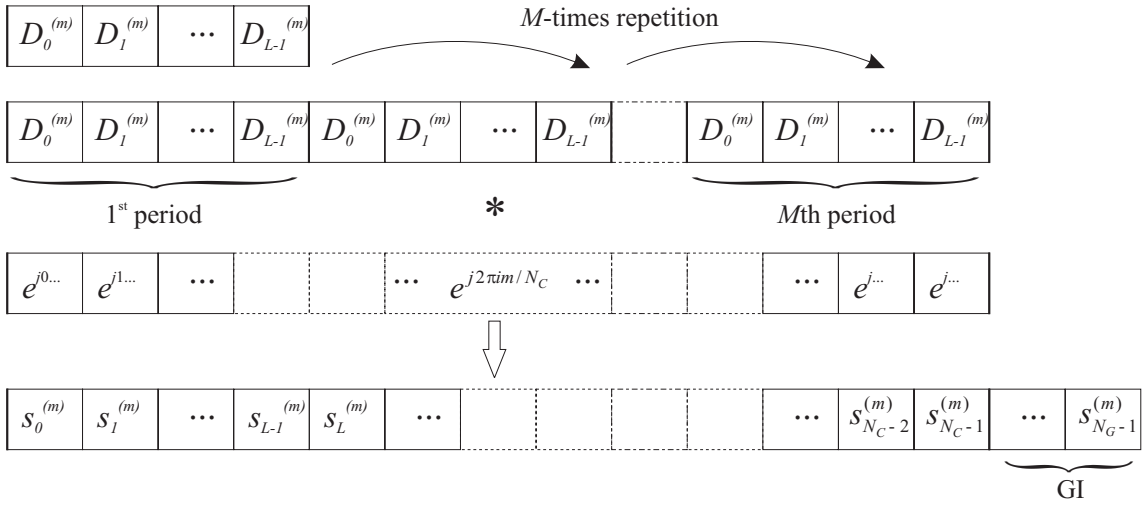


Figure 20. OFDM-FDMA uplink transmit signal for an arbitrary user m



$$S_l^{(m)} \cdot \mathrm{d}(l-m) \quad \bullet\!-\!\circ \quad s_i^{(m)} \cdot e^{j2\pi im/N_C} \qquad (23)$$

This yields a phase rotation of the transmit time symbols $s_i^{(m)}$ with the constant frequency $f_0 = m / N_C$. But this has no significant impact on the complexity of the transmitter structure. Also, the signal envelope of a single OFDM symbol is still constant. The simplified synthesis of the transmit time signal for the multi-user case is depicted in Figure 20. First, the vector of L modulation symbols $\vec{D}^{(m)}$ is calculated and repeated M-times on the time axis. Then, the time sequence is elementwise multiplied by the user-specific, phase rotating sequence $e^{j2\pi im/N_C}$. In the last step, the guard interval is added. Figure 20 describes the simple transmit signal processing and the low computation complexity at the transmitter side and in the mobile terminal for the uplink case.

Figure 19 and Figure 20 show that the time signal of the OFDM-FDMA based uplink scheme with DFT spreading can be considered as a blockwise single carrier periodic transmission system where a cyclic prefix is integrated into a single block as a guard interval. Therefore, the signal envelope is nearly constant and additional techniques like $\pi / 4$ – QPSK can be employed to even reduce the resulting small PAR for this single carrier system. An additional advantage of this OFDM-FDMA based uplink system is the flexible use of sub-carrier allocation process and data rate adaptation for a certain user.

OFDM-BASED AND SYNCHRONIZED CELLULAR NETWORK

In the preceding sections, several uplink- and downlink-schemes for the connection between mobile terminals and base stations were discussed. In this section, the focus will be broadened from

individual links to the overall cellular network. In this context, resource allocation and synchronization of the network play an important role.

As before, the OFDM receiver in a cell has to deal with ISI effects, which occur in multi-path propagation situations in broadband radio channels. In a sufficiently designed OFDM system, these effects can be completely avoided.

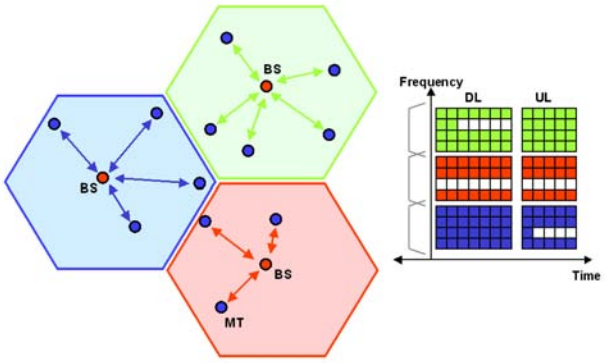
Consequently, the OFDM receiver can also deal with superimposed signals which have been transmitted by several distinct and adjacent base stations (BS) in a cellular environment, if the cellular network is synchronized in time and carrier frequency. All adjacent BS operate simultaneously in the same frequency band which leads to a reuse factor of 1.

In current cellular radio networks each base station assigns resources independently and exclusively to its users. To be able to use a TDMA or FDMA multiple access scheme in a cellular environment, an off-line radio resource planning is required to avoid co-channel interference situations between adjacent cells. As a consequence only a small fraction of the available resource determined by the spatial reuse is assigned to each cell which can dynamically be accessed by its users (Zander & Kim, 2001). However, due to this fixed resource distribution among adjacent cells, a dynamic and flexible shift of resources between cells is technically difficult. Such a conventional cellular network with a fixed frequency planning is shown in Figure 21 for a time division duplex (TDD) system as an example.

By introducing the OFDM transmission technique in such a cellular environment, the limitations of fixed resource allocation can be overcome. Since the OFDM transmission technique is robust in multi-path propagation situations, a synchronized network can be established. All BS and MT are synchronized in this case and the signals from adjacent BS will be received with a mutual relative delay no longer than the guard interval. Under these synchronized network conditions each BS can use all available resources simultaneously. With this technique at hand it is possible to add an additional "macro" diversity to a cellular environment by transmitting the same signal from synchronized BS.

Synchronized networks have been intensively studied, for example, for DVB-T broadcast systems as a single frequency network (SFN). In this case the same information signal is transmitted

Figure 21. Conventional cellular network with fixed resource allocation



on the same resources from different BS. In the communication case and in a synchronized network different information signals are transmitted by the adjacent BS but all received signals can be considered as co-sub-carrier interferer which allows in general the allocation of all available resources for each BS.

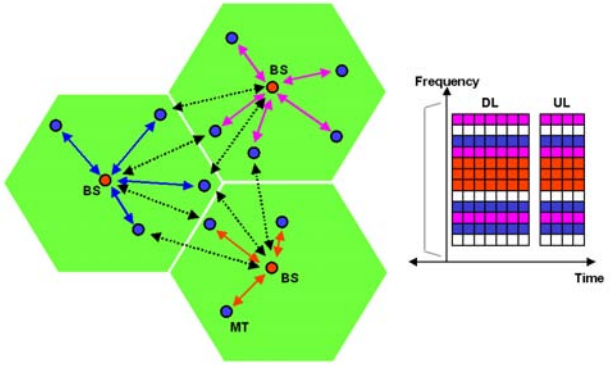
A synchronized network will also be considered to implement a dynamic resource allocation scheme by assigning different sub-carriers of an OFDM-FDMA multiple access scheme to users in adjacent cells. Since the OFDM sub-carriers remain orthogonal in a synchronized network, no MAI between different user in adjacent cells will occur as long as the sub-carriers have been allocated in an exclusive way.

All resources can be accessed in this case by the MT inside a cellular network. Synchronized networks can be used to provide the needed flexibility inside a cellular environment to allocate system resources in those cells where this bandwidth is needed. Especially for non-uniformly distributed users inside a cellular environment or for hot spot situations the system capacity can be largely increased in synchronized networks. The synchronization concept is shown in Figure 22 for an OFDM-FDMA and TDD system as an example. In this case the resource management could be based on co-channel interference (CCI) measurements processed in each BS.

Self-Organized Cell Synchronization

The dynamic sharing of all available resources between adjacent cells requires a tight time and frequency synchronization of all BS and all MT inside the cellular environment. All MTs are synchronized to a single BS using a specific test signal which is transmitted in a downlink preamble. It is assumed in this paragraph that the required network synchronization is achieved without any assistance of a central controller but in a totally decentralized and self-organized way. Furthermore a TDD system is assumed. Synchronization between adjacent BS can be achieved not in a direct way but indirectly if all MT inside a single cell transmit a specific test signal at the end of each frame (or super frame) in an uplink postamble. These different test signals transmitted from all MT in adjacent cells will be received in all BS inside a local environment. Each BS can process this information to synchronize

Figure 22. Flexible radio resource management by making all resources available to all BS in a synchronized OFDM-based cellular environment



clock and carrier simultaneously to establish a synchronized network.

Test signals will therefore be used in the down- and uplink to synchronize all BS in a local environment and all MT inside a single cell, see Figure 23. The test signal itself is designed to allow an almost interference-free time and frequency offset estimation.

To generate the test signal structure, each BS selects a single pair of two adjacent sub-carriers for each frame inside the preamble as it is shown in Figure 24. The sub-carriers inside the test signal are chosen randomly and independently by each BS from a set of allowed sub-carrier pairs placed equidistantly in the frequency band and separated by a guard band of unused sub-carriers to reduce interference in a non-synchronized situation. During the downlink preamble each BS transmits the specific test signal on the individually selected pair of sub-carriers.

In the uplink all MT inside a single cell transmit a test signal which is identical to that one they have received in the preamble at the beginning of the data frame. Each BS receives these test signals in

Figure 23. Test signal structure which is used for the synchronization of MT to a single BS during downlink and between all BS during the uplink phase

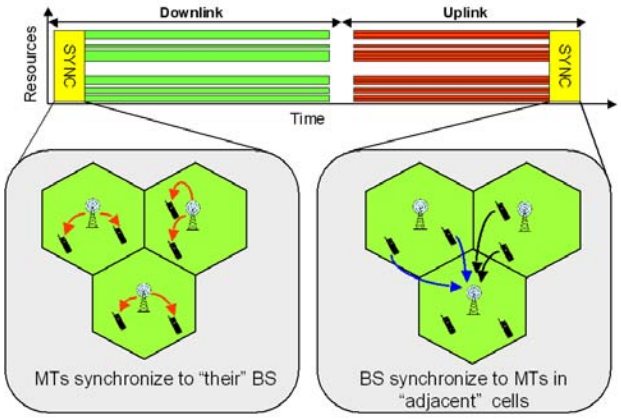
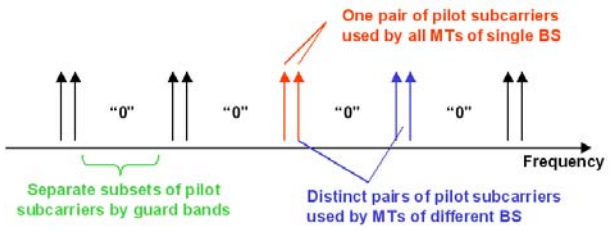


Figure 24. Sub-carrier allocation of test signals



a superimposed form from all MT inside the cell on the same sub-carrier pair. Test signals from MT in adjacent cells will be observed by the BS on distinct pairs of sub-carriers and can therefore be distinguished and processed separately. Each BS selects randomly the sub-carrier pair for the test signal in the preamble of each data frame. Therefore data collisions between test signals of adjacent BS will only occur rarely but do not influence the synchronization process at all. All received test signals are evaluated in the frequency domain as shown in Figure 25.

The signal processing and test signal evaluation is identical in the downlink and uplink. To avoid ISI and ICI during the fine synchronization procedure the test signals are designed to be phase continuous for the duration of N_p consecutive OFDM symbols as it is shown in Figure 25.

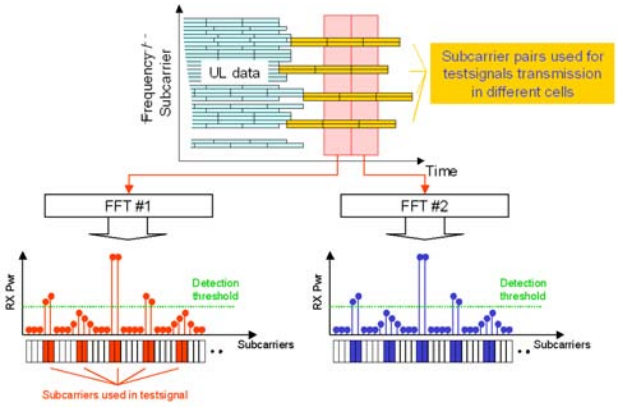
In the downlink case and in the MT synchronization phase a single FFT output signal already contains the time offset information between BS and MT in a certain phase rotation between the two considered adjacent sub-carriers. The carrier frequency offset between BS and MT is given by the phase rotation between the FFT output signals of the same sub-carrier but the two consecutive FFT.

Using this synchronization technique in each BS and MT which is based on phase difference measurements the time and frequency offset estimates are obtained simultaneously for each possible received sub-carrier pair. But only those measurements which exceed a certain amplitude threshold will be used for the subsequent adjustment of the BS time and frequency offsets.

Self-Organized Resource Management

One additional important design aspect for a 4G system is the capability to serve the time-varying data rate demands of all MT efficiently, incorporating high traffic peaks at isolated BS. Therefore dynamic channel allocation (DCA) is considered as an important feature for future networks. Centralized resource management schemes in which a central unit has the complete knowledge about the resource allocation in all cells have been

Figure 25. Time interval free of ISI and ICI between different test signals is used for the estimation of fine time and frequency offsets



investigated with respect to OFDM systems in Wahlqvist et al. (1997) and Wang et al. (2003), for example.

In the following, however, it is assumed that each BS decides in the radio resource management (RRM) procedure and in the sub-carrier allocation process in a self-organizing (SO) way without any cooperation and communication between adjacent cells and without a central management unit. Therefore, the proposed system concept is termed SO-DCA.

The assumed OFDM-FDMA and TDD scheme shown in Figure 22 is only one possible way of arranging the resource management. The SO-DCA concept can be applied for any orthogonal multiple access scheme with a TDMA and/or an

Figure 26. Each BS determines the resource allocation process by measuring the signal power on all sub-carriers inside the available bandwidth

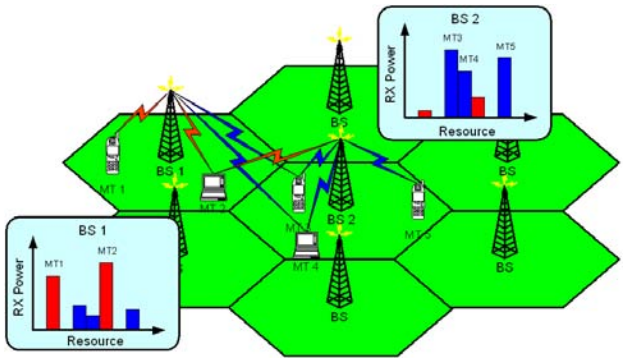
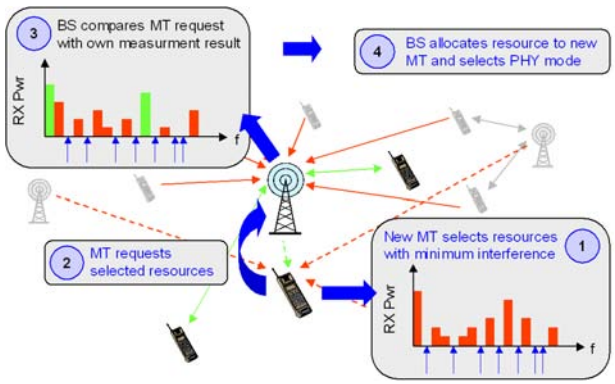


Figure 27. Resource allocation and PHY Mode selection based on the interference measurements at the MT and BS



FDMA component. A MAC frame consists of one down link (DL) and one uplink (UL) period and has a total duration of $T_F = T_{UL} + T_{DL}$.

This paragraph introduces a suitable DCA algorithm which strongly benefits from the tight synchronization between all BS and MT inside the cellular environment. The RRM process is mainly based on continuous CCI measurements in each frame. Based on the available CCI measurements (see Figure 26), each BS decides independently of other BS which resources will be covered for a new MT.

In order to increase the system throughput, a link adaptation (LA) procedure is further introduced. Each BS makes decisions about the modulation scheme and channel code rate (PHY mode) which can be used currently on the individual link. The choice of the applied PHY mode is derived from the radio channel measurement. The main task for the DCA algorithm is to assign a sufficient number of sub-carrier resources to a specific user (MT) to satisfy the current quality of service (QoS) demand. The sub-carrier selection process in the DCA procedure is important to allocate those sub-carriers which are less attenuated by the radio channel. This selection process is mainly based on CCI measurements in the MT and BS. The resource allocation process is summarized in Figure 27.

CONCLUSION

Some aspects for future mobile communication networks have been considered in this chapter. The OFDM transmission technique itself has a large potential due to the robust behaviour in wideband frequency selective and time variant radio channels. The combination with multiple access schemes showed good performance under realistic channel assumptions. A system proposal for an air interface structure for downlink and uplink has been discussed. Future cellular networks require high flexibility for data sources with different and time variant data rate in multi-path propagation environments. Therefore a synchronized cellular network has been proposed and the completely decentralized and self-organized time and carrier synchronization aspects have been discussed. Finally, a self-organized RRM has been proposed to establish a totally decentralized organization inside each BS. All these different techniques and technical concepts can be combined in a way to establish a future powerful and flexible mobile communications network for the 4G.

REFERENCES

Aldinger, M. (1994). Multicarrier COFDM scheme in high bitrate radio local area networks. *Proc. of Wireless Computer Networks 94*, Den Haag, Netherlands (pp. 969-973). New York: IEEE.

Bello, P. A. (1963). Characterization of randomly time-variant linear channels. *IEEE Transactions on Communications, 11,* 360-393.

Bingham, J. (1990, May). Multicarrier modulation for data transmission: An idea whose time has come. *IEEE Communications Magazine, 28*, 5-14.

Brüninghaus, K., & Rohling, H. (1997). On the duality of multi-carrier spread spectrum and single-carrier transmission. *Zweites OFDM-Fachgespräch*, Braunschweig, Germany (pp. 210-215). Braunschweig: TU Braunschweig.

Brüninghaus, K., & Rohling, H. (1998). Multicarrier spread spectrum and its relationship to single carrier transmission. *Proc. of the IEEE VTC'98*, Ottawa, Canada (pp. 2329-2332). New York: IEEE.

Chang, R. W. (1966). Synthesis of band-limited orthogonal signals for multichannel data transmission. *Bell Syst. Tech. J., 45*, 1775-1796.

Classen, F., & Meyr, H. (1994). Frequency synchronization algorithms for OFDM systems suit-

able for communication over frequency selective fading channels. *Proc. IEEE VTC 94*, Stockholm, Sweden (pp. 1655-1659). New York: IEEE.

Galda, D., Rohling, H., Costa, E., Haas, H., & Schulz, E. (2002). A low complexity transmitter structure for the OFDM-FDMA uplink. *Proc. IEEE VTC'02 Spring*, Birmingham, Alabama, May (pp. 1024-1028). New York: IEEE.

Gross, J., Karl, H., Fitzek, F., & Wolisz, A. (2003). Comparison of heuristic and optimal subcarrier assignment algorithms. *Proc. of Intl. Conf. on Wireless Networks (ICWN)*, Las Vegas, Nevada (pp. 249-255). Las Vegas: CSREA Press.

Hanzo, L. et al. (2003). *OFDM and MC-CDMA for broadband multi-user communications, WLANs and broadcasting.* New York: Wiley.

Kaiser, S. (1998). *Multi-carrier CDMA mobile radio systems: Analysis and optimization of detection, decoding and channel estimation.* Fortschritt-Berichte VDI, Reihe 10, Nr. 531, VDI-Verlag, Düsseldorf, Germany.

Kaiser, S. (2002). OFDM code-division multiplexing in fading channels. *IEEE Trans. on Communications, 50*, 1266-1273.

Linnartz, J. P. (2000). Synchronous MC-CDMA in dispersive, mobile rayleigh channels. *Proc. of 2nd IEEE Benelux Signal Processing Symposium (SPS-2000)*, Hilvarenbeek, The Netherlands (pp. 1-4). New York: IEEE.

Mizoguchi, M. et al. (1998). A fast burst synchronization scheme for OFDM. *Proc ICUPC 98*, Florence, Italy (pp. 125-129). New York: IEEE.

Pätzold, M. (2002). *Mobile fading channels.* New York: Wiley.

Peled, A., & Ruiz, A. (1980). Frequency domain data transmission using reduced computational complexity algorithms. *Proc. IEEE ICASSP,* Denver, Colorado (pp. 964-967). New York: IEEE.

Rohling, H., Galda, D., & Schulz, E. (2004). An OFDM based cellular single frequency communication network. *Proc. of the Wireless World Research Forum '04*, Beijing, China (pp. 254-258). Zurich: WWRF.

Rohling, H., & Grünheid, R. (1997). Performance comparison of different multiple access schemes for the downlink of an OFDM communication system. *Proc. IEEE VTC'97,* Phoenix, Arizona (pp. 1365-1369). New York: IEEE.

Saltzberg, B. R. (1967). Performance of an efficient parallel data transmission system. *IEEE Trans. on Communications, 15*, 805-811.

Shen, M., Li, G., & Liu, H. (2004). Design tradeoffs in OFDMA traffic channels. *Proc. of IEEE ICASSP '04,* Montreal, Canada (pp. 757-760). New York: IEEE.

Toufik, I., & Knopp, R. (2004). Channel allocation algorithms for multi-carrier systems. *Proc. of the IEEE VTC '04*, Los Angeles, CA, September (pp. 1129-1133). New York: IEEE.

Wahlqvist, M. et al. (1997). Capacity comparison of an OFDM based multiple access system using different dynamic resource allocation. *Proc. IEEE VTC'97*, Phoenix, Arizona (pp. 1664-1668). New York: IEEE.

Wang, W. et al. (2003). Impact of multiuser diversity and channel variability on adaptive OFDM. *Proc. IEEE VTC 2003 Fall*, Orlando, Florida, October (pp. 547-551). New York: IEEE.

Weinstein, S. B., & Ebert, P. M. (1971). Data transmission by frequency-division multiplexing using the discrete fourier transform. *IEEE Transactions on Communication Technology, 19*, 628-634.

Zander, J., & Kim, S. L. (2001). *Radio resource management for wireless networks.* London: Artech House Publishers, Mobile Communications Series.

This work was previously published in Mobile Multimedia Communications: Concepts, Applications, and Challenges, edited by G. Karmakar and L. Dooley, pp. 151-177, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Chapter 8.25
Malicious Software in Mobile Devices

Thomas M. Chen
Southern Methodist University, USA

Cyrus Peikari
Airscanner Mobile Security Corporation, USA

ABSTRACT

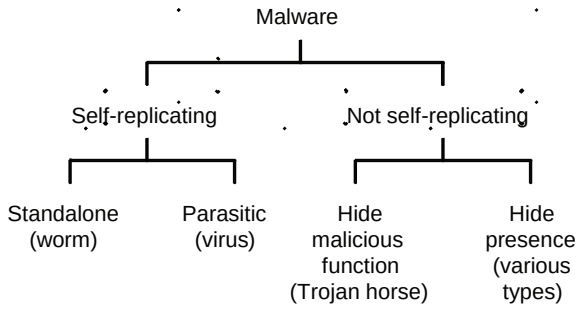
This chapter examines the scope of malicious software (malware) threats to mobile devices. The stakes for the wireless industry are high. While malware is rampant among 1 billion PCs, approximately twice as many mobile users currently enjoy a malware-free experience. However, since the appearance of the Cabir worm in 2004, malware for mobile devices has evolved relatively quickly, targeted mostly at the popular Symbian smartphone platform. Significant highlights in malware evolution are pointed out that suggest that mobile devices are attracting more sophisticated malware attacks. Fortunately, a range of host-based and network-based defenses have been developed from decades of experience with PC malware. Activities are underway to improve protection of mobile devices before the malware problem becomes catastrophic, but developers are limited by the capabilities of handheld devices.

INTRODUCTION

Most people are aware that malicious software (malware) is an ongoing widespread problem with Internet-connected PCs. Statistics about the prevalence of malware, as well as personal anecdotes from affected PC users, are easy to find. PC malware can be traced back to at least the Brain virus in 1986 and the Robert Morris Jr. worm in 1988. Many variants of malware have evolved over 20 years. The October 2006 WildList (www.wildlist.org) contained 780 viruses and worms found to be spreading "in the wild" (on real users' PCs), but this list is known to comprise a small subset of the total number of existing viruses. The prevalence of malware was evident in a 2006 CSI/FBI survey where 65% of the organizations reported being hit by malware, the single most common type of attack.

A taxonomy to introduce definitions of malware is shown in Figure 1, but classification is

Figure 1. A taxonomy of malicious software



sometimes difficult because a piece of malware often combines multiple characteristics. Viruses and worms are characterized by the capability to self-replicate, but they differ in their methods (Nazario, 2004; Szor, 2005). A virus is a piece of software code (set of instructions but not a complete program) attached to a normal program or file. The virus depends on the execution of the host program. At some point in the execution, the virus code hijacks control of the program execution to make copies of itself and attach these copies to more programs or files. In contrast, a worm is a stand-alone automated program that seeks vulnerable computers through a network and copies itself to compromised victims.

Non-replicating malware typically hide their presence on a computer or at least hide their malicious function. Malware that hides a malicious function but not necessarily its presence is called a Trojan horse (Skoudis, 2004). Typically, Trojan horses pose as a legitimate program (such as a game or device driver) and generally rely on social engineering (deception) because they are not able to self-replicate. Trojan horses are used for various purposes, often theft of confidential data, destruction, backdoor for remote access, or installation of other malware. Besides Trojan horses, many types of non-replicating malware hide their presence in order to carry out a malicious function on a victim host without detection and removal by the user. Common examples include bots and spyware. Bots are covertly installed software that secretly listen for remote commands, usually sent through Internet relay chat (IRC) channels, and execute them on compromised computers. A group of compromised computers under remote control of a single "bot herder" constitute a bot net. Bot nets are often used for spam, data theft, and distributed denial of service attacks. Spyware collects personal user information from a victim computer and transmits the data across the network, often for advertising purposes but possibly for data theft. Spyware is often bundled with shareware or installed covertly through social engineering.

Since 2004, malware has been observed to spread among smartphones and other mobile devices through wireless networks. According to F-Secure, the number of malware known to target smartphones is approximately 100 (Hypponen, 2006). However, some believe that malware will inevitably grow into a serious problem (Dagon, Martin, & Starner, 2004). There have already been complex, blended malware threats on mobile

devices. Within a few years, mobile viruses have grown in sophistication in a way reminiscent of 20 years of PC malware evolution. Unfortunately, mobile devices were not designed for security, and they have limited defenses against continually evolving attacks.

If the current trend continues, malware spreading through wireless networks could consume valuable radio resources and substantially degrade the experience of wireless subscribers. In the worst case, malware could become as commonplace in wireless networks as in the Internet with all its attendant risks of data loss, identity theft, and worse. The wireless market is growing quickly, but negative experiences with malware on mobile devices could discourage subscribers and inhibit market growth. The concern is serious because wireless services are currently bound to accounting and charging mechanisms; usage of wireless services, whether for legitimate purposes or malware, will result in subscriber charges. Thus, a victimized subscriber will not only suffer the experience of malware but may also get billed extra service charges. This usage-based charging arrangement contrasts with PCs which typically have flat charges for Internet communications.

This chapter examines historical examples of malware and the current environment for mobile devices. Potential infection vectors are explored. Finally, existing defenses are identified and described.

BACKGROUND

Mobile devices are attractive targets for several reasons (Hypponen, 2006). First, mobile devices have clearly progressed far in terms of hardware and communications. PDAs have grown from simple organizers to miniature computers with their own operating systems (such as Palm or Windows Pocket PC/Windows Mobile) that can download and install a variety of applications. Smartphones combine the communications capabilities of cell phones with PDA functions. According to Gartner, almost 1 billion cell phones will be sold in 2006. Currently, smartphones are a small fraction of the overall cell phone market. According to the *Computer Industry Almanac*, 69 million smartphones will be sold in 2006. However, their shipments are growing rapidly, and IDC predicts smartphones will become 15% of all mobile phones by 2009. Approximately 70% of all smartphones run the Symbian operating system, made by various manufacturers, according to Canalys. Symbian is jointly owned by Sony Ericsson, Nokia, Panasonic, Samsung, and Siemens AG. Symbian is prevalent in Europe and Southeast Asia but less common in North America, Japan, and South Korea. The Japanese and Korean markets have been dominated by Linux-based phones. The North American market has a diversity of cellular platforms.

Nearly all of the malware for smartphones has targeted the Symbian operating system. Descended from Psion Software's EPOC, it is structured similar to desktop operating systems. Traditional cell phones have proprietary embedded operating systems which generally accept only Java applications. In contrast, Symbian application programming interfaces (APIs) are publicly documented so that anyone can develop applications. Applications packaged in SIS file format can be installed at any time, which makes Symbian devices more attractive to both consumers and malware writers.

Mobile devices are attractive targets because they are well connected, often incorporating various means of wireless communications. They are typically capable of Internet access for Web browsing, e-mail, instant messaging, and applications similar to those on PCs. They may also communicate by cellular, IEEE 802.11 wireless LAN, short range Bluetooth, and short/multimedia messaging service (SMS/MMS).

Another reason for their appeal to malware writers is the size of the target population. There were more than 900 million PCs in use worldwide

in 2005 and will climb past 1 billion PCs in 2007, according to the *Computer Industry Almanac*. In comparison, there were around 2 billion cellular subscribers in 2005. Such a large target population is attractive for malware writers who want to maximize their impact.

Malware is relatively unknown for mobile devices today. At this time, only a small number of families of malware have been seen for wireless devices, and malware is not a prominent threat in wireless networks. Because of the low threat risk, mobile devices have minimal security defenses. Another reason is the limited processing capacity of mobile devices. Whereas desktop PCs have fast processors and plug into virtually unlimited power, mobile devices have less computing power and limited battery power. Protection such as antivirus software and host-based intrusion detection would incur a relatively high cost in processing and energy consumption. In addition, mobile devices were never designed for security. For example, they lack an encrypting file system, Kerberos authentication, and so on. In short, they are missing all the components required to secure a modern, network-connected computing device.

There is a risk that mobile users may have a false sense of security. Physically, mobile devices feel more personal because they are carried everywhere. Users have complete physical control of them, and hence they feel less accessible to intruders. This sense of security may lead users to trust the devices with more personal data, increasing the risk of loss and appeal to attackers. Also, the sense of security may lead users to neglect security precautions such as changing default security configurations.

Although mobile devices might be appealing targets, there are certain drawbacks to malware for mobile devices. First, mobile devices usually have intermittent connectivity to the network or other devices, in order to save power. This fact limits the ability of malware to spread quickly. Second, if malware is intended to spread by Bluetooth, Bluetooth connections are short range. Moreover, Bluetooth devices can be turned off or put into hidden mode. Third, there is a diversity of mobile device platforms, in contrast to PCs that are dominated by Windows. Some have argued that the Windows monoculture in PCs has made PCs more vulnerable to malware. To reach a majority of mobile devices, malware writers must create separate pieces of malware code for different platforms (Leavitt, 2005).

EVOLUTION OF MALWARE

Malware has already appeared on mobile devices over the past few years (Peikari & Fogie, 2003). While the number is still small compared to the malware families known for PCs, an examination of prominent examples shows that malware is evolving steadily. The intention here is not to exhaustively list all examples of known malware but to highlight how malware has been developing.

Palm Pilots and Windows Pocket PCs were common before smartphones, and malware appeared first for the Palm operating system. Liberty Crack was a Trojan horse related to Liberty, a program emulating the Nintendo Game Boy on the Palm, reported in August 2000 (Foley & Dumigan, 2001). As a Trojan, it did not spread by self-replication but depended on being installed from a PC that had the "liberty_1_1_crack.prc" file. Once installed on a Palm, it appears on the display as an application, Crack. When executed, it deletes all applications from the Palm (www.f-secure.com/v-descs/lib_palm.shtml).

Discovered in September 2000, Phage was the first virus to target Palm PDAs (Peikari & Fogie, 2003). When executed, the virus infects all third-party applications by overwriting them (http://www.f-secure.com/v-descs/phage.shtml). When a program's icon is selected, the display turns gray and the selected program exits. The virus can spread directly to other Palms by infrared beaming or indirectly through PC synchronization.

Another Trojan horse discovered around the same time, Vapor is installed on a Palm as the application "vapor.prc" (www.f-secure.com/v-descs/vapor.shtml). When executed, it changes the file attributes of other applications, making them invisible (but not actually deleting them). It does not self-replicate.

In July 2004, Duts was a proof-of-concept virus, the first to target Windows Pocket PCs. It asks the user for permission to install. If installed, it attempts to infect all EXE files larger than 4096 bytes in the current directory.

Later in 2004, Brador was a backdoor for Pocket PCs (www.f-secure.com/v-descs/brador.shtml). It installs the file "svchost.exe" in the Startup directory so that it will automatically start during the device bootup. Then it will read the local host IP address and e-mail that to the author. After e-mailing its IP address, the backdoor opens a TCP port and starts listening for commands. The backdoor is capable of uploading and downloading files, executing arbitrary commands, and displaying messages to the PDA user.

The Cabir worm discovered in June 2004 was a milestone marking the trend away from PDAs and towards smartphones running the Symbian operating system. Cabir was a proof-of-concept worm, the first for Symbian, written by a member of a virus writing group 29A (www.f-secure.com/v-descs/cabir.shtml). The worm is carried in a file "caribe.sis" (Caribe is Spanish for the Caribbean). The SIS file contains autostart settings that will automatically execute the worm after the SIS file is installed. When the Cabir worm is activated, it will start looking for other (discoverable) Bluetooth devices within range. Upon finding another device, it will try to send the caribe.sis file. Reception and installation of the file requires user approval after a notification message is displayed. It does not cause any damage.

Cabir was not only one of the first malware for Symbian, but it was also one of the first to use Bluetooth (Gostev, 2006). Malware is more commonly spread by e-mail. The choice of Bluetooth meant that Cabir would spread slowly in the wild. An infected smartphone would have to discover another smartphone within Bluetooth range and the target's user would have to willingly accept the transmission of the worm file while the devices are within range of each other.

In August 2004, the first Trojan horse for smartphones was discovered. It appeared to be a cracked version of a Symbian game Mosquitos. The Trojan made infected phones send SMS text messages to phone numbers resulting in charges to the phones' owners.

In November 2004, the Trojan horse—Skuller—was found to infect Symbian Series 60 smartphones (www.f-secure.com/v-descs/skulls.shtml). The Trojan is a file named "Extended theme.SIS," a theme manager for Nokia 7610 smartphones. If executed, it disables all applications on the phone and replaces their icons with a skull and crossbones. The phone can be used to make calls and answer calls. However, all system applications such as SMS, MMS, Web browsing, and camera do not work.

In December 2004, Skuller and Cabir were merged to form Metal Gear, a Trojan horse that masquerades as the game of the same name. Metal Gear uses Skulls to deactivate a device's antivirus. This was the first malware to attack antivirus on Symbian smartphones. The malware also drops a file "SEXXXY.SIS," an installer that adds code to disable the handset menu button. It then uses Cabir to send itself to other devices.

Locknut was a Trojan horse discovered in February 2005 that pretended to be a patch for Symbian Series 60 phones. When installed, it drops a program that will crash a critical system service component, preventing any application from launching.

In March 2005, ComWar or CommWarrior was the first worm to spread by MMS among Symbian Series 60 smartphones. Like Cabir, it was also capable of spreading by Bluetooth. Infected phones will search for discoverable Bluetooth devices within range; if found, the infected phone

will try to send the worm in a randomly named SIS file. But Bluetooth is limited to devices within 10 meters or so. MMS messages can be sent to anywhere in the world. The worm tries to spread by MMS messaging to other phone owners found in the victim's address book. MMS has the unfortunate side effect of incurring charges for the phone owner.

Drever was a Trojan horse that attacked antivirus software on Symbian smartphones. It drops non-functional copies of the bootloaders used by Simworks Antivirus and Kaspersky Symbian Antivirus, preventing these programs from loading automatically during the phone bootup.

In April 2005, the Mabir worm was similar to Cabir in its ability to spread by Bluetooth. It had the additional capability to spread by MMS messaging. It listens for any arriving MMS or SMS message and will respond with a copy of itself in a file named "info.sis."

Found in September 2005, the Cardtrap Trojan horse targeted Symbian 60 smartphones and was one of the first examples of smartphone malware capable of infecting a PC (www.f-secure.com/v-descs/cardtrap_a.shtml). When it is installed on the smartphone, it disables several applications by overwriting their main executable files. More interestingly, it also installs two Windows worms, Padobot.Z and Rays, to the phone's memory card. An autorun file is copied with the Padobot.Z worm, so that if the memory card is inserted into a PC, the autorun file will attempt to execute the Padobot worm. The Rays worm is a file named "system.exe" which has the same icon as the system folder in the memory card. The evident intention was to trick a user reading the contents of the card on a PC into executing the Rays worm.

Crossover was a proof-of-concept Trojan horse found in February 2006. It was reportedly the first malware capable of spreading from a PC to a Windows Mobile Pocket PC by means of ActiveSync. On the PC, the Trojan checks the version of the host operating system. If it is not Windows CE or Windows Mobile, the virus makes a copy of itself on the PC and adds a registry entry to execute the virus during PC rebooting. A new virus copy is made with a random file name at each reboot. When executed, the Trojan waits for an ActiveSync connection, when it copies itself to the handheld, documents on the handheld will be deleted.

In August 2006, the Mobler worm for Windows PCs was discovered (www.f-secure.com/v-descs/mobler.shtml). It is not a real threat but is suggestive of how future malware might evolve. When a PC is infected, the worm copies itself to different folders on local hard drives and writable media (such as a memory card). Among its various actions, the worm creates a SIS archiver program "makesis.exe" and a copy of itself named "system.exe" in the Windows system folder. It also creates a Symbian installation package named "Black_Symbian.SIS." It is believed to be capable of spreading from a PC to smartphone, another example of cross-platform malware.

At the current time, it is unknown whether Crossover and Mobler signal the start of a new trend towards cross-platform malware that spread equally well among PCs and mobile devices. The combined potential target population would be nearly 3 billion. The trend is not obvious yet but Crossover and Mobler suggest that cross-platform malware could become possible in the near future.

INFECTION VECTORS

Infection vectors for PC malware have changed over the years as PC technology evolved. Viruses initially spread by floppy disks. After floppy disks disappeared and Internet connectivity became ubiquitous, worms spread by mass e-mailing. Similarly, infection vectors used by malware for mobile devices have changed over the past few years.

Synchronization: Palm and Windows PDAs were popular before smartphones. PDAs install

software by synchronization with PCs (Foley & Dumigan, 2001). For example, Palm applications are packaged as Palm resource (PRC) files installed from PCs. As seen earlier, Palm malware usually relied on social engineering to get installed. This is a slow infection vector for malware to spread between PDAs because it requires synchronization with a PC and then contact with another PC that synchronizes with another PDA. Much faster infection vectors became possible when PDAs and then smartphones started to feature communications directly between mobile devices without having to go through PCs.

E-mail and Web: Internet access from mobile devices allows users away from their desktops to use the most common Internet applications, e-mail and the World Wide Web. Most mobile devices can send and receive e-mail with attachments. In addition, many can access the Web through a microbrowser designed to render Web content on the small displays of mobile devices. Current microbrowsers are similar in features to regular Web browsers, capable of HTML, WML, CSS, Ajax, and plug-ins. Although e-mail and the Web are common vectors for PC malware, they have not been used as vectors to infect mobile devices thus far.

SMS/MMS messaging: Commonly called text messaging, SMS is available on most mobile phones and Pocket PCs. It is most popular in Europe, Asia (excluding Japan), Australia, and New Zealand, but has not been as popular in the U.S. as other types of messaging. Text messaging is often used to interact with automated systems, for example to order products or services or participate in contests. Short messages are limited to 140 bytes of data, but longer content can be segmented and sent in multiple messages. The receiving phone is responsible for reassembling the complete message. Short messages can also be used to send binary content such as ringtones or logos. While SMS is largely limited to text, MMS is a more advanced messaging service allowing transmission of multimedia objects—video, images, audio, and rich text. The ComWar worm was the first to spread by MMS (among Symbian Series 60 smartphones). MMS has the potential to spread quickly. ComWar increased its chances by targeting other phone owners found in the victim's address book. By appearing to come from an acquaintance, an incoming message is more likely to be accepted by a recipient. MMS will likely continue to be an infection vector in the future.

Bluetooth: Bluetooth is a short-range radio communication protocol that allows Bluetooth-enabled devices (which could be mobile or stationary) within 10-100 meters to discover and talk with each other. Up to eight devices can communicate with each other in a piconet, where one device works in the role of "master" and the others in the role of "slaves." The master takes turns to communicate with each slave by round robin. The roles of master and slaves can be changed at any time.

Each Bluetooth device has a unique and permanent 48-bit address as well as a user-chosen Bluetooth name. Any device can search for other nearby devices, and devices configured to respond will give their name, class, list of services, and technical details (e.g., manufacturer, device features). If a device inquires directly at a device's address, it will always respond with the requested information.

In May 2006, F-Secure and Secure Networks conducted a survey of discoverable Bluetooth devices in a variety of places in Italy. They found on average 29 to 154 Bluetooth devices per hour in discoverable mode in the different places. In discoverable mode, the devices are potentially open to attacks. About 24% were found to have visible OBEX push service. This service is normally used for transfer of electronic business cards or similar information, but is known to be vulnerable to a BlueSnarf attack. This attack allows connections to a cellular phone and access to the phone book and agenda without authorization. Another vulnerability is BlueBug, discovered

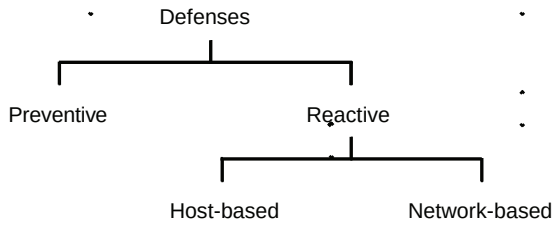
in March 2004, allowing access to the ASCII Terminal (AT) commands of a cell phone. These set of commands are common for configuration and control of telecommunications devices, and give high-level control over call control and SMS messaging. In effect, these can allow an attacker to use the phone services without the victim's knowledge. This includes incoming and outgoing phone calls and SMS messages.

The Cabir worm was the first to use Bluetooth as a vector. Bluetooth is expected to be a slow infection vector. An infected smartphone would have to discover another smartphone within a 10-meter range, and the target's user would have to willingly accept the transmission of the worm file while the devices are within range of each other. Moreover, although phones are usually shipped with Bluetooth in discoverable mode, it is simple to change devices to invisible mode. This simple precaution would make it much more difficult for malware.

MALWARE DEFENSES

Practical security depends on multiple layers of protection instead of a single (hopefully perfect) defense (Skoudis, 2004). Fortunately, various defenses against malware have been developed from decades of experience with PC malware. A taxonomy of malware defenses is shown in Figure 2. Defenses can be first categorized as preventive or reactive (defensive). Preventive techniques help avoid malware infections through identification and remediation of vulnerabilities, strengthening security policies, patching operating systems and applications, updating antivirus signatures, and even educating users about best practices (in this case, for example, turning off Bluetooth except when needed, rejecting installation of unknown software, and blocking SMS/MMS messages from untrusted parties). At this time, simple preventive techniques are likely to be very effective because there are relatively few threats that really spread in the wild. In particular, education to raise user awareness would be effective against social engineering, one of the main infection vectors used by malware for mobile devices so far.

Figure 2. A taxonomy of malware defenses



Host-Based Defenses

Even with the best practices to avoid infections, reactive defenses are still needed to protect mobile devices from actual malware threats. Reactive defenses can operate in hosts (mobile devices) or within the network. Host-based defenses make sense because protection will be close to the targets. However, host-based processes (e.g., antivirus programs) consume processing and power resources that are more critical on mobile devices than desktop PCs. Also, the approach is difficult to scale to large populations if software must be installed, managed, and maintained on every mobile device. Network-based defenses are more scalable in the sense that one router or firewall may protect a group of hosts. Another reason for network-based defenses is the possibility that the network might be able to block malware before it actually reaches a targeted device, which is not possible with host-based defenses. Host-based defenses take effect after contact with the host. In practice, host-based and network-based defenses are both used in combination to realize their complementary benefits.

The most obvious host-based defense is antivirus software (Szor, 2005). Antivirus does auto-

matic analysis of files, communicated messages, and system activities. All commercial antivirus programs depend mainly on malware signatures which are sets of unique characteristics associated with each known piece of malware. The main advantage of signature-based detection is its accuracy in malware identification. If a signature is matched, then the malware is identified exactly and perhaps sufficiently for disinfection. Unfortunately, signature-based detection has two drawbacks. First, antivirus signatures must be regularly updated. Second, there will always be the possibility that new malware could escape detection if it does not have a matching signature. For that case, antivirus programs often include heuristic anomaly detection which detects unusual behavior or activities. Anomaly detection does not usually identify malware exactly, only the suspicion of the presence of malware and the need for further investigation. For that reason, signatures will continue to be the preferred antivirus method for the foreseeable future.

Several antivirus products are available for smartphones and PDAs. In October 2005, Nokia and Symantec arranged for Nokia to offer the option of preloading Symbian Series 60 smartphones with Symantec Mobile Security Antivirus. Other commercial antivirus packages can be installed on Symbian or Windows Mobile smartphones and PDAs.

In recognition that nearly all smartphone malware has targeted Symbian devices, a great amount of attention has focused on the vulnerabilities of that operating system. It might be argued that the system has a low level of application security. For example, Symbian allows any system application to be rewritten without requiring user consent. Also, after an application is installed, it has total control over all functions. In short, applications are totally trusted.

Although Windows CE has not been as popular a target, it has similar vulnerabilities. There are no restrictions on applications; once launched, an application has full access to any system function including sending/receiving files, phone functions, multimedia functions, and so forth. Moreover, Windows CE is an open platform and application development is relatively easy.

Symbian OS version 9 added the feature of code signing. Currently all software must be manually installed. The installation process warns the user if an application has not been signed. Digital signing makes software traceable to the developer and verifies that an application has not been changed since it left the developer. Developers can apply to have their software signed via the Symbian Signed program (www.symbiansigned.com). Developers also have the option of self-signing their programs. Any signed application will install on a Symbian OS phone without showing a security warning. An unsigned application can be installed with user consent, but the operating system will prevent it from doing potentially damaging things by denying access to key system functions and data storage of other applications.

Network-Based Defenses

Network-based defenses depend on network operators monitoring, analyzing, and filtering the traffic going through their networks. Security equipment include firewalls, intrusion detection systems, routers with access control lists (ACLs), and antivirus running in e-mail servers and SMS/MMS messaging service centers. Traffic analysis is typically done by signature-based detection, similar in concept to signature-based antivirus, augmented with heuristic anomaly based detection. Traffic filtering is done by configuring firewall and ACL policies.

An example is Sprint's Mobile Security service announced in September 2006. This is a set of managed security services for mobile devices from handhelds to laptops. The service includes protection against malware attacks. The service can scan mobile devices and remove detected malware automatically without requiring user action.

In the longer term, mobile device security may be driven by one or more vendor groups working to improve the security of wireless systems. For instance, the Trusted Computing Group (TCG) (www.trustedcomputinggroup.org) is an organization of more than 100 component manufacturers, software developers, networking companies, and service providers formed in 2003. One subgroup is working on a set of specifications for mobile phone security (TCG, 2006a). Their approach is to develop a Mobile Trusted Module (MTM) specification for hardware to support features similar to those of the Trusted Platform Module (TPM) chip used in computers but with additional functions specifically for mobile devices. The TPM is a tamper-proof chip embedded at the PC board level, serving as the "root of trust" for all system activities. The MTM specification will integrate security into smartphones' core operations instead of adding as applications.

Another subgroup is working on specifications for Trusted Network Connect (TCG, 2006b). All hosts including mobile devices run TNC client software, which collects information about that host's current state of security such as antivirus signature updates, software patching level, results of last security scan, firewall configuration, and any other active security processes. The security state information is sent to a TNC server to check against policies set by network administrators. The server makes a decision to grant or deny access to the network. This ensures that hosts are properly configured and protected before connecting to the network. It is important to verify that hosts are not vulnerable to threats from the network and do not pose a threat to other hosts. Otherwise, they will be effectively quarantined from the network until their security state is remedied. Remedies can include software patching, updating antivirus, or any other changes to bring the host into compliance with security policies.

FUTURE TRENDS

It is easy to see that mobile phones are increasingly attractive as malware targets. The number of smartphones and their percentage of overall mobile devices is growing quickly. Smartphones will continue to increase in functionalities and complexity. Symbian has been the primary target, a trend that will continue as long as it is the predominant smartphone platform. If another platform arises, that will attract the attention of malware writers who want to make the biggest impact.

The review of malware evolution suggests a worrisome trend. Since the first worm, Cabir, only three years ago, malware has advanced steadily to more infection vectors, first Bluetooth and then MMS. Recently malware has shown signs of becoming cross-platform, moving easily between mobile devices and PCs.

Fortunately, mobile security has already drawn the activities of the TCG and other industry organizations. Unlike the malware situation with PCs, the telecommunications industry has decades of experience to apply to wireless networks, and there is time to fortify defenses before malware multiplies into a global epidemic.

CONCLUSION

Malware is a low risk threat for mobile devices today, but the situation is unlikely to stay that way for long. It is evident from this review that mobile phones are starting to attract the attention of malware writers, a trend that will only get worse. At this point, most defenses are common sense practices. The wireless industry realizes that the stakes are high. Two billion mobile users currently enjoy a malware-free experience, but negative experiences with new malware could have a disastrous effect. Fortunately, a range of

host-based and network-based defenses have been developed from experience with PC malware. Activities are underway in the industry to improve protection of mobile devices before the malware problem becomes catastrophic.

REFERENCES

Dagon, D., Martin, T., & Starner, T. (2004). Mobile phones as computing devices: The viruses are coming! *IEEE Pervasive Computing, 3*(4), 11-15.

Foley, S., & Dumigan, R. (2001). Are handheld viruses a significant threat? *Communications of the ACM, 44*(1), 105-107.

Gostev, A. (2006). *Mobile malware evolution: An overview.* Retrieved from http://www.viruslist.com/en/analysis?pubid=200119916

Hypponen, M. (2006). Malware goes mobile. *Scientific American, 295*(5), 70-77.

Leavitt, N. (2005). Mobile phones: The next frontier for hackers? *Computer, 38*(4), 20-23.

Nazario, J. (2004). *Defense and detection strategies against Internet worms.* Norwood, MA: Artech House.

Peikari, C., & Fogie, S. (2003). *Maximum wireless security.* Indianapolis, IN: Sams Publishing.

Skoudis, E. (2004). *Malware: Fighting malicious code.* Upper Saddle River, NJ: Prentice Hall.

Szor, P. (2005). *The art of computer virus research and defense.* Reading, MA: Addison-Wesley.

Trusted Computing Group (TCG). (2006a). *Mobile trusted module specification.* Retrieved from https://www.trustedcomputinggroup.org/specs/mobilephone/

Trusted Computing Group (TCG). (2006b). *TCG trusted network connect TNC architecture for interoperability.* Retrieved from https://www.trustedcomputinggroup.org/groups/network/

KEY TERMS

Antivirus Software: Antivirus software is designed to detect and remove computer viruses and worms and prevent their reoccurrence.

Exploit Software: Exploit software is written to attack and take advantage of a specific vulnerability.

Malware Software: Malware software is any type of software with malicious function, including for example, viruses, worms, Trojan horses, and spyware.

Smartphone: Smartphones are devices with the combined functions of cell phones and PDAs, typically running an operating system such as Symbian OS.

Social Engineering: Social engineering is an attack method taking advantage of human nature.

Trojan Horse: A Trojan horse is any software program containing a covert malicious function.

Virus: A virus is a piece of a software program that attaches to a normal program or file and depends on execution of the host program to self-replicate and infect more programs or files.

Vulnerability: Vulnerability is a security flaw in operating systems or applications that could be exploited to attack the host.

Worm: A worm is a stand-alone malicious program that is capable of automated self-replication.

This work was previously published in Handbook of Research on Wireless Security, edited by Y. Zhang, J. Zheng, and M. Ma, pp. 1-10, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Index

A

aacPlus codec 2847
absence management system 3461
abstract data types (ADT) 341, 343
acceptance 258
acceptance model for mobile technology and services 91
access 1220
access and device management server 6
access control 2347, 2573
access cost 3238
access point agent (APA) 1230
access points (AP) 2707
access router 3135
access trees 3091
accessibility 375
accessibility and availability 253
accountability 2130, 2133, 2139, 2141
accounting 1226
achievable competitive advantages 2345
acoustic data channel 1135, 1136
active and programmable networks 643
active application 649
active badge system 3405
active distraction 2059
active mobile group 3430
active networks (AN) 642, 643, 649
active tags 3375
active/programmable node 644
activity theory (AT) 1783
ad hoc arrangements 2137
ad hoc audience response systems 1396
ad hoc coordination 2132
ad hoc environments 1110
ad hoc networks 1110, 1404, 3069
ad hoc networks, wireless 957
adaptation 792
adaptive algorithms 3209
adaptive beamfoming algorithm 562
adaptive communication environment (ACE) 601
adaptive conjoint analysis (ACA) 1872
adaptive information delivery 191
adaptive signal processing 3367
adaptivity 851
added-value services (AVS) 1573, 1579
AdminMaster agent 1448
adoption behaviour 84
adoption processes 1596
adoption strategies 84
advanced audio coding (AAC) 2847
advanced encryption standard (AES) 2681, 2769, 2787
advanced encryption standard (AES) encryption mode 2775
advanced mobile technologies 85
advanced resource discovery protocol 2957
advanced video coding (AVC) 2849
advertising 790, 791
advertising service 3214, 3219
Aether systems 255
affordability 383
agents 304, 312, 631, 1226, 1236
agent attributes 304
agent butler 883
agent charger 885
agent classification 298
agent construction tools 301
agent data integrity, protection of 307
agent evolution 887
agent fabrication 887
agent factory 884, 1649
agent immigration 886
agent integrity 2715
agent learning and maturing 627
agent migration 722
agent mobility 2345
agent monitoring protocol (AMP) 305
agent platform (AP) 851
agent security 2741, 2745, 2747, 2749, 2751
agent technologies 633, 2600
agent-based e-marketplace (AEM) 1228, 1236
agent-based marketplaces 1227
agent-based software 825
agent-based system (AS) 1228
agents-enabled electronic commerce 2344
aggregation 359, 360, 1115
aggregation relationship 3107
aggregators 703
agile software development 3325
agile usability approach 3320
aglets 1719, 1720, 2347
AirG 1680
AIRS algorithm 2897
Alexandria, Virginia 1535
Aliant Mobility 1676
ambient intelligence 499, 850, 3405, 3421

AMEC 1236
American customer satisfaction model (ACSM) 1930
American National Standard Institute (ANSI) 1446, 1453, 1462
American Standard Code for Information Interchange (ASCII) 2845
Amstrad's PenPad 144
analogue-to-digital converter (ADC) 1142
analytic hierarchy process (AHP) 377, 797, 1942
analytical model (AM) 3556
angle of arrival (AOA) 575, 1756, 1761
anonymity 2617, 2662, 2799
anonymous dynamic source routing protocol (AnonDSR) 2703
anonymous on-demand routing (ANODR) 2705
anonymous on-demand routing, discount (discount ANODR) 2705
anonymous routing protocol for mobile ad hoc networks (ARM) 2706
ANOVA tests 2055
ANSI/NISO Z39.19 1451, 1458
antecedent reference model 2813
anti-join 357
antivirus software 3598
Anycast 120
AP monitoring 3162
aphasia 3529, 3531
Apple Newton 144
application developer 1759
application distribution 459, 462, 463, 464, 465, 468, 469
application domains 1765, 3404
application domains, healthcare communities 1765
application kernel 460
application ontology 3407, 3408
application perspective layer (APL) 2854
application programming interfaces (APIs) 730, 998, 1203
application programs 1221
applications 1593, 1595, 1597, 1693
appointments and reminders application 3502
ApproximateLinearOrder algorithm 3095
architectural partitioning 1572
architectural usability 1572
area variability 3054
AR-PAD project 985, 986
AR-Phone project 986
ARPU 1338
Arthur Business Consulting 1558
artificial intelligence (AI) 599, 1683, 3453
artificial intelligence (AI) in mobile computing 3446
artificial intelligence (AI) techniques in mobile applications 3447
artificial neural networks (ANN) 3488
ask the audience 1397
ASP 1667
assignment algorithm 671
assisted global positioning system (A-GPS) 1756, 1761
assisted global positioning system (A-GPS) roadside 1686
assisted GPS (AGPS) 386
assistive technology (AT) 3270, 3362, 3367
asymmetric crypto system 1610
asymmetric digital subscriber line (ASDL) 1548
asymmetric traveling salesman problem 3116
asynchronous invalidation report 3013
asynchronous service discovery 1109
AT&T Wireless 1331, 1676
auction database 1650
auction host 1649
auction system architecture 1646
audiences 240
audio 813
audio applications 2850
audio coding 2846
audio coding schemes 2847
audio conferencing 3130
audio data transmission 3132
audio memo 1002
audio streaming 3474
audio streaming, speech recognizer 3475
Audiovox 8450 or 8455 1683
augmentative and alternative communication (AAC) 3362, 3367
augmentative and alternative communication (AAC) devices 3530, 3532
augmentative and alternative communication (AAC) tools 3532
augmented reality (AR) 987, 3329
AR metaphor, tangible 988
AR working planes techniques 943
AR, face-to-face 992
authentication 1239, 2346, 2577, 2654, 2662, 2663, 2728, 2787, 3239
authentication and key agreement (AKA) 2726, 2771
authentication centre (AuC) 2753
authentication key (AK) 2774
authentication service 1965
authentication, authorization, accounting (AAA) 2675, 2792, 2806, 3243
authentication, subscriber 2793
authenticity 2617
authorities' mobile interface architecture 1564
authorization 2347, 2579, 2654
authorization service 1965
automated product ordering 2150
automatic follow-me service 660
automatic service discovery 1576
automatic summarization 2421
automatic teller machine (ATM) 1559, 1700
automatic, speech recognition (ASR) request 3475
automation technologies 1483
autonomic computing 649, 3209
autonomic computing system 3206
autonomic infrastructures 642
autonomic manager 3206
autonomy 851, 2070, 2076, 2099
Autonomy Portal-in-a-Box 192
autonomy, Buddhist concept of 2076
autonomy, concrete version of 2073
autonomy, mobile phone and 2066
availability 383
available support 1470, 1479
available-to-promise (ATP) 1493
AvantGo 137
average revenue per user (ARPU) 87, 507, 701, 708, 2510
awareness 1360, 1366

awareness 1774, 2130, 2133, 2138, 2139, 2141

B

backend layer 460
bandwidth 595, 1557, 1652, 1698, 2584
bandwidth constraints 3243
bandwidth cost 3238
bandwidth efficiency 3138
bandwidth skimming 3072
bandwidth, low 3241
Bangeman, Martin 3456
bank agent (BA) 1230
banks as content provider 1249
base stations (BSs) 491, 651, 2774, 3014, 3134, 3138
basic telecom services (BTS) 2308
batching 3072
batteries 1189, 1191
battery power constraints 3240
beampattern 562, 563
behavior engine 603
behaviorist learning 111
behavioural control 93
Beijing 1668
belief-desire-intention (BDI) 824, 851
belief-desire-intention (BDI) agent 3450
belief-desire-intention (BDI) architecture 3454
Bell 1676
Bell Canada 1677
Bell Mobility 1676
Bell Mobility (CDMA) 1680
Bell Mobility (CDMA 1X) 1683
Bell Mobility's roadside assistance 1686
Bell-LaPadula model 2831
billing 1226, 1575
billing mechanism 708
binary runtime environment for wireless (BREW) 918, 1337
biometric user authentication 2663
bit error rate (BER) 3572
bitmap image 2848
bizware 768
BlackBerry 34, 1681, 1682
Blackboard® 1385
black-box protection 2586
Blister Entertainment Inc. 1683
blogs 821
blood donor recruitment (BDR) project 444
Bluetooth 29, 85, 150, 575, 822, 1068, 1189, 1244, 1259 1403, 2040, 2147, 2180, 2187, 2375, 2661, 3594, 3595
Bluetooth Point of Sale (B-POS) 1237, 1245
Bluetooth Point of Sale (B-POS) architecture 1239
Bluetooth Point of Sale (B-POS) secure mobile payment system 1237
Bluetooth, devices 3322
body-relative plane techniques 940
Bologna City Hall (Italy) 1565
bookmarking feature 1382, 1386
boosted trees 2881
Botfighter 2451
boundary interworking unit 655
bridging technology 1397
broadband 595, 1698
broadband Internet access 1532
broadband wireless access (BWA) 2772
broadband wireless technology 2159
broadcast 1105, 1115
broadcast disk 3033
broadcast model 1104
broadcast-based approaches 1110
broker module 569
browser 464, 748
browser software 42
building trust 2807
built-in memory model 660, 661, 671
bundling 705
Buongiorno! 1738, 1746
Buongiorno! MyAlert 1738, 1745
Buongiorno! MyAlert business model 1744
business ecosystems 700
business environment 2343
business intelligence (BI) 2369
business lines 1751
business logic 460
business logic layer 460
business models 1338, 1594, 1596, 2178, 2464
business models types 2178
business models, constructional view of 2466
business networks 701
business opportunities 84
business oriented location models (BOLMs) 2535, 2538, 2563
business priorities 1751
business processes 198, 1603, 2392, 2402, 2506
business processes reengineering 2400
business reductionism 698
business transaction 2149
business unit (BU) 2535
business-to-business (B2B) 18, 26, 39, 1226, 1615, 2093, 2369, 3424
business-to-business (B2B) application domain 2103
business-to-consumer (B2C) 18, 26, 39, 1226, 1246, 1615, 2371, 2807, 3424
business-to-consumer (B2C) e-commerce 2807, 2817
business-to-consumer (B2C) mobile commerce 2807
business-to-employee (B2E) 2203, 2204
buying policy (BP) 1231
byair.com 1667
Byzantine faults 2829

C

cache invalidation 3012, 3013, 3014, 3038
cache prefetching 3037
cache replacement 3033, 3038
cache replacement policies 3036
cache state information (CSI) 3014, 3015
cache validation 3033
caching 3013, 3032
caching for location-based services 3034
caching-efficiency-based method (CEB) 3034, 3035
calculation aggregation 360
call admission control policy 2858
call drops (CD) 791, 1173
call for proposals (CFP) 721
call forwarding 87
call push architecture 1682
call waiting 87
call-back locking (CBL) 3023, 3029
caller ID 87
Cambridge positioning systems (CPS) 1671
camera apparatus 3175

camera phones 2040
camera phones in social contexts 2027
camera phones, situated use of 2030
camera phones, social uses of 2032
camera phones usage, different spaces 2030
campaign designer 742
campus-wide wireless networks 1538
Canada 1675
Canada's mobile sector 1684
Canada's mobile sector, successful services 1679
Canadian mobile desert 1680
candidate memory cell 2900
candidate set 3209
capability 383
capable-to-promise (CTP) 1493
capsule 644
captive value networks 702
CASBA 1642
cascading style sheet (CSS) 799
case based reasoning (CBR) 3421, 3422
CBC-MAC protocol (CCMP) 2681, 2685
cell 457
cell broadcast 35
cell identifier (CID) 389
cell phone 589
cell regrouping algorithm (CEREAL) 687
cell tower technology 1533, 1686
cell-global-identity (CGI) 1761
cell-global-identity (CGI) methods 1756
cell-global-identity with timing advance (CGI-TA) 1761
cell-ID (CID) 1049
cellular architecture 1176
cellular networks 595, 704
cellular phones, smart 813, 917, 1187
cellular systems 2766
cellular systems, next generation 3204
central design record (CDR) 3326, 3327
central processing unit (CPU) 818, 834, 1190, 1443
central repository 1576
centralized service directory model 1106
Centre for Public Service Innovation 766
certificates 2828
certificate authorities (CAs) 1714, 2654, 2689, 2717, 2728, 2787
certification intermediation 1727
chaining 3072, 3076
channel capacity 595
channel choices 2463
channel code rate 3586
channel state information (CSI) 3574
charging mechanisms 2347
Chaska, Minnesota 1538
China 1665
China Mobile 1669
China Unicom 1669
China, m-payment in 1670
China's rapid mobile diffusion 1668
China's Wi-Fi market 1668
Chinese mobile market 1666, 1669
chunk information for efficient processing 779
Cingular 1331
cipher block chaining (CBC) 2774
circuit switching 1254
circuit-switched networks 700
city government wireless network initiatives 1533
city news broadcasting service 1565
City of Stockholm Executive Office 3457
City of Stockholm, Sweden 3456
civic structure 282, 283
clamshell phone designs 1337
class trust properties 2831
classification and regression trees (CART) 2864
classification and regression trees (CART) model confusion matrix 2881
classification of packets within a flow 3241
clients 240
client caching model 3033
client computers 1213
client device flexibility 384
client disconnection 3019
client energy 3018
client unit (CU) 2535
client-server 1443
client-server model 304
client-side programming 1213
clinical work practices 1429
CLIP 29
closed and proprietary model 1336
cluster analyses 700
cluster integration 788
cluster-based solutions 1111
clustering 1112, 1115, 2841
clustering critical-path (CCP) algorithm 3100
CMG interoperability 1680
co-channel interference (CCI) 3562
code division multiple access (CDMA) 29, 574, 1332, 1676, 2187, 2852, 3561
code division multiple access (CDMA) 1X 1676
code division multiple access (CDMA) 1xEV/DO 1676
code division multiple access 2000 (CDMA 2000) 1265, 1666, 2374
code division multiple access 2000 (CDMA 2000) 1xRTT 1676
code on demand 2568
coding techniques 2843
cognitive dimensions of notations 1944
coherent conceptualization 1570
coiled antenna 3374
collaboration services 1363, 1366
collaborative augmented reality 984, 986, 991, 992
collaborative capability 520
collaborative learning 111, 3551
collaborative mobile applications field study 3251, 3254, 3256, 3257, 3269
collaborative modelling scenario 948
collaborative practices, University of Lapland 1969
collaborative supply chain management 2384
collaborative systems 2021
commercial exploitation 1563
commercial short message service 86
common gateway interface (CGI) 1270
communication 240, 790
communication channel 1575
communication infrastructure technologies 85
communication primary for migration 482

communication services 1363, 1366
communication subsystem 2739
communication technologies 968
communication, securing of 2726
communicative rationality 2069
communities of practice (CoP) 1966
community snapshot generation 1776
compact flash (CF) card 1049, 2375
compact HTML (cHTML) 35, 508, 798
COMPASS system 3407
COMPASS2008 project 3406
compatibility 93
compatibility services 1727
compensatory communication devices 3532
compensatory software 3531
competition forces 1757, 2306
competitive local exchange carrier (CLECs) 1678
complementary product market 1721
complex good 697
complex system 697
component agent system 3300, 3319
composition 892
compression format 1173
compression ratio 2426
compression techniques 2843
computational optimization 3487
computational power 3471
computer aided design (CAD), desktop 945
computer aided software engineering (CASE) 2295
computer industry 701
computer supported collaborative learning (CSCL) 1968, 1974
computerized patient record system (CPRS) 1444
computer-supported collaborative learning (CSCL) 3540, 3551
computer-supported collaborative work (CSCW) 280, 287, 2021, 2026, 3540, 3551
computer-supported intentional learning environments (CSILE) 3540, 3551
concept map 782
concept match 800
Conference Assistant, The 3406
conferencing service provider 965
confidentiality 2617, 2662, 2787
conformity assessment 3184
congestion window 496
connected device configuration (CDC) 914
connected limited device configuration (CLDC) 914
connectedness 2068
connectivity 791
constancy 2810
constrained mobile environments 3031
constraint database model 342
construction at a distance (CAAD) 938
constructivist learning 111
consumer adoption barriers 1632
consumer adoption drivers 1629
consumer side antecedents 2817
consumer trust 2807
consumer trust, antecedents of 2807, 2817
consumer-driven agent-based e-marketplace 1230
consuming services 3214, 3218
contacts application 3498
content adaptation 3218
content aggregation service 1574, 1576
content delivery 241
content delivery service 1965
content distribution network (CDN) 3073, 3077
content management 1068
content management agent 583
content ontology 3409
content owners 703
content provider 1249, 1759
content provision services 1574, 1576
content repurposing 2851
content-focused market strategy 2526
context 1068, 3454
context aggregation service 1574
context and situation awareness 3414
context changes 3412
context discovery 1371
context information 1064
context metadata 1065, 1067
context mobile context-aware applications 3225
context of use 239
context provision service 1574
context subsystem 3416
context-aware applications 3405
context-aware computing 1047
context-aware retrieval (CAR) 567
context-awareness mechanisms 1567
contextual awareness 2020
contextual information 566, 568, 3430
contextualization 194, 567
continuous location change 1759
continuous proximity indexes 320
continuous queries (CQs) 315, 337, 578
continuous-loop personalization 31
control module 3364
control packets 3138
convenience 1714
conventional bank account 1708
converged network, and AAA 2794
converged network, and authentication 2797
COO 35
cooperative browsing 498
cooperative browsing, between-document 500, 501
cooperative browsing, between-image 503
cooperative browsing, between-page 502
cooperative browsing, within-document 500
cooperative browsing, within-image 503
cooperative browsing, within-page 502
coordination 2132, 2134, 2135
coordinator database 1650
co-present interaction 2040
COPS for provisioning (COPS-PR) 3241
CORBA Trading Object Service 1576
core network (CN) 2753, 2771
corporate social responsibility (CSR) 13
corporate strategy 404
correspondent node (CN) 3132, 3138
cost analysis 2988
cost structure 684
count aggregation 360
counter mode (CM) 2769

country-to-country communication 87
credibility 372, 373
credibility assurance 376
credibility evaluation 376
credibility of mobile applications 372
credibility, active 373
credibility, limitations of addressing 378
credibility, passive 373
crime, prevention 249
cross-case analysis 2453
cross-layer design 2841
cross-market differentiation by tying 2512
Crossroads Copenhagen in Denmark 3455
cryptographic protection 2830
cryptographic watermarks 2741
cryptography 312
CSG 941, 943
CSG operations 942
customer delight 1892
customer differentiation 1862
customer motivations, proposed model of 1978
customer relations 241, 3461
customer relationship management (CRM) 788, 1226, 1881, 1886, 1911, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2226, 2227, 2228, 2229, 2231, 2258, 2369
customer satisfaction 1892
customisation 1557
customised presentation data 1404
customized agent software 817, 823
Cyberguide project 3406
cyber-PR 242
cyberspace 247
cyborg 1173
cyclic redundancy code (CRC) 2661

D

DAB-IP 1144, 1145, 1146
daily data (DD) 591
data broadcasting 3079
data cabinets 1538
data collection 98, 2114
data communication services 1547, 1548
data communication, interruption of 3132
data consistency 871
data dissemination 3068
data distance 3036
data encryption standard (DES) 2774
data formats 270, 1570
data guard 877
data indexing 3111
data integrity 2787
data integrity protection protocol 309
data jumps 589
data management for mobile computing 3029
data mining, spatial 1759
data networking technologies 85
data preparation 2878
data preprocessing 2903
data protection 2691
data protection (802.11i standard) 2681
data retrieval algorithms 3118
data security 1974
data standardization 1385
data tampering 2741
data transfer 3132
data transmission speed 792, 871, 3132
data transmission, latency in 3132
data warehouse 1606
databases (DBs) 351, 354, 355, 3556
database 2 1445
database layer 460
database management system (DBMS) 313, 460, 537, 910
database operations on mobile devices 350, 355
database queries 334, 353
database server 5, 1220
database service discovery model 858, 869
DataJoiner 1444
Datang 1666
DataSearchMaster 1449
DataSearchWorker agent 1448
DAVINCI 730, 755
DAVINCI integration framework 732
DAVINCI mobile work concepts 733
DBSync 739, 739, 740
DBSync architecture 739
debit cards 1559
decentralized hospital computer program (DHCP) 1444
decision support 377, 1942
decision support framework 3422
decomposed theory of planned behaviour 90
defuzzification 2999
degree of freedom (DOF) 988
delay 2843, 2844
delay Doppler function 3563
delay variance 2843, 2844
delayed duplicate acknowledgement (DDA) 493
delayed uplink (DU) 3018
delight 1889
delivery context 373, 380, 804, 1944
demodulation 1139
demographics 1858
denial of service (DoS) attacks 647, 649, 2661, 2770, 3243
denial of service (DoS) attacks threat management 646
Department of Defense 39
Department of Health and Human Services (DHHS) 1463
derived variables 2865
description logics (DL) 800, 2959, 2968
design process 821
Desire2Learn 1385, 1386
desktop computers 2850
detailed design 1229
developing nations 1858
device convergence 1687
device heterogeneity 3218
device independent Web engineering (DIWE) 2290
device inherent speech recognition 3483
device mobility 758
device mobility during a usability evaluation 3184
device platform 295
device type 441
dhamma 2076
dialogue campaigns 1660
Differentiated Services (Diffserv) 3241
Diffserv networks 3133
Diffserv-based wireless network 3242

Diffie-Hellman (DH) key exchange 2727
Diffie-Hellman (DH) protocol 2718
diffusion 91
diffusion of innovations 1627
digital assistant 854
digital cameras 1259
digital cash 2621
digital content and services 697
digital convergence 2475
digital divide 777, 1532, 1858, 2065, 3271, 3280, 3462
digital documents 968
digital encoding 87
digital government 776
digital government information 777
digital imaging and communications in medicine (DICOM) 1270
digital information 778
digital libraries 1544
digital mobile broadcasting 2522
digital mobile networks 698
digital mobile systems, evolution of 2852
digital multimedia broadcasting (DMB) 152, 168, 2522, 2523
digital phone (GSM and PCS) 2463
digital photography 2040
digital radio 2850
digital radio mondiale (DRM) 2849
digital rights management (DRM) 1124, 1622, 2725, 2286, 2287
digital rights management (DRM) solutions, features of 1119
digital rights management (DRM) technology, mobile multimedia 1117
digital service 776
digital signatures 547, 2590
digital technologies 144
digital text 813
digital video broadcasting (DVB) 2847
direct object placement techniques 940
direct relationship (DR) 2536
direct sequence spread spectrum (DSSS) 2768
directed graph 2829
direction bit (DIRECTION) 2757
directory access protocol (DAP) 3242
directory agent (DA) 1106, 1116
direct-to-home 1675
disconnection time, client 3013
discovery algorithm 670
discovery services via search engine model (DSSEM) 859, 862
DiscoveryLink 1445
discrete event simulation 1227
discrete fourier transformation (DFT) 3480
discrete fourier transformation (DFT)-based spreading technique 3578
discrete reusable information objects 1388
display module 3364
disruptive technologies 697
distance education 1381, 1384
distance learning (d-learning) 1350, 1384
distance learning instructional materials 1382
distilled statecharts (DSC) 1229, 1236
distributed anonymous secure routing protocol (ASRP) 2707
distributed approach 3448
distributed multimedia systems 3134
distributed service directories 1107
distributed speech recognition 3476, 3551
distributed speech recognizer (DSR) 3477
distributed speech recognizer (DSR), backend 3478
distributed speech recognizer (DSR), front-end 3477
distributed system 1203
distributed user interface 498
distribution level 460
distribution models 2464
distributive context 567
distributive profile 572
divide phenomena 1563
domain name system (DNS) 2686
domain ontologies 3408
dominant players 710
Doppler frequency 3564
dot.com bust 43
double buffering 1200
Double Donut 700
Double Helix model 699
down link (DL) 595, 3586
downloading capabilities 87
downloads to portable devices 86
drill-and-practice 3529
drill-and-practice software 3531
dual-task 239
Dublin Core metadata standard 1387
DUPACK 492, 496
duplicated video packets 3138
DVB-H 1144, 1145, 1146
Dynabook 144
dynamic channel allocation (DCA) 3584
dynamic channel allocation (DCA) techniques 3562
dynamic semantic location modeling (DSLM) 2531
dynamic semantic location modeling (DSLM) ontology 2533
dynamic semantic location modeling (DSLM), framework of 2533
dynamic service offers 3404
dynamic source routing (DSR) 2700, 2997
dynamic time warping (DTW) 3484
dynamic topological changes 3243
dynamic touch 247
dynamic type warping 3484
dynamic voice 3461

E

Earcon 2059
earliest deadline first (EDF) 2998
early-stage identification 1869
ease of adoption 2017
ease of use (EOU) 1888, 2819, 2206
EasyPark 1681
e-auctions 1652
eavesdropping 2588
e-banking 1627, 2767
e-business 386, 712, 1544, 2323
e-business model 2378
ebXML 1236
e-commerce 18, 19, 24, 26, 39, 41, 44, 204, 412, 546, 1117, 1118, 1192, 1226, 1544, 1585, 1591, 1615, 1625, 1626, 1641, 1652, 1690, 1691, 1695, 1713, 1720, 1886, 1894, 2170, 2178, 2212, 2213, 2214, 2215, 2216, 2218, 2219, 2220, 2221, 2222, 2223, 2226,

2229, 2232, 2324, 2767, 2807
e-commerce adoption 1616
e-commerce applications 1209, 1618
e-commerce applications adoptions 1615
e-commerce applications developers 1823
e-commerce customer relationship management (eCRM) 2258
e-commerce legislation 1546
e-commerce platforms 2345
e-commerce requirements 2347
e-commerce services 1226
e-commerce studies, trust antecedents in 2808
e-commerce system structure 1207
e-commerce systems 1204
e-commerce technologies 1204
e-commerce, global 2325
economic context 567
economical sustainability 1565
economies of scale 2306, 2308, 2310, 2314, 2316, 2318, 2320, 2321
ecosystem 1332
e-coupon 1979
e-customer 1858
edge-counting algorithm 1452
education 249
education on virtual organization 1974
educational design 817
educational tasks 125
eEurope 2002 Action Plan 1549
effectiveness 239
efficiency 239
e-government 249, 253, 756, 782, 1544, 1563, 2767, 3463
e-government contact centre 1550
e-government efforts 248
e-government implementations 254
e-government information portal 1551
e-government initiative 1549
e-government of Jordan 1549
e-government services 1558, 1562
e-health 416, 431, 1442
e-health record (EHR) 1443
e-healthcare system 1450
e-learning 835, 1350
e-learning advancement 1344
e-learning objects navigator (eLON™) 1382, 1383, 1386, 1388, 1389
electric telecommunications 2852
electrocardiogram (ECG) 415
electronic auction service framework 1640
electronic check (e-check) 1238
electronic data interchange (EDI) 1470, 1625
electronic medical record (EMR) 1434
electronic money (e-money) 1237
electronic payment system 1714
electronic performance support systems (EPSS) 112
electronic public relations 242
electronic service guide (ESG) 1152
electronic services 1550
electronic wallets 1691
eligible rate estimate (ERE) 493
Elisa 705
elliptic curve (EC) 1242
elliptic curve integrated encryption scheme (ECIES) 1244
e-mail 790, 824
e-marketplace 1227, 1236, 1544
eMate 144
embedded approach 3448
embedded devices, limitations of 3471
EMC 1666
emergency medical services 249
emergency response support 3407
emotion recognition 3512
encapsulating security payload (ESP) 2691
encrypted function 2586
encryption 1715, 1720
end handoff 3134
end user device 7
end-user needs 3458
e-negotiation 19, 26
energy constrained nodes 3243
energy-efficient indexing 3104
English-language thesaurus WordNet 1444
enhanced data rates for GSM evolution (EDGE) 35, 1265, 1270, 1334, 1676, 2187
enhanced messages service (EMS) 133
enhanced observed time difference (E-OTD) 1756, 1761
enterprise application integration (EAI) 2368, 2369
enterprise applications 1339
enterprise architecture (EA) 2368, 2369
enterprise architecture (EA), benefits to mobility 2383
enterprise business architecture (EBA) 2369
enterprise information architecture (EIA) 2369
enterprise mobility 2147
enterprise model (EM) 2371
enterprise portal 196, 1366
enterprise resource planning (ERP) 788, 2369
enterprise solution architecture (ESA) 2369
enterprise technology architecture (ETA) 2369
entertainment services 2446
entity 1068
environment properties 1938
E-OTD 35
e-portfolio 1259
EPSRC 409
equi-join 357, 358
equipment identity register (EIR) 2753
Erickson 1667, 1680, 2133, 2140, 2142, 3460
error rate 215, 2843, 2844
e-services 3552
e-signature 744
ESRC 409
e-strategies 1690
e-supply chain management 2378
ethnographic action research 818, 834
ethnography 2126, 3333, 3348
Europe, design of mobile television 1143
European Commission (EC) 1570
European Commission's Information Society Technologies (IST) initiative 112
European Telecommunications Standards Institute (ETSI) 2677
evaluation subsystem 3416
evaluation targets 226
evolution of mobile Internet technologies 253
evolutionary paradigms 2827
evolutionary psychology (EP) 1954

e-work 2065
e-work, mobility lifestyle 2062
execution environment 649
execution environment network (EEN) 644
execution tampering 2741
expectancy 1887
expectancy theory 1887, 1892
experienced credibility 373
expert system 790
expertise 373
explicit bad state notification (EBSN) 493
exploit software 3598
exploratory factor analysis (EFA) 1979
Extended ASCII codes 2846
Extended ASCII sets 2846
extensible authentication protocol (EAP) 2768, 2787
extensible authentication protocol method for GSM subscriber identity modules (EAP-SIM) 2676
extensible hypertext markup language (XHTML) 798
extensible hypertext markup language mobile profile (XHTML-MP) 1350
extensible markup language (XML) 9, 798, 821 1341, 1445, 2370
extensible markup language (XML) functionality 3460
extensible markup language linking language (XLink) 800
extensible markup language/simple object application protocol (XML/SOAP) 2299
extensible markup langugage (XML) security 2633, 2652
extensible markup langugage (XML) security standards 2637
extensible stylesheet language transformations (XSLT) 799
Extramadura Association (Spain) 1565
extreme programming (XP) 3326

F

face-to-face sessions 3529
facial expression reconstructor 3516
facilitating conditions 96
facilitators 2165
fairness 2102
fast broadcasting (FB) 3071
fast fourier transform (FFT) 1136, 1142
faster information exchange 249
fast-Fourier-transform (FFT) algorithms 3480, 3570
Federal Communications Commission (FCC) 1048
Federation of Student Unions in Stockholm (SSCO) 3460
FedEx 1340
Fido 1676
field device management 633
field service applications 1340
fieldwork data collection 98
filtering 358, 366
filtering process 569
filtering track 1460
financial design 1153, 1163
financial markets 2191
financial model 2831
financial news delivery 2435
financial news, fractal summarization of 2434
finder 2449
finders fee 705
fingerprinting 2588
Finland 696
Finnish mobile market 704
fire fighting 249
firmographics 1863
first generation (1G) game 292
first-year programming class 1399
Fishbein and Ajzen's theory of reasoned action 1823
Fishmarket Project 1642
Fitts' Law 206, 216, 224
Fitts' Law, applying 210
fixed hosts (FHs) 3014
fixed infinite planes 942
fixed server component 742
fixed-line service providers (FSPs) 30
flash memory 1189, 1192
flexibility 312, 962
floating point 3471
focus groups 3280, 3460
fold-down screen (FDS) 2054
FOMA 701
foreign agent (FA) 3132, 3133, 3138
foreign ownership is limited to 25 percent 1678
Föreningssparbanken 3460
formal learning 1966
formalism, escape from 1940
fourth generation (4G) game 293
fractal summarization 2424
fractal summarization model 2420, 2424
fractal summarization, cue feature in 2430
fractal summarization, summarization features in 2427
fractal summarization, visualization of 2432
frame check sequence (FCS) 2686
Freeze-TCP 490
fequency division multiple access (FDMA) 3561
frequency Doppler function 3564
frequency resolution 1142
frequency response 1142
frequency shift keying (FSK) 1137, 1140, 1142
frequency shift keying, differential (DFSK) 1138
frequently asked questions (FAQs) 1362, 1366
frontware 3306
fuel cells 1189
functional and architectural partitioning 1572
functional module 910
functional partitioning 1572
functionalities 578
fuzzification 2999, 3002
fuzzy cognitive map (FCM) 3422
fuzzy logic-based priority scheduler (FLPS) 2999

G

gadgetware architectural style (GAS) 74
Gagne's events of instruction 1384, 1385, 1390
GALILEO 1049
Game Boy 1333
game industry 702
Garmin 34
gateway general packet radio service support node (GGSN) 2676, 2754
Gdynia City Hall (Poland) 1565
gender difference 1975
gender effect 1984

general packet radio service (GPRS) 9, 35, 85, 86, 88, 120, 457, 509, 1260, 1332, 1471, 1622, 1757, 2187, 2677, 2725, 2752, 2770
general packet radio services (GPRSs) ciphering algorithm 2756
general packet radio services (GPRSs) network architecture 2753
general packet radio services (GPRSs) networks 2853
general packet radio services encryption algorithm (GEA) 2756
general packet radio services encryption algorithm supports and the network (SGSN) 2757
general packet radio services encryption algorithm using the encryption key (GPRS-Kc) 2756
general packet radio services support nodes (GSN) 2754
general packet radio services tunneling protocol (GTP) 2754
general practitioner (GP) 407
general public information 1565
general public information services 1566
general trends of mobile text communication 2134
generic log adapter (GLA) 3206
geo-coded datasets 30
geographic awareness 2020
geographic information systems (GISs) 519, 1691, 3321
geographic service location 1111
geographic service location approaches 1111
geometric model 3032
geostationary 595
GFSAG model, knowledge base 1605
GFSAG model, mobile computing functions 1611
GFSAG model, requirements 1612
GFSAG model, virtual reality concepts 1607
GFSAG, activities of 1604
Glenayre 34
global availability 1544
global cellular NetVillage 1670
global differences 1858
global enterprise, benefits of mobility 2384
global location management scheme 660, 663
global mobile personal communications by satellite (GMPCS) 1547, 1548
global networks 590
global positioning systems (GPSs) 9, 34, 85, 150, 386, 574, 806, 813, 1048, 1409, 1683, 1686, 1691, 1761 1861, 2373, 3032, 3509
global positioning systems, assisted (A-GPS) 1049
global positioning systems (GPS) game 1683
global positioning systems (GPS)-equipped 1683
global positioning systems (GPS) receiver 30
global roaming 87
global supply chain management systems 2368
global system for mobile communication (GSM) 29, 85, 86, 420, 457, 704, 1260, 1270, 1607, 1608, 1757, 2131, 2187, 2374, 2677, 2752, 2767, 2852, 3562
global system for mobile communication (GSM) network standard 1688, 2131
global system for mobile communications (GSM) network 87
global system for mobile communications (GSM) network operators 87
global system for mobile communications (GSM) popularity 87
global system for mobile communications (GSM) technology, advantages of 87
globalization 1488
goodput 496
Google 1334
government applications 790
government information, design of 776
government services delivery channels assessment 1551
governmental e-services 1558
graBBit 805
granularity 1383, 1387, 1388
graphical user interface (GUI) 559, 561, 562, 1201, 1444, 1458, 3503
graphical user interface (GUI) design 1201
graphical user interface (GUI) elements 2856
graphics engineering 2023
graphics interchange format (GIF) 2848
green pages 1576
grid service 3220
GRiDPaD 144
group key handshakes 2681
group support systems (GSS) 518
group temporal key (GTK) 2682
Guangzhou 1668
GW 1667
GWcom 1667
GWcom's web portal 1667

H

hand measurements 1984
handheld computing 918
handheld devices 174, 1185, 1694, 1878, 2615
handheld firewall 2661
handheld, client-side 910, 918
handheld, server-side 910, 918
handoff latency 3132, 3141
handoff latency, high 3132
handoff latency, large 3132
handoff performance 3132
handoff schemes 3130
handoff schemes for multimedia transmission 3130
handoff with mobile node 3131
handover management 651
handset subsidies 1338
Handspring 34
handwriting recognition (HR) 999
hardware-based speech recognition 3483
harmonic broadcasting (HB) 3072
harmonic distortion 1142
hash 1115
head mounted display (HMD) 985, 986, 2059
head mounted display (HMD) configuration 988
heading feature in fractal summarization 2429
health care service 1580
health hazards 792

Health Insurance Portability and Accountability Act of 1996 (HIPAA) 404, 1463
healthcare organizations 404, 413
hedonic elements 1616
hedonic value 2237, 2238, 2249
Heidegger 1955
heterogeneous hardware 1570
heterogeneous technologies 191
hidden Markov models (HMMs) 3485, 3509
hierarchical display 2419
hierarchical indexing 3104
hierarchical mobile IPv6 2984, 2988
High BER 489
high frequency range (long range) 3376
high loss rate 3240
high performance 1187
high performance scheduling mechanism 3151
high signaling traffic 3132
high signaling traffic with the HA 3132
high speed packet access (HSPA) 2771
high-end devices 914
high-level modeling 1228
highly mobile devices 150
high-quality multimedia transmission 2845
high-speed mobile networks 1117
hi-tech mobile phones 2067
home agent (HA) 3132, 3138
home and device controllers 2022
home location register (HLR) 591, 593, 652, 2676, 2753, 2771
home network 3132
home network address 3132
home station (HS) 594
home subscriber service (HSS) 2676
Hong Kong culture 2124
horizontal differentiation 2511
horizontal industry structure 700
horizontal mobile business application 2167
horizontal/modular configuration 706
hospitals mobile applications 2391, 2414
host computers 1218
host mobility 3132
host security 2741
host system 2376
host-based defenses 3595
HostMaster 1448
Hotmail 1686
hotspot 2959
HP Web Services 2347
HSCSD 35
HTML 464
human computer interaction (HCI)
human immune system 2897
human mobile computing performance 206, 208, 210, 211, 216
human operator 938
human perception of quality 2844
human properties 1938
human resources (HR) 789, 1226
human-computer interaction (HCI) 210, 226, 228, 230, 231, 233, 234, 283, 288, 1399, 3321, 1772
Hummingbird Enterprise Portal 192
HUT Dynamics 3138
hybrid content location protocol (HCLP) 1112
hypertext transfer protocol (HTTP) 798, 1619
hyper-text transfer protocol (HTTP) Server Adapter 1575
hyperwave information portal 192

I

IBM Aglet Workbench SDK 2.0 1448
IBM DB2 1220
IDABC (former IDA Framework) 1569
ideal communication 1947
identification number (ID) 1242, 2717
identification portal 3463
identifiers 1124
identities of mobile users (IMSI) 2762
identity management 386
Id-Synch and P-Synch 3239
IEEE 802.11 standard 1532
IEEE 802.11 standard 2849
IEEE 802.11 WiFi standard 976
IETF Transport Area Working Group (TSVWG) 3136
image collection 3519
image-based applications 2851
IMAGO system, service discovery 864
i-Mate 1260
i-menu 1975
immediacy of action 2150
immediate access 1692
immune systems approach 2896
i-mode© 48, 134, 507, 701, 1250, 1339, 1601, 1886
implemented auction system overview 1647
implicit scope information (ISI) 3034
imprecision 1758
IMT-2000 1666
incentive-based marketing 281
increased data encryption 2786
index of performance 224
indexes supporting range queries 326
indexes supporting soundness-enriched queries 327
indexes, aggregating/enumerating 324
indexes, reporting 323
indexing 314
indexing mobile objects 313
indirect relationship (IR) 2536
individual view 282
individual-based target marketing 258
individualism 2809
individuation 2070
industry clockspeed 710
industry configuration 707
industry evolution 710
industry structure 699
infection vectors 3593, 3593, 3594
inferential capability 580
Informa Telecoms 1676
informal learning 151, 1966
information and communication technologies (ICTs) 10, 40, 188, 785, 1547, 1549, 2289, 2368
information and computing technologies (ICT), implications for future 2140
information appliances 3270, 3280
information availability requirement 452
information delivery 2418
information exchange 1785
information filtering (IF) 566, 572
information management 970
information management system

(IMS) 1341, 1181
information object 782
information on demand (IOD) 1667
information processing, models of 3192
information retrieval (IR) 566
information security requirement 453
information services 2, 1362, 1366
information shopping 1652
information systems (ISs) 2093
information systems (ISs) integration 729, 729, 755
information technologies (ITs) 11, 296, 1351, 1466, 2093
information technology (IT) infrastructure 404
Information Technology Association of Jordan (INT@J) 1546
information visualisation 2019, 2026
informational added values (IAVs) theory 205
information-based applications 2154
information-intensive 1469, 1479
information-logistical evaluation 3414
infrared (IR) technology 1142, 1189, 2374
infrastructure de-regulation 1547
infrastructure liberalization 1547
infrastructure-based environments 1104
inheritance relationship 3107
inhibitors 2165
innovation diffusion theory (IDT) 40, 89, 91, 2206, 2017
input (INPUT) 2757
input and output devices 1191
input device 779, 1188
input methods 1191
input modality 1568
input module 3363
input time 216
INSEAD 812
inspection 376
instant information release 249
instant talk 1681, 1684
institutional banking 2191
integrated circuit (IC) 1251, 1254
integrated service architecture 700
integrated services digital network (ISDN) 1548
integration architecture 734
integration framework overview 734
integration middleware 735, 735, 736
integrity 312, 792, 2346, 2603
integrity protection 309
intellectual resources 968
intelligent agents 304, 579, 782, 3454
intelligent agent paradigm 3449
intelligent install and update 384
intelligent migration 624
intelligent mobile agents, implementation of 627
intelligent mobile computing 589
intelligent software agents, integration 3373
intelligent user interfaces (IUIs) 3454
intelligent user interfaces (IUIs) for mobile computing 3442
intentional name resolver (INR) 1107
intentional naming system (INS) 1107
interaction design (ID) 288
interaction trajectory 282, 283, 283
interactivity 1557
inter-carrier interference (ICI) 3566
interconnectivity framework 589
interdependencies 2440, 2443
interface 778, 782
interface delivery service 1965
interfering distraction 2059
intermediate industry structures 700
internal communication 3461
international financial services 1602
international long distance (ILD) services 1555
international mobile subscriber identity (IMSI) 2755, 2770
International Telecommunications Union (ITU) 84, 1666, 1669, 1676, 2849
Internet Engineering Task Force (IETF) 2846, 3131
Internet paradigm, model based on 2830
Internet protocol (IP) 1666, 3132, 3136
Internet protocol (IP) networks 3131
Internet protocol version 6 (IPv6) 420, 430
Internet purchasing behaviour 90
Internet service providers (ISPs) 30, 43, 1547
Internet Streaming Media Alliance (ISMA) 2847
Internet-based learning 1344
Internet-enabled therapy 3537
Internet-enabled therapy devices 3537
interoperability 962, 1109, 1385, 1569
interoperable platform 1564
interruption 2125, 3131
intersection set operation 364
inter-symbol interference (ISI) 3565
intralocation-area location update 660
invalid access prevention policy 3029
invalidation report (IR) 3013, 3033
inverse discrete Fourier transform (IDFT) 3568
Investment Promotion Incentives (IPI) 1546
IP multimedia system (IMS) 420, 430, 465, 466, 467
iPod 1170, 1174, 1332
IPSec 119
IPv6 Stateless Address Autoconfiguration (SAA) 3135
IPv6-based mobility protocols 2982, 2985
IrDA Data 1189
IST initiatives 1563
IT Council 3456
IT Department, Stockholm 3457
items of interest (IOIs) 2698
iterative model refinement 939

J

JAMES 3300, 3303
Japan 697
Japan, mobile services industry in 701
Japanese Digital Cellular (JDC) 2852
Japanese mobile market 701
Japanese model 709
Java 1670, 1683, 1685
Java 2 Micro Edition (J2ME) platform 390, 391, 392, 394, 395, 396, 397, 402, 456, 852, 909, 918, 1244

Java virtual machine (JVM) 1107
Java-enabled phones 2464
JINI 1107, 3220
Joint Photographic Experts Group (JPEG) 2000 2848
Joint Photographic Experts Group (JPEG) 2000 Wireless (JPWL) 2849
Joint Photographic Experts Group (JPEG) 2000 Wireless (JPWL) methods 2850
Joint Photographic Experts Group (JPEG) 2000, motion 2849
Joint Photographic Experts Group (JPEG) standard 2848
Joint Photographic Experts Group (JPEG) type compression techniques 2851
Jomotel 1548
Jordan Mobile Telephone Services (Fastlink) 1548
Jordan Telecom (JT) 1548
Jordan, prospects of mobile government 1543
Jordanian House of Parliament 1546
Jordanian IT strategic plan (REACH) 1545, 1556
Jordanian telecommunication industry/market 1545
Jordanian telecommunication market 1556
J-Phone 698
JTCP 492
jukebox 1174
junk data (JD) 591
JXTA 1197, 1198

K

Kant 1954
KASBAH 1642
keiretsu 702
Keitai 1975
Kellogg, W. A. 2133, 2140, 2142
key certificate authority 548, 549, 551, 552, 553, 554, 555
key encryption key (KEK) 2774
key performance indicators (KPIs) 707
key seed negotiation protocol 308
keypad design factors 1984
killer app 1698
King Abdullah the 2nd 1546
kiosk 1197
knowledge management (KM) 789, 1359, 1366, 2496
knowledge management systems (KMSs) 188, 196, 197, 797
knowledge repository 1606
knowledge representation 804, 1571
knowledge space 1948
knowledge-based system 1604
Korea economy 1699
KSACI 854
Kyocera 34

L

lab environment 2042
lab evaluations 2059
LAMP stack 1218
language acquisition device (LAD) 1954
laptop computers 998, 1002, 1430, 2843, 3130, 3460
largest object first (LOF) algorithm 3100
law enforcement 249
layer models 700
layers of encryption 3383
LCD 985
LDAP access engine and directory structures 3239
Leacock and Chodorow algorithm 1452
lead firms 702
leadership 1468, 1469, 1473, 1474
LEAP 852
learning management system (LMS) 1350, 1383, 1385
learning objects (LO) 1382, 1387
learning portal 1966
learning, face-to-face 820
least mean squares (LMS) 558, 559
legal context 374
Lehman Brothers Inc. 1666
leisure 790
LEMP 3076
lexical decoding 3487
LFS 35
license management 1124
license, digital 1124
licensing 2472
Likert scale 3184
limited memory 3238
limited security 3243
linear and time invariant (LTI) 3567
linear ordering 3087
link adaptation (LA) procedure 3586
link signal strength 3133
linkcell size determination 3058
linkcell size optima 3051, 3054, 3056
little smart users 1666
Livelink Wireless 192
Livetunes 2850
local area networks (LANs) 85, 838, 1532
local mobility 760
local service provision (LSP) 405
local wireless interface (LWI) 1135
local-based service (LBS) game 295
locales foundation 282
locales framework 282, 288
localization 175, 241, 1591, 1594
localization services 1574, 1576
location and time ontologies 3404
location application 3499
location area design algorithms 682
location area forming algorithm (LAFA) 685
location awareness 2020, 3509
location based services (LBSs) 1955, 2530, 2531
location based services (LBSs) directory 2450
location dependent query processing 3197
location feature in fractal summarization 2428
location granularity level 340
location information 1754
location knowledge 3320
location management 339, 651
location management agent 583
location model 339, 3032
location model platform (LMP) 2546
location modeling 2532, 2563
location ontology 3409
location orientation 2505
location pattern matching (LPM) 575
location search 652
location service 340
location tracking 1585
location transparency and dependency 1585
location update 652
location update costs, analysis of 2991
location-aware computing 574, 2167

location-aware intelligent agent (LIA) 575
location-aware method (LAM) 3064
location-aware queries (LAQs) 335, 336
location-aware query resolution 3040
location-aware service 2017
location-based information exchanges 254
location-based mobile commerce 3040
location-based queries 315
location-based service offering 7
location-based services (LBSs) 88, 386, 1181, 1564, 1601, 1685, 1687, 1754, 2017, 3031, 3038
location-based services (LBSs), caching for 3034
location-based services (LBSs) provider 1759
location-dependent cache invalidation 3034, 3038
location-dependent information processing 351, 355, 370
location-dependent invalidation 3033
location-dependent IR 3034
location-dependent queries (LDQs) 335, 336, 339
location-measurement units (LMUs) 1761
location-orientation 196
location-oriented information delivery 191
location-related operators, processing of 340
location-sensing technologies 578
Locknut 3592
login service 1964
logistic regression 1980
long range navigation (LORAN) 3321
look-at-this (LAT) applications 2851
loss leader model 2472
lost work 792
low density parity check (LDPC) 2773
low frequency range (short range) 3376
low power consumption 1187
Lowrance 34
loyalty program customers 1911
lurking 1774

M

M&M project 3301
macro/micro-payment 1244
Magellan 34
MAGO 1745
mail system 3463
maintenance services 639
maintenance services, semantic peer-to-peer discovery of 639
malware 3588, 3589, 3598
malware defenses 3595, 3595, 3596
malware in mobile devices 3588
malware, evolution of 3591
malware, non-replicating 3589
MAMDAS 1445, 1446, 1448, 1450, 1458, 1462
managing mobile environment 621
MapMe™ 1686
market configuration 696
market forces 1757
market power 710
market transactions 759
marketability 1860
Markov chain 3133
MARS 2865
MASK 2704
massively multi-player online game (MMOG) 295
master device 499
Matrix software 1671
m-business 2784
m-commerce operational circumstances 3058
m-commerce studies, trust antecedent in 2810
means-end chain (MEC) 1869
media 244, 247
media access 2725
media access control (MAC) Filtering 2786
media content 2844
media gateway discovery protocol (MeGaDiP) 1109
media model 2473
media phones 1757
media player 1002
media richness theory 1892
media services language (MSL) 3389
mediated communication 1785
medical subject heading (MeSH) 1444, 1450, 1451, 1453, 1459, 1462
medium access control (MAC) 954, 962
MEDLINE 1451, 1459
MEDTHES 1444, 1451, 1453, 1454, 1459, 1461, 1462
Mel cepstrum 3481
Mel spectrum 3480
memory 790, 1188, 1191, 3471
memory card 2660
memory cell introduction 2903
mental context 1065
mentoring program 1353
merchant adoption barriers 1634
merchant adoption drivers 1630
merchant preferred applications 1635
message authentication code (MAC) 2587
message authentication code (MAC) layer protocols 3004
message encryption 2741
message format 1888
message integrity check (MIC) 2786
message latency 454
messaging 2
messaging purpose 2135
messaging purpose, text communication 2135
messaging service 1575
metadata 1124, 1386, 1391, 1443
meta-directories 3239
meta-information 2850
methodology perspective 2199
metrics 377
metropolitan area networks (MANs) 1532, 2660
MexE 35
Michigan Internet AuctionBot 1642
microbrowsers 1187, 1190, 1191, 1193, 1601
Microcell 1676
Microcell (GSM/GPRS) 1680
Microcell telecommunications wireless 1676
micro-coordination 2132, 2136
micro-grooming 2130, 2138, 2140, 2141
micropayment services 708
micropayments 40, 1627
Microsoft 34, 1118

middleware 619
middleware for robots (Miro) 601, 602
MIDlet applications, as a gateway to mobile CRM 1913
MIDlet technologies, managerial implications 1923
minimum delay 594
minimum mean square error (MMSE) equalization 3573
minimum variance distortionless response (MVDR) 558, 559
Ministry of Information and Communications Technology (MoICT) 1550
MINPATH algorithm 2971
Mint Inc. 1681
minutes of use (MoU) 708
missing neighbor pilot 3209
mission statement of TRC 1547
MIT Laptop Project 842
mix route algorithm (MRA) 2709
m-Mode 1335
MobiAgent 853
mobile (emerging) technologies 2373
mobile access 191, 372, 770, 1360, 1366
mobile access adaptation 1588
mobile accessibility 1944
mobile ad hoc networks (MANETs) 952, 1115, 2833, 2858, 2996
mobile ad hoc networks (MANETs) QoS models 2836
mobile ad hoc networks (MANETs) QoS routing 2837
mobile ad hoc networks (MANETs) QoS, developments in 2835
mobile ad hoc networks (MANETs), QoS 2833, 2858
mobile added values (MAVs) 204
mobile ads 1879
mobile advertising (m-advertising) 22, 288, 1653, 1661, 1656, 1658, 1878, 1883, 1884, 1893, 1905
mobile advertising (m-advertising), European perspective 1653
mobile advertising (m-advertising) privacy concerns 1904
mobile advertising (m-advertising) technology 1894, 1896
mobile advertising (m-advertising), permission-based 1878
mobile agents (MA) 296, 302, 304, 305, 312, 618, 630, 642, 643, 649, 714, 1227, 1236, 1713, 1715, 1719, 1720, 2344, 2568, 2739, 2740, 2742, 2743, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2943
mobile agents (MA), advantages of 299
mobile agents (MA), behaviors of 720
mobile agents (MA), disadvantages of 299
mobile agents (MA), e-commerce services 1226
mobile agents (MA), integration 2936
mobile agents (MA), technology 2600, 2616
mobile agents (MA), strongly 2584
mobile agents (MA), weakly 2584
mobile agents (MA), models 300
mobile agents (MA), reality systems 937
mobile agents (MA)-based payment protocol 1715
mobile agents (MA)-based restaurant ordering system 1713
mobile agents (MA)-based services 1231
mobile agents (MA)-based systems 1228
mobile and distributed brains architectures 632
mobile application part (MAP) 2688
mobile applications 19, 796, 1594, 1601, 1721, 1937, 2807
mobile applications in knowledge management 2496
mobile applications market, case of 1725
mobile applications, clinical environment 2394
mobile applications, credibility of 372, 374
mobile artificial autonomous agents programming language (3APL-M) 852
mobile assistance 1564
mobile audio commercials 2851
mobile authorization 1135
mobile automotive cooperative services (MACS) 1499, 1500, 1512, 1501, 1515, 1503, 1515, 1505, 1515, 1506, 1515, 1508, 1515, 1510, 1515
mobile automotive cooperative services (MACS) design framework 1501, 1502
mobile automotive cooperative services (MACS) design framework, application of 1503
mobile automotive cooperative services (MACS) development process model 1504
mobile automotive cooperative services (MACS) live service 1512
mobile automotive cooperative services (MACS) prototype and platform 1509
mobile automotive cooperative services (MACS) safety aspects 1510
mobile automotive cooperative services (MACS) service network 1506
mobile automotive cooperative services (MACS) service scenarios 1505
mobile automotive cooperative services (MACS) technologies 1508
mobile banking 47, 790, 1246, 1249, 1252, 1254, 1699
mobile banking life cycle 1248
mobile banking systems and technologies 1246
mobile banking technologies 1249
mobile broadband 2766
mobile broadband wireless access (MBWA) 2767
mobile broker (m-broker) 875
mobile business (m-business) 1359, 1587, 1592, 1738, 2257, 2273, 2323, 2418, 2784
mobile business (m-business) application 2167, 2168
mobile business (m-business) customer focus 2263, 2279
mobile business (m-business) models 2170
mobile business (m-business) process reengineering 2391
mobile business (m-business) strate-

gic focus 2271, 2279
mobile business (m-business) strategies 2265, 2279
mobile business (m-business) intelligence 205
mobile caching 3031
mobile care units 3462
mobile central processing 1185
mobile central processing units 1187, 1191
mobile channels 1566, 2212
mobile city (mCity) 3455, 3456, 3457, 3458, 3465
mobile city (mCity) experiences 3462
mobile city (mCity) project manager 3457
mobile city (mCity) project, organization of the 3458
mobile city (mCity), competence network 3458
mobile city (mCity), focus groups of the 3459
mobile city (mCity), m-government 3465
mobile city (mCity), municipal and national strategy 3464
mobile city (mCity), Stockholm as an IT capital 3464
mobile city (mCity), working process 3458
mobile clients (MCs) 335, 388, 389, 452, 3038
mobile clients (MCs) profile 1589
mobile clinical learning tools 1256
mobile code 3316
mobile code technology 2568
mobile collaboration 518
mobile collaboration in learning environments 3547
mobile collaboration in learning environments paper prototype testing 3542
mobile collaboration, human factors 518
mobile collaborative reading 3540
mobile commerce (m-commerce) 2, 18, 19, 24, 25, 26, 38, 47, 56, 204, 249, 279, 526, 534, 790, 805, 1118, 1135, 1175, 1183, 1191, 1193, 1246, 1252, 1254, 1466, 1467, 1468, 1470, 1479, 1480, 1481, 1584, 1585, 1592, 1593, 1594, 1595, 1598, 1601, 1615, 1625, 1626, 1690, 1691, 1692, 1693, 1694, 1698 1780, 1787, 1840, 1841, 1844, 1851, 1853, 1855, 1878, 1884, 1892, 1983, 2169, 2257, 2338, 2614, 2654, 2807, 3422
mobile commerce (m-commerce) adoption 1593, 1594, 1616, 1619
mobile commerce (m-commerce) adoption studies 1595
mobile commerce (m-commerce) application adoptions 1615
mobile commerce (m-commerce) applications 1210, 1593
mobile commerce (m-commerce) in Canada 1676
mobile commerce (m-commerce) in China 1665
mobile commerce (m-commerce) in South Africa 1690, 1692
mobile commerce (m-commerce) multimedia messaging peer 1194
mobile commerce (m-commerce) security 1587
mobile commerce (m-commerce) system structure 1208
mobile commerce (m-commerce) systems 1204
mobile commerce (m-commerce) technologies 1204
mobile commerce (m-commerce) transaction processing 1208
mobile commerce (m-commerce), applications of 21
mobile commerce (m-commerce), benefits of 1692
mobile commerce (m-commerce), challenges facing 1692
mobile commerce (m-commerce), growth potential 1694
mobile commerce (m-commerce), multimedia messaging peer for 1197
mobile commerce (m-commerce), reference model 1596
mobile commerce (m-commerce), technology required to enable 1693
mobile commerce (m-commerce), trends in 1694
mobile commerce (m-commerce), uses of 1691
mobile communications 253, 1562, 1947, 2323
mobile communications infrastructure 441
mobile communications system design 3563
mobile communications systems 2844, 2852, 2853
mobile communications technologies 88
mobile communications industry 1754
mobile communications markets 1296
mobile computers 3130
mobile computing 18, 224, 313, 589, 650, 651, 817, 981, 1175, 1584, 1592, 1602, 1607, 1641, 1960, 2081, 3031, 3081, 3151, 3236, 3414, 3454
mobile computing device adoption and diffusion 2092
mobile computing environments 1961, 3187
mobile computing in healthcare 1430
mobile computing work, sociotechnical nature of 2079
mobile computing, challenges of 1584
mobile computing, intelligent agents 3450
mobile connectivity 1585
mobile consumers 1780
mobile consumer agents (MCA) 1230
mobile content management 1371
mobile content services 269
mobile content services models 270
mobile content services, costs of 273
mobile content services, parties in 270
mobile credibility 373
mobile credibility, initiatives for improving 374
mobile credit card billing 1627
mobile customer relationship management (mCRM) 1912, 2259
mobile customer relationship man-

agement (mCRM) applications, benefits to the firm 1914
mobile data services (MDS) 507, 696, 1296, 1309, 1741
mobile data services (MDS) usage patterns 1305, 1311
mobile data solutions 1466, 1467, 1470, 1479
mobile database 370
mobile database environment 352
mobile decision support 3553, 3560
mobile design for older adults 3270, 3281, 3273, 3274, 3276
mobile devices 1, 19, 249, 618, 777, 778, 780, 782, 1190, 1246, 1248, 1250, 1584, 1592, 1626, 2418, 2435, 3212, 3220, 3509, 3560
mobile devices, acoustic data communication 1135
mobile devices, input performance of 207
mobile devices, interfaces for 3320
mobile devices, Internet access from 3594
mobile devices, movement mechanism 3161
mobile devices, movement process 3163
mobile devices, penetration 253
mobile devices, ranking value 3160
mobile devices, trends 780
mobile devices, user interface evaluation 3169
mobile devices, voice-enabled user interfaces 3494
mobile dialogue campaigns 1659
mobile digital communication and therapy tools 3531
mobile digital music collection portal 1168
mobile digital rights management (MDRM), problems and issues 1120
mobile DSS 3557
mobile ecosystem 28
mobile education (m-education) 123
mobile ELDIT 1377
mobile e-learning 3349, 3360
mobile e-mail 86
mobile emergency assistant 3407
mobile enterprises 2530, 2531, 2533, 2564
mobile entertainment services 2464
mobile environments 498, 621, 1103
mobile environments architecture 3188
mobile environments, data broadcasting 3079
mobile environments, portal design in 1962
mobile environments, data dissemination in 3068
mobile environments, database queries in 334
mobile e-services 2440
mobile evaluation 2042, 2043
mobile evaluation in a lab environment 2045, 2046, 2052
mobile e-work 2062, 2065
mobile e-work, social innovation 2061
mobile e-work, supporting regional/rural communities 2061
mobile e-working conditions 2063
mobile feedback, speed of 1921
mobile financial applications 21
mobile games 173, 289, 290, 295, 2463, 2464, 2518, 2518, 2529
mobile games, characteristics of 290
mobile games, distribution of 2469
mobile games, generations of 292
mobile games, limitations of 290
mobile games, platforms of 291, 295
mobile games, product strategy for 2468
mobile games, revenue logic of 2471
mobile games, service types 292
mobile games, types of 2464
mobile geographic information viewer (mobile GIS) 748
mobile goods 759
mobile government (m-government) 248, 253, 756, 757, 770, 1563, 3455, 3463
mobile government (m-government), applications and services 249
mobile government (m-government), characteristics of 248, 250
mobile government (m-government), concept of 248
mobile government (m-government), designing 756
mobile government (m-government), drivers of 253
mobile government (m-government), issues in 253
mobile government (m-government), major issues of 253
mobile government (m-government), services 249, 1563
mobile government (m-government) software 255
Mobile Government Consortium International (MGCI) 256
mobile government (m-government) in Jordan 1543
mobile government (m-government), usability driven open platform 1562
mobile governmental services 1543
mobile graphical user interface 1525
mobile handheld devices 1183, 1185, 1189, 1193
mobile hardware technology 3012
mobile human computer interaction 2019, 2026
mobile health (m-health) 411, 420, 1409, 1425
mobile health (m-health) analytical hierarchy process approach 1413, 1415
mobile health (m-health) case study 1415, 1424
mobile health (m-health) case study proposed actions 1424
mobile health (m-health) reference model 432, 450
mobile health (m-health) system 1411, 1413
mobile health (m-health), background 434
mobile healthcare delivery system networks 436
mobile hosts (MHs) 490, 3014, 3130, 3134
mobile human-computer interaction 225
mobile iMode services 2464
mobile information 968, 2446
mobile information and communication technologies (M-ICTs) 1359, 1429
mobile information device profile

(MIDP) 390, 396, 464, 465, 914, 1346
mobile information exchange 205
mobile information processing 3185
mobile information system (MIS) 451
mobile infrastructure 253
mobile innovations 93
mobile input performance 209
mobile intelligent agent-based architecture 712
mobile interface architecture 1564
mobile interfaces 1564
mobile Internet 696, 1691, 3459
mobile Internet access 1530, 1531
mobile Internet adoption, gender differences in 1978
mobile Internet applications, adoption of 253
mobile Internet providers 651
mobile Internet site 1627
mobile Internet usage 1975
mobile inventory management 22
mobile IP (MIP) 496, 1181, 3130, 3131, 3132, 3133
mobile IP (MIP) handoff 3141, 3142, 3143
mobile IP (MIP) handoff instant 3142
mobile IP (MIP) handoff latency 3132
mobile IP (MIP) testbed 3140
mobile IP network architecture 660, 669
mobile IPv6 119
mobile knowledge management (m-KM) 188, 196, 197, 205, 1360, 1366
mobile knowledge management (m-KM) approaches 2505
mobile knowledge management (m-KM) portals 2498, 2500
mobile knowledge management (m-KM) services 1360, 1366
mobile knowledge portal 189
mobile learning (m-learning) 108, 109, 122
mobile learning (m-learning) 1350, 1383
mobile learning (m-learning) 1766, 2023, 2026
mobile learning (m-learning) 3283, 3299, 3551
mobile learning (m-learning) 835
mobile learning (m-learning), environments 3350
mobile learning (m-learning), in museums 3282, 3299
mobile learning (m-learning) environment, pedagogic design in 1961
mobile learning (m-learning) environment, portals supporting 1960
mobile learning management system (M-LMS) 1346, 1371
mobile location dependent query processing 3198
mobile marketing (m-marketing) 258, 280, 288, 1125, 1132, 1885, 1893
mobile marketing (m-marketing), acceptance of 263
mobile marketing (m-marketing), applications 260
mobile marketing (m-marketing), key issues in 257
mobile marketing (m-marketing), technological platform 258
mobile marketing (m-marketing), technology intervention perspective 279
mobile marketing (m-marketing) and advertising, permission-based 1885
mobile marketing (m-marketing) as technology invention 283
mobile marketplace (m-marketplace) 2968
mobile markets segmentation 1300, 1311
mobile markets user categorization 1298, 1311
mobile markets, differences 706
mobile middleware 20, 26, 35, 535, 1125, 1215, 2659
mobile multimedia 3529
mobile multimedia communications 2844
mobile multimedia content 2844
mobile multimedia services (MMS) 696, 2140, 2675
mobile multimedia services (MMS) and contents 1104
mobile multimedia speech and language therapy devices 3529
mobile multimedia systems 2844
mobile multimedia transmission 2843
mobile multimedia, critical issues of DRM 1119
mobile multimedia, DRM solution 1121
mobile multimedia, DRM technology 1117
mobile multimedia, mobile DRM framework 1120
mobile music services 2515, 2515, 2529
mobile network operator (MNO) 1699, 2487
mobile networked text communication 2130
mobile networks 420, 682, 2843, 3069, 3133, 3237
mobile newspaper (m-newspaper) 791
mobile nodes (MNs) 2996, 3132, 3138
mobile number portability 697
mobile office (MO) 2203
mobile office (MO) messaging adapter 1575
mobile ontologies 2908, 2913, 2916
mobile operating systems 1185, 1191
mobile operator 1247, 1248, 1563
mobile operator network 1581
mobile payment (m-payment) 1245, 1586, 1592, 1670, 1701, 2280, 2517, 2517, 2529
mobile payment (m-payment) adoption 1627
mobile payment (m-payment) adoption, barriers to 1632
mobile payment (m-payment) adoption, drivers for 1629
mobile payment (m-payment) issues 1699
mobile payment (m-payment) pilots 1628
mobile payment (m-payment) policy implications 1699
mobile payment (m-payment) solutions 1626, 1627
mobile payment (m-payment) systems 1706
mobile payment (m-payment), characteristics of 1702
mobile people 3455
mobile phones 3, 84, 86, 87, 151, 712, 871, 984, 986, 987, 988,

990, 992, 1757, 2073, 2125, 2131, 2132, 2134, 2139, 2140, 2141, 2143, 3460
mobile phones and autonomy 2066
mobile phones and SMS 2132
mobile phones communication innovation 2124
mobile phones customer type discrimination 2871
mobile phones keypad design 1985
mobile phones messaging satisfaction 1984
mobile phones, advertising 1654
mobile phones/Internet 2125
mobile phones usage 2896
mobile phones use, across cultures 2110
mobile portal (m-portal) 2, 30, 187, 196, 703, 805, 810, 1983
mobile portal solution 2496
mobile portlet 196, 1366
mobile providers' location-based system (LBS) 1683
mobile public relations strategies 240
mobile pull campaigns 1658
mobile push campaigns 1658
mobile queries scheduling 347
mobile query processing 353, 370, 3194
mobile querying broadcasted data 3195
mobile querying with cache data 3197
mobile resource 1944
mobile satellite networks 2374
mobile security (m-security) 1592
mobile server 335
mobile services (m-services) 84, 171, 892, 1255, 1565, 1566, 1586, 1593, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2320, 2233, 2251, 2936
mobile services (m-services) adoption 84
mobile services (m-services) background 2235, 2256
mobile services (m-services) delivery 1567
mobile services (m-services) evaluation 1918
mobile services (m-services) industry 696
mobile services (m-services) industry matrix 700
mobile services industry matrix (MOSIM) 700
mobile services (m-services) markets 699
mobile services (m-services) provision 1567
mobile services switching centre (MSC) 2753
mobile services (m-services) temporal and spatial context 2238, 2256
mobile services (m-services), acceptance model method 97
mobile services (m-services), earlier approaches 2000
mobile services (m-services), future research 101
mobile services (m-services), proposed model of acceptance 98
mobile services (m-services), qualitative stage 97
mobile services (m-services), quantitative stage 99
mobile services (m-services), success of 696, 707
mobile services (m-services), suitability of 767
mobile services (m-services), user acceptance of 1996
mobile services (m-services), user focused development of 3455
mobile society 2146
mobile software agents 1640
mobile speech recognition 3468
mobile station (MS) 591, 594, 962
mobile stock trading 86
mobile supply chain management (m-SCM) 2378
mobile supply chain management (m-SCM) technologies 1486
mobile support 452
mobile support stations (MSSs) 388, 452, 3014
mobile switching centers (MFCs) 591
mobile systems 2843
mobile systems, and access control 2792, 2806
mobile systems, and authentication 2792, 2806
mobile systems, and authorisation 2792, 2806
mobile systems, evolution of 2852
mobile systems, quality of service 2853
mobile techniques 198
mobile technologies (MT) 85, 146, 147, 230, 350, 408, 756, 1353, 1466, 1473, 1475, 1477, 1479, 1480, 1487, 1967, 1974, 2290, 2368, 2400, 3455
mobile technologies (MT) healthcare support 1408, 1425
mobile technologies (MT), application in 2377
mobile technologies (MT) and services 84
mobile technologies (MT) and services, acceptance model 91
mobile technologies (MT), managerial implications 99
mobile technologies (MT), user acceptance 101
mobile telecommunications 808, 1757
mobile telephone service 1397, 1554, 1556, 1760, 2852
mobile telephone, prospects of 1556
mobile telephony industry 698
mobile television 1143, 1164
mobile television, design of 1143, 1146, 1167
mobile terminals 87, 3563
mobile text communication 2140
mobile text communication, general trends of 2134
mobile ticket reservations 86
mobile transactions 249, 1585
mobile transition (m-transition) 2198
mobile transmission systems 2845
mobile units, localization of 339
mobile users 335, 1585, 3404
mobile user interfaces 3168
mobile user interface designer 738
mobile user queries 3191
mobile users, dynamic demands 3404
mobile value-added services 2509, 2526, 2525, 2529
mobile value-added services characteristics 2519, 2529
mobile value-added services char-

acteristics (connectedness) 2521, 2529
mobile value-added services characteristics (contemporariness) 2522, 2529
mobile value-added services characteristics (content-focus) 2521, 2529
mobile value-added services characteristics (customization) 2521, 2529
mobile value-added services content-focused market strategy 2526, 2529
mobile value-added services development strategy 2525
mobile value-added services in Korea 2509
mobile value-added services motivations 2511, 2529
mobile value-added services real-time market-responding strategy 2525, 2529
mobile value-added services various bundling strategy 2526, 2529
mobile value-added services, motivations of 2511
mobile value-added services, new strategies for 2525
mobile video services 2516, 2516, 2529
mobile video streaming 1128
mobile video telephony 2851
mobile virtual communities of commuters 1771, 1779, 1772
mobile virtual communities of commuters, collaboration model 1773
mobile virtual communities of commuters, sociability 1774
mobile virtual communities, characteristics 1765
mobile virtual network operator (MVNO) 2489
mobile virtual network operator (MVNO) business cases 2484
mobile virtual network operator (MVNO) business models 2478
mobile virtual network operator (MVNO) impact 2487
mobile Web 1754, 1937
mobile Web engineering 380, 1944
mobile Web navigation 2969
mobile wireless devices 3130
mobile work 2065
mobile worker 733, 743
mobile working environments 729, 729, 755
mobile world 242
mobile/ICT devices 2125
mobile/wireless applications 3463, 3465
mobile-agent technology 1444
mobile-application architecture 731
mobile-application architecture guides 731
MOBIlearn 1960, 1966
mobile-based healthcare setting 404
mobile-based learning systems 2023
mobile-client data caching 3021
mobile-computing application 731
mobile-computing application servers 731
mobile-enabled organization 2294
mobile-enabled process 2294
mobileIF 565
mobileIF, contextualization in 567
mobile-phone marketing 2924
mobile-phone user 1761
mobile-PR 247
mobility 197, 565, 631, 756, 823, 851, 1047, 1185, 1564, 1592, 1607, 2083, 3032, 3082, 3130, 3216
mobility management 650
mobility management models 660
mobility pattern history 655
mobility perspectives 757
mobility problem 3132
mobility simulator 688
mobility, economical perspective on 759
mobility, patterns of 2155
mobility, sociological perspective on 760
mobility, technological perspective on 758
mobility, variations in 3007, 3009
mobilization 2074
mobilize emergency services (EMS) 1784
model variables 2866
ModelAccess 3414
moderating variables 97
modified discrete cosine transform (MDCT) 2847
modular application ontology 3404
modular ontology architecture 3408
modular service ontology schema 3410
modular services architecture 700
modulation scheme 3586
MONA system 3407
Morse code 3362, 3367
MOSQUITO project 2859
motivational theories 91
motivations 1975
MotoHealth 411
Motorola 34
moving image coding 2849
moving images 2844
moving object database queries (MODQs) 337
moving object location update 345
moving object representation 314
moving objects 314
moving objects, modeling of 341
MP3 (MPEG audio Layer 3) compression format 108, 1174, 2847
MP3 download 1333
MP3 players 813, 1168, 1169
MP4 108
MPEG 1118
MPEG video compression standard 2847
MPEG-4 AAC standard 2847
MPEG-4 AVC 2849
MPEG-4 transmission 3133
MPEG-4 video application 3133
MPEG4-IP player 3146
MPEG4-IP player playing streaming video 3147
MRCP 3474
MRCP ASR request 3476
MRCP TTS request 3476
MS-Mobile 1337
mStudent 3460
Mtone Wireless Corp 1670
MTS Mobility 1676
multi-agent framework 3425
multi-agent system (MAS) 304, 3425, 3438, 3454
multicast backbone (MBone) 3073
multicast group size 3009
multicast overlay networks 3073
multicasting 578, 1116
multicasting protocols 3009
multi-channel service delivery 1565

multiframe coder 3133
multiframe video coding 3133
multiframe-block motion compensation (MF-BMC) approach 3133
Multi-Hop Communications 3243
multilayered evaluation approach 3171, 3172, 3175
multi-layered evaluation approach 3184
multimedia applications 2844, 2850
multimedia capability 1187
multimedia communication 3130
multimedia communication session 2856
multimedia content 2845
multimedia content and applications 2843
multimedia data transfer 3146
multimedia database 3029
multimedia in wireless networks 3130
multimedia information 778, 2850
multimedia message service (MMS) 9, 30, 85, 86, 88, 590, 1001, 1002, 1064, 1128, 1130, 1132, 1134, 1195, 1352, 1601, 1757, 1761 1889, 2513, 2529, 2662, 2846, 2851
multimedia message service (MMS) composer 1199
multimedia message service (MMS) encoder 1200
multimedia message service (MMS) kiosk 1198
multimedia message service (MMS) peer and kiosk architecture 1197
multimedia message service (MMS) peers 1197, 1198
multimedia message service (MMS) peers, design and implementation of 1198
multimedia message service (MMS) player 1200
multimedia message service (MMS) sender 1200
multimedia message service (MMS) solutions 1201
multimedia over wireless networks 3130, 3133
multimedia prototypes 3533
multimedia services 426, 1110, 2140
multimedia services and contents 1103
multimedia streaming 3069
multimedia system 2845
multimedia technologies 56
multimedia transmission 3130, 3132
multimedia units (MMUs) 3134, 3138
multimodal interface 3509
multimodal services 1567
multimodality 385
multinational corporation (MNC) 280
multipath collector module 3134
multipath distributor module 3134
multi-path propagation 3563
multi-path propagation scenario 3563
multiplayer network games 1683
multiple access interference (MAI) 3561, 3574
multiple displays 498
multiple network interfaces 3130
multiple non-collaborative sources 3185
multiple relationship (MR) 2536
multiple remote databases 366
multiple servers 351
multiple-choice question 1401
multi-point distribution systems (DTH/MDS) 1675
multipurpose Internet mail extension (MIME) 3475
multipurpose Internet mail extensions (MIME) 2846
multipurpose remote controls 1691
multi-sensory feedback 993
multithreaded auction server 1647
multithreading 731
multi-tier information transmission processes 2850
multi-user case 3579
multi-user detection's (MUD) techniques 3573
multi-user diversity (MUD) 3571
multi-user uplink system 3578
municipal organizations of Stockholm 3461
museum mobile educational activities 3288, 3299
music collection 1174
music download 173
music industry 702
musical instrument digital interface (MIDI) 1136, 1142
mutual authentications 549, 1239
mutuality 282, 283, 284
My Babes 2447
MyAlert 1738
MySQL 1220

N

Napster 1333
narrowband signals 3568
narrowband sub-channels 3568
NASA dataset 2976
NASDAQ exchange 3097
National e-Government Initiative 1550
National Health Service (NHS) 404, 406, 412
National Information Standards Organization (NISO) 1446
National Programme for Information Technology (NPfIT) 404, 406, 407, 411
national regulatory authorities (NRAs) 699
national service provision (NSP) 405
nationwide patient database 404
naturalistic observations 1404
navigation systems 1691
navigational cues, comparison of 2046
N-cube 589, 592
negotiation process 721
negotiation support systems (NSSs) 3422, 3436
neighbor set 3210
nervous shuffling 1396
net present value (NPV) 1620
NetEase 1669
Netlogo platform 3433
netset 1667
netware 768
network access servers (NAS) 2675
network address translation (NAT) 2757
network bandwidth 3474
network command language (NCL) 619
network components, cost evolution of 2494
network connectivity 871
network convergence 2794

network dependency 3474
network externalities 698
network infrastructure 1693
network latency 3084
network layer 2776
network layer-based industry 3130
network mobility (NEMO) 2801
network model 2987
network operators 697
network provider 1759
network security 1540
network security requirement 453
network service provider (NSP) 1698
network size, variation in 3006
network society 761
network survivability requirement 453
network topology 2952
networkability 291
network-based approach 3447
network-based defenses 3596
networked industries 698
networking environments 650
neural network-based agent 624
neural network-based mobile architecture 618
neutral data format 3218
new economy 1690
new income 1692
new marketing medium 1692
New Orleans, Louisiana 1536
next generation mobile communications 3561
Nextel 1331
next-generation networks (NGNs) 2292
n-grams prediction models 2969
niche time 1692
Nippon Telegraph & Telephone Corp. (NTT) 2853
node migration 622, 623
node migration, challenges of 623
node virtual environment (NVE) 644
node virtual environment network (NVEN) 644
NodeManager agent 1448
Nokia 34, 703
Nokia's Youth Text 2004 program 1687
nomadic lifestyle 2065
nomadicity 651
non face-to-face gaming 992
non-AR, face-to-face 992
non-bank financial institutions 1699
nonformal learning 1966
non-line-of-sight (NLOS) 2376
non-location-related queries (NLRQs) 335, 336
non-mobile group 3430
non-repudiation 2787
non-SMS data revenue 708
nonspecialist devices 1396
nonstandard-dimension hierarchies 1758
non-textual information 2846
non-value-added 1496
nonvoice services 696, 697
Nordic Mobile Telephone Group 2852
Nordic Mobile Telephony (NMT) 704
not invented here (NIH) syndrome 406
notebook computers 85, 1185
notifications 1566
novel fuzzy scheduler 2996
NTT DoCoMo 1339
NTT DoCoMo 32
NTT DoCoMo 507, 701
NTT DoCoMo i-Mode approach 1334
NUS Mobile IP (MIP) 3138

O

OA&M approach 2494
object manipulation 988
object-oriented database 1445
object-oriented indexing 3104
observed behavior, model based on 2830
observed time difference of arrival (OTDOA) 1756, 1761
occasional data (OC) 591
occupational health and safety 2155
occurrence rate variability 3056
offered data services 1339
offered services 1339
off-line capable 384
off-line customer 42
off-line phase 2930
OHSUMED 1459
on-demand distance vector routing, ad hoc (AODV) 2700
on-demand-based 3080
one-stop government 1549
one-time two-factor authentication (OTTFA) 1271
online e-business 2620
online environment 776
online governmental services 253
online learning 1393
online phase 2932
online public relations 242
on-mobile join operations 356
on-mobile join processing 355
on-mobile location-dependent information processing 355, 370
on-mobile location-dependent operations 362
ontologies 804, 2533, 2534, 2564 2968
ontologies in computer science 2911
ontology and epistemology 2909
ontology changes 3412
ontology shopping 2918
ontology structure 3407
ontology universally unique identifier (OUUID) 2959
ontology usage 3412
ontology-based diagnostics services integration 638
ontology-based diagnostics, based on maintenance data 637
ontology-based service provision 3406
ontology-based service selection 3407
ontology-based standardization of maintenance data 636
OntoServ.Net implementation issues 636
open auction 1641, 1652
open infrastructure 3414
open mesh network technology 1541
Open Mobile Alliance (OMA) 2846
open service platform 1563
open standards 698
OpenGL ES 987
operating systems (OS) 1190, 1570
operator-driven business models 506, 710
optimal linkcell size determination 3045
optimistic two-phase locking (O2PL) 3023, 3029
Oracle 10g 1221
Oracle databases 1221
Oracle7.2 1221

Oracle8i 1221
Oracle9i 1221
order entry systems 2405
ordered access list 3119
organisation design 1149, 1154
organizational interoperability 1571
organizational memory 968
organizational readiness 1615
organizational willingness 765
orientation infinite planes 941
original unit (OU) 2535
orthogonal frequency division multiplexing (OFDM) 3561
orthogonal frequency division multiplexing (OFDM) block diagram 3570
orthogonal frequency division multiplexing (OFDM) spectrum 3568, 3569
orthogonal frequency division multiplexing (OFDM) structure 3571
orthogonal frequency division multiplexing (OFDM) symbols 3567
orthogonal frequency division multiplexing (OFDM) system parameters 3567, 3572
orthogonal frequency division multiplexing (OFDM) system structure 3569
orthogonal frequency division multiplexing (OFDM) systems 3565
orthogonal frequency division multiplexing (OFDM) transmission technique 3561, 3562, 3563, 3567, 3568, 3581
orthogonal frequency division multiplexing (OFDM) transmission technique, advantages of 3566
orthogonal frequency division multiplexing (OFDM) transmit signal 3566
orthogonal frequency division multiplexing (OFDM) uplink signal processing 3574
orthogonal frequency division multiplexing (OFDM) uplink transmission scheme 3573
orthogonal frequency division multiplexing (OFDM) with multiple access schemes 3571
orthogonal frequency division multiplexing (OFDM)-based and synchronized cellular network 3580
orthogonal frequency division multiplexing (OFDM)-based cellular environment 3582
orthogonal frequency division multiplexing (OFDM)-based modulation scheme 3574
orthogonal frequency division multiplexing (OFDM)-based systems 3561, 3566
orthogonal frequency division multiplexing (OFDM)-based uplink systems 3573
orthogonal frequency division multiplexing-code divison mulitple access (OFDM-CDMA) 3571
orthogonal frequency division multiplexing-code divison mulitple access (OFDM-CDMA) systems 3571, 3573
orthogonal frequency division multiplexing-fequency division multiple access (OFDM-FDMA) 3571
orthogonal frequency division multiplexing-fequency division multiple access (OFDM-FDMA)-based systems 3575
orthogonal frequency division multiplexing-fequency division multiple access (OFDM-FDMA) multi-user uplink system design 3575
orthogonal frequency division multiplexing-fequency division multiple access (OFDM-FDMA) multiple access scheme 3582
orthogonal frequency division multiplexing-fequency division multiple access (OFDM-FDMA) schemes 3561
orthogonal frequency division multiplexing-fequency division multiple access (OFDM-FDMA) technique 3572
orthogonal frequency division multiplexing-fequency division multiple access (OFDM-FDMA) uplink system 3574
orthogonal frequency division multiplexing-fequency division multiple access (OFDM-FDMA) uplink transmit signal 3580
orthogonal frequency division multiplexing-time division multiple access (OFDM-TDMA) 3571
orthogonal frequency division multiplexing-time division multiple access (OFDM-TDMA) curve 3573
orthogonal frequency division multiplexing-time division multiple access (OFDM-TDMA) structure 3572
OSA/ParlayX 1574
outdoor structures 937
output devices 1188
overlay network 3078

P

P2Cast 3074
packaging and synchronization 1371
packed data gateway (PDG) 2676
packed-based IP-networks 700
packet data network (PDN) 2754
packet data protocol (PDP) 2754
packet delivery ratio (PDR) 3009
packet switch (PS) 2674
packet switching 1255
packet switching networks 2844
packet tunneling costs, analysis of 2992
PacketWriter 255
pagers 813
paging 1547
paging cost function 684
pairwise transient key (PTK) 2682
Palm 34
Palm OS 535, 538
Palm OS 912, 918
Palm OS (Cobalt) 912
Palm OS (Garnet) 912
paper-based material 3530
paper-based therapy material, sample 3531
paradox 1947
parallel air channels 3111
parametric stereo (PS) 2847
PartiallyLinearOrder algorithm

3096
participation levels 1397
partner relationship (PR) 2536
passive distraction 2060
passive mobile group 3430
passive tags 3375
path management module 3134
patient to doctor profiling 3371
payment authentication 49
payment instruction 2617
payment mechanism 1714, 1717
payment mechanism, impact of 1717
payment protocol 1716
payment system 1714, 1716, 1720
Paymint (parking service) 1681
PC business 699
PC industry 699
PE I 3518, 3520
PE II 3519, 3521
PE III 3519, 3522
peak-to-average ratio (PAR) 3561, 3574
Pebbles applications 1403
pedagogical agent 778
pedagogy 1966
peer-to-peer (P2P) model 1194, 1196, 1203, 2281, 2282
peer-to-peer (P2P) network 953
peer-to-peer (P2P) paradigm 2703
pen computing 1003
pen-based interface 999
penetration rate of mobile phones 708
perceived characteristics of innovating (PCI) 187, 2206
perceived ease of adoption 2012
perceived ease of use (PEOU) 95, 1615, 1618, 1625, 1888, 2008, 2017, 2094
perceived ease of use (PEOU) of general advertisements 1887
perceived enjoyment 93, 1618, 1625
perceived knowledge 373
perceived risk (PR) 2207
perceived switching cost 1932
perceived usefulness (PU) 94, 1615, 1618, 1625, 1888, 1889, 2094
perceived usefulness (PU) of personalized message 1887
perceived usefulness (TAM), 2017
perceived value 2006
performance 375
performance evaluation 3002
performance metrics 3003
performance nature of lectures 1405
periodic broadcasting 3070
peripheral location areas 655
permission 33, 258
permission-based marketing 281
permutation-based pyramid broadcasting (PPB) 3071
persistence layer 460
persistent non-congestion detection (PNCD) 493
personal area networks (PANs) 150, 590, 2728
personal communications services (PCS) 652, 1677
personal computer (PCs) 406, 730, 1185, 1620
personal digital assistants (PDAs) 2, 9, 28, 85, 87, 92, 497, 522, 590, 712, 796, 813, 837, 871, 1003, 1185, 1187, 1194, 1202, 1256, 1260, 1367, 1382, 1395, 1397, 1430, 1620, 1701, 1754, 1761, 1857, 1878, 1886, 1938, 1947, 2023, 2026, 2189, 2373, 2419, 2464, 2653, 2843, 3130, 3270, 3460, 3530
personal digital assistants (PDAs) applications 3536
personal digital assistants (PDAs) interface design 1257
personal digital assistants (PDAs) messaging 1202
personal digital assistants (PDAs) MMS peer design 1199
personal digital assistants (PDAs) models 3532
personal digital assistants (PDAs), courses 1392
personal digital assistants (PDAs), learning environment 1389
personal identification number (PIN) 2655, 2663
personal information management (PIM) 2, 1003
personal innovativeness 93
personal technologies 1406
personal tours 3460
personal virtual environment (PVLE) 1962
personalization 33, 241, 380, 804, 1886, 1888, 1892, 3414
personalized call-ring service 2513, 2513, 2529
person-to-person communication 708
pervasive computing (PC) 589, 782, 1570, 3029 3212
Philadelphia, Pennsylvania 1533
phone quality audio signals 2846
phoneme-based speech recognizer 3470
PHY mode 3586
physical context 1065
physical mobility 760
Piconet 2968
pivot point displacement 2856
platform certification, market effects of 1726
platform certification, structural effects of 1721
platform design 1573
platform management 1575
platform ontologies 1579
platform services 1574
playback disruption 3138
plucker 137
plug-in 1203
Plumtree Wireless Device Server 193
Pocket Blue 255
Pocket Rescue 255
PocketPCs 85, 913, 1185, 1260
podcasting 837, 1174
point of sale (POS) 1245, 2280, 2282, 2283, 2288
point-to-point (p2p) 3074
policy computing 3236
policy decision point (PDP) 3241
policy warehouse 3238
policy-based approach 3237
policy-based architecture 3238, 3243
policy-based architecture for security 3238
policy-based management 3210
polyphonic ringtone 174
portability 290, 375, 1243, 3082
portable battery-powered devices 85
portable console (device) 295
portable devices 2850, 3130
portable devices, downloads 86
portable game devices 108
portable handheld devices 87
portable media center 913

portable media players 108
portable network graphics (PNG) 2848
PORTABLEPKI 2788
portal service 1963
portals 1563, 1961, 2498
portlet (miniportal) 189
position infinite planes 941
positioning 1756
positioning accuracy 1759
positioning aggregation 360
positioning approach 1756
positioning system 1756
positioning techniques 1756
positioning technology 1758
postal code 1570
posthuman 1174
post-it notes 1401
post-join 371
post-processing operations 359
power consumption 3084, 3471
power of pull 253
power of push 253
power of reach 253
power spectrum 3480
PQL query language 1445
practical indexes 316
pragmatics 2788
pre-buffered data 3133
pre-buffering time 3133
predicted future location (PFL) 583
prediction model learning 2972
prediction set data 2875
preferred applications 1635
pre-join 371
premium rate SMS 708
pre-paid card services 1547
pre-processing operations 358
presence 1181
presence awareness 385
presentation layer 460
presumed credibility 373
price offer 3430
price sensitivity 94
price tolerance 1932, 1933
primary notation 1940
primary service provider (PSP) 407
prior knowledge 92
privacy 96, 255, 792, 1066, 2346, 3218
privacy issues 1887
privacy preserving routing (PPR) 2707
privacy protection 2820
private data 3080
proactive information delivery 191
proactive scheduling 3070
proactivity 851
process requirements 764
procurement 789
produced classification model 2903
product architecture 699
product differentiation 2150
product/industry configuration 699
production function 2306, 2308, 2310, 2320, 2321, 2322
profile management agent 583
profile matching 3373
profiles 1592
profiles management 1589
program manager 3503
programming language 3312
projection 361
projection carving 943
projection colouring 945
promoters 2831
protected extensible authentication protocol (PEAP) 2787
protocol 312, 1203
protocol description unit (PDU) 3361
prototyping 235, 920, 3281
prototyping, with storyboards 924, 936
provider stationary agent (PSA) 2944
proxies 3073, 3077, 3220
proximity 1066
proximity-based LAD 656
Psion 34
Psion I 144
PSP 1333
public data 3080
public e-services 3463
public information 1565
public information services 1564
public key cryptography standards (PKCS12) 1245
public key infrastructure (PKI) 35, 50, 1714, 2656, 2715, 2774, 2787, 3136
public key infrastructure (PKI) SIM card 2662
public land mobile network (PLMN) 2675
public mobile services 1565
public mobile telecommunications 1547
public mobile telephony (Cellular) 1547
public procurement act 3463
public relations (PR) 240, 247
public switched telephony network (PSTN) 1547, 1555
publish/subscribe paradigm 575
pull campaigns 1660
pull model 575
pull strategy 585
pull, push, and tracking services 1754
pull-based scheduling 3070
pulse code modulation (PCM) 2846
purchase order 2617
push campaigns 1659
push content mechanisms 3
push model 575
push strategy 584
push to talk over cellular (PoC) 469
push-based scheduling 3070
pyramid broadcasting (PB) 3070

Q

qualitative data analysis 2041
qualitative study 2130, 2133, 2134
quality attributes 375, 1940
quality control 174
quality function deployment (QFD) 377, 797, 1942
quality of data (QoD) 3554, 3560
quality of service (QoS) 420, 430, 802, 1181, 1786, 2773, 2833, 2841, 2843, 2998, 3027, 3078, 3237, 3571
quality of service (QoS) adaptation 2841
quality of service (QoS) classifications 3411
quality of service (QoS) concepts and models 2854
quality of service (QoS) demand 3586
quality of service (QoS) factor 2851
quality of service (QoS) in mobile networks 2853
quality of service (QoS) MAC for MANETs 2836
quality of service (QoS) management 2856
quality of service (QoS) management interface, dynamic

2856
quality of service (QoS) management user interface, dynamic 2857
quality of service (QoS) management, dynamic 2856
quality of service (QoS) mechanisms 2838
quality of service (QoS) negotiation user interface, static 2857
quality of service (QoS) negotiations 2854
quality of service (QoS) protocols 2843, 2855
quality of service (QoS) provision ring 2858
quality of service (QoS) provisioning 2858, 2999, 3133
quality of service (QoS) request 2854
quality of service (QoS) requirements 2851, 2856
quality of service (QoS) routing protocol 2858
quality of service (QoS), policy based architecture for 3240
quality of service (QoS), policy-based 3241
quality of service (QoS), soft 2842
quality of service on-demand routing, ad-hoc (AQOR) 2838
quality, cost, temporal triangle (QCTT) model 2855, 2856
quality, cost, temporal triangle (QCTT) threshold line 2858
quantitative analysis 688
queries 353, 354, 356, 360
query integration system (QIS) 1445
query latency 3016
query processing 366, 367
query processing strategies 3192
query types 315
querying fixed-object databases 338
querying in abstract data type model 343
querying in constraint database model 343
querying in MOST model 344
querying moving-object databases 341, 343
querying with uncertainty 345
query-processing techniques 1759
question-answer relationships (QAR) 3542, 3551
quick information collection 249

R

radio access technology (RAT) 961
radio channel 3564
radio channel behaviour 3563
radio channel state information (CSI) 3574
radio channel transmission 3567
radio frequency identification (RFID) 29, 439, 759, 2020, 2026, 2376, 3387
radio frequency identification (RFID) description 3373
radio frequency identification (RFID) transponder programmers 3378
radio frequency identification (RFID) transponders, types 3374
radio network controller (RNC) 2770
radio resource management (RRM) 3562
radio resource management (RRM) procedure 3585
radio resource management (RRM) techniques 3562
radio trunking 1547
Radio538 2448
railway mobile terminals 1516
random access memory (RAM) 1189
random mobile movement paths 3162
random movement path generation 3161
rank metric map 3159
rapidly prototyping mobile interactions 922, 936
rate control modules 3134
rationality 2070
rationalization 2073
REACH 1545
REACH 1.0 1546
REACH 2.0 1546
REACH 3.0 1546
REACH initiative 1546
REACH, background 1545
reachability 241
reactive scheduling 3070
reactivity 851
read-once write-all 3023
read-only memory (ROM) 1189
real time enterprise 2146
really simple syndication (RSS) 9
real-time 3D design modelling 937
real-time application 3131
real-time capability 1187
real-time data (RTD) 591, 594
real-time market-responding strategy 2525
real-time multimedia applications 3132
real-time order 1467
real-time strategy (RTS) game 291
real-world scenario 3132
recognition 2102
recognition algorithm 1139
recognizer parameters 1139
recontextualisations 1171
records management 968
reductionism 698
redundant recoding 1940
reference model 1594, 1597, 1598
referent models 228, 239
regional director (RD) 2292
registration algorithm 669
registration authority (RA) 2787
registration cancellation 654
registration latency 3141
registration notification 654
registration request 3132
registration time 3141
registry stationary agent (RSA) 2946
regulated industry 699
regulatory forces 1757
regulatory framework 704, 2480
related signed response (GPRS-SRES) 2756
relation definitions 1451
relational algebra set operations 363, 365
relational database 1445
relationships 240, 244
relationships, text communication 2135
relative advantage 1466, 1468, 1472, 1479
reliability 375, 2821
reliability of service 2151
remediation 1947
remote communication 2158
remote computing 776
remote databases 351

remote evaluation (REV) 619, 2568
remote monitoring 1641
remote procedure calls (RPCs) 296, 304, 619
remote servers 367
remote supervision 790
repository management problem 3042
repository size variability 3051
representation languages 1570
reputed credibility 373
request processing flow 3158
request reports (RRs) 3013
Research in Motion (RIM) 34
Research in Motion's (RIM) BlackBerry device 1681
reservation protocol (RSVP) 3241
residual computation dependency problem 481
resource allocation protocol (RAP) 3241
resource description framework (RDF) 799
resource reservation 2858
resource sharing 1563
responsiveness 3474
result demonstrability (RD) 2207
reusability 3312
reusable information objects (RIOs) 1384, 1395
reusable learning objects (RLOs) 1384, 1395
revenue generation 1531, 1540
revenue logics 2463, 2464
revenue model 1594
revenue model 2178
revenue sharing 2472
revenue sharing models 1338
rights enforcement 1124
rights insertion 1124
ringtone 174
roaming 730
robot control paradigm 599
robust intelligent control 597, 617
robustness 375, 600
Rogers' "navigate mobile internet" 1686
Rogers (TDMA, GSM/GPRS) 1680
Rogers AT&T Wireless 1676
Rogers Cable 1682
Rogers Communications 1676
Rogers Wireless 1676
roleplaying game (RPG) 290, 295
rotated component matrix 1980
route reply (RREP) 2701
route request (RREQ) 2701
router advertisement (RA) cache 3134, 3138
router advertisements (RA) 3133
router advertisements (RA) entry 3134
routing and tunneling algorithm 670
routing area identity (RAI) 2755
row scan 3118
Royal Institute of Technology (KTH) 3458
RTK GPS 941
rule evaluation 3002

S

SAFER architecture 305, 882, 2715
SAFER architecture, agent 884
SAFER architecture, agent evolution 887
SAFER architecture, agent fabrication 887
SAFER architecture, agent roaming 888
sales force automation 2150
sampling frequency 1142
Samsung 34
Samsung SPH-A600 1683
Sanyo 8100 phone 1683
Sartre 1955
SaskTel Mobility 1676
satellite communications 589
satellite connections 2853
satellite-based augmentation systems (SBAS) 1050
satellite-based communication systems 2853
satisfaction 239, 1889
scalability 1112, 1116, 3013
scalability requirements 3238
scalable vector graphics (SVG) 798
scenario-based design (SBD) 3321, 3326
scheduler performance 3009
scheduling algorithms 2998
scheduling protocol 3091
scheduling services 3461
scope distribution 3032
scope numbers (SN) 3034
Scottsburg, Indiana 1538
SCTP multi-homing 3134
S-DMB 1145, 1145, 1146
sealed-bid auction 1641, 1652
seamless access 3132
seamless IP-diversity based generalized mobility architecture (SIGMA) 3132, 3134
seamless IP-diversity based generalized mobility architecture (SIGMA), timing diagram of 3137
seamless IP-diversity based generalized mobility architecture (SIGMA) handoff 3134, 3144, 3145, 3146
seamless MPEG-4 streaming 3133
seamless multimedia over mobile networks 3133
seamless multimedia transmission 3134
seamless video 3134
search services 1363, 1366
secure agent data integrity shield (SADIS) 305
secure collaborative learning practices 1967
secure digital (SD) chip 1383
secure distributed anonymous routing protocol (SDAR) 2704
secure electronic transaction (SET) standard 1714
secure government network (SGN) 1550
secure socket layer (SSL) 874, 1714, 2655
secured trust 2831
security 96, 255, 312, 454, 740, 792, 1238, 1610, 1714, 2785, 3218, 3312
security architecture 2767, 2773
security assertion markup language (SAML) 802
security association identifier (SAID) 2774
security parameter establishment (SPE) 2703
security protection 2820
security protocols 2741, 2771
security threats 2600
security, black-box 2586
security, in home networks 2800
security, in wireless environment 2800
security, tamper resistant storage 2728
segmentation of workers 2155

self-confidence 1871
self-configuration 643, 647
self-expression 2073
self-extension 2073
self-governance 2074
self-healing 597, 643, 647
self-identity 2077
self-optimization 643, 647
self-organized cell synchronization 3582
self-organized resource management 3584
self-organizing (SO) way 3585
self-protection 643, 647
self-ranking algorithm 3151
self-starting/proactive 580
selling change management 788
semantic (cognitive) context 567
semantic analysis 3487
semantic balance 631
semantic distance 2958
semantic interoperability 1571
semantic layer 2958
semantic location modeling 2530
semantic matching 2950
semantic matchmaking 2968
semantic peer-to-peer discovery 639
semantic profile 572
semantic reference model 3404
semantic service discovery 1109, 1110, 2959
semantic service platform 3415
Semantic Web 381, 630, 796, 804, 1945, 2936
Semantic Web services registry (SWSR) 2946, 2947
Semantic Web technologies 1562
semantically annotated resource 2968
semantic-distance metric (SDM) 1446
semantic-enabled m-commerce 2957
semi-active tags 3375
semiotic levels 375, 1939
semiotics 381, 804, 1945
sensitivity analysis (SA) 3552
sensor networks 649
sensory voice recognition module 3484
sensory-aided mobile computing 2021
serializability 3021, 3030
server load 1239, 3474
server log preprocessing 2972
server-based two-phase locking (S2PL) 3030
servers 355, 356, 357, 358, 366, 367
service 1116
service agents (SAs) 1107
service architecture 700
service client plug-in feature 1201
service composition 1571
service consumption phase 1151, 1167
service delivery phase 1150, 1150, 1151
service dependent speech recognition 3474
service description header (SDH) 622
service design 1152, 1161
service development phase 1150
service directory 1105, 1116
service discovery 1104, 1116
service discovery protocols (SDPs) 860, 2968
service discovery service (SDS) 1108
service discovery, issues in 1109
service execution 1577
service improvements 84
service location protocol (SLP) 1106, 1576, 2958
service location, scalability issues in 1112
service matching 3412
service mobility 758
service ontology 3410, 3411
service provider 808, 1576, 1755, 1757, 1759, 3239
service repository 1116, 1573, 1576
service requester 1576
service requirements 2841
service retrieval 3410
service roaming 3413
service set changes 3412
service subsystem 3416
service usability 1558
service usage description 2948
service-oriented architectures (SOA) 1569, 3212
service-oriented information-logistical platform 3404
service-oriented system 1966
services adoption 1593
services, deployment of 1564
serving GSN (SGSN) 2676, 2754
serving network (SN) 2756
servlet specification 3313
session description protocol (SDP) 424
session initiation protocol (SIP) 420, 430, 468, 470
session mobility 758
Set-Primary-Address action 3136
setup handoff 3134
Shanghai 1668
shareable content object reference model (SCORM) 119, 1386, 1388
shared data 3080
shopping application 3501
short message service center (SMSC) 2513
short message services (SMSs) 9, 18, 30, 38, 85, 187, 279, 708, 836, 1000, 1003, 1134, 1194, 1249, 1363, 1366, 1548, 1601, 1625, 1755, 1761, 1863, 1886, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2294, 2513, 2513, 2529, 2846, 3361, 3367
short message services (SMSs) channel 1558
short message services (SMSs) Chat 1398
short message services (SMSs) Poll-Center 1398
short message services (SMSs), case of 2130
short message services (SMSs), social context of 2136
short message service (SMS) management systems 3461
shortest time 589, 594
shyness, overcoming 2136
signal processor 3469, 3474
signal quality 3133
signaling system 7 (SS7) 652, 2757, 2797
signalling costs 682
signalling protocol 428
signal-to-noise ratio (SNR) 1137, 3473
signal-to-noise ratio (SNR) level 3133
signature-based indexing 3085
signed response (GPRS-SRES)

2756
similarity index (SI) 3428
simple object access protocol (SOAP) 798
simple object application protocol/ hyper text transfer protocol (SOAP/HTTP) 2300
simple proximity indexes 319
simple service discovery protocol (SSDP) 1105
simplified techniques 940
simulation environment and methodology 3003
simulations 1171, 1174
Sina 1669
single air channel 3104
single carrier (SC) modulation schemes 3565
single frame-block motion compensation (SF-BMC) 3133
single sign-on (SSO) protocol 2794
single tree CART models 2879
single user detections (SUD) 3573
single-class indexing 3104
SIRIUS Satellite Radio 2850
situated learning 111
situation detection 3412
situation ontology 3411
situation-dependent mobile application 2168
Skype 697
skyscraper broadcasting (SB) 3071
slave device 499
small form factors 499
small/medium enterprises (SMEs) 1551, 3460
smart algorithms 3322, 3325
smart cards 1627, 2733
smart clients 383
smart devices 630
smart phones 9, 85, 87, 92, 108, 497, 913, 1003, 1185, 1238, 1245, 1245, 1397, 1430, 1565, 1601, 1947, 3598
smart space 3212
SNACK-NS (New snoop) 491
SNACK-TCP 491
snapshot queries 578
sniffing algorithms 3323
social behaviour 2125
social cognitive theory (SCT) 89
social engineering 3598
social identity 2125
social influence 95
social integration 1871, 2041
social mobility 760
social networking 190, 3281
social relationships 2141
social shaping 762, 768
social translucence 2133, 2139, 2142
social upkeeping 2138
social usability 1974
sociotechnical action perspective 2086
sociotechnical analysis 2087
sociotechnical ensemble 2081
Söderhallarna, Stockholm 3460
software agents 892, 1641, 2583
software business models 2464
software customization 3305
software development toolkits (SDK) 1942
software tools 1606
Sohu 1669
Sonera 704, 3467
Sony 1333
Sony-Ericsson 34, 705
sophisticated service discovery protocols (SDPs) 2957
SoundHelper 3534
SoundHelper interface 3533
soundness-enriched queries 315
space zones 2124
spatial and temporal dimensions 1430
spatial data 3031
spatio-temporal data 313
spatio-temporal queries 337
speaking mode 3473
speaking style 3473
spectral band replication (SBR) 2847
spectrum availability 1338
speech and language therapists (SLTs) 3530
speech and language therapy (SLT) 3529, 3530
speech and language therapy (SLT) application design 3536
speech and language therapy (SLT), accessibility 3532
speech and language therapy (SLT), application design 3536
speech and language therapy (SLT), conventional aids 3530
speech and language therapy (SLT), Internet-enabled 3537
speech and language therapy (SLT), multimedia prototypes 3533
speech and language therapy (SLT), research and development 3537
speech and language therapy (SLT), socio-cultural issues 3536
speech and language therapy (SLT), SoundHelper 3533
speech disorder 3529
speech recognition 3497, 3509
speech recognition on embedded devices 3470, 3471
speech recognizers in UC 3473
speech signal 3468
speech sounds 3529
speech synthesis 3498, 3510
speech technology 191, 3497
speed of delivery 2151
Sphinx 3.5 3479
spontaneity 2810
Sprint 1331
SRA 3159
SSM 1444
St. Cloud, Florida 1538
standard conformity assessment 3171
standard query languages 347
standards-based technical reductionism 699
Stanford Mosquito (MNET) 3138
static quality of service specification 2856
statistical time division multiplexing (STDM) 2844
Steering Committee, Stockholm 3457
still images 2844
still images applications 2851
still images coding 2848
still images transmission 2851
still picture 813
stochastic gradient boosting 2871
Stockholm 3455
Stockholm Academic Forum 3460
Stockholm IT Council 3457
Stockholm Visitor's Board 3459
Stockholm's e-strategy 3456, 3464
stocks application 3502
stovepipe solutions 1340
Strategic Health Authority (StHA) 407
strategic IT initiative 1546
stream control transmission protocol

(SCTP) 3134
stream control transmission protocol, Linux kernel (LKSCTP) 3138
streaming 1126, 1128, 1134, 3069, 3073, 3078, 3221
streaming multimedia 3130
streaming technology 1128
structured query language (SQL) 1109, 1220
structures supporting nearest neighbor queries 319
structures supporting range queries 316
structures supporting soundness-enriched inquires 321
students 3460
stylesheets 1388
sub-carrier allocation 3577
sub-carrier allocation of test signals 3583
sub-carrier allocation process 3575, 3585
sub-carrier frequency 3567
sub-carrier selection process 3576
sub-carrier signals 3567
sub-carrier spreading technique 3576
sub-carrier transmit symbols 3576
sub-channel signals 3567
subject categories (SCs) 1453
subjective workload 214, 224
subnets 3130
subprofile 569
subscriber identification module (SIM) 758
subscriber identification module (SIM)-lock 705
subscriber identity module (SIM) card 87
subscriber identity module (SIM) card 2661, 2662, 2755
subscriber station (SS) 2774
substitute management 3462
Sudan, mobile phone use 2110
SUEDE 920
Sumit Mobile Systems Ltd. 1671
supplement physical capture limitations 938
supply chain 1226
supply chain integration 1466
supply chain management (SCM) 788, 1483, 1485, 1497, 2368, 2369
supply chain management (SCM), existing model of 2372
supply-side market effects 1729
support of the current serving GSN (SGSN) 2771
support software 1221
surface of revolution 945
Swedish PTT Televerket 2852
Swedish Road Administration 3461
Swiftel 1548
switching cost 1932
Swordfish 1681, 1682
Sybil attack 2710
Symbian 1337
Symbian operating system (OS) 912, 917
Symbianphone.com 1666
symbolic model 3032
synchronization 35, 1003, 1189, 1193
synchronized cellular network 3580
synchronized multimedia integration language (SMIL) 798
syntactics 2788
system interface 3516
system requirements 1206
system structures 1206

T

tablet PCs 108, 813, 817, 834, 1000, 1003, 1190, 1430
Tandy Radio Shack 34
Tandy's Zoomer 144
target publics 242, 243
targeted customer segment 2172
task of learning description 475
task ontologies 3408
task scheduling 671
task-oriented mobile distance learning 473, 487
task-oriented seamless mobility, agent-based approach 476
taxonomy of database operations 355
TCP enhancements 488, 490
TCP enhancements, classification of 490
TCP for 3G cellular networks 490
TCP for WLAN 493
TCP/IP network infrastructure 3131
TCP/IP networks 2846
TCP-Reno 491
TCP-Veno 490
TCPW-A 493
TCP-Westwood with agile probing (TCPW-A) 493
TD-CDMA (time division-CDMA) network 2852
TDMA 29
TD-SCDMA 1666
technical interoperability 1570
technoeconomic evaluations 2483
techno-economic evaluations, structure of 2483
technoeconomic methodology 2483
techno-economic terms 2495
technology acceptance 2017
technology acceptance literature, review 89
technology acceptance model (TAM) 1297, 1615, 1625, 1627
technology acceptance model (TAM) 1887, 1888, 1892, 1975, 1976, 1977, 1983, 1998, 2017, 2093, 2094
technology acceptance model (TAM) 2206
technology acceptance model (TAM) 89, 90
technology acceptance model (TAM) propositions, extended 2099
technology acceptance model (TAM) propositions, traditional 2098
technology acceptance model for mobile services (TAMM) 2001, 2018
technology adoption 1976
technology constraints 1868
technology design 1151, 1160
technology determinism 3271, 3281
technology diffusion 1627
technology domain 576
technology forces 1757
technology intervention 288
technology perspective 2199
teenagers 2131, 2132, 2133
TeleCard 1548
Telecom Finland 704
telecom industry 2308
telecom operators 696
TelecomCity 3455
telecommunications 2327
telecommunications regulatory commission (TRC) 1547, 1758
telecommunications, evolution of 2852
teledensity 1698
tele-health 431
telemedicine 420, 430, 1443

telephony 2130, 2132, 2139, 2140
teleradiology 1271
tele-work 2061
Telia 3460
TeliaSonera 3467
TeliaSonera 697
TELUS (CDMA and iDen) 1680
TELUS Mobility 1676
TELUS's national CDMA 2000 1X network 1684
template description language (TDL) 3237
temporal continuity/long-Lived 580
temporal key integrity protocol (TKIP) 2681, 2786
temporal-dependent invalidation 3033
temporary disconnection 3217
temporary logical link identity (TLLI) 2755
temporary mobile subscriber identities (TMSI) 2755, 2770
terrestrial digital multimedia broadcasting (T-DMB) 1145, 1145, 1146
test pilots 3460
test signal structure 3583
Testbed Botnia 3455
text applications 2850
text coding 2845
text database 1606
text messaging 1003, 1402, 1601, 1889, 2130, 2131, 2132, 2133, 2135, 2140, 2141, 2143
text messaging for Young Adults 2133
text-messaging studies 1406
text-to-speech (TTS) request 3475
theory of planned behaviour (TPB) 89, 90
theory of portable PKI (PORTABLEPKI) 2788
theory of reasoned action (TRA) 89, 90, 1977, 1983, 2094
thesaurus maintenance 1451
thesaurus structure 1451
ThesMaster agent 1448
thin client 461
thin client architecture 460
three dimensional (3D) network game of 295
three layer quality of service (TRAQS) model 2854
three-tier architecture 2420
throughput 2843, 2844, 3141
thumb board text interface 1000
Thuraya Satellite Telecommunications Company 1548
tickets, printing and validating 1526
tickets, validation and vending 1516
time consciousness 1863
time difference of arrival (TDOA) 575
time division duplex (TDD) system 3581
time division multiple access (TDMA) 2187
time division multiple access (TDMA) 2774
time division multiple access (TDMA) 3561
time division synchronous code division multiple access (TD-SCDMA) 2187
time management systems (TMS) 568
time of arrival (TOA) 35, 1761
time of arrival (TOA) positioning method 1761
time ontology 3409
time zones 2124
time-critical decision making 3552
time-critical decision making problems 3552
time-critical decision support 3554
time-critical decisions 3560
time-critical information 1565
time-critical situations 3552
time-division-duplex (TDD) 559
time-parameterized queries 322
timing advance (TA) 1756
T-Mobile 1331
tool-being 2075
topiary 920
topoi 1951
Toronto Parking Authority (TPA) 1681
total signaling costs, analysis of 2993
touch screens 1188
tourism 3406, 3459
tourism, Stockholm 3459
tourist digital assistant (TDA) 3387
traffic information system 791
train set data 2875
training 792
trajectory-based queries 315
transceivers 3378
transducer 3473
transferring delay 480
transferring failure problem 479
transformation 699
transitive trust 2830
transmission degradation 3474
transmission efficiency 3134
transmission error 454
transmission medium 595
transmission of control packets 3134
transmission of duplicated video packets 3134
transmission of multimedia information 2843
transmission perspective layer (TPL) 2854
transmission protocols 1570
transparency 375
transparent access to multiple bioinformatics information sources (TAMBIS) 1445
transport encryption key (TEK) 2774
transport layer handoff schemes 3130
transport layer protocol 489
transport layer security (TLS) 2655, 2774, 2787
transportation 249
traVcom service 1775
travel services 2
TREC 1460
TREC9 1460
tree-based indexing 3085
TreeNet confusion matrix 2883, 2884
TreeNet models 2881
TreeNet variable importance 2886, 2892
triangle routing 3132
triangulation 2041
trilateration 1048
Trojan Horse 3598
Trojan horse, Drever 3593
Trojan horse, Locknut 3592
Trojan horse, Skuller 3592
trust 1615, 1632, 1974, 2010
trust center 1720
trust classes, model based on 2831
trust graph 2829
trust models 2827
trust policy language (TPL) 3237
trust properties 2831
trust, antecedents in 2808

trust, defining 2828
trust, mathematical model for 2829
trusted third party (TTP) 1715, 1717, 1720, 2828, 2831
trustworthiness 373
tunnel establishment protocol 658
two-level browsing scheme 499
two-phase locking 3022

U

U.S. city government wireless networks 1530
ubiquitous access 3
ubiquitous applications 3320
ubiquitous computing (UC) 782, 2024, 2026, 3029, 3405, 3454, 3468
ubiquitous education 123
ubiquitous mobile systems 2827
ubiquity 241, 805
U-Know Campus Navigator 2504
U-Know yellow pages 2503
U-Know, features of 2503
uncertainty 3560
uncertainty avoidance 2809
uncomfortable silences 1396
unicasting (one-to-one) technology 1868
unicode 2846
unicode standard 798
unification of approaches 346
unified medical language system (UMLS) 1450
unified modeling language (UML) 1572, 2195, 2291
unified process 1572
unified theory of acceptance and use of technology (UTAUT) 90
uniform resource identifier (URI) 798
uniformity 1864
uninterrupted multimedia transmission 3130
union set operation 364
unique institutions 1666
unit matching 3486
United Kingdom, mobile phone use 2110
United States of America 1331
universal description, discovery and integration (UDDI) 1576, 2370
universal mobile telecommunications system (UMTS) 9, 420, 696, 1125, 1134, 1144, 1265, 1271, 1622, 1757, 2374, 2676, 2726, 2766, 2767, 2770, 2853
universal mobile telecommunications system (UMTS) subscribers identity module (USIM) 2677
universal mobile telecommunications system (UMTS) systems 3562
universal plug and play (UPnP) 2958
universally unique identifier (UUID) 2958
University of Richmond, Virginia 1538
unmanned autonomous vehicle (UAV) 2842
unpredictable delay 489
unpredictable object movement modeling 347
unsecured network 1536
update reports (URs) 3013
update reports (URs) caching model 3014
updated invalidation report (UIR) 3013
uplink (UL) 596, 3586
uplink time of arrival (TOA) 1756
upper ontology 3407
UPS 1340
up-to-date traffic information 3461
usability 239, 375, 1068, 1243, 1564, 1860
usability requirements 1564
usage of mobile phone 1558
use of short message services 2133, 2136, 2137, 2142
USE-ME.GOV 1562
USE-ME.GOV applications 1579
USE-ME.GOV general architecture 1573
USE-ME.GOV platform 1573
USE-ME.GOV platform interfaces 1574
USE-ME.GOV project 1562, 1563
USE-ME.GOV system 1562, 1573
user acceptance 767
user agents (UAs) 424, 1107
user assistant agent (UAA) 1230
user context 374
user datagram protocol (UDP) 2688
user device 2850
user equipment 3205
user experience 2041
user interfaces (UIs) 732, 910, 947, 1565, 1648, 1757, 2856, 3217, 3556
user knowledge 1377
user login 3239
user management service 1575
user manager module 569
user mobility 758, 1104
user mobility during the usability evaluation 3184
user mobility model 2986
user notification 3
user performance measurement 3172, 3184
user perspective layer (UPL) 2854
user predisposition 91
user preference 2850
user profile 381, 804
user profile changes 3412
user profile ontology 3409
user registration service 1965
user satisfaction measurement 3171, 3184
user service requestor (USR) 2942
user subsystem 3416
user-centered mobile computing 2019
user-end module 569
user-friendly mobile services 3455
uses and gratifications theory 1983
USSD 35
utilitarian elements 1616
utilitarian value 2237, 2238, 2249
utility function 2960

V

valid scope 3032, 3039
valid scope area 3036
valid scope distribution 3032, 3039
validity queries 322
value chain 759, 1563, 1595, 2511, 2512, 2523
value chain analyses 700
value networks 2440
value-added services (VASs) 173
value-added services (VASs), characteristics of 2519
various bundling strategy 2526
varying precision 1758
V-Card 1125, 1126, 1127, 1132
V-Card core architecture 1127
V-Card examples 1128
V-Card streaming technology 1128
V-Card, evaluation of 1131
V-Card, legal aspects 1129

vector graphics 3533
vector quantization (VQ) 3485
vehicle movement 1759
vendor agent (VA) 1230
Verisign 43
Verizon 1331
vertical differentiation 2512
vertical handoff 964
vertical industry structure 700
vertical mobile business application 2168
vertical/integrated configuration 706
Veterans Health Administration (VHA) 1444
video applications 2851
video capture devices 813
video conferencing 2844, 2850, 2852, 3130, 3131
video data transmission 3132
video on demand (VoD) 2793
video phones 2851
video streaming 88, 1128, 2850, 3069
video transmission 2852
Vila Nova de Cerveira (Portugal) 1565
viral marketing 2183
Virgin Mobile 1679
Virginia Tech (VT) 3328
virtual active set 3210
virtual check 1239
virtual clock (VC) 2998
virtual communities 1772, 1781, 1974
virtual community for mobile agents 881, 890
virtual democracy 772
virtual mobility 760
virtual network computing (VNC) 3184
virtual online classroom 3352
virtual private networks (VPNs) 85, 822, 1245, 2677, 2758, 2776
virtual stores 1544
virus 3598
virus intrusion 87
visibility 2133, 2139, 2141
visitor location register (VLR) 591, 652, 2753, 2770
VistA 1444
Vodafone 705
Vodafone Japan 708
Vodafone Live! 509
vodcasts 837
voice communication 2132, 2133, 2135, 2137, 2139
voice driven emotion recognizer mobile phone 3511
voice input 1568
voice input channel 1565
voice over Internet protocol (VoIP) 659, 2793
voice transmissions 590
voice-centric business paradigm 697
voice-enabled user interfaces 3494
voice-over-WLAN 697
VoiceUI 748
voluntary cooperation 50
VR techniques 939
vulnerability 3598
VXML 35

W

walk-up-and-use 3271, 3281
walled garden 32
Walsh-Hadamard matrix 3577
watermarking 1124, 2588
wavelet transform 2851
WBT-based cours 812
weakest link 697
wearable computer 2060
wearable computing devices 783
wearable devices 813
wearable displays, comparison of 2052
weather application 3500
Web browsing 2419
Web clipping 35
Web interface 735
Web mining system 2924, 2927
Web mining system, case study of 2929
Web mining system, framework 2929
Web ontology language (OWL) 800
Web ontology language (OWL-S) 1578
Web service access 3322
Web service modeling ontology (WSMO) 3411
Web service provider (WSP) 2947, 2948
Web service retrieval 2951
Web servers 4, 1219
Web services 404, 416, 891, 1573, 2299, 3221, 3510
Web services architecture (WSA) 1572
Web services description language (WSDL) 2370
Web sites 2932
Web sites database 1606
Web sites, adaptive 2971
Web usage mining 2971
Web-based Internet applications 248
Web-based seamless migration 473, 487
Web-based services 1563
WebCT 1385
Web-enabled phones 85
WebQuest tool 3175
weighted trust graph, model based on 2829
wide area network (WAN) 838, 2205
wideband code division multiple access (WCDMA) 1264, 1271, 1666, 2180, 2187, 2374
wideband code division multiple access (WCDMA) (3GSM4) 1666
wideband code division multiple access (WCDMA) 3G mobile network 2853
wideband code division multiple access (WCDMA) technology 2853
WiMobile 1182
windowing 3480
Windows Mobile 912, 919
wired communication systems 2853
wired equivalent privacy (WEP) 2768, 2786
wired Internet 1978, 1981
wired networks 1217
wireless access gateway (WAG) 2675
wireless access points (WAPs) 2785, 3130
Wireless Alexandria 1535
wireless application environment (WAE) 2655
wireless application protocol (WAP) 30, 48, 86, 87, 88, 192, 389, 390, 399, 400, 452, 508, 698, 798, 806, 1250, 1255, 1364, 1366, 1588, 1601, 1619, 1666, 1693, 1755, 1762, 2418, 2655, 2675, 2846
wireless application protocol (WAP) stack 1203
wireless application protocol (WAP), benefits of 88
wireless application protocol (WAP),

challenges 88
wireless application protocol 2.0 (WAP 2.0) 1350
wireless application protocol (WAP) 2020, 2026
wireless application protocol (WAP) GET command 2846
wireless cell 3039
wireless channel utilization 3017
wireless channels 3133
wireless clients 3134
wireless communications 1585, 1691, 1754, 1759, 1840, 1842, 2170, 3012
wireless communications systems, fixed 2853
wireless computing 354
wireless connections 2845
wireless connectivity 1003
wireless data networks 3130
wireless devices 409, 3552
wireless environment, privacy 97
wireless environment, security 97
wireless fidelity (Wi-Fi) 29, 85, 151, 730, 837, 1181, 1181, 1182, 1197, 1592, 1666, 3322
wireless fidelity (Wi-Fi) blanketing 29
wireless fidelity (Wi-Fi) market 1668
wireless fidelity (Wi-Fi) technology 1532
wireless fidelity (Wi-Fi) security protocol 2774
wireless fidelity protected access (WPA) security standard 1538, 2768, 2786
wireless footprint 1535
wireless information security 1540
wireless Internet 1978
wireless Internet network 3131
wireless link 3132
wireless local area networks (WLANs) 21, 29, 120, 1622, 1968, 2187, 2375, 2674, 2785, 2849, 3133
wireless local area networks (WLANs) operator 2476, 2489
wireless local area networks (WLANs) environment 493
wireless local area networks (WLANs) networks 1868
wireless local area networks-access gateway (WLAN-AG) 2675
wireless local area networks-access point name (W-APN) 2676
wireless local community (WLC) 1780
wireless local community (WLC) business partner 1781
wireless local community (WLC) conceptual design 1781
wireless local loop (WLL) standards 3561
wireless marketing 260
wireless markup language (WML) 452, 798, 2023, 2026, 2418
wireless media 3130, 3082
wireless metropolitan area networks (WMANs) 2767
wireless middleware 1588
wireless mobile applications, development of 388
wireless mobile data networks 3130
wireless mobile environment 3012
wireless mobile environments, models for trust 2829
wireless mobile Internet (WMI) 591
wireless mobile networks 3132
wireless mobile networks, trust in 2828
wireless mobile technologies 153, 170, 776, 777
wireless network architectures 1176
wireless networks 817, 822, 1216, 1584, 2848, 2850, 3132, 3133
Wireless Philadelphia 1533
wireless private area networks (WPANs) 2785
wireless service providers (WSPs) 30, 173
wireless session protocol (WSP) 452
wireless technologies, evolution of 87
wireless technology, beyond third generation (B3G) 2674
wireless technology, first generation (1G) 86
wireless technology, fourth generation (4G) 29, 430, 438, 2766, 2843, 3205, 3561
wireless technology, fourth generation (4G) air interface 3572
wireless technology, fourth generation (4G) downlink interface 3572
wireless technology, fourth generation (4G) mobile communication systems 3562
wireless technology, fourth generation (4G) systems 3562
wireless technology, fourth generation (4G) uplink interface 3573
wireless technology, pre-generation (Pre-G) 292
wireless technology second generation systems (2G/2.5G) 29, 86, 436, 806, 1676, 1685, 1686, 2752, 2845
wireless technology, second generation (2G) wireless system 2180
wireless technology, third generation (3G) 1, 29, 85, 88, 279 420, 438, 465, 490, 526, 590, 697, 1126, 1128, 1133, 1134, 1264, 1270, 1601, 1666, 1886, 2374, 2674, 2845, 2852, 2862, 3208, 3457, 3562
wireless technology, third generation (3G) mobile medical image viewing 1261
wireless technology, third generation (3G) network 806
wireless technology, third generation (3G) telecom market 2475
wireless technology, third generation (3G) wireless systems 2158, 2180
wireless technology, third generation mobile virtual network operators (3G MVNOs) 2475,. 2486
wireless technology, third generation mobile virtual network operators (3G MVNOs) business strategies 2483
wireless technology, types of 1532
wireless telecommunications 18
wireless transport layer security (WTLS) 1238 2655
wireless Web 1690, 1691
wireless wide-area networks (WWANs) 1622, 2785
wire-line based video phones 2851
wire-line networks 2850
Wisemax 1671
word of mouth (WOM) 1981

word-based speech recognizers 3470
WordDial 805
WordNet 1444, 1446, 1451, 1462
work settings 1430
workflow management 1767
working memory 779
workplace-based learning 820
World Trade Organization (WTO) 1555
World Wide Web Consortium (W3C) 802
worldwide interoperability for microwave access (WiMAX) 438, 697, 976, 1182, 2375
Wi-Max technology 1532
Worm 3598
Worm, Cabir 3592
Worm, Mabir 3593
WOz approach 926
WSPs 1687
Wu and Palmer algorithm 1453

X

X.509 1245

Y

Yahoo! 1685
yellow pages agent (YPA) 1230
young adults 2133
youth market, how to target 1662

Z

zero window adjustment (ZWA) 490
Zettair 1460, 1461
ZigZag 3074